Le concept de biais en épidémiologie

Nicolas Brault

Université Sorbonne Paris Cité

Thèse de doctorat

de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

Ecole doctorale 400 : « Savoirs scientifiques»

Laboratoire SPHERE / UMR 7219

Le concept de biais en épidémiologie

Par Nicolas Brault

Thèse de doctorat d'Épistémologie, histoire des sciences et des techniques

Dirigée par Alain Leplège et Joël Coste

Présentée et soutenue publiquement le 12 décembre 2017

Présidente du jury : Laurence Meyer, Professeur des universités-Praticien hospitalier de santé publique, Université Paris-Sud.

Rapporteurs : Isabelle Drouet, Maître de conférences en philosophie, Université Paris-Sorbonne.

Maël Lemoine, Maitre de conférences (HDR) en philosophie des sciences biologiques et médicales, Université de Tours.

Examinateur : Elodie Giroux, Maître de conférences en philosophie, Université Lyon III-Jean Moulin

Directeur de thèse : Alain Leplège, Professeur d'histoire et de philosophie des sciences, Université Paris-Diderot.

Co-directeur de thèse : Joël Coste, Professeur des universités-Praticien hospitalier de biostatistique et d'épidémiologie, Université Paris-Descartes.

Titre : **Le concept de biais en épidémiologie.**

Résumé : **Cette thèse, qui s'inscrit dans la tradition méthodologique de l'épistémologie historique, porte sur l'histoire et la formation du concept de biais dans l'épidémiologie moderne. Elle montre que la fonction opératoire du concept de biais est essentiellement critique, au sens où ce concept, que les épidémiologistes opposent au cours de l'histoire aux concepts d'objectivité, de preuve et de causalité, joue un rôle décisif dans la constitution de l'épidémiologie comme science, mais aussi dans l'avènement d'une médecine scientifique. Un éclairage historique et critique est apporté à la définition actuelle du biais, conçu comme une erreur ou un écart systématique par rapport à la vérité, ainsi qu'aux différentes taxinomies des biais qui jalonnent l'histoire de ce concept, dont l'origine se situe chez les fondateurs de la statistique mathématique. Le biais apparait ainsi comme une menace aussi bien à la validité du plan d'expérience d'une étude épidémiologique qu'à la validité de l'inférence statistique et du raisonnement médical. En d'autres termes, ce sont les conséquences que la révolution probabiliste a eues sur l'épidémiologie et sur la médecine qui sont ici étudiées, et qui ont conduit les épidémiologistes et les médecins à une forme de scepticisme et même de criticisme envers leurs propres inférences, ce qui donnera naissance au mouvement de la médecine fondée sur des preuves.**

Mots clefs : **Histoire et épistémologie de la médecine, Histoire et épistémologie de l'épidémiologie, Biais, Erreur aléatoire et erreur systématique, Vérité, Validité, Niveaux de preuve, Causalité, Plan d'expérience, Etude cas-témoins, Etude de cohorte.**

Title : **The concept of bias in epidemiology**

Abstract: **This PhD thesis, belonging to the tradition of historical epistemology, deals with the history and the formation of the concept of bias in epidemiology. It shows that the operational function of the concept of bias is essentially critical, in the sense that this concept, used by epidemiologists throughout history as an antonym to both objectivity, causality and evidence, is central to both the construction of epidemiology as a scientific discipline and the advent of scientific medicine. An historical and critical account is given of the actual definition of bias, conceived as a systematic error or deviation from the truth, and to the various taxonomies of bias which marked the history of this concept, whose origin goes back to the founders of mathematical statistics. Bias thus appears as a threat to the validity of the design of an epidemiological study, and to the validity of statistical inference and medical reasoning. In other words, what is studied here is the consequences of the probabilistic revolution on both epidemiology and medicine, which led epidemiologists and physicians to a kind of scepticism or even criticism about their own inferences, which would ultimately give birth to the evidence-based medicine's movement.**

Keywords : **History and Philosophy of Medicine, History and Philosophy of Epidemiology, Bias, Random Error and Systematic Error, Truth, Validity, Hierarchy of Evidence, Causation, Design of Experiment, Case-control Study, Cohort Study.**

SOMMAIRE

REMERCIEMENTS

Je tiens tout d'abord à remercier mes deux directeurs de thèse : Alain Leplège et Joël Coste. Cette direction bicéphale s'est révélée complémentaire et fructueuse et m'a permis de m'orienter dans cette discipline encore récente qu'est l'histoire et l'épistémologie de l'épidémiologie. Leurs encouragements renouvelés, leurs conseils précieux et leurs critiques toujours pertinentes m'ont permis de surmonter les difficultés que j'ai rencontrées et de mener à bien ce travail. Je leur en suis infiniment reconnaissant. Enfin le fait de m'avoir permis de participer à leurs enseignements, soit à l'Université Paris-Diderot pour M. Leplège, soit à l'Ecole pratique des hautes études pour M. Coste, soit au séminaire de l'ED400 consacré à l'histoire et l'épistémologie de l'épidémiologie m'a permis de circonscrire mais aussi d'enrichir ma réflexion, ainsi que de faire le point sur l'état des connaissances dans le domaine de l'épistémologie de l'épidémiologie.

Je remercie Elodie Giroux d'avoir bien voulu faire partie d'abord de mon comité de thèse puis de mon jury de thèse. Les discussions que nous avons pu par ailleurs avoir, dans le cadre notamment du séminaire Philmed, sur des thématiques connexes à mon sujet ont été riches d'enseignements.

Je remercie aussi Philippe Bizouarn d'avoir bien voulu faire partie de mon comité de thèse. Ses remarques pertinentes ont aiguillé ma réflexion.

Je remercie Laurence Meyer tout d'abord pour m'avoir permis de suivre en 2014 l'Ecole d'été de santé publique et d'épidémiologie du Kremlin-Bicêtre, dont elle est la responsable, et bien sûr pour avoir bien voulu faire partie et présider le jury de ma thèse. Les quinze jours d'enseignement de cette école d'été furent denses et j'y ai appris beaucoup sur l'épidémiologie, discipline qui m'était à la fois familière et étrangère, familière car l'épidémiologie était une discipline que j'enseignais, parmi d'autres, aux étudiants de PACES, étrangère car il me manquait les bases statistiques fondamentales de cette discipline. Je remercie à cette occasion tous les enseignants de cette école d'été, qui savent se montrer très pédagogues pour enseigner une discipline très technique.

Je remercie Maël Lemoine et Isabelle Drouet d'avoir accepté d'être les rapporteurs de cette thèse.

Je remercie les organisateurs du séminaire Philmed organisé à l'IHPST, notamment Hélène Richard, Jean-Baptiste Trabut et Delphine Olivier, ainsi que les nombreux participants à ce séminaire. Ils m'ont permis d'aiguiser ma réflexion et de découvrir le vaste et riche champ disciplinaire qu'est la philosophie de la médecine.

Je remercie le comité organisateur des 6èmes rencontres doctorales internationales en philosophie des sciences organisées en septembre 2017 à Grenoble, ainsi que les participants. Ils m'ont permis, en acceptant ma contribution, de présenter une partie de ma thèse et de confronter mon point de vue avec celui de mes pairs, qu'ils soient étudiants ou chercheurs confirmés. Les discussions que nous avons pu avoir pendant et après ces rencontres ont été les bienvenues. Je remercie d'ailleurs le laboratoire SPHERE, et son directeur Pascal Crozet, pour son soutien financier à cette occasion.

Je remercie Jean Gayon et Pietro Corsi, qui ont dirigé mon mémoire de Master et m'ont inculqué la méthode et la rigueur propres à l'épistémologie historique.

Je remercie Adeline et Nathalie pour leur soutien moral et leur aide précieuse, notamment dans la mise en page du présent travail.

Je remercie mes amis Pierre et Arnaud pour leur soutien et pour tout le reste. Je remercie mes parents, bien sûr, pour avoir toujours été à mes côtés.

Enfin, mes remerciements vont à mon épouse, Nada, qui m'a soutenu dans mon projet de faire une thèse et tout au long de sa réalisation. Sans elle rien de tout cela n'aurait été possible. Je remercie mon fils Nouh, dont le sourire et la joie de vivre m'ont permis de relativiser et constituent une source inépuisable de motivation et de bonheur.

AVERTISSEMENT

Dans le cadre de cette thèse, nous avons travaillé sur des textes (ouvrages ou articles) qui sont majoritairement écrits en langue anglaise. De nombreux extraits de ces textes, quand ils ont été traduits, l'ont été par nos soins, faute de traduction disponible en français. Lorsque c'est le cas, nous avons en général mis en note le texte original, ou bien, quand ce n'est pas le cas, renvoyé aux passages concernés dans le texte original. Pour ne pas alourdir inutilement notre propos et ne pas multiplier les notes, nous n'avons pas précisé à chaque fois qu'il s'agissait de notre traduction.

Ainsi, sauf si le contraire est explicitement indiqué, les extraits d'ouvrages ou d'articles qui sont traduits le sont par nos soins**.**

Nous assumons donc la responsabilité de ces traductions, en espérant ne pas avoir commis d'erreurs et en restant ouvert à la discussion sur certains concepts dont la traduction pose véritablement problème. Comme c'est souvent le cas quand le langage est extrêmement technique, le vocabulaire statistique et épidémiologique recèle ainsi de nombreuses difficultés de traduction, et l'absence de standards internationaux en la matière ne simplifie guère la tâche du traducteur.

« *La philosophie est une réflexion pour qui toute matière étrangère est bonne, et nous dirions volontiers pour qui toute bonne matière doit être étrangère*. »[1]

INTRODUCTION

1. Vérité et erreur en épistémologie

La question de la vérité et de l'erreur constitue sans nul doute une des questions les plus anciennes de la philosophie et de la science, et la plus centrale de l'épistémologie, aussi bien dans son sens anglo-saxon de théorie de la connaissance que dans son sens francophone de philosophie des sciences. Ainsi Aristote définit-il dans *La métaphysique* ce que l'on appelle la conception classique de la vérité et de l'erreur :

> « Dire de ce qui est qu'il est, ou de ce qui n'est pas qu'il n'est pas, c'est dire vrai ; dire de ce qui n'est pas qu'il est ou de ce qui est qu'il n'est pas, c'est dire faux. »[2].

S'il n'est certes pas question ici d'erreur mais bien de vérité et de fausseté, il reste néanmoins que l'erreur consiste bien, selon le *Centre National de Ressources Textuelles et Lexicales*, dans l' « action » ou le « fait de se tromper » c'est-à-dire « de tenir pour vrai ce qui est faux et inversement »[3]. Cette notion d'erreur renvoie ainsi à un des sens courants du verbe « errer » qui signifie « s'écarter, s'éloigner de la vérité »[4]. Ainsi l'erreur consiste pour un ou plusieurs sujet(s) à dire ou à tenir pour vrai quelque chose qui n'est pas vrai objectivement c'est-à-dire indépendamment du ou des sujets qui la tiendrai(en)t pour ou l'énoncerai(en)t comme vrai.

S'il ne s'agit pas ici de retracer l'histoire des conceptions philosophiques de la vérité, il peut être utile de noter que ces conceptions se divisent, par-delà les diverses nuances propres à chaque position épistémologique, en deux catégories en fonction du critère de vérité qui est proposé :

---

[1] Canguilhem, Georges, *Le normal et le pathologique*, 7ème éd., Paris, Quadrige/PUF, 1998, p. 7.
[2] Aristote, *Métaphysique*, III, 7, 1011b25, trad. fr. Jules Tricot, Paris, Librairie J. Vrin, 2000.
[3] http://www.cnrtl.fr/definition/erreur
[4] *Ibid.*

- D'un côté, la conception de la vérité comme correspondance à la réalité, c'est-à-dire de l'adéquation ou de la conformité entre la chose (ou le fait, ou l'objet, etc.) et l'énoncé (ou l'idée, ou la représentation, ou le jugement) ou, comme le dit Aristote :« Ce n'est pas parce que nous disons la vérité en t'appelant blanc que tu l'es, mais c'est parce que tu es blanc qu'en le disant, nous disons la vérité »[5] . Cette conception de la vérité, que l'on peut qualifier de sémantique ou de réaliste, sera clarifiée d'un point de vue logique par Tarski[6] en 1944, notamment à travers la distinction entre métalangage (auquel appartient le prédicat « vrai ») et langage-objet (duquel relève la proposition ou l'énoncé que l'on peut qualifier de « vrai »), afin d'éviter les paradoxes logiques comme celui d'Epiménide le crétois.
- De l'autre côté, la conception de la vérité comme cohérence, c'est-à-dire comme la définit David Hilbert (1862-1943) dans une correspondance avec Gottlob Frege (1848-1925) au début de l'année 1900, dans le cadre d'une interrogation sur les fondements des mathématiques : « Si des axiomes arbitrairement posés ne se contredisent pas l'un l'autre ou bien avec une de ses conséquences, ils sont vrais et les choses ainsi définies existent. Voilà pour moi le critère de la vérité et de l'existence »[7]. Cette conception peut être qualifiée par opposition à la première, avec toutes les précautions d'usage, d'idéaliste ou de syntaxique.

En ce sens, corrélativement à ces deux conceptions de la vérité, il y aurait deux conceptions ou plutôt deux types d'erreur : une erreur de raisonnement ou erreur logique, qui consisterait en une contradiction soit entre des axiomes (ou des prémisses), soit entre des axiomes et leur(s) conséquence(s), et que l'on qualifie en général de sophisme ou de paralogisme ; et une erreur de perception, d'observation ou de jugement sur la réalité, qui consisterait à dire ou à percevoir quelque chose qui ne correspond pas à la réalité.

Cependant, la permanence de cette interrogation sur la vérité ne saurait masquer la profonde méfiance que provoque désormais cette notion non seulement au sein du

[5] Aristote, *Métaphysique,* θ 10, 1051b 6-9, trad. fr. Lukasiewicz (2000), p. 55.
[6] Tarski, Alfred, « The semantic conception of truth and the foundations of semantics » , Philosophy and Phenomenological Research, 4, 1944, p. 341-376.
[7] Rivenc, François et de Rouilhan, Philippe (éds.), *Logique et fondements des mathématiques*, Paris, Payot, 1992, p. 215-235)

grand public, mais aussi chez les scientifiques eux-mêmes[8]. Les philosophes de l'ère du soupçon, Nietzsche en particulier, ont été parmi les premiers à souligner la démonétisation du concept de vérité. Nietzsche dit ainsi :

> « Les vérités sont des illusions dont on a oublié qu'elles le sont (...), des pièces de monnaie qui ont perdu leur empreinte et qui entrent dès lors en considération, non plus comme pièces de monnaie, mais comme métal »[9].

Chez les épistémologues, ces dernières décennies (depuis les années 1960-1970) ont vu triompher certaines formes de relativisme : de *La structure des révolutions scientifiques*[10] de Thomas Kuhn, qui parait en 1962, au « programme fort » de David Bloor[11] et son épigone français Bruno Latour, en passant par les *Leçons sur la volonté de savoir*[12] de Michel Foucault de 1971, tous ces ouvrages ont sérieusement distendu, et parfois rompu, le lien entre la vérité et la réalité en refusant ou en rendant problématique la possibilité d'une correspondance entre la pensée ou le langage d'un côté et la réalité de l'autre. La version radicale de cette position épistémologique, qui est certes beaucoup plus nuancée que cette présentation ne le suggère, est incarnée par l'anarchisme épistémologique de Paul Feyerabend et sa célèbre formule : « Toutes les méthodologies ont leurs limites, et la seule "règle" qui survit, c'est "tout est bon" [« *Anything goes* »] »[13]

Ainsi la notion centrale de « paradigme » chez Kuhn réduit la vérité à un consensus au sein d'une communauté, consensus qui est susceptible de changer à travers une révolution scientifique, mais qui ne rapproche pas les scientifiques de la vérité :

---

[8] Le témoignage de Jacques Bouveresse est intéressant à lire sur ce point : « Je me souviens du premier colloque de rentrée du Collège de France, dont j'avais assuré la direction avec Changeux et dont l'intitulé était « La vérité dans les sciences ». J'eus à cette occasion la possibilité d'entendre certains scientifiques éminents se demander sérieusement si on peut parler réellement de *vérités* dans les sciences, puisqu'il n'est pas du tout certain que, parmi toutes les propositions que la science est amenée à formuler, on puisse en trouver beaucoup dont elle serait prête à affirmer sans réserve ni hésitation qu'elles sont vraies. » in Bouveresse, J., « Tyrannie de la science ou liberté par la science ? », Juin 2015. http://www.opuscules.fr/tyrannie-de-la-science-ou-liberte-par-la-science/

[9] Nietzsche, Friedrich, *Le Livre du philosophe*, trad. fr. A. K. Marietti, Paris, Aubier-Flammarion, 1969 [1873], p. 182-183.

[10] Kuhn, Thomas S, *The Structure of Scientific Revolutions*, 2nd ed., University of Chicago Press, 1970; trad. fr. Laure Meyer, *La structure des révolutions scientifiques*, Paris, Flammarion, 1983

[11] Bloor, David, *Knowledge and social imagery*, 2nd ed, Chicago, University of Chicago Press, 1991; trad. fr. Dominique Ebnöther, *Socio-logie de la loqique ou les limites de l'epistémologie*, Paris, Pandore, 1983

[12] Foucault , Michel, *Leçons sur la volonté de savoir*, suivi de *Le Savoir d'Œdipe*, Cours au Collège de France (1970-1971), édition établie sous la direction de François Ewald et Alessandro Fontana par Daniel Defert, Paris, Gallimard/Seuil, 2011,

[13] Feyerabend, Paul, Against *Method*, Londres, Verso, 1975; trad. fr. Baudouin Jurdant et Agnès Schlumberger, *Contre la méthode*, Paris, Seuil, 1979. La citation est à la page 20

> « Nous devrons peut-être abandonner la notion, explicite ou implicite, selon laquelle les changements de paradigmes amènent les scientifiques, et ceux qui s'instruisent auprès d'eux, de plus en plus près de la vérité ». (Kuhn, 1983, p. 232).

Foucault, quant à lui, voit à l'origine de la connaissance une « falsification » : selon lui, « si la connaissance se donne comme connaissance de la vérité, c'est qu'elle produit la vérité par le jeu d'une falsification première et toujours reconduite qui pose la distinction du vrai et du faux » (Foucault, 2011, p. 4). Latour, enfin, fait de la notion de « véridiction » le « fil rouge »[14] de son œuvre. Selon lui, la connaissance scientifique n'a pas de privilèges sur les autres modes de connaissance et « réduire la connaissance à sa seule dimension scientifique, c'est rendre mensongers tous les autres modes de connaissance : le droit, la politique, l'art, la religion ou… le journalisme ». Il ajoute : « Il nous faut apprendre à respecter les autres modes de connaissance, qui nous sont tout aussi utiles, mais que nous avons eu tendance à minorer au profit d'une idée de la science » (Fossier et Gardella, 2006, p. 113-129). Ainsi, comme le montre Bouveresse :

> « dans une conception comme celle de Latour, la notion de vérité ne comporte pas de dimension ontologique. Il n'y a rien qui soit vrai indépendamment de ce que nous disons. Il y a seulement des façons différentes de dire-vrai, de « véridiction » (comme on les appelle), qui ont chacune leur mode d'opération et leur légitimité ; et il y a autant de modes de connaissance différents qu'il y a de modes de véridiction »[15].

En ce sens, pour certains de ces auteurs, la vérité, la connaissance, la raison ne seraient que des mythes commodes, des effets de discours ou de pouvoir, des notions qui relèveraient de l'arbitraire d'une culture ou d'une époque, quand ils ne nous conduiraient pas tout droit à l'horreur totalitaire[16].

Pourtant, il ne s'agit pas ici de défendre une vision naïve de la vérité ou du progrès des sciences qui consisterait à considérer ce progrès comme une simple accumulation

---

[14] Fossier, Arnaud et Gardella, Édouard, « Entretien avec Bruno Latour », *Tracés*, février 2006, p. 113-129.

[15] Bouveresse, J., « Tyrannie de la science ou liberté par la science ? », Juin 2015. http://www.opuscules.fr/tyrannie-de-la-science-ou-liberte-par-la-science/

[16] « La raison est totalitaire », disaient Theodor W. Adorno et Max Horkheimer en 1947 dans *La Dialectique de la Raison* (le mot raison traduisant ici le mot allemand « Aufklärung », c'est-à-dire les Lumières). Voir Adorno , Theodor W. et Horkheimer, Max, *Dialektik der Aufklärung: philosophische Fragmente*, Amsterdam, 1947 ; trad. fr. Éliane Kaufholz-Messmer, *La dialectique de la raison: fragments philosophiques*, Paris, Gallimard, 2013.

qui nous rapprocherait, par correction successive des diverses erreurs, d'une vérité éternelle ; tout comme il faudrait être aveugle pour ne pas voir les profondes révolutions qui ont émaillé l'histoire des sciences. Il est aussi parfaitement évident, quoique paradoxal, que le progrès considérable qui a lieu dans toutes les sciences au cours des deux derniers siècles est inversement proportionnel au degré de confiance qu'on accorde à ce progrès : le progrès scientifique est ainsi en quelque sorte victime de son succès, car si telle ou telle découverte ou nouvelle théorie invalide les anciennes théories, alors la théorie en vigueur est elle-même susceptible d'être invalidée tôt ou tard. De même, l'impossibilité qu'il y a parfois de trancher entre deux hypothèses contradictoires, comme par exemple la dualité onde-corpuscule en mécanique quantique, incite à la plus grande prudence d'un point de vue épistémologique, quand elle ne conduit pas au scepticisme. Enfin, il ne s'agit pas non plus de nier la pertinence des explications que l'on qualifie généralement d'« externalistes » de l'activité scientifique, qui se concentrent non sur le contenu des connaissances mais sur les hommes qui élaborent ces contenus, hommes dont l'étude psychologique et sociologique permettrait justement d'expliquer ledit contenu. Il est fort possible que les deux perspectives doivent être réconciliées, ou au moins tenues ensemble pour comprendre la marche même de la science.

Surtout, et peut-être faut-il voir dans ce fait un début d'explication à cet esprit relativiste, il apparait que c'est la notion même de vérité qui a effectivement été relativisée dans la pensée scientifique en général, au sens où précisément son corrélat, à savoir l'erreur, a pénétré au cœur même de la vérité : ce changement est dû essentiellement à l'avènement de la pensée probabiliste dans les sciences, qui a profondément changé aussi bien le visage de la science que celui de la réalité. Pascal Engel dit ainsi que la révolution probabiliste (1800-1930) montre que « les phénomènes naturels pourraient ne pas être parfaitement prévisibles et que le hasard pourrait s'insérer dans les choses », ce qui conduit à « une nouvelle conception de la connaissance objective, qui autorise l'erreur jusque dans la mesure des phénomènes »[17]. Ce terme de « révolution probabiliste » renvoie, comme l'indique en note P. Engel, à un ouvrage[18], désormais classique, paru dans les années 1980 qui est

[17] Engel, Pascal, « La tragi-comédie des erreurs », in Rousseau, Dominique et Morvan, Michel, *L'erreur.*, Paris, Odile Jacob, 2000., p.13-16.

[18] Krüger, Lorenz, Daston, Lorraine et Heidelberger, Michael (eds), *The Probabilistic Revolution*, Volume 1: Ideas in History; Volume 2: Ideas in the Sciences, Cambridge, Mass., MIT Press, 1987.

le produit du travail d'un groupe de chercheurs issus de disciplines différentes, qui ont travaillé ensemble en 1982-1983 à l'université de Bielefeld en Allemagne, afin de déterminer s'il y a eu une révolution probabiliste dans les sciences en général, ou en d'autres termes si le paradigme probabiliste est devenu le paradigme dominant, pas simplement d'ailleurs dans une discipline particulière mais dans les sciences en général. De façon fort intéressante pour notre propos, Alain Desrosières dit ceci à propos des travaux de ce groupe :

> « Cette entreprise, initiée par un physicien et philosophe, Lorenz Krüger, était inscrite dans le prolongement des questions soulevées par T. Kuhn sur les « révolutions scientifiques » et leurs « changements de paradigme ». Y a-t-il eu, au XIX^{e} siècle, diffusion d'un « paradigme probabiliste » dans les diverses sciences, tant naturelles que sociales ? Le fait de rassembler des spécialistes de ces deux types de sciences est fort original et s'est révélé fructueux, en montrant notamment que les circulations et échanges de schèmes cognitifs ont eu lieu *dans les deux sens*, et non pas seulement des premières vers les secondes, comme il est souvent dit : le cas exemplaire était celui de la *moyenne*, qui est passée de l'astronomie à la « science de l'homme », puis est revenue de celle-ci vers la physique, *via* A. Quetelet, John Herschel et James Clerk Maxwell. L'hypothèse centrale du groupe est celle d'un basculement d'un modèle « déterministe » de la science, caractéristique du XVIII^{e} siècle, pour lequel la probabilité, dite « épistémique », était liée à une insuffisante connaissance des états du monde, vers un modèle « probabiliste », ou l'aléa est consubstantiel à ces états du monde ». Il ajoute que « par la multiplicité des approches historiques portant sur des disciplines très différentes, cette question en apparence théorique a évolué vers une sociologie des sciences, dans laquelle les contenus des énoncés scientifiques circulent aussi naturellement que leurs énonciateurs, ce qui est la meilleure façon d'abolir l'opposition entre internalisme et externalisme. » [19].

Or, au tout début du troisième article de cet ouvrage, I. Bernard Cohen, se demandant s'il y a bien eu une « révolution scientifique » probabiliste souligne ainsi que l'introduction des statistiques et des probabilités a conduit à « une transformation

[19] Desrosières, Alain, « L'histoire de la statistique comme genre : style d'écriture et usages sociaux », *Genèses*, vol. 39 / 1, 2000, p. 121-137.

radicale et fondamentale de la pensée dans les domaines de la médecine et de la santé publique »[20].

2. Vérité et erreur en épidémiologie : le concept de biais

Ainsi, à travers l'étude du concept de biais en épidémiologie, il s'agit d'étudier l'impact de cette révolution probabiliste dans le domaine de la médecine et de la santé publique. En ce sens, ce travail prolonge les multiples études sur l'histoire et l'épistémologie des statistiques qui ont été menées depuis les années 1980 par les chercheurs du « groupe de Bielefeld » mais aussi par d'autres historiens et philosophes des sciences. En effet, les travaux de ces chercheurs se sont en général arrêtés à l'orée du vingtième siècle, avec Francis Galton, George Udny Yule, Karl Pearson et Ronald Aylmer Fisher, considérés comme les fondateurs de la statistique mathématique. Or, l'épidémiologie telle qu'elle se présente au milieu du vingtième siècle est essentiellement le produit des avancées permises par Pearson et Fisher, notamment la notion de « plan d'expérience » telle qu'elle est thématisée par Fisher dans son ouvrage, cette notion étant étroitement liée à celle de biais.

Dès lors, c'est l'histoire de l'application des outils forgés par ces pères fondateurs de la statistique dans le champ des statistiques médicales, des difficultés qu'ils posent en pratique et des solutions mises en œuvre pour les résoudre qu'il convient de faire. Autrement dit, il s'agit d'étudier comment le paradigme probabiliste va modifier la manière dont les épidémiologistes mais aussi les médecins vont pratiquer leur discipline, mais aussi comment ils vont utiliser ce paradigme pour faire de leur discipline une science : il faut souligner que ces disciplines s'y prêtent particulièrement bien car les statistiques sont tout autant un moyen de connaitre et de faires des inférences scientifiques (notamment causales) qu'un moyen d'agir et de prendre des décisions en situation d'incertitude. Or, l'épidémiologie, comme instrument de la santé publique, et la médecine constituent sans nul doute les disciplines où ces deux besoins, connaitre et agir, sont indissociables, mais aussi où l'incertitude est tout à la fois maximale et vitale, soit pour des individus dans le cadre de la clinique, soit pour des populations dans le cadre de la santé publique. En ce sens, le concept de biais, tel qu'il est hérité des

[20] Cohen, I. Bernard, « Scientific Revolutions, Revolution in Science, and a Probabilistic Revolution 1800-1930 », in Krüger, Lorenz, Daston, Lorraine et Heidelberger, Michael (eds), *The Probabilistic Revolution*, Volume 1: Ideas in History, p. 23-44

statistiques, et tel qu'il va être utilisé par les épidémiologistes et les médecins au cours de la deuxième moitié du vingtième siècle, joue un rôle décisif dans cette histoire, en tant précisément qu'il incarne, pour un médecin ou pour un épidémiologiste, le risque de se tromper, et de se tromper de façon systématique.

Or, si l'histoire de la statistique est maintenant relativement bien connue, l'histoire de son impact sur la médecine et la santé publique est encore largement à faire. Les ouvrages qui traitent de ce sujet se sont d'ailleurs essentiellement centrés sur l'histoire et la philosophie de l'essai clinique, à de rares exceptions[21]. Et pourtant, l'épidémiologie est une discipline tout à fait particulière qui constitue un objet d'étude stimulant pour l'épistémologue. En effet, comme le montre Alex Broadbent, dans son ouvrage *Philosophy of Epidemiology*[22], l'épidémiologie se distingue des autres sciences par sa « non-conformité aux images philosophiques classiques de la science », au sens où « ni l'expérimentation ni la théorie n'occupent une place prépondérante en épidémiologie » (Broadbent, 2013, p. 3-4). De plus, en épidémiologie, plus qu'en physique ou en biologie, l'enjeu est de taille (« *stakes are high »* in Broadbent, 2013, p. 3-4) : dans la mesure où en effet l'épidémiologie a pour but de découvrir des relations causales entre des comportements, des substances (aliments, virus, médicaments, etc.) et des maladies, se tromper dans une inférence causale, c'est-à-dire inférer à partir une étude épidémiologique que tel aliment par exemple cause telle maladie alors que ce n'est pas le cas ou inversement que tel aliment ne cause pas une maladie alors que c'est le cas, a des répercussions sanitaires immédiates. L'enjeu de l'épidémiologie est de taille au sens où il en va de la santé mais aussi de la vie même des populations : le problème est donc indissociablement épistémologique et moral.

Or, dans la mesure où une des fonctions de la statistique consiste à diminuer les erreurs, notamment les erreurs d'observation, grâce par exemple à la courbe de Laplace-Gauss, afin de parvenir par exemple à une estimation correcte de la position d'une planète, il existe néanmoins un certain type d'erreurs – les erreurs non-aléatoires – que le calcul ne peut pas estimer ou diminuer, d'où la nécessité d'un autre concept que celui d'erreur. Ce nouveau concept d'erreur, c'est – du moins espérons-nous le montrer – celui de biais, conçu comme une erreur systématique. C'est ce concept qu'il

[21] Les exceptions sont essentiellement Morabia, Alfredo, (éd.) *A History of Epidemiologic Methods and Concepts*, Basel, Birkhäuser Basel, 2004 ; et Leplège, Alain, Bizouarn, Philippe et Coste, Joël, *De Galton à Rothman: les grands textes de l'épidémiologie au XXe siècle*, Paris, Hermann, 2011.

[22] Broadbent, Alex, *Philosophy of epidemiology*, Basingstoke, Hampshire; New York, Palgrave Macmillan, 2013.

s'agit ici d'étudier, non pas d'ailleurs dans le domaine des statistiques mais dans celui de l'épidémiologie. Pourquoi avoir donc choisi précisément ce concept de biais ? Et pourquoi en épidémiologie et pas par exemple en statistiques, ou bien dans les sciences en général ?

Qu'est-ce que l'épidémiologie ?

Pour savoir ce que signifie le concept de biais en épidémiologie, il n'est pas inutile de déterminer au préalable ce qu'est l'épidémiologie. Le mot « épidémiologie » provient du grec *epi* = « au-dessus », « parmi » ; *demos* = « peuple », « district » ; *logos* = « mot », « discours ». Il s'agit donc littéralement d'étudier ce qui tombe sur le peuple, c'est-à-dire essentiellement à l'origine les épidémies. Le *Dictionnaire d'épidémiologie* de Last[23] définit l'épidémiologie comme « l'étude de la distribution et des déterminants des évènements ou des états en relation avec la santé dans des populations spécifiées, et l'application de cette étude au contrôle des problèmes de santé ».

Les historiens de l'épidémiologie[24] considèrent que les premiers travaux d'épidémiologie expérimentale sont ceux de James Lind (1716-1794) : en 1747, il étudie les épidémies de scorbut à bord des longs courriers et découvre qu'un régime comportant de la vitamine C (en l'occurrence des citrons) permet de contrôler ces épidémies. De même, ils distinguent un général deux pères fondateurs de l'épidémiologie : un Français, Pierre Charles Louis qui créa la « méthode numérique » en médecine ; et un anglais, William Farr (1807-1883), lui-même élève de Pierre C. Louis, qui précise, dès 1838, la notion de risque, et montre l'importance des analyses longitudinales (« cohortes ») pour évaluer les risques[25].

[23] Last, John M., et International Epidemiological Association (éds), *A dictionary of epidemiology*, 4ème edition, New York, Oxford University Press, 2001, p. 14.

[24] Voir Morabia, Alfredo, (éd.) *A History of Epidemiologic Methods and Concepts*, Basel, Birkhäuser Basel, 2004. Peut-être faut-il préciser ici que si nous saluons le travail effectué par Morabia, nous sommes loin de partager ses choix historiographiques et notamment son approche positiviste, et parfois triomphaliste, de l'histoire de l'épidémiologie.

[25] Sur cette question de la distinction entre la notion de risque et celle de taux par William Farr, et la redécouverte de cette distinction par l'épidémiologie moderne, voir Vandenbroucke, Jan P., « On the rediscovery of a distinction », *American Journal of Epidemiology*, vol. 121 / 5, 1985, p. 627–628. Voir aussi les différents articles consacrés William Farr dans Morabia, 2004, p. 149-199. Pour une rapide biographie de William Farr en français, voir par exemple Dupâquier, Michel, « William Farr, démographe », *Population (French Edition)*, vol. 39 / 2, mars 1984, p. 339-355.

Enfin, les débuts de l'épidémiologie analytique, où l'on recherche les déterminants des maladies à partir d'observations faites sur des populations, datent du milieu du XIXe siècle[26], époque de la fondation de la *Société anglaise d'épidémiologie* (1850). Trois travaux fondateurs sont habituellement cités : celui du médecin danois Peter Panum, qui étudie en 1846 la dynamique de la rougeole aux îles Féroé, identifie le mode de transmission direct de personne à personne, et fournit une estimation de la durée d'incubation ; celui du médecin obstétricien hongrois Ignace Semmelweis, en 1847, à Vienne, qui découvre que la fièvre puerpérale est une maladie transportée, en l'absence d'hygiène appropriée, par les mains des soignants des femmes pendant leurs accouchements ; et enfin celui du médecin britannique John Snow, qui, en 1854, mène une étude épidémiologique lors de l'épidémie de choléra à Londres, à l'issue de laquelle, grâce à la comparaison des fréquences de la maladie dans des quartiers desservis par des réseaux d'eau différents, il conclut qu'il y a un agent transmissible à l'origine du choléra et que celui-ci est véhiculé par l'eau.

Depuis une cinquantaine d'années, on distingue entre trois types ou trois branches de l'épidémiologie, qui représentent les trois temps ou les trois phases d'une action de santé publique : l'épidémiologie descriptive, l'épidémiologie explicative ou analytique, et l'épidémiologie évaluative. Ainsi, on décrit tout d'abord les problèmes de santé dans la population c'est-à-dire qu'on les fait apparaitre ; puis on cherche à comprendre les causes ou les facteurs de risque de ces problèmes de santé, et à agir sur eux ; et enfin on évalue si l'action de santé a été efficace ou non.

L'épidémiologie descriptive a pour objectif d'étudier la fréquence et la répartition des problèmes de santé dans les populations.

L'épidémiologie analytique (ou explicative) a pour but de mettre en évidence une relation causale entre les facteurs de risque et la maladie. On distingue deux types d'enquêtes explicatives : les enquêtes de cohorte, qui sont en général prospectives, et les enquêtes cas/témoins, en général rétrospectives.

Enfin, l'épidémiologie évaluative vise à évaluer les actions de santé et à comparer les différentes stratégies et leur efficacité avec pour objectif l'aide à la décision et l'amélioration de la qualité des soins. Dans le domaine médical, on distingue

[26] Pour être tout à fait précis, cette recherche des causes ou des déterminants des maladies, par exemple les déterminants géographiques ou climatiques, remonte historiquement au moins à Hippocrate, et notamment au traité *Airs, eaux, lieux*, rédigé probablement dans la deuxième moitié du Ve siècle avant notre ère. Voir Hippocrate, *Airs, eaux, lieux*, trad. fr. Jacques Jouanna, Paris, Les Belles Lettres, 2003.

traditionnellement trois champs d'application de l'évaluation : les structures et organisations sanitaires (dans les hôpitaux par exemple, pour mesurer la qualité des soins) ; les stratégies médicales, qu'elles soient préventives, diagnostiques, ou thérapeutiques ; l'évaluation de l'impact des activités de soins sur l'état de santé d'une population. Dans ce travail, nous nous focaliserons sur l'épidémiologie analytique, car c'est bien dans ce domaine nous semble-t-il que les problèmes épistémologiques (notamment celui de la causalité) et que la question du biais se pose avec le plus d'acuité.

Pourquoi le concept de biais en épidémiologie ?

Pour répondre à cette question il peut paraitre pertinent de citer les mots de Iain Chalmers, qui définissent assez correctement, bien que dans un domaine connexe mais distinct, le but poursuivi dans notre étude :

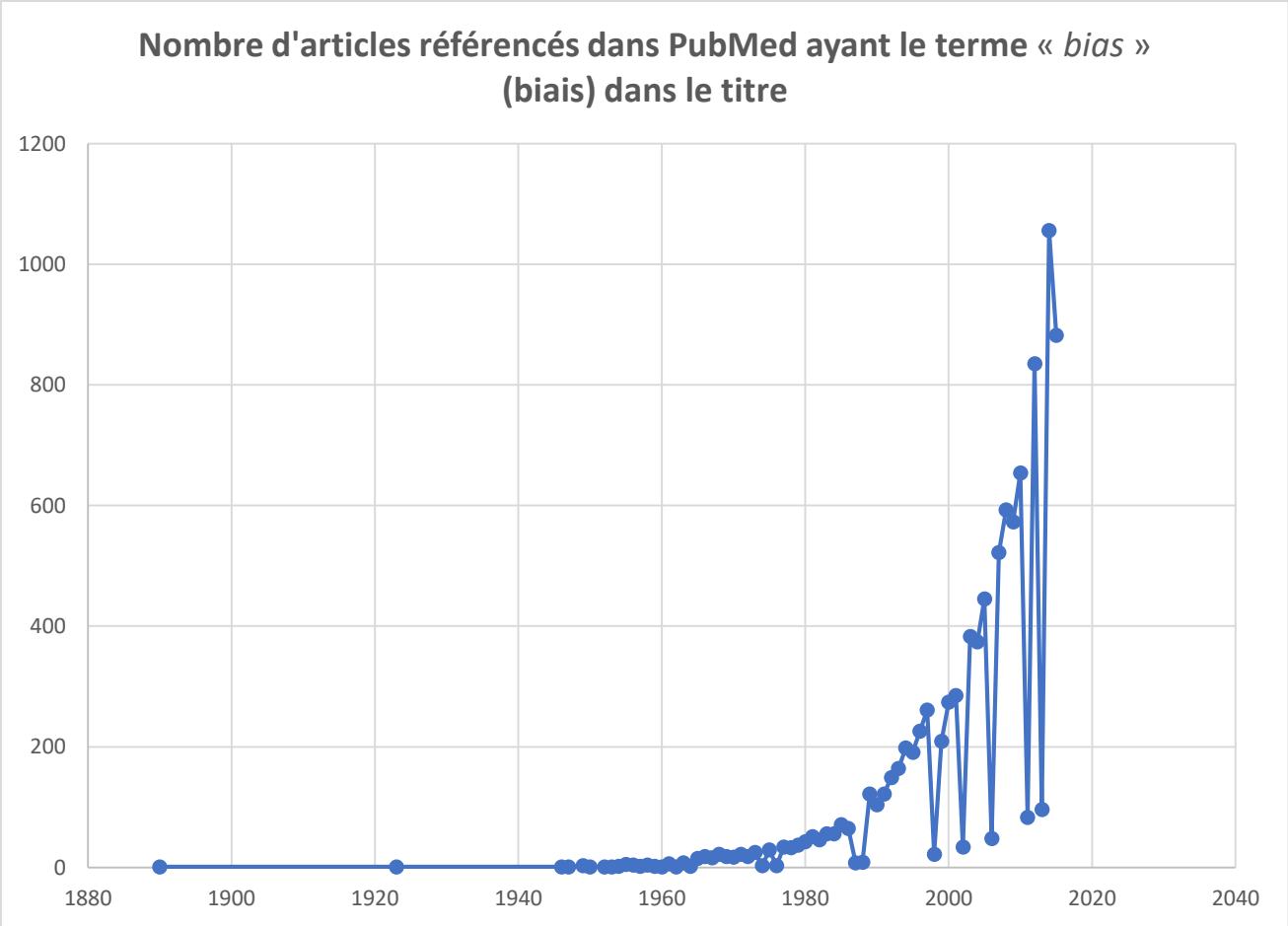
> « Les histoires des essais cliniques ont rapporté et analysé le développement de la quantification dans l'évaluation thérapeutique, l'émergence de la pensée probabiliste, l'application de la théorie et des méthodes statistiques, et de la sociologie, de l'éthique et de la politique des essais cliniques ; mais ce qui est surprenant est qu'elles n'ont que rarement identifié comme un thème à part entière le développement des efforts pour contrôler les biais »[27]

En effet, nous considérons comme Chalmers que cet effort pout « contrôler les biais » est fondamental pour comprendre l'histoire de la discipline épidémiologique, et pas seulement ni essentiellement d'ailleurs l'histoire des essais cliniques, et que cela n'a pas été traité comme un « thème à part entière ». Pourtant, un simple coup d'œil à un dictionnaire d'épidémiologie suffit à montrer l'importance qu'a prise ce concept au sein de l'épidémiologie : John Last, dans la quatrième édition de son *Dictionnaire d'épidémiologie*, distingue ainsi pas moins cinq sources de biais ainsi que vingt-sept biais différents, à chacun desquels il consacre une définition particulière. La définition de ce concept occupe donc une quantité d'espace non négligeable dans ce dictionnaire, tout comme dans l'*Encyclopédie des biostatistiques*[28], qui contient pas moins de 3752

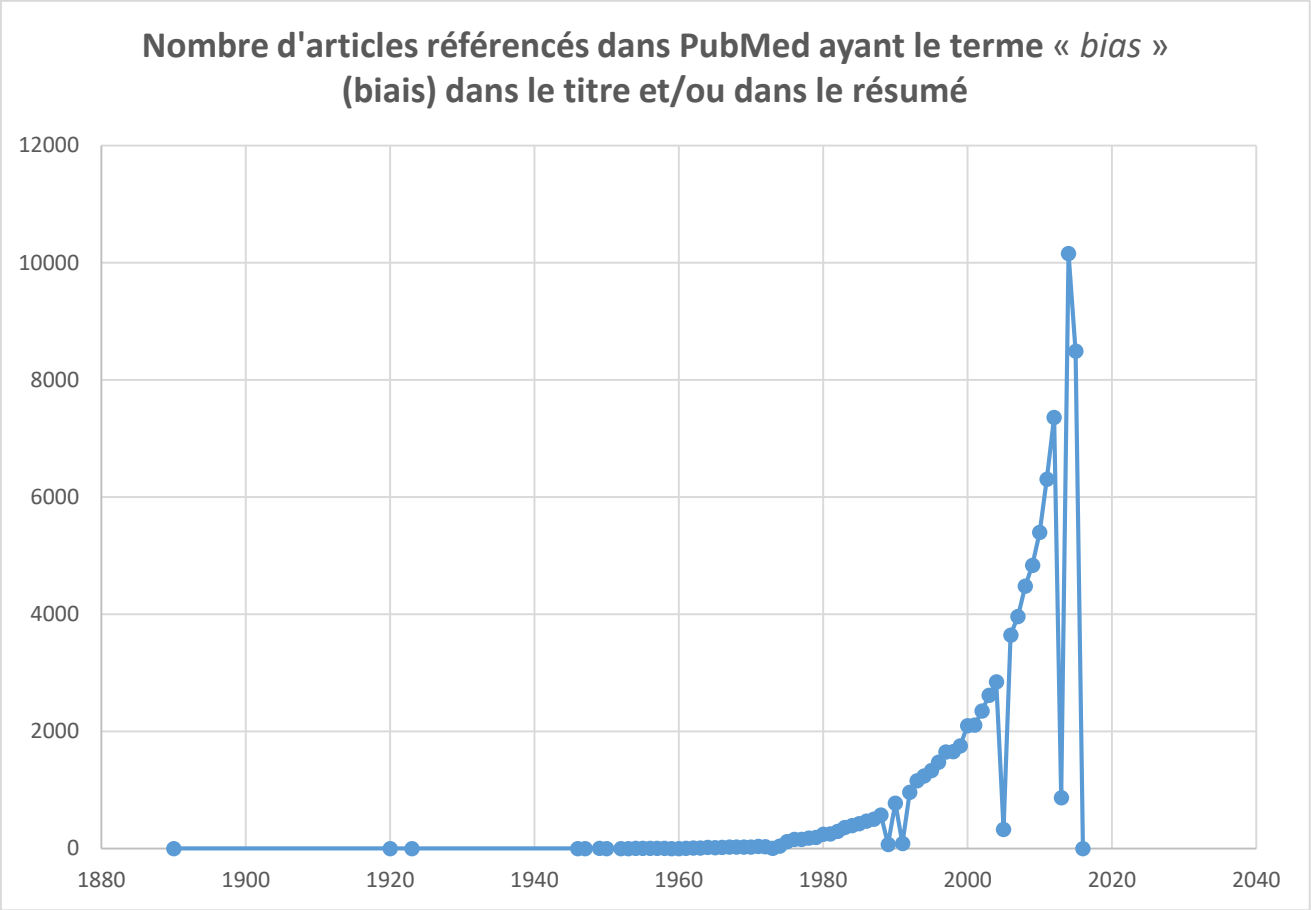
[27] Chalmers I. "Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments". *International Journal of Epidemiology,* 2001, vol. 30, p. 1156-1164.
[28] Armitage, Peter, et Colton, Theodore (éds.), *Encyclopedia of Biostatistics*, Wiley, 2005.

occurrences sur un total de 6267 pages, soit plus d'une occurrence toutes les deux pages.

De même la notion de biais semble aujourd'hui centrale en matière de recherche médicale en général et épidémiologique en particulier. Une simple recherche sur Pubmed[29] visant à recenser le nombre de fois où le mot « biais » apparait dans le titre d'un article suffit, à titre d'indice, à en manifester l'importance : la recherche donne ainsi 12 182 résultats au total et fournit surtout une illustration graphique de l'intérêt croissant pour cette notion. Ainsi, en 2014, 1085 articles comportaient la notion de biais dans leur titre, contre 522 en 2007, 274 en 2000, 104 en 1990, 43 en 1980, 17 en 1970, et seulement 1 en 1960 et 1950. Voici les résultats exprimés graphiquement, avec pour le graphique 1 le nombre d'articles référencés dans *PubMed* ayant le terme « *bias* » dans le titre ; pour le graphique 2 le nombre d'articles référencés dans *PubMed* ayant le terme « *bias* » dans le titre et/ou dans le résumé :



[29] http://www.ncbi.nlm.nih.gov/pubmed?term=bias%5BTitle%5D. Recherche effectuée le 8 juin 2015 à 13H



Qu'est-ce qu'un biais en épidémiologie ?

D'un point de vue épidémiologique et selon la définition du *Dictionnaire d'épidémiologie* de Last, édité pour l'Association Internationale d'Epidémiologie, qui fait autorité dans le sens où cette définition est à la fois descriptive (description d'un état des lieux ou d'un consensus sur le sens d'un mot) et normative (fixation d'un sens pour ceux qui pratiquent la discipline en question) , le biais se définit comme une « déviation des résultats ou des inférences par rapport à la vérité »[30]. Avant de commenter cette définition et pour bien saisir l'apparente l'unité du concept de biais, nous pouvons regarder la suite de la définition, où sont décrites cinq différentes « manières selon lesquelles la déviation par rapport à la vérité peut se produire :

1. « *Variation* systématique (unilatérale) des mesures par rapport aux *vraies* valeurs
2. *Variation* des mesures statistiques résumées (moyennes, taux, mesures d'association, etc.) par rapport à leurs *vraies* valeurs résultant de variations

[30] « *Deviation of results or inferences from the truth », in* Last, John M., et International Epidemiological Association (éds), *A dictionary of epidemiology*, 4ème edition, New York, Oxford University Press, 2001, p. 14.

systématiques dans les mesures, ou autres défauts dans le recueil des données, ou défauts dans le plan de l'étude ou dans l'analyse.

3. *Déviation* par rapport à la *vérité* dans les inférences résultant du plan de l'étude, du recueil des données, ou de l'analyse ou l'interprétation des résultats.
4. Une tendance des procédures (dans le plan de l'étude, le recueil des données, l'analyse, l'interprétation, ou la publication) à produire des résultats ou des conclusions qui s'*écartent* de la *vérité*.
5. Des préjugés qui conduisent à une sélection consciente ou inconsciente des procédures d'étude qui s'*écartent* systématiquement de la *vérité* vers une direction particulière ou à une interprétation unilatérale des résultats. » (Last, 2001, p. 14-15)[31].

Ce qui fait l'unité du concept de biais est donc cet écart, ou cette déviation, ou cette variation par rapport à la vérité ; écart ou variation qui est « systématique ». Dès lors il y a deux façons d'analyser cet énoncé : qu'est-ce qu'un *écart* par rapport à la vérité ? Ou bien qu'est-ce qu'un écart par rapport à la *vérité* ? De même il s'agit de déterminer ce qu'ajoute le prédicat « systématique » à la notion de variation ou d'erreur. Que signifie un « écart systématique » ? En quoi l'adjectif « systématique » vient-il modifier la notion d'erreur ou de variation ? A quoi s'oppose une erreur ou une variation systématique ?

En d'autres termes, il s'agit de déterminer tout d'abord ce que signifie cette notion d'écart, et si les différents sens qui sont donnés à cette notion sont identiques ou bien cohérents, si en d'autres termes l'unité du concept n'est pas illusoire : en effet, le problème est qu'une erreur de mesure (1. et 2.) n'est pas du même ordre qu'une erreur de raisonnement (3.), qui n'est pas non plus du même ordre qu'une erreur de procédure (4.), qui n'est pas non plus du même ordre qu'une erreur de jugement (5.). Si l'on peut noter qu'il s'agit bien dans tous les cas d'une erreur, et que certaines erreurs peuvent

[31] « 1. *Systematic (one-sided) variation of measurements from the true values (syn: systematic error).*
2. Variation of statistical summary measures (means, rates, measures of association, etc.) from their true values as a result of systematic variation of measurements, other flaws in data collection, or flaws in study design or analysis.
3. Deviation of inferences from the truth as a result of flaws in study design, data collection, or the analysis or interpretation of results.
4. A tendency of procedures (in study design, data collection, analysis, interpretation, review, or publication) to yield results or conclusions that depart from the truth.
5. Prejudice leading to the conscious or unconscious selection of study procedures that depart from the truth in a particular direction or to one-sidedness in the interpretation of results. », in Last, 2001, p. 14-15. Nous soulignons.

en entrainer d'autres, l'épistémologue est en droit de s'interroger sur l'unité d'un concept qui englobe aussi bien la notion statistique d'erreur systématique (par opposition à l'erreur aléatoire), la notion logique d'erreur de raisonnement (ce que le logicien appelle un sophisme ou un paralogisme), ou encore la notion psychologique de préjugé. Ainsi, pour citer Canguilhem, « un même mot n'est pas un même concept »[32]. Il ajoute : « il faut reconstituer la synthèse dans laquelle le concept se trouve inséré, c'est-à-dire à la fois le contexte conceptuel et l'intention directrice des expériences ou observations ». C'est précisément cette reconstitution qu'il convient d'effectuer dans cette thèse. En ce sens, il s'agit, pour reprendre le mot de Bachelard, d'écrire l'histoire de la vérité en épidémiologie, c'est-à-dire de décrire la manière dont l'épidémiologie s'est constituée comme une science, c'est-à-dire comme une discipline capable et fondée à produire des connaissances qui soient vraies.

Ainsi conformément à la tâche qu'assignait Canguilhem à l'histoire des sciences, il s'agit d'écrire « une histoire de la formation, de la déformation et de la rectification des concepts scientifiques »[33], et en l'espèce du concept de biais, car il s'agit de démontrer que ce concept est bien un concept scientifique, en étudiant sa formation, sa déformation et sa rectification au sein de l'épidémiologie et plus largement de la médecine.

3. Epistémologie du concept de biais

Notre problématique est donc à la fois épistémologique et historique. En effet si l'on définit avec Canguilhem, un concept comme « une dénomination et une définition, autrement dit un nom chargé d'un sens, capable de remplir une fonction de discrimination dans l'interprétation de certaines observations ou expériences »[34], ou autrement dit si on définit la « valeur cognitive » ou l' « efficacité théorique » d'un

[32] Canguilhem, Georges. « L'histoire des sciences dans l'œuvre épistémologique de Gaston Bachelard », Dans : *Études d'histoire et de philosophie des sciences*. Problèmes & Controverses, 3ème édition, Paris, Vrin, 1975, p. 173-186.

[33] Canguilhem, Georges. « Le concept de réflexe au XIXe siècle », Dans : *Études d'histoire et de philosophie des sciences*. Problèmes & Controverses. 1ère éd. 1968, 3ème édition Paris, Vrin, 1975, p. 295-304.

[34] Canguilhem, Georges. « La constitution de la physiologie comme science », Dans : *Études d'histoire et de philosophie des sciences*. Problèmes & Controverses, 3ème édition, Paris, Vrin, 1975 p. 226-273. La citation est à la page 235

concept par sa « fonction d'opérateur »[35], on peut alors se demander quelle est la valeur cognitive du concept de biais.

Autrement dit, quelle fonction de discrimination remplit-il au sein de l'épidémiologie ? Quelle(s) opération(s) permet-il de faire qu'un autre concept, comme celui d'erreur par exemple, ne permet pas ? En d'autres termes, pourquoi les épidémiologistes, mais aussi les médecins, ont eu besoin de ce concept dans la pratique de leur discipline et l'ont formé, déformé et rectifié au cours de l'histoire de cette discipline ?

Pour répondre à ces questions, il est nécessaire de considérer l'histoire de l'épidémiologie durant la deuxième moitié du vingtième siècle, afin de déterminer qui, quand, comment et pourquoi la notion de biais a pu devenir un concept scientifique à part entière et un concept central de la théorie épidémiologique. Pour se faire, il faut étudier et comparer les différentes définitions qui en ont été données et établir s'il y a ou non une unité et une continuité de ces définitions, ce qui se maintient et ce qui se défait, et si finalement les différents acteurs de cette histoire parlent bien de la même chose quand ils parlent de biais. En effet, « un concept, parce qu'il renferme une norme opératoire ou judicatoire, ne peut varier dans son extension sans rectification de sa compréhension »[36] : c'est ce lien dialectique entre compréhension et extension, entre sens et dénotation qu'il convient d'analyser en premier lieu, et cela ne peut se faire qu'en retraçant l'histoire du concept de biais, à travers non seulement les articles, mais aussi les manuels d'épidémiologie, en identifiant les principaux acteurs de cette histoire mais aussi les évènements qui l'ont jalonnée.

Intension du concept de biais

Or, l'étude des textes considérés comme fondateurs de l'épidémiologie moderne, qui apparaissent dans les années 1940-1950 , montre que le concept de biais, quand il est employé n'est en général pas défini, ou bien de façon très vague (« quelque chose », dit laconiquement Mainland en 1953[37]), ou bien encore que le sens statistique du mot

[35] Canguilhem, Georges. « Le concept et la vie », in *Études d'histoire et de philosophie des sciences*. Problèmes & Controverses, 3ème édition, Paris, Vrin, 1975, p. 335-364.
[36] Canguilhem, Georges. « L'objet de l'histoire des sciences », Dans : *Études d'histoire et de philosophie des sciences*. Problèmes & Controverses, 3ème edition, Paris, Vrin, 1975. P. 9-23.
[37] Mainland, Donald, « The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease », *American Heart Journal*, vol. 45 / 5, Mai 1953, p. 644-654.

(« distorsion systématique par rapport à un résultat statistique attendu »[38]) coexiste bien souvent avec le sens commun du mot (qui renvoie à la notion de préjugé), voire s'y substitue. De plus la définition donnée par le *Dictionnaire d'épidémiologie* ne correspond pas tout à fait non plus à celle qui sont données, quand elles sont données, dans les textes considérés comme fondateurs. Enfin l'origine même de la définition donnée dans le Dictionnaire d'épidémiologie n'est pas en fait la vraie origine : l'article renvoie en effet à un article de David Sackett de 1979[39]. Or, cette définition du mot « biais » énoncée par Sackett renvoie elle-même à une définition donnée par Edmund A. Murphy en 1976[40], définition (Murphy définit le biais comme un « processus qui à toute étape du raisonnement tend à produire des résultats qui s'écartent systématiquement des vraies valeurs »[41]) qui en réalité ne correspond pas tout à fait aux différents sens du mot biais qui jalonnent l'épidémiologie depuis les années 1950. Surtout elle intervient dans un contexte complètement différent : le projet de Murphy ne porte pas en effet sur l'épidémiologie mais bien sur la médecine[42] car il s'agit pour lui de faire de la médecine, et non de l'épidémiologie, une science : le raisonnement duquel il parle est bien le raisonnement médical, et non le raisonnement épidémiologique. Le contexte d'emploi du concept n'est donc plus le même : ainsi le concept de biais apparait dans le chapitre consacré à la notion de preuve en médecine.

Néanmoins, si le concept de biais, comme le contexte dans lequel il est employé, est différent, il s'inscrit tout de même dans une certaine continuité par rapport au concept épidémiologique de biais : la définition épidémiologique du *Dictionnaire d'épidémiologie* reprend ainsi des éléments de la définition de Murphy, sans pour autant l'y réduire. Mais la question se pose de savoir si le sens épidémiologique du mot « biais », et particulièrement l'inflexion particulière qui est donné à son intension par Murphy s'inscrit dans la continuité du sens statistique, présent depuis le début du siècle et que l'*Oxford English Dictionary* définit comme « une distorsion systématique d'un résultat statistique attendu, en raison d'un facteur non pris en compte dans sa dérivation ; aussi, une

[38] Il est intéressant de noter dans l'immédiat le saut entre le concept statistique de biais compris comme un écart systématique par rapport à un « résultat statistique attendu (*expected*) » et le concept épidémiologique de biais comme écart systématique par rapport à la vérité.

[39] Sackett, David L., « Bias in Analytic Research », *Journal of Chronic Diseases*, vol. 32 / 1-2, Février 1979, p. 51-63.

[40] Murphy, Edmund M., *The Logic of Medicine*, Baltimore, MD, Johns Hopkins University Press, 1976.

[41] Murphy, 1976, p. 239

[42] « *The book (…) is not concerned with epidemiology (…) nor statistics* ». Dans: Murphy, 1976, p. 9

tendance à produire de telles distorsions »[43]. Ce changement de sens mérite ainsi une étude plus approfondie, et modifie en tout cas nécessairement l'extension du concept de biais.

Extension du concept de biais

En second lieu, il s'agit donc de déterminer précisément quelle est son extension ou sa dénotation, c'est-à-dire à quel « objet du monde » il renvoie : cela ne peut se faire qu'en analysant le concept d'un point de vue épistémologique dans la relation étroite qu'il entretient avec le concept de vérité. Plus précisément, il s'agit de montrer que, dans la mesure où le concept de biais est issu des statistiques, qui le conçoivent comme une erreur particulière, ou plutôt comme un genre d'erreur particulier, à savoir l'erreur systématique, ce concept va symboliser les difficultés qu'a posées l'introduction de la méthode statistique en médecine, et donc les difficultés de l'épidémiologie à se constituer comme une discipline à part entière, mais aussi comme une source de connaissances problématique pour les médecins.

4. Enjeux épistémologiques du concept de biais et plan de la thèse

Deux enjeux sont ainsi à distinguer : le premier est interne à l'épidémiologie, le deuxième est externe à l'épidémiologie mais interne à la médecine.

Le premier enjeu réside dans la définition même de l'étude épidémiologique, c'est-à-dire la question de savoir ce qu'est un bon « plan d'expérience » ou un bon « plan d'étude », c'est-à-dire un plan d'expérience qui permette, conformément au projet de Ronald Fisher, d'appliquer un test statistique et donc d'effectuer une inférence sur le monde. Ainsi, la question, qui va occuper les épidémiologistes entre la fin des années 1940 et le milieu des années 1960, porte sur la possibilité de faire une inférence statistique, et donc causale, à partir d'enquêtes épidémiologiques (études cas-témoins ou études de cohorte) qui sont essentiellement observationnelles, et justement non-expérimentales. Cette question de la validité des inférences épidémiologiques est vivement débattue au sein de la communauté des épidémiologistes et des statisticiens durant ces années et ce n'est pas un hasard si c'est autour de la notion de biais que le

[43] *The New Oxford Dictionary of English*, éd. Judy Pearsall, Oxford, Clarendon Press, 1999, p. 169.

débat va se nouer : le biais va ainsi rapidement devenir tout à la fois l'instrument principal pour critiquer la validité d'une étude épidémiologique, mais aussi le moyen pour améliorer la méthodologie des études épidémiologiques. En d'autres termes, la thèse défendue ici est que l'histoire de l'épidémiologie consiste précisément en une histoire de la lutte contre les biais, ou autrement dit que cette lutte contre les biais constitue le moteur de l'histoire de l'épidémiologie.

C'est en effet en mobilisant le concept de biais que les épidémiologistes de l'époque cherchent à supprimer toute trace de subjectivité dans l'étude à travers les procédures de randomisation, puis celle d'aveugle et de double-aveugle, et enfin éventuellement celle de placebo. Ces procédures visent à donner une objectivité à l'épidémiologie, objectivité que l'on pourrait qualifier, pour reprendre les catégories avancées par Lorraine Daston et Peter Galison dans leur ouvrage[44], de « mécanique ». C'est en effet seulement en dépersonnalisant autant que possible la procédure d'enquête qu'une inférence statistique, concernant un lien de causalité, a pu être considérée comme possible : si la randomisation est directement héritée de Fisher, c'est bien aux épidémiologistes, notamment Hill, que l'on doit par exemple l'introduction de l'aveugle et du double aveugle en médecine. En ce sens, la fonction opératoire du concept de biais, qui explique en partie le succès de ce terme au sein de la discipline épidémiologique, consiste dans sa fonction, essentiellement méthodologique, de discrimination entre une bonne et une mauvaise enquête épidémiologique, non pas tant au sens où la bonne enquête serait totalement exempte de biais, mais au sens où elle les minimise, au sens où elle évite autant que possible les erreurs méthodologiques. En d'autres termes, la fonction du concept de biais est essentiellement méthodologique, et si l'on considère comme Broadbent, que l'épidémiologie est essentiellement une méthode, ou un ensemble de méthodes, au sens où « l'expertise d'un épidémiologiste est méthodologique » (Broadbent, 2013, p.5), alors le concept de biais acquiert une importance cruciale dans le devenir de l'épidémiologie comme discipline scientifique à part entière.

C'est ce premier enjeu dont traite la première partie de notre étude : elle a pour thème central la naissance du concept de biais en épidémiologie. Cette naissance est relativement longue – elle s'étend sur plus d'une dizaine d'années (1940-1950) – et

[44] Daston, Lorraine, et Galison, Peter, *Objectivity*, New York, Zone Books, 2010; trad. fr. Hélène Quiniou et Sophie Renaut, *Objectivité*, Dijon, Les Presses du réel, 2012.

difficile – le concept statistique de biais coexistant avec son jumeau : la notion traditionnelle de préjugé, pour des raisons qui ne sont pas simplement sémantiques mais qui tiennent aussi à la logique même de l'étude épidémiologique, ce qui explique en partie sa difficile reconnaissance par les épidémiologistes. Dans cette partie, il est nécessaire de revenir sur sa gestation, encore plus longue et tout aussi confuse, en identifiant les moments de sa conception dans l'histoire du calcul des probabilités : nous avons donc procédé à une sorte d'archéologie du concept de biais, en remontant à ses premières occurrences, notamment dans l'œuvre de Francis Galton et de ses successeurs dans l'histoire des probabilités, Karl Pearson, Ralph Weldon et Ronald Fisher, ceci dans le contexte d'une discussion sur la théorie de Darwin sur l'évolution par sélection naturelle. De même, un détour par l'utilisation de la méthode statistique dans les sciences sociales de l'époque, notamment la psychologie sociale naissante qui apparait avec les sondages d'opinion, permet de mieux cerner le concept de biais et ses enjeux épistémologiques dans la mesure où c'est dans ce contexte épistémologique que la notion moderne de biais est d'abord définie.

Le second enjeu est plutôt quant à lui interne à la médecine, ou plutôt il est interne à la médecine dans la mesure où précisément l'épidémiologie s'installe progressivement comme une discipline à part entière avec ses méthodes et ses concepts au sein même de l'enseignement et de la pratique de la médecine. L'enjeu ici concerne l'impact qu'a l'introduction des statistiques en médecine. En effet, l'épidémiologie vient bouleverser les méthodes de recherche en médecine et par conséquent les modes de raisonnement médicaux. Ainsi, dans cette médecine du XXe siècle qui apparait en crise[45], les médecins que Marks qualifient de « réformateurs » (Marks, 2000) vont progressivement considérer comme des « objets de méfiance (…) les patients, mais aussi les médecins généralistes, les médecins hospitaliers, les infirmières et, surtout, les sociétés pharmaceutiques » (Marks, 2000, p.12) et ainsi imposer de nouvelles méthodologies comme la randomisation, le double-aveugle ou encore le médicament contre placebo.

[45] Edmond Murphy parle par exemple d'une « crise des soins médicaux » (Murphy, 1976, p. 8). Harry Marks souligne la « méfiance sociale » qui s'est installée vis-à-vis de l'industrie pharmaceutique mais aussi des praticiens inexpérimentés (Voir Marks, Harry M., « Confiance et méfiance dans le marché : les statistiques et la recherche clinique (1945-1960) », *Sciences sociales et santé*, vol. 18 / 4, 2000, p. 9-27). C'est aussi à cette époque qu'apparait le mouvement dit de « l'anti-médecine », fondé notamment par Ivan Illich (Voir par exemple à ce sujet une conférence de Michel Foucault faite en 1974 au Brésil : Foucault, Michel, « Crise de la médecine ou crise de l'anti-médecine », in Foucault, Michel, *Dits et écrits,* Paris, Gallimard, coll. Quarto, 1994, Tome III, n° 170, p. 40-58.

En réaction à cette crise, qui est essentiellement liée au caractère non-scientifique de la médecine clinique[46], d'autres médecins, comme Alvan Feintein, Edmond Murphy ou David Sackett, vont alors tenter de fonder scientifiquement la médecine en s'appuyant notamment sur des méthodes quantitatives. Ils vont alors accorder une place centrale au concept de biais et en faire un instrument critique essentiel : il s'agit de mettre en garde les médecins actuels comme futurs contre les erreurs systématiques qui peuvent affecter la littérature scientifique et médicale et les conduire à commettre des erreurs diagnostiques ou thérapeutiques. C'est pourquoi ces médecins vont modifier le concept épidémiologique de biais (entendu comme menace à la validité de l'étude) pour en faire l'antithèse de la vérité. A travers ce concept de biais, les médecins des années 1970 invitent ainsi leurs étudiants et leurs collègues au scepticisme et à une forme de criticisme vis-à-vis de leurs connaissances mais aussi de leurs raisonnements

La deuxième partie de notre enquête est donc consacrée à cet enjeu : alors que l'épidémiologie commence à être considérée, par les épidémiologistes eux-mêmes mais aussi par les médecins, comme une discipline scientifique à part entière, précisément parce que les épidémiologiste ont résolu dans les années 1960 les trois problèmes (le plan de l'étude, la nature de la preuve épidémiologique, et la question de l'inférence causale) auxquels était confrontée leur discipline, les médecins commencent à considérer l'épidémiologie comme un moyen de rendre la médecine plus scientifique. Ils font alors du concept de biais un élement central de cette médecine scientifique qu'ils entendent fonder, sous la figure par exemple de l'épidémiologie clinique, théorisée par Feinstein et Sackett. Sackett, dont la définition du biais en 1979[47] est restée dans l'histoire, en fera de même un élément central de sa « méta-méthodologie »[48] qu'est l'*Evidence-Based Medicine*, qui vise à permettre aux médecins d'évaluer et de critiquer une étude ou un article mais aussi de hiérarchiser les niveaux de preuve en fonction justement de la quantité de biais présents dans l'étude.

[46] Voir par exemple Feinstein, Alvan R., « The basic elements of clinical science », *Journal of Chronic Diseases*, vol. 16 / 11, 1963a, p. 1125–1133; ou aussi Feinstein, Alvan R., « Boolean Algebra and Clinical Taxonomy: Analytic Synthesis of the General Spectrum of a Human Disease », *New England Journal of Medicine*, vol. 269 / 18, 1963b, p. 929-938. Edmond Murphy parle lui d'un « manque d'académisme » (Murphy, 1976, p. 5).

[47] Sackett, David L., « Bias in Analytic Research », *Journal of Chronic Diseases*, vol. 32 / 1-2, février 1979, p. 51-63.

[48] J'emprunte le terme à Anne Fagot-Largeault, Cours donné au Collège de France le 21 novembre 2001, intitulé « Méthodologie de la preuve et évaluation du niveau de preuve »

Cette rectification du concept de biais par les médecins va en retour modifier le concept proprement épidémiologique de biais, au point de rendre ces deux concepts difficiles à distinguer. C'est ce mouvement dialectique du concept de biais entre médecine et épidémiologie qui constitue donc l'objet de notre seconde partie, et qui fait selon nous du biais un concept véritablement scientifique en tant qu'il constitue la principale menace non pas simplement à la validité du plan d'étude ou de la mesure, mais à la validité même de l'inférence.

PARTIE 1 : DE L'IDEE DE BIAIS AU CONCEPT DE BIAIS

Dans cette première partie, il s'agit tout d'abord, après une rapide étude lexicographique, de remonter aux origines du concept de biais, c'est-à-dire de décrire le contexte originel dans lequel il intervient. Si le contexte est celui des statistiques au sens large, la notion de biais va apparaitre à l'occasion de deux expériences différentes : d'abord, l'expérience de Weldon qui, vers 1894, lance 12 dés à 26 306 reprises, expérience qui sera commentée par Pearson (dans son fameux article de 1900 où il introduit le test du χ^2) puis par Fisher ; ensuite, les expérimentations agronomiques menées par Fisher comme statisticien à la station expérimentale de Rothamsted, et qui vont le conduire à définir ce que doit être un bon plan d'expérience. La question à laquelle il faudra alors répondre est de savoir pourquoi cette notion de biais apparait à ce moment précis de l'histoire des probabilités, sur un problème pourtant classique des probabilités (le lancer de dés), puis sur un problème nouveau posé par l'application des méthodes statistiques à l'expérimentation. Il conviendra alors de resituer ces deux expériences dans l'histoire des probabilités, en faisant notamment le lien avec la théorie des erreurs et la courbe de Gauss, appelée aussi courbe des erreurs.

Puis il conviendra de décrire et d'expliquer comment et pourquoi la notion de biais va progressivement investir le champ de l'épidémiologie, les difficultés et les résistances que cela va provoquer, mais aussi de montrer et de justifier le vague autour de sa définition, flou qui va perdurer assez longtemps et qui pourrait être tout aussi bien lié à une volonté de certains épidémiologistes de faire accepter leurs nouvelles méthodologies par le champ médical, qu'à une certaine confusion dans l'esprit même des épidémiologistes sur ce qu'est réellement un biais. En effet, nous verrons que le concept central de biais, en tant qu'il est hérité des statistiques, renvoie essentiellement à la notion de randomisation, randomisation qui a précisément pour fonction, dans le cadre d'une expérience, d'égaliser les facteurs ou les variables qui ne sont pas soumis à l'étude afin de pouvoir effectuer un test statistique de signification. Or, égaliser les facteurs, c'est les rendre équipossibles ou équiprobables, au sens où par exemple la face d'un dé a autant de chances de sortir que les cinq autres : un biais est donc quelque chose qui fait qu'une face (ou plusieurs) a plus de chances de sortir qu'une autre, ce qui fait dire à un anglais que le dé est « *biased* » et à un français qu'il est « pipé ».

Pourtant, à ce concept statistique de biais va venir se greffer la signification commune, qu'on peut qualifier de psychologique, du mot « biais » en anglais, à savoir celle de préjugé, de partialité qui va – c'est en tout cas notre hypothèse – venir modifier, enrichir mais aussi rendre plus problématique la notion statistique de biais, pour former le concept épidémiologique de biais. Il sera alors temps de faire un détour par les sciences sociales, notamment la psychologie, la sociologie et la psychologie sociale où l'application des méthodes statistiques pose des problèmes similaires à ceux rencontrés en épidémiologie, pour des raisons qui tiennent pour l'essentiel à ce qu'il s'agit dans les deux cas d'appliquer les méthodes statistiques, en l'espèce des plans d'expérience, non plus à un dé ou à des blettes mais à des êtres humains ; et surtout au fait qu'il s'agit non pas de faire des expérimentations où l'expérimentateur peut contrôler toutes les variables ou presque, mais des études observationnelles, où cela s'avère beaucoup plus difficile. La naissance des sondages d'opinion, à la fin des années 1930, donne ainsi lieu à une discussion intense sur le concept de biais, vive discussion qui n'aura lieu en épidémiologie que dans les années 1950-1960. A travers cette discussion sur les biais, c'est la méthode des études épidémiologiques, et donc leur valeur épistémique qui est mise à l'épreuve durant les années 1950. Autrement dit, il s'agit de déterminer si l'épidémiologie peut légitimement être considérée ou non comme une science.

CHAPITRE 1 :
ARCHEOLOGIE DU CONCEPT DE BIAIS

En guise d'introduction à cette première partie il peut être utile de déterminer plus précisément les différents sens du mot « biais » en anglais comme en français, ainsi que son étymologie, afin de circonscrire plus précisément le sens et la dénotation de ce mot.

Originellement, d'après le *Centre National de Ressources Textuelles et Lexicales*[1], le mot français[2] « biais » a environ trois sens principaux (nous excluons le sens technique qu'il a en couture):

1) Au sens propre il désigne une « direction, forme, [ou] position oblique ».
2) Au sens figuré et péjoratif, il désigne une « déformation, un travers » (au sens psychologique ou moral du terme, d'après la citation donnée de Gide en exemple : « Je ne sais même plus si je fus bien avisé d'écrire que ces doctrines purent, à leur époque, être utiles à notre pays, tant le biais qu'elles donnent aux esprits peut devenir redoutable. Gide, *Journal,*1934, p. 1209); ou encore un « détour, subterfuge ».
3) Au sens figuré mais non péjoratif, il désigne un « moyen de résoudre un problème », ou alors le « côté, aspect, point de vue sous lequel une chose se présente »

Le *Littré* fait remonter son étymologie au latin « *bifax* »[3] c'est-à-dire celui ayant un double regard, louche, celui qui, littéralement, a deux faces (*bis* : deux, et *facies* : face). Le CNRTL penche quant à lui plutôt pour une étymologie dérivée du latin « *biaxius* » : « qui a deux axes ».

L'*Oxford English Dictionary*[4] fait quant à lui remonter l'étymologie du mot anglais « *bias* » au mot français « biais ». Le sens originel du mot renvoie ainsi, comme en français, à une ligne oblique, par exemple la diagonale d'un carré *(« An oblique or slanting line »*) : le mot intervient donc d'abord dans le contexte de la géométrie (le dictionnaire d'Oxford renvoie à une citation d'Oresme du XIVème siècle, citée aussi dans le Littré).

[1] http://www.cnrtl.fr/definition/biais (Consulté le 15 janvier 2015)
[2] Dans la mesure où le mot anglais « *bias* » est directement emprunté au mot français « biais », il nous a paru utile de faire ce détour par la définition du mot en français.
[3] https://www.littre.org/definition/biais (Consulté le 15 janvier 2015)
[4] *The New Oxford Dictionary of English*, éd. Judy Pearsall, Oxford, Clarendon Press, 1999, p. 169-170

Par extension, le mot va s'appliquer au jeu de boules (sens attesté dès 1570) : il renvoie à la construction ou à la forme de la boule qui induit un mouvement oblique, la ligne oblique qu'elle suit, ou la sorte d'impulsion qui lui est donnée et qui la fait rouler de façon oblique (« *A term at bowls, applied alike to: The construction or form of the bowl imparting an oblique motion, the oblique line in which it runs, and the kind of impetus given to cause it to run obliquely* »)

Par transfert, le mot « *bias* » aurait ensuite pris deux significations différentes : l'une très précoce, l'autre beaucoup plus tardive. Le premier sens dérivé du sens géométrique du mot « biais » réfère, dès le XVIe siècle, à une « inclination, un penchant, une tendance ou à une aptitude » ; ou encore à « une disposition prépondérante » ou à « une propension » ; à une « prédisposition envers » quelque chose ou quelqu'un ; et enfin à une « prédilection » ou à un « préjugé » (« *An inclination, leaning, tendency, bent; a preponderating disposition or propensity; predisposition towards; predilection; prejudice* »)

Le second sens dérivé du sens géométrique original renvoie quant à lui à la notion statistique de « biais » et est, de ce fait, beaucoup plus tardive. Le biais statistique désigne alors « une distorsion systématique d'un résultat statistique attendu en raison d'un facteur non pris en compte dans sa dérivation ; aussi, une tendance à produire de telles distorsions » (« *A systematic distortion of an expected statistical result due to a factor not allowed for in its derivation; also, a tendency to produce such distortion* »)

Le quatrième sens, que le dictionnaire désigne comme obsolète renvoie à une direction ou une manière de faire ou de penser ordinaire ou habituelle, tandis que le cinquième sens, assez proche du quatrième renvoie à une « impulsion » ou une « influence », ou, dans un sens plus rare, au « centre de gravité » (par exemple d'un corps qui chute). Une addition au dictionnaire de janvier 2005 signale qu'aux Etats-Unis, le mot « biais » peut désigner une attaque ou un crime motivés par la haine ou l'intolérance envers un autre groupe social, habituellement fondée sur la race ou la sexualité, et renvoie à la notion de crime de haine *(« U.S. Designating an attack or (violent) crime motivated by hatred or intolerance of another social group, usually on the basis of race or sexuality »*). Ce sens existerait depuis les années 1950.

Le sens essentiel qui nous intéresse est bien évidemment le sens statistique du terme, puisque c'est essentiellement en ce sens-là que les épidémiologistes vont, souvent implicitement, employer le mot « biais » ; essentiellement car il apparait que

le sens psychologique du mot (inclination, tendance, préjugé…) est lui aussi utilisé par les épidémiologistes, les deux sens étant même parfois utilisés dans le même article comme si les deux significations étaient parfaitement identiques et interchangeables. Nous laissons de côté dans l'immédiat la question de savoir si cette confusion est volontaire ou fortuite. Ce qui est intéressant dans la définition du dictionnaire d'Oxford est que la première référence qui est donnée pour attester de l'utilisation du mot « biais » en un sens statistique renvoie à un article sur lequel le dictionnaire ne nous donne que peu d'informations : *London, Edinb. & Dublin Philos. Mag.* 50, 167, qui date de 1900. Or il se trouve que l'article en question est en fait l'article, célèbre dans l'histoire des statistiques[5], où Karl Pearson introduit la notion de χ^2 et qui est intitulé : « *On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling* »[6] (« Sur le critère qui permet de décider si, dans le cas d'un système de variables en corrélation, un ensemble donné de déviations par rapport à la valeur probable est tel qu'il peut être raisonnablement supposé avoir été obtenu par un échantillonnage au hasard ») .

Dans cet article, Pearson va s'appuyer sur une expérience réalisée par Ralph Weldon, son ami et collègue, qui a consisté à lancer 26 306 fois douze dés. Il convient à présent d'expliquer les raisons qui ont conduit Weldon à mener cette expérience, et de la situer dans son contexte, qui est en fait celui d'une tentative de démonstration de la sélection naturelle de Charles Darwin, que Jean Gayon a qualifiée, dans son ouvrage qui retrace l'histoire de l'hypothèse de sélection naturelle[7], de « stratégie de la preuve directe » et qui est celle des biométriciens, c'est-à-dire Walter Franck Weldon et Karl Pearson. Mais auparavant, il faut étudier la signification originale que le mot « biais » prend chez Francis Galton : nous soutenons que Galton, dans son ouvrage sur l'*Hérédité naturelle,* va utiliser le mot « biais » dans un sens qui fait office de moyen terme logique entre son sens trivial de préjugé ou de tendance et son sens scientifique

[5] La revue *Science* considère cet article comme une des vingt ruptures scientifiques majeures du vingtième siècle. Cf. Hacking, I. (1984). « Trial by number ». *Science 84* **5** 69–70.

[6] Pearson, Karl, « On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling », in Kotz, Samuel et Johnson, Norman L. (éds.). *Breakthroughs in Statistics*, New York, NY, Springer New York, 1992, p. 11-28.

[7] Gayon, Jean, *Darwin et l'après Darwin: une histoire de l'hypothèse de sélection naturelle*, Paris, Editions Kimé, 1992. Gayon consacre son Chapitre VIII (p. 204-260) à l'étude du lien entre l'hypothèse de sélection naturelle et la biométrie.

ou statistique de distorsion systématique, et qui va selon nous déterminer le sens qu'il va prendre par la suite au sein de l'épidémiologie.

1.1 La notion de biais chez Francis Galton.

1.1.1 Galton et la « Loi suprême de la Déraison ».

Francis Galton (1822-1911), le plus célèbre des cousins de Charles Darwin, reste dans l'histoire comme étant le fondateur de l'eugénisme[8]. Il ne s'agit pas ici néanmoins de nous intéresser à cet aspect de la pensée de Galton mais plutôt à la place qu'il occupe dans l'histoire de la statistique mathématique et surtout à l'influence considérable qu'il exerça sur l'école biométrique anglaise, notamment sur W.F.R. Weldon et K. Pearson. Plus précisément, il s'agit de montrer que Galton est l'un des premiers, si ce n'est le premier, à donner un sens particulier au concept de « biais » dans un contexte statistique, et en quelque sorte celui qui va permettre à ce sens statistique d'advenir sous la plume de Karl Pearson puis de Ronald Fisher qui fourniront le cadre conceptuel au concept proprement épidémiologique de biais. Galton va en fait essentiellement l'utiliser dans le cadre particulier d'une discussion sur la théorie darwinienne de l'évolution par sélection naturelle, et plus spécifiquement dans le cadre de l'élaboration d'une théorie statistique de l'hérédité, à une époque où en fait les mécanismes mêmes de l'hérédité étaient encore inconnus, les théories de G. Mendel n'ayant pas encore été redécouvertes. Pour faire court, la grande thèse de Galton est la « loi d'hérédité ancestrale »[9], c'est-à-dire l'idée que, comme le formule Jean Gayon, « les enfants ne « tendent » pas à ressembler à leurs parents, mais au type moyen de la race » (Gayon, 1992, p.125). Cette idée d'une « tendance centripète » renvoie à l'idée galtonienne de « réversion », concept que Galton emprunte à Darwin et qu'il remplacera plus tard par celui de « régression », ce dernier concept jouant ultérieurement un rôle important dans l'histoire des statistiques, notamment suite à K. Pearson qui en fera un concept central de son épistémologie. Pour les besoins de notre propos, il est nécessaire de rentrer un tant soit peu dans le détail de la théorie de Galton et plus précisément ses postulats de départ, que nous ne chercherons pas par ailleurs à évaluer

[8] Pour une mise au point synthétique sur l'histoire de l'eugénisme voir: Gayon, Jean, « L'eugénisme, hier et aujourd'hui », in *Médecine/Sciences*, 15, n°6-7 (juin-juillet 1999), I-VI

[9] L'expression est en fait de K. Pearson. Voir Gayon,1992, p. 142

ou à critiquer. Ce que nous avons en vue est le chapitre IX de l'ouvrage de Galton, *Hérédite naturelle*[10], où Galton, s'intéressant à la « Faculté artistique » et à son caractère héréditaire, se pose la question de « l'effet du biais dans le mariage » (« *Effect of Bias in Marriage* », in Galton,1889, p. 162-163). Nous soutenons l'hypothèse que ce chapitre condense, bien que de façon assez obscure, la conception qu'a Galton du biais, et surtout comment cette notion va se retrouver étroitement liée à celle de hasard et de sélection, et finalement au cœur même la théorie eugéniste de Galton. L'analyse éclairante de Jean Gayon, bien qu'effectuée d'un point de vue complètement différent du nôtre, nous sera ici d'une aide précieuse. Mais précisons d'abord quelques points saillants de la pensée galtonienne.

Tout d'abord, Galton est proprement fasciné par ce que nous appelons aujourd'hui la loi normale ou loi de Laplace-Gauss, et que Galton, qui l'a trouvée chez Adolphe Quételet en 1863, appelle « loi des écarts à la moyenne » ou « loi de fréquence de l'erreur ». Selon lui, en effet, cette loi révèle l'ordre derrière le chaos. Cette fascination est attestée aussi bien à travers son fameux quinconce (aussi appelée « planche de Galton »), définie par Galton lui-même comme « une illustration mécanique de la cause de la courbe de fréquence » (Galton, 1889, p. 63. Citée et traduite dans Gayon, 1992, p. 131), que dans cette célèbre citation de Galton :

> « Je connais peu de choses capables d'impressionner l'imagination autant que la merveilleuse forme d'ordre cosmique exprimée dans la « loi de fréquence de l'erreur ». La loi eût été personnifiée et déifiée par les Grecs, s'ils l'avaient connue. Plus grande est la cohue, plus grande est l'apparente anarchie, plus parfait est son empire. C'est la loi suprême de la Déraison. » (Galton, 1889, p. 63. Citée et traduite dans Gayon, 1992, p. 128).

1.1.2 Loi normale, hérédité et sélection naturelle :

Pourtant, comme les historiens des statistiques et du calcul des probabilités l'ont montré, l'utilisation que fait Galton de cette courbe des erreurs est différente de celle qu'en faisaient Laplace et les mathématiciens de l'époque qui se concentraient sur des problèmes liés à l'astronomie et à la géodésie : pour eux, il s'agissait ainsi de minimiser les erreurs de mesure afin de déterminer la meilleure estimation de la position d'un point,

[10] Galton, Francis, *Natural Inheritance*, Londres, MacMillan, 1889.

donc comme le dit Galton de « se débarrasser ou bien de fournir une juste estimation de l'erreur » : alors que pour Galton, « ces erreurs ou ces déviations étaient précisément les choses qu'[il] voulai[t] préserver afin de les connaitre. »[11]. Cette nouvelle façon d'utiliser la loi normale est d'ailleurs liée à un changement de « signification méthodologique » qui a été opéré auparavant par Adolphe Quételet par rapport à la conception de Laplace. Gayon montre ainsi que Quételet se sert de la loi normale comme d'une « instrument pour détecter des ensembles d'objets réels homogènes » (Gayon, 1992, p. 127) : elle devient dès lors un « outil pour identifier des « populations » comme des entités objectives » (Gayon, 1992, p. 127). La loi normale est ainsi successivement passée d'un moyen de diminuer et d'estimer les erreurs d'observation et de mesure à un moyen de déterminer des objets réels, par exemple des populations homogènes, pour enfin devenir un instrument pour connaitre les erreurs c'est-à-dire les variations au sein d'une population biologique. Surtout, Galton va considérer que et l'hérédité et la sélection naturelle vont se conformer à cette fameuse loi des écarts. Ces deux points méritent d'être précisés.

Concernant tout d'abord l'hérédité, Galton va comparer, dans son article de 1871 sur les liens de parenté (« *On blood-relationship* »[12]), la formation d'un enfant à « un échantillonnage aléatoire » et se référer « très explicitement à l'image de l'urne » (Gayon, 1992, p. 124). Pour résumer la pensée de Galton, les deux parents peuvent être représentés comme deux urnes pleines de boules, et l'enfant comme « un échantillon tiré au hasard dans une urne commune rassemblant toutes les boules parentales » (Galton, 1871, p. 400. Citée et traduite dans Gayon,1992, p. 124). Dès lors, l'enfant ne va pas nécessairement ressembler à ses parents, car les parents ont des caractères patents mais aussi des caractères latents, ce qui explique que l'hérédité se manifeste aussi bien comme variabilité (les enfants diffèrent des parents : « la variabilité familiale est un facteur de dispersion ») que comme réversion (les enfants diffèrent des parents mais ressemblent aux grands-parents ou arrières-grands-parents, etc., et finalement au « type » moyen de la « race » : « la réversion est un facteur d'homogénéisation » (Gayon, 1992, p. 135), variabilité et réversion allant ainsi en sens contraire. Dès lors, chaque génération peut être analysée par la loi normale, au sens où la distribution des caractères va effectivement suivre une loi normale : Galton montre

[11] Galton, Francis, *Memories of my Life*, Londres, Methuen and Co, 1908, p. 305.

[12] Galton, Francis, « On blood-relationship », *Proceedings of the Royal Society of London*, vol. 20 / 130-138, 1871, p. 394–402. La métaphore de l'urne se situe à la page 400.

ainsi avec son expérience sur les pois de senteur que la distribution des graines-filles (classées en fonction du critère de poids) suit une loi normale, mais que la moyenne des poids des graines-filles s'écarte de la moyenne parentale pour revenir à celle du type ancestral moyen.

Concernant à présent la sélection naturelle, Galton va montrer que la sélection naturelle se conforme à la « loi des écarts » en utilisant une analogie avec un artilleur qui tire de manière répétée sur un rempart :

> « Ainsi n'est-il pas déraisonnable de considérer la nature comme un tireur dont l'action est sujette à la loi des écarts [« *law of deviation* »], tout comme les impacts disposés sur une cible de part et d'autre du point qui était visé »[13].

La sélection naturelle va ainsi favoriser le type moyen et écarter les extrêmes : comme le dit Galton, « ceux qui dévient beaucoup de la moyenne, par défaut ou par excès » ne laisseront qu'une « toute petite contribution » aux « générations futures » (Galton, 1877, p. 532).

1.1.3 Le biais dans le mariage et la sélection sexuelle.

Venons-en maintenant au problème du biais : dans le chapitre IX de l'ouvrage *Hérédité naturelle*, Galton s'intéresse donc à la faculté artistique afin de déterminer si cette faculté, au même titre que la taille ou la couleur des yeux dont il a traité dans les chapitres précédents, est héritable et suit les lois de l'hérédité dont il cherche à démontrer le fonctionnement. Ce chapitre se conclut par la question de l'effet du biais dans le mariage (« *Effect of Bias in Marriage* »), et plus précisément, comme il le formule dans le chapitre VI, par la question de « l'influence du biais dans la sélection matrimoniale sur la race »[14]. En effet, Galton vient de montrer dans la partie de ce même chapitre consacré à la « sélection matrimoniale » (« *Marriage Selection* ») que les artistes éprouvent une « légère réticence » (« *slight disinclination* » in Galton, 1889, p. 162) à se marier entre eux : alors qu'en cas d' « indifférence parfaite », c'est-à-dire au hasard (Galton utilise, dans le tableau 9b, le terme « *Chance combinations*»[15]), donc selon la loi de probabilité théorique, 42 mariages homogamiques (c'est-à-dire

[13] Galton, Francis, « Typical Laws of Heredity », *Nature*, vol. 15, 1877, p. 492-495, 521-514, 532-533. La citation est à la p. 514. Citée et traduite dans Gayon, 1992, p. 164.
[14] *« the influence on the race of Bias in Marriage Selection* », in Galton, 1889, p. 87.
[15] Galton, 1889, p. 207.

entre artistes) auraient dû avoir lieu, seuls 36 ont été observés. La question qu'il se pose alors dans cette dernière partie est de savoir ce qu'il se passerait s'il y avait un « biais » dans le mariage, c'est-à-dire si les artistes choisissaient de ne se marier qu'entre eux ou si, à l'inverse, les personnes douées de la faculté artistique choisissaient de n'épouser que des personnes qui ne sont pas justement douées de cette faculté. Deux choses sont à noter dans l'argumentation de Galton à ce moment précis : tout d'abord, il est évident que Galton entend le mot « biais » en son sens trivial et psychologique de préférence, de tendance ou d'inclination, comme le recours à son antonyme anglais (« *disinclination* ») l'atteste ; ensuite, Galton ne va pas en fait parler de la faculté artistique, mais bien de la taille pour expliquer ce qu'il entend par l'effet du biais. S'il est bien évidemment plus simple de mesurer ou du moins d'exprimer sous la forme de mesures la taille que la faculté artistique, il est néanmoins possible d'interpréter ce recours au critère de la taille comme une conclusion générale sur ce qu'il vient de dire sur l'hérédité. Que se passe-t-il alors en cas de biais ?

Deux phénomènes opposés vont se passer : tout d'abord si les plus grands se marient avec les plus petits, ou plus précisément si le « degré d'écart » 99 (soit le plus grand degré d'écart par rapport à la moyenne sur une échelle statistique, ou, d'un point de vue géométrique, le point d'une courbe normale qui se situe juste avant l'intersection entre l'axe des abscisses et celui des ordonnées) se marie avec le degré 1 (à l'autre extrémité de l'échelle ou de la courbe), le degré 98 avec le degré 2, et ainsi de suite, alors, logiquement, la taille des inter-parents ou des parents-moyens (« *mid-parents* »[16]) sera identique pour tout le monde : leur erreur probable Q sera donc égale à zéro, tout comme celle du système des co-fraternités-moyennes, et l'erreur probable Q de la génération suivante sera égale à celle de la co-fraternité[17], c'est-à-dire 1,5 pouce. A l'inverse, si les personnes de grande taille se marient exclusivement ou presque avec des personnes de grande taille, alors l'erreur probable dans la génération suivante, c'est-à-dire l'écart ou la déviation par rapport à la moyenne, sera augmentée. Galton conclut alors :

[16] Comme l'explique Gayon, il s'agit d'un artifice statistique utilisé par Galton qui consiste à créer un « progéniteur imaginaire de sexe composite, dont le caractère est mesuré par la moyenne des deux parents, après que le caractère de la mère ait été ramené à l'échelle paternelle (dans le cas de la taille par exemple, on multiplie la taille maternelle par 1.08 pour obtenir la taille paternelle) », in Gayon, 1992, p. 139.

[17] La co-fraternité désigne la progéniture d'une groupe d'inter-parents qui ont la même taille, alors que la fraternité désigne les enfants d'un même inter-parent.

> « Quel que soit le caractère ou la force du biais dans la sélection au sein du mariage, aussi longtemps qu'il reste constant, la Q [l'erreur probable] de la population aura elle aussi tendance à rester constante, et la ressemblance statistique entre les générations successives de la population future sera assurée » (Galton, 1889, p. 163).

Ainsi, le biais au sein du mariage va en fait affecter la moyenne même de la population qui va dévier par rapport à sa moyenne statistique c'est-à-dire par rapport à son « type racial » : la distribution des caractères dans les générations suivantes ne sera donc pas conforme à une distribution normale. En somme ce biais va conduire à l'apparition d'une nouvelle « race ». Galton disait ainsi en 1877, dans son article consacré aux « Lois typiques de l'hérédité » :

> « Pour que la loi de la sélection sexuelle puisse coopérer avec les conditions d'une population typique, il est nécessaire que cette sélection soit *nulle*, autrement dit, qu'il n'y ait pas la moindre tendance pour les hommes grands à se marier préférentiellement à des femmes plus grandes plutôt qu'à des petites » (Galton, 1877, p. 514. Citée et traduite dans Gayon,1992, p. 173)[18].

Clairement, comme le montre Gayon, la sélection sexuelle correspond ici à « ce que nous appelons aujourd'hui « homogamie » ou encore « accouplements assortis », ou autrement dit « le choix d'un partenaire sexuel de même caractère que soi » (Gayon, 1992, p. 173). Or, si intuitivement nous pensons que si le même s'unit au même, leurs enfants seront aussi similaires, Galton « dit exactement le contraire : s'il existe dans une population une tendance à l'appariement d'individus semblables, la « condition typique » ne peut se maintenir, et la moyenne *doit* changer » (Gayon, 1992, p. 173[19]). Cette conception de Galton ne peut en fait se comprendre que dans le cadre de son idéologie eugéniste. Il dit ainsi :

> « Si l'on mariait les hommes de talent à des femmes de talent, *de même caractère physique et moral qu'eux-mêmes*, on pourrait, génération après génération, produire une race humaine supérieure ; cette race n'aurait pas davantage tendance à faire retour aux types ancestraux plus médiocres que ne le font nos races désormais bien établies de chevaux de course ou de chiens de chasse »[20]

[18] C'est Gayon qui souligne.
[19] C'est Gayon qui souligne.
[20] Cité dans Gayon, 1992, p. 174. Nous soulignons.

Il dit d'ailleurs sensiblement la même chose en conclusion de *L'hérédité naturelle* :

> « La valeur d'un bon stock pour le bien-être des générations futures est donc évidente, et il est bon de rappeler le signe précoce par lequel nous pouvons nous assurer qu'une nouvelle variété bien dotée possède la stabilité nécessaire pour produire facilement un nouveau stock. [Ce signe] se trouve dans son refus de se mélanger librement avec d'autres formes » (Galton, 1889, p. 198).

En d'autres termes, le biais au sein du mariage, entendu ici au sens de préférence ou de tendance soit vers le même, soit vers l'exact opposé, modifie l'hérédité au sens où elle modifie les deux effets de l'hérédité, à savoir la variabilité familiale comme facteur de dispersion (qui devient beaucoup trop grande) et la régression comme facteur d'homogénéisation (qui va se faire mais sur une autre moyenne ou un autre « type racial »), ce qui va conduire à l'apparition d'une nouvelle « race » ; ou plus précisément, c'est ce processus de mariage assorti, comme instrument d'eugénisme positif, qui va permettre l'avènement d'une « race » supérieure. Par-delà l'idéologie eugéniste propre à Galton dans laquelle ces propos s'insèrent, il reste que le mot « biais », encore entendu en son sens trivial, commence à acquérir une certaine épaisseur scientifique en tant qu'il intervient dans des controverses scientifiques autour de la question de l'évolution. Ainsi, Edwin Ray Lankester (1847-1929), zoologiste et biologiste évolutionnaire darwinien, étudiant de Thomas Huxley et titulaire de la chaire Jodrell de zoologie à l'*University College* de Londres (et qui aura pour successeur en 1891 un certain W.F.R. Weldon), utilise-t-il en 1896 le terme en un autre sens pour signifier que les variations ne se font pas complètement au hasard :

> « Ainsi, chaque partie d'un animal qui varie, ne varie pas « de façon égale autour d'une norme », mais varie en accord avec la tendance constitutionnelle de l'organisme, que l'on pourrait appeler son biais ancestral ou son biais de groupe »[21]

Dès lors, il semble possible d'avancer que dans les années 1880-1890, le biais, par-delà son sens psychologique et trivial de préjugé ou de tendance, devient, dans le cadre d'un développement de la théorie statistique lié à un débat sur la théorie

[21] Cité dans Bowler, Peter J., *Darwin deleted: imagining a world without Darwin*, Chicago, The University of Chicago Press, 2013, p. 164.

darwinienne de l'évolution par sélection naturelle, une sorte d'antonyme de la notion de hasard (« *random* ») et une sorte de synonyme de la notion de sélection (en l'occurrence chez Galton la sélection sexuelle), sans pour autant qu'aucun des auteurs qui l'utilise ne prenne la peine de le définir, ce qui montre implicitement que le sens auquel ils se réfèrent est le sens commun, alors que précisément il ne l'est plus tout à fait. Il est temps à présent d'étudier plus précisément la première occurrence du sens statistique du mot « biais » qui serait apparu, d'après l'*Oxford English Dictionary*, en 1900 dans un article de Karl Pearson, suite à l'expérience du lancer de dés de W.F.R. Weldon, restée célèbre dans l'histoire de la statistique mathématique, et qui intervient précisément dans le même contexte d'une discussion sur la théorie de Darwin.

1.2 L'expérience du lancer de dés de Weldon :

1.2.1 Le contexte scientifique de l'expérience de Weldon.

Le zoologiste et biologiste évolutionnaire, Walter Frank Raphael Weldon (1860-1906) est un ami et collègue de Karl Pearson (1857-1936) à l'*University College* de Londres, mais aussi un disciple, comme Pearson, de Francis Galton, les trois étant les fondateurs de la revue *Biometrika* en 1901. Il est connu pour avoir appliqué la notion de corrélation, héritée de Galton, à la biologie. En effet, après avoir lu l'ouvrage de Galton, *Natural Inheritance*, publié en 1889, Weldon aurait, d'après K. Pearson[22], pris conscience qu'il fallait « établir les preuves » de cette « hypothèse de travail » qu'est la théorie darwinienne. Il va alors appliquer la méthode anthropométrique héritée de Galton aux espèces sauvages, et par là passer de l'anthropométrie à la biométrie : pour Weldon en effet, selon la formule célèbre dans l'histoire de la biologie et des statistiques, « le problème de l'évolution animale est essentiellement un problème statistique » (Cité dans Gayon,1992, p. 207). Dès lors, il va comparer la largeur de la carapace de 400 crevettes dans un premier article publié en 1889[23], puis, dans un second article de 1892[24], étudier la corrélation entre la taille de quatre organes, toujours chez les crevettes, et établir un « degré de corrélation » entre deux organes chez le même

[22] Pearson, Karl, « Walter Frank Raphael Weldon, 1860-1906 », *Biometrika*, **5**, p. 1-52, 1906.
[23] Weldon, W. F. R., « The Variations Occurring in Certain Decapod Crustacea.-- I. *Crangon vulgaris* », *Proceedings of the Royal Society of London*, vol. 47 / 286-291, janvier 1889, p. 445-453.
[24] Weldon, W. F. R., « Certain Correlated Variations in *Crangon vulgaris* », *Proceedings of the Royal Society of London*, vol. 51 / 308-314, janvier 1892, p. 1-21.

individu (il établira d'autres « variations corrélées » chez le « crabe enragé » dans un article de 1893[25]). Son expérience de lancer de dés intervient ainsi dans ce contexte d'une tentative de démonstration empirique de la sélection naturelle.

Or, le problème qu'il rencontre en étudiant les populations de crabes (il étudie une population de la baie de Naples et une autre issue de la baie de Plymouth) est qu'en étudiant le coefficient de corrélation pour vingt-trois paires de mesures sur différents caractères du crabe, un caractère (la largeur frontale) se comporte de manière atypique dans la population napolitaine, au sens où la distribution de ce caractère est clairement non gaussienne et asymétrique (c'est une courbe à double bosse[26]). La question centrale qui occupe donc Weldon, et qui va aussi occuper Pearson et leurs collègues, est donc la suivante, comme le résumera Egon Pearson en 1965 :

> «Est-ce que la loi normale s'ajuste à cette distribution et si ce n'est pas le cas, qu'est-ce que cela signifie ? » [27]

En effet, comme le dit Gayon, « ce genre de distribution n'était pas analysable avec les moyens de la statistique galtonienne » (Gayon, 1992, p. 213), Galton étant proprement fasciné par ce qu'il appelle la « loi des écarts à la moyenne ». C'est pourquoi Weldon va se tourner vers le mathématicien Karl Pearson, ce qui va marquer le début d'une collaboration mais aussi d'une amitié entre les deux hommes. La résolution du problème par Pearson apparait dans ses « *Contributions to the Mathematical Theory of Evolution* »[28], consacrées à la « dissection des courbes de fréquence asymétriques » et où il va introduire la notion de « moment » pour justement procéder à cette dissection. Comme le dit Gayon :

> « la question première qui se pose à un statisticien lorsqu'il est confronté à une distribution non-gaussienne est de savoir s'il a affaire à une population dans laquelle les mesures sont homogènes, c'est-à-dire affectées par le même genre de causes » (Gayon, 1992, p. 214).

---

[25] Weldon, Walter Frank Raphael, « On certain correlated variations in *Carcinus maenas* », *Proceedings of the Royal Society of London*, vol. 54 / 326-330, 1893, p. 318–329

[26] Pour un dessin de cette courbe, voir Magnello, M. Eileen, « Karl Pearson and the Establishment of Mathematical Statistics », *International Statistical Review*, vol. 77 / 1, avril 2009, p. 3-29. L'image est à la page 14.

[27] « *The question 'does a Normal curve fit this distribution and what does this mean if it does not?' was clearly prominent in their discussions*. », in Pearson, E. S., « Studies in the History of Probability and Statistics. XIV Some Incidents in the Early History of Biometry and Statistics, 1890-94 », *Biometrika*, vol. 52 / 1/2, juin 1965, p. 3-18

[28] Pearson, K., « Contributions to the Mathematical Theory of Evolution », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 185 / 0, janvier 1894, p. 71-110

Or, c'est précisément le problème que rencontre Weldon : il avoue même à Karl Pearson, dans une lettre datée du 23 avril 1893, qu'il a « eu un choc » (cité dans Pearson, E.S.,1965, p. 9). En effet, Weldon était convaincu que « l'apparente symétrie de la variation chez les animaux montrait que tout « accident » survenait à peu près aussi souvent que n'importe quel autre, et qu'il n'y avait pas, chez tous les animaux [qu'il avait] rencontrés, de « tendance », comme les biologistes le disent, « à varier dans une direction plutôt qu'une autre » (cité dans Pearson, E.S.,1965, p. 9). Weldon va alors explorer le problème « d'une manière empirique, à la fois en jetant des dés et en calculant les termes d'un nombre de distributions binomiales, $N (q+p)^n$ avec $p \neq q$. » (cité dans Pearson, E.S.,1965, p. 9). Weldon va alors calculer l'expansion $(0,6+0,4)^{20}$ puis $(0,7+0,3)^{20}$, ce qui va effectivement provoquer un choc chez lui :

> « J'espérais que si un organe varie dans une direction particulière – c'est-à-dire si p est plus grande que q – l'asymétrie de la courbe donnerait une sorte de mesure de la différence entre les deux ; et qu'une sorte de cinétique de la variation [« *kinetic of variation* »] pourrait être construite. Mais si p est deux fois plus grande que q, avec les résultats abominables ci-joints, ce maigre espoir tombe en lambeaux. » (Cité dans Pearson, E.S.,1965, p. 9).

Parallèlement à ses calculs d'équation, Weldon lance aussi des dés : plus précisément, en bon empiriste qui entend disposer de données quantitatives solides, il réalise une expérience qui consiste à lancer 26 306 fois 12 dés afin de « juger si les différences entre une série de fréquences de groupe et une loi théorique, prise comme un tout, étaient ou n'étaient pas supérieures à ce qui peut être attribué aux fluctuations hasardeuses d'un échantillon aléatoire » (« *to judge whether the differences between a series of group frequencies and a theoretical law, taken as a whole, were or were not more than might be attributed to the chance fluctuations of random sampling.* »[29]), car, à l'époque, il n'existe pas de test, comme celui du χ^2, pour répondre simplement et rapidement à cette question. Dès lors Weldon se retrouve face à trois explications possibles, explicitées ainsi par E. S. Pearson :

> « - La divergence entre la théorie et l'observation n'est pas supérieure à celle que l'on pourrait attendre d'un échantillonnage aléatoire.
>
> - Les données sont hétérogènes, et composées de deux ou de plusieurs distributions normales.

[29]Cité dans Pearson, E.S.,1965, p. 10-11.

> - Les données sont homogènes, mais il y a une réelle asymétrie dans la distribution des variables mesurées. » (Pearson, E.S.,1965, p. 9).

La troisième option est sans doute « la plus difficile à accepter, étant donné le prestige qui entourait à l'époque la loi normale » (Pearson, E.S.,1965, p. 9). Or, une des avancées essentielles de K. Pearson dans ses « *Contributions to the Mathematical Theory of Evolution* », est justement de présenter une « méthode pour analyser des courbes de fréquence « anormales, c'est-à-dire non-gaussiennes » (Gayon,1992, p. 214). C'est ainsi qu'en disséquant la courbe asymétrique en deux courbes normales, Pearson montre qu'il faut retenir la deuxième option, c'est-à-dire que les données sont hétérogènes et composées de deux courbes normales, ou en d'autres termes, qu'il y a bien deux populations homogènes (deux « races » de crabes) distinctes dans la baie de Naples. Comme le montre Eileen Magnello :

> « après que Pearson eut examiné les courbes asymétriques de Weldon dérivées de ses données sur les crabes de Naples, il réalisa qu'une méthode objective pour mesurer la qualité de l'ajustement [« *goodness-of-fit* »] manquait encore pour les distributions qui ne se conformaient pas à la loi normale » (Magnello, 2009, p. 18).

Ce test est celui du χ^2, que Pearson expose en 1900, et qui est précisément fait pour tester la qualité d'ajustement des courbes dissymétriques que l'on retrouve souvent en biologie et en économie (Magnello, 2009, p. 18). Autrement dit, c'est un test que les statisticiens d'aujourd'hui qualifieraient de non-paramétrique, au sens où il ne fait aucune hypothèse sur la loi de probabilité sous-jacente à la distribution des données, ce qui permet d'élargir considérablement le champ des méthodes statistiques, puisque les fréquences peuvent être connues ou supposées connues *a priori*, comme dans le cas d'un lancer de dés, ou au contraire inconnues. Si le texte de Pearson que nous allons maintenant étudier a été maintes fois analysé par les statisticiens et les historiens de la statistique, il n'intéresse néanmoins notre propos que dans la mesure où le mot « biais » apparait ou réapparait après avoir été utilisé par Galton en 1889. Pourquoi ce terme a-t-il disparu pendant une dizaine d'années ?

Il parait bien difficile de répondre à cette question, dont la réponse se situe de toute façon en dehors de notre sujet. Néanmoins, il peut être intéressant de noter qu'un autre terme très proche va constituer un des objets centraux de l'étude de Pearson, au

moins de 1895[30] à 1916[31] : celui de « *skewness* », ou sous sa forme adjectivale « *skew* ». Ainsi Pearson parle-t-il de « *skew variation* », de « *skew curve* » ou encore de « *skew correlation* ». Or, en anglais, si son sens statistique renvoie à la notion d'asymétrie (d'une courbe), « *skew* », au sens commun du mot, renvoie au caractère oblique d'un angle, ou encore à un parti-pris, un préjugé, ou bien à une distorsion et est considéré par l'*Oxford English Dictionary*[32] comme un synonyme du mot « *bias* ». En ce sens, on peut émettre l'hypothèse que si le mot « biais » n'apparait plus dans la littérature statistique de l'époque, c'est parce qu'un mot similaire occupe en quelque sorte déjà l'espace conceptuel de cette notion. Mais ceci mériterait une étude à part entière qui serait hors de propos ici, et il est temps d'aborder le texte de Karl Pearson consacré au test du χ^2.

1.2.2 L'analyse de l'expérience de Weldon par Pearson.

L'objectif de l'article de Pearson, qui parait en 1900, soit six ans après l'expérience du lancer de dés, qui n'apparait d'ailleurs dans l'article qu'à titre d'illustration[33], est d'exposer sa découverte du critère qui permet de déterminer si les déviations des valeurs observées d'une ou de plusieurs variables (ou d'un ensemble de variables) en corrélation, par rapport à leurs valeurs probables, sont dues ou non au hasard. Ce test, le test du χ^2, aussi appelé test de la qualité d'ajustement ou d'adéquation (« *goodness of fit* ») permet en d'autres termes de vérifier si un échantillon d'une variable aléatoire Y donne des observations comparables à celles d'une loi de probabilité P définie *a priori* dont on pense, pour des raisons théoriques ou pratiques, qu'elle devrait être la loi de Y. Autrement dit, il s'agit de comparer le théorique avec l'empirique, l'*a priori* avec l'*a posteriori*, afin de déterminer si l'observation ou les données observées s'accordent ou s'ajustent (« *fit* ») aux hypothèses théoriques. Le terme « biais » va alors apparaitre dans la sixième partie

[30] Pearson, Karl, « Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 186, 1895, p. 343-414.

[31] Pearson, Karl, « Mathematical contributions to the theory of evolution. XIX. Second supplement to a memoir on skew variation », *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 216, 1916, p. 429–457.

[32] https://en.oxforddictionaries.com/definition/skew. Accédéle 6 juin 2017.

[33] L'expérience du lancer de dés occupe les deux premières illustrations parmi les trois illustrations qui concernent la « fréquence connue ou supposée connue *a priori* », tandis qu'il y a quatre illustrations pour la « fréquence de population générale inconnue *a priori*».

de l'article, intitulée : « Fréquences connues ou supposées connues *a priori* ». Voici le tableau en question :

Figure 1-1: Valeur observée et valeur théorique de 26306 lancers de 12 dés de Weldon par Pearson[34]

No. of Dice in Cast with 5 or 6 Points	Observed Frequency, m'	Theoretical Frequency, m	Deviation, e
0	185	203	− 18
1	1149	1217	− 68
2	3265	3345	− 80
3	5375	5576	−101
4	6114	6273	−159
5	5194	5018	+176
6	3067	2927	+140
7	1331	1254	+ 77
8	403	392	+ 11
9	105	87	+ 18
10	14	13	+ 1
11	4	1	+ 3
12	0	0	+ 0
	26306	26306	

Pearson commente les résultats ainsi:

« *The results show a bias from the theoretical results, 5 and 6 points occurring more frequently than they should do. Are the deviations such as to forbid us to suppose the results due to random selection? Is there in apparently true dice a real bias towards those faces with the maximum number of points appearing uppermos*t? » (Pearson, 1900, in Kotz et Johnson, 1992, p. 21).

La question est donc de savoir si ces écarts ou ces déviations par rapport à la fréquence théorique sont dus au hasard ou bien à un biais dans le dé. On peut noter que le biais apparait ici d'abord comme une déviation « par rapport aux résultats théoriques », donc comme une propriété de la distribution, puis comme une propriété physique du dé, au sens où il aurait été mal fabriqué (volontairement ou non), comme c'était le cas pour la boule dans le jeu de boules. L'hypothèse de Pearson est que dans

[34] Pearson, 1900, in Kotz et Johnson, 1992, p. 21.

la mesure où les points sur chaque face du dé sont obtenus en enlevant de la matière, alors il est logique que les faces 5 et 6 soient plus légères que les autres. De même, il faut souligner que la fréquence théorique du dé est connue : si on ne peut pas prédire quelle face du dé va apparaître à chaque lancer, on sait néanmoins que chaque face a en moyenne une chance sur six d'apparaitre, et donc que deux faces ont 1 chance sur 3 d'apparaitre à chaque lancer. Pearson va alors mettre les différentes déviations au carré (e^2) et les diviser par la moyenne *m* de la fréquence théorique, donc faire la somme des écarts des carrés à la moyenne :

Figure 1-2 : Test du χ^2 appliqué par Pearson aux 26306 lancers de 12 dés de Weldon [35]

Group	e^2	e^2/m
0	324	1.59606
1	4624	3.79951
2	6400	1.91330
3	10201	1.82945
4	25281	4.03013
5	30976	6.17298
6	19600	6.69628
7	5929	4.72807
8	121	0.30903
9	324	3.72414
10	1	0.07346
11	9	9.00000
12	0	.00000
Total	...	43.87241

La somme totale des écarts des carrés à la moyenne (soit 43,87241) est donc égale au χ^2 et χ est donc égal à 6,623625. Or il y a seulement une chance sur 62 499 d'arriver à un tel résultat. Pearson va alors refaire le calcul en se fondant non plus sur la fréquence théorique mais sur la fréquence réelle de l'échantillon, très large, de 26 306 lancers, c'est-à-dire 0,3377 :

> « Le Professeur Weldon m'a suggéré que nous devrions prendre la 26 306 x (0,337 + 0, 6623) plutôt que la binomiale 26 306 x (1/3+2/3) pour représenter la

[35] Pearson, 1900, in Kotz et Johnson, 1992, p. 21.

> distribution théorique, la différence entre 0,337 et 1/3 représentant le biais du dé » [« …*the difference between .3377 and 1/3 representing the bias of the dice.* »]. (Pearson, 1900, in Kotz et Johnson, 1992, p. 22).

Effectivement si l'on prend la fréquence observée et non la fréquence théorique, à ce moment-là les chances que les dés soient pipés ne sont plus que d'1 sur 8.

Nous tenons donc là une première approximation du sens du mot « biais », qui n'apparait d'ailleurs pas comme un concept important ou décisif dans la pensée de Karl Pearson : le biais est donc la différence entre la distribution d'une fréquence qui est connue ou supposée connue *a priori* et sa fréquence réelle ou observée, c'est-à-dire sa fréquence *a posteriori*. Le problème vient du fait que dans le cas du dé, la distribution du dé est bien connue *a priori*, puisque chaque face a théoriquement 1 chance sur 6 de sortir, ce qui en retour permet de calculer ou de quantifier précisément le biais en question (ici, Biais = 0,3377 – 0,3333). La question qui reste en suspens est donc de savoir comment détecter un tel biais si la fréquence n'est pas connue ou supposée connue *a priori*, question à laquelle l'article de Pearson ne donne malheureusement pas de réponse.

1.2.3 L'analyse de l'expérience de Weldon par Fisher

L'expérience du lancer de dés de Weldon va être à nouveau analysée, quoique d'une manière différente, dans l'ouvrage de Fisher, intitulé, *Statistical Methods for Research Workers*[36], publié en 1925, dans le chapitre III, consacrée aux « *Distributions* », c'est-à-dire aux distributions de fréquence. Les distributions de fréquence permettent en effet d'étudier la variation, cette étude de la variation étant définie par Fisher comme une des trois définitions ou comme un des trois objectifs possibles de la science des statistiques (Fisher, 1950, p. 1). Plus précisément, nous dit Fisher, les distributions de fréquence (Fisher va étudier successivement la loi normale, les séries de Poisson et la loi binomiale, où apparait l'expérience de Weldon) permettent de déterminer, grâce aux tests statistiques, si un échantillon donné d'une population est semblable à un autre échantillon, et donc de déterminer à quelle

[36] Fisher, Sir Ronald Aylmer, *Statistical Methods For Research Workers*, Edinburgh, Oliver and Boyd, 1925. Nous utilisons pour notre étude sur la 11ème édition: Fisher, Sir Ronald Aylmer, *Statistical Methods For Research Workers*, 11ème, Edinburgh, Oliver and Boyd, 1950.

population les conclusions tirées à partir d'un échantillon vont pouvoir s'appliquer. Il dit ainsi au début du chapitre III :

> « L'idée d'une population infinie distribuée dans une distribution de fréquence en fonction d'une ou de plusieurs caractéristiques est fondamentale pour tout travail statistique. A partir d'une expérience limitée, par exemple les individus d'une espèce, ou du temps qu'il fait dans une certaine localité, nous pouvons obtenir quelque idée d'une population infinie hypothétique d'où notre échantillon est tiré, et par là de la nature probable des futurs échantillons auxquels nos conclusions seront appliquées. Si un second échantillon contredit cette attente [« *expectation* »], nous concluons, en langage statistique, qu'il est tiré d'une population différente ; que le traitement auquel le second échantillon d'organismes a été exposé a réellement fait une différence matérielle, ou que le climat (ou les méthodes pour le mesurer) s'est matériellement altéré. Les tests critiques de ce genre peuvent être appelés des tests de signification [« *tests of significance* »], et quand de tels tests sont disponibles nous pouvons découvrir si un second échantillon est ou n'est pas significativement différent du premier » (Fisher, 1950, p. 31).

Parmi ces tests, il y a bien évidemment le test du χ^2, inventé par K. Pearson, mais aussi le test inventé par Fisher, à savoir l'analyse de la variance (ANOVA pour *ANalysis of VAriance*) , outil qui permet de vérifier que plusieurs échantillons sont issus d'une même population par la comparaison de leur moyenne ou de leur espérance. Fisher pose le problème ainsi : si le dé est bien équilibré, alors p=1/3 (en effet, si on jette un dé équilibré, il y a une chance sur trois qu'il fasse plus de 4), la loi binomiale étant normalement : $(2/3 + 1/3)^{12}$. Or,

> « si un ou plusieurs dés ne sont pas équilibrés, et si tous conservent le même biais tout au long de l'expérience, les fréquences devraient être approximativement déterminées par $(q+p)^{12}$, où p est une fraction à déterminer d'après les données » (Fisher, 1950, p. 63-64).

Fisher reprend alors le tableau donné par Pearson, en intégrant toutes les données (c'est-à-dire avec la fréquence théorique de 1/3 et celle observée sur l'échantillon de 0,3377). Voici le tableau en question :

Figure 1-3: Analyse de la variance appliquée par Fisher aux 26306 lancers de 12 dés de Weldon[37]

TABLE 10

Number of Dice with 5 or 6.	Observed Frequency.	Expected True Dice.	Expected Biased Dice.	Measure of Divergence $\frac{x^2}{m}$.	
				True Dice.	Biased Dice.
0	185	202·75	187·38	1·554	·030
1	1149	1216·50	1146·51	3·745	·005
2	3265	3345·37	3215·24	1·931	·770
3	5475	5575·61	5464·70	1·815	·019
4	6114	6272·56	6269·35	4·008	3·849
5	5194	5018·05	5114·65	6·169	1·231
6	3067	2927·20	3042·54	6·677	·197
7	1331	1254·51	1329·73	4·664	·001
8	403	392·04	423·76	·306	1·017
9	105	87·12	96·03	3·670	·838
10	14	13·07	14·69		
11	4	1·19	1·36	·952	·222
12	...	·05	·06		
	26306	26306·02	26306·00	35·491	8·179
				$n = 10$	$n = 9$

Fisher conclut immédiatement que « les observations ne sont pas compatibles avec l'assertion que les dés ne sont pas pipés [« *unbiased* »]. » (Fisher, 1950, p. 64). Fisher va alors appliquer le test du χ^2 hérité de Karl Pearson et défini par Fisher comme « un moyen de tester l'accord entre l'observation et l'hypothèse » (Fisher, 1950, p. 79). L'hypothèse nulle est donc ici que l'observation est proche de la théorie, c'est-à-dire que les dés ne sont pas truqués. Fisher refait le raisonnement de Pearson pour montrer qu'il n'y a que 0,001% de chances que le χ^2 dépasse 35.49 (Fisher, 1950, p. 65). Fisher va alors utiliser son propre test de la variance pour montrer que les dés sont pipés, c'est-à-dire que les valeurs observées dévient des valeurs théoriques :

> « La variance des séries binomiales est pqn. Ainsi avec un dé équilibré [« *true dice »]* et 315 672 essais, le nombre attendu [« *expected »]* de dés marquant plus de 4 est de 105,224 avec une variance de 70149,3 et un écart-type de

[37] Fisher, 1950, p. 64

264,9; le nombre observé dépasse l'espérance [« *exceeds expectation »*] de 1378, soit 5,20 fois son écart-type » (Fisher, 1950, p. 65).

Or « une déviation normale ne dépasse cinq fois son écart-type qu'une fois sur 5 millions » (Fisher, 1950, p. 65). Pour Fisher, ce test est le « test le plus sensible de la présence d'un biais » [38], c'est-à-dire plus sensible et donc plus efficace que le test du χ^2, au sens où il donne des « cotes beaucoup plus élevées »[39] que le test de Pearson, en se concentrant non pas sur les « divergences de toute sorte »[40], mais en testant séparément la divergence liée à la valeur de *p*.

Le biais apparait donc ici comme une déviation par rapport à la normale, ou plus précisément comme une déviation par rapport à l'écart-type qui est beaucoup trop élevée (5 fois plus que l'écart-type) pour être normale, et qui, en tant que telle, est hautement improbable (1 chance sur 5 million que cela arrive). Dès lors, cette distribution de la fréquence d'apparition des numéros 5 et 6 dans une longue série de lancers de dés doit permettre de conclure, étant donné son caractère hautement improbable, que les dés sont pipés, c'est-à-dire que le 5 et le 6 ont plus de chances d'apparaitre que les autres faces des 12 dés, pour une raison à déterminer. Et c'est précisément le test statistique, en l'espèce ici l'analyse de la variance, qui permet de déceler la présence d'un biais.

1.3 Le biais comme erreur systématique du plan d'expérience :

1.3.1 Le biais et la loi normale.

Fisher va de nouveau utiliser la notion de « biais » dans le chapitre VI, consacré au « Coefficient de corrélation », aux sections 35 et 36, intitulées respectivement : « *Transformed Correlations »* et « *Systematic Errors »*. Dans la section 35, Fisher va proposer une technique afin de pouvoir tester correctement s'il existe une association entre deux variables :

> «En plus de tester la significativité d'une corrélation, afin de déterminer s'il y a ou non une preuve substantielle d'une association, il est aussi fréquemment requis d'effectuer une ou plusieurs des opérations suivantes, et il est possible

[38] « *the most sensitive test of the bias*», in Fisher, 1950, p. 65.
[39] « *so much higher odds* », in Fisher, 1950, p. 65.
[40] « *discrepancies of any kind* », in Fisher, 1950, p. 66.

pour chacune d'entre elles d'utiliser l'écart-type dans le cas d'une quantité normalement distribuée »[41]

L'idée consiste à effectuer une certaine transformation mathématique afin que la distribution des observations suive une loi normale, donc susceptible d'être traitée avec les tests statistiques : c'est ce que l'on appelle en statistiques la « transformation de Fisher » (ou la « transformation Z de Fisher »[42]). En effet la distribution d'échantillonnage du coefficient de corrélation *r* de Pearson ne suit pas la distribution normale car elle a une variance inégale et est fortement asymétrique : la transformation de Fisher permet de convertir le *r* de Pearson en une variable Z distribuée normalement (symétrique et de variance égale). Cette transformation permet alors de calculer les intervalles de confiance du coefficient de corrélation de Pearson, et donc de tester la significativité des différences entre des coefficients de corrélation. Fisher va ainsi donner la représentation graphique de *r* (Figure 1-4) et de Z (Figure 1-5) pour 8 paires d'observations faites sur des populations qui ont un coefficient de corrélation de 0 et de 0.8 :

[41] « *In addition to testing the significance of a correlation, to ascertain if there is any substantial evidence of association at all, it is also frequently required to perform one or more of the following operations, for each of which the standard error would be used in the case of a normally distributed quantity* », in Fisher, 1950, p. 197.

[42] La formule de cette transformation remarquable est : $Z = \frac{1}{2} ln\,[(1 + r)/(1 - r)]$. Fisher explique cette transformation et son rapport aux autres distributions de fréquence dans son article : Fisher, Sir Ronald Aylmer, « On a Distribution Yielding the Error Functions of Several Well Known Statistics », *Proceedings of the International Congress of Mathematics*, vol. 2, 1924, p. 805–813.

Figure 1-4 : Courbes de distributions du coefficient de corrélation r pour 8 paires d'observation (Fisher) [43]

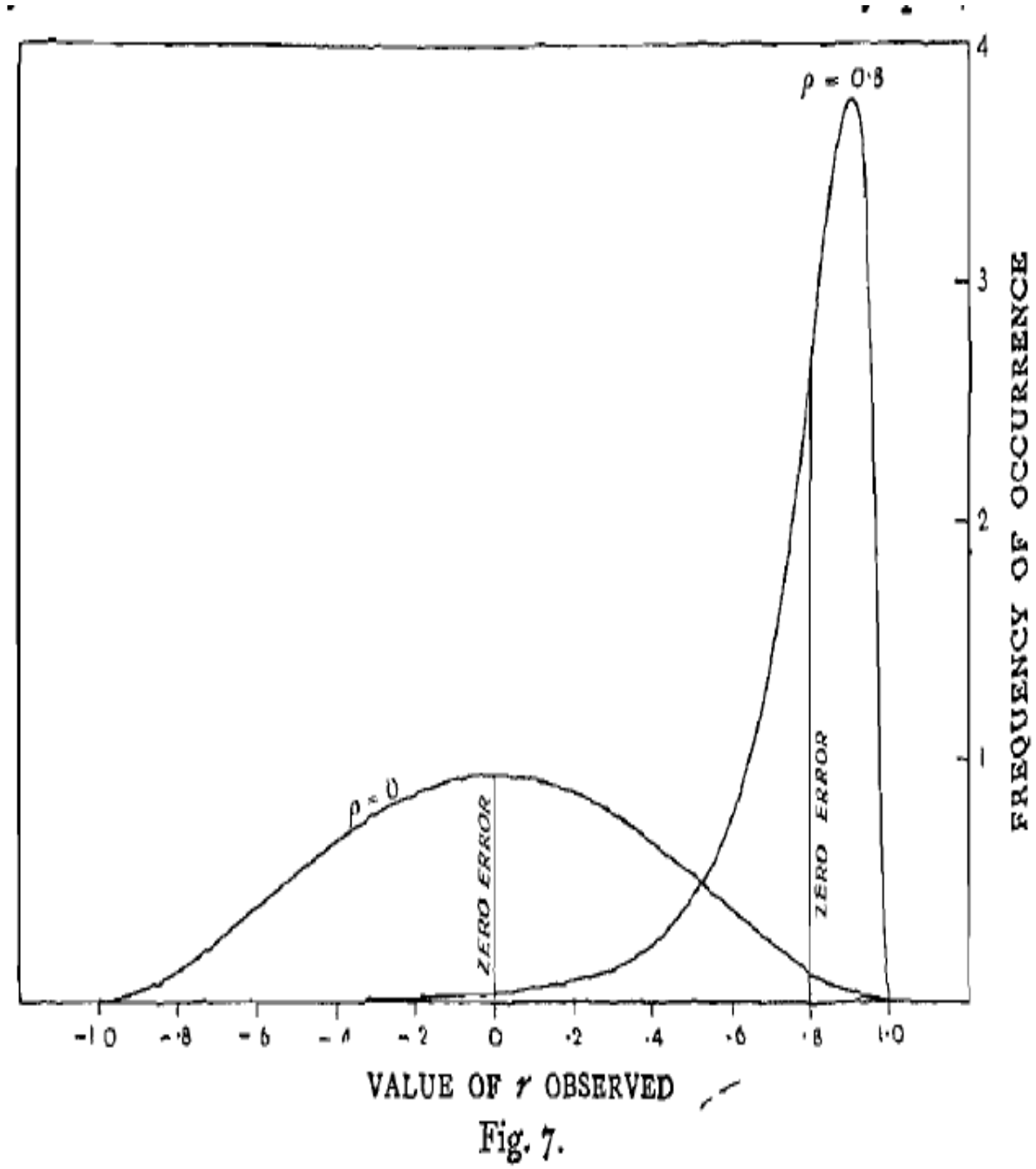


Fig. 7.

[43] Fisher, 1950, p. 200.

Figure 1-5: Courbes de distributions de r transformées en Z pour 8 paires d'observations (Fisher) [44]

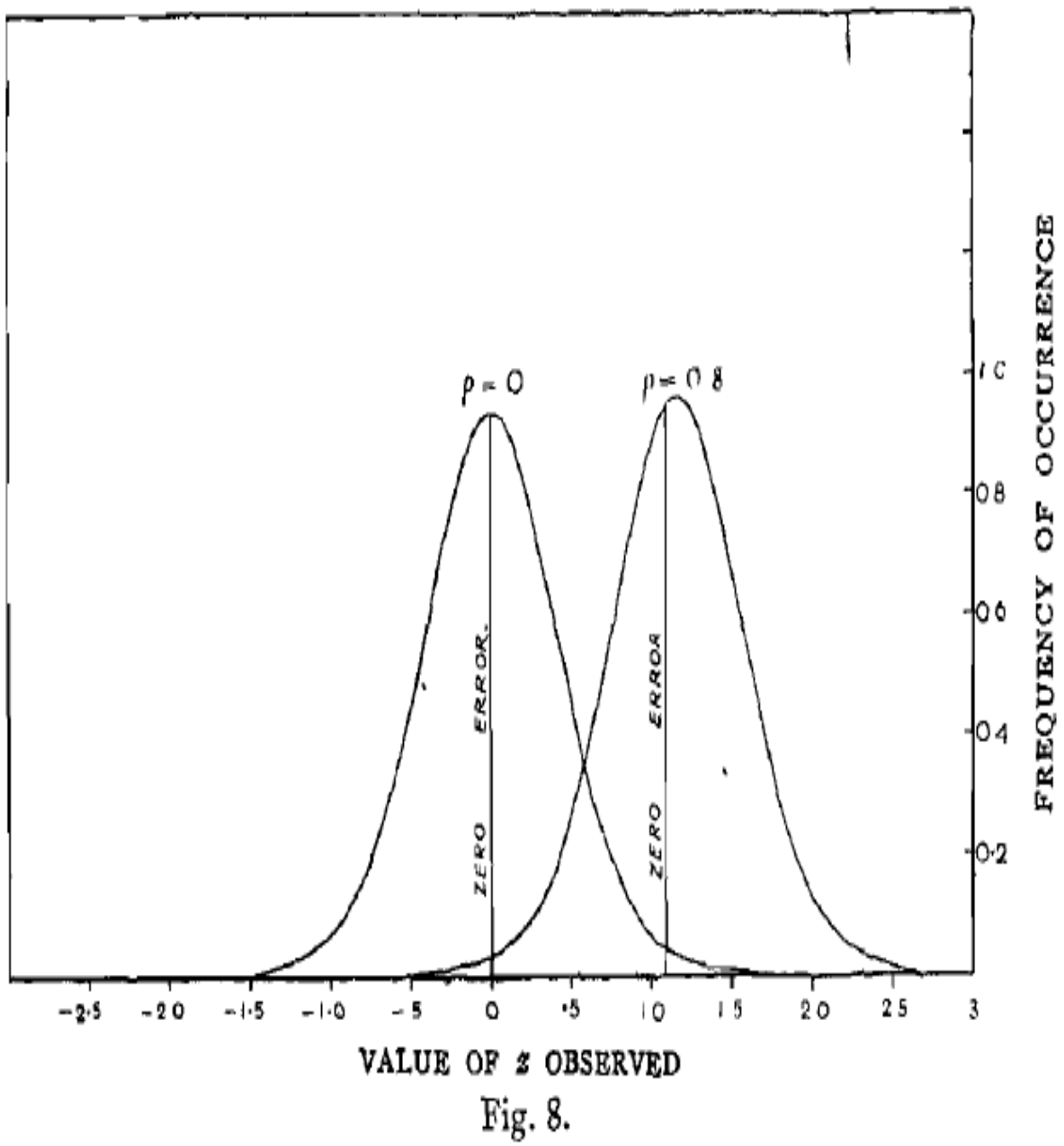


Fig. 8.

[44] Fisher, 1950, p. 200.

La différence entre les deux figures est très marquée. Ainsi, selon Fisher :

« *The two curves in Fig. 7 are widely different in their modal heights; both are distinctly non-normal curves; in form also they are strongly contrasted, the one being symmetrical, the other highly unsymmetrical. On the contrary, in Fig. 8 the two curves do not differ greatly in height; although not exactly normal in form, they come so close to it,(…), although the curve itself is as symmetrical as the eye can judge of, yet the ordinate of zero error is not centrally placed.* » (Fisher, 1950, p. 200-201).

Pour Fisher cet écart par rapport à l'erreur zéro, c'est-à-dire par rapport à l'estimation de la moyenne des valeurs observées, révèle un « petit biais introduit dans l'estimation du coefficient de corrélation tel qu'on le calcule habituellement »[45], biais dont il va traiter dans la section suivante consacrée aux « erreurs systématiques ». Dans cette section, Fisher va donner deux méthodes pour corriger le biais de l'estimation, biais dont la valeur augmente en fonction du nombre d'échantillons. Voici la première méthode, purement mathématique :

« La valeur de Z obtenue à partir de n'importe quel échantillon est une estimation de la vraie valeur, ζ, qui appartient à la population échantillonnée, tout comme la valeur de *r* obtenue à partir d'un échantillon est une estimation d'une valeur de la population, *p.* Si la méthode pour obtenir la corrélation était libre de biais, les valeurs de Z seraient normalement distribuées autour d'une moyenne $\bar{Z}$, dont la valeur s'accorderait avec ζ. En fait, il y a un léger biais qui fait que la valeur moyenne de Z est d'une manière ou d'une autre plus grande numériquement que ζ ; ainsi, la corrélation, qu'elle soit positive ou négative, est légèrement exagérée. Ce biais peut être efficacement corrigé en soustrayant de la valeur Z la correction: $\frac{p}{2\,(n'-1)}$ »[46]

[45] « *The figure, in fact, reveals the small bias which is introduced into the estimate of the correlation coefficient as ordinarily calculated »,* in Fisher (1950), *Op.cit.,* p. 200-201

[46] « *The value of Z obtained from any sample is an estimate of a true value, ζ, belonging to the sampled population, just as the value of r obtained from a sample is an estimate of a population value, p. If the method of obtaining the correlation were free from bias, the values of Z would be normally distributed about a mean $\bar{Z}$, which would agree in value with ζ. Actually there is a small bias which makes the mean value of Z somewhat greater numerically than ζ ; thus the correlation, whether positive or negative, is slightly exaggerated. This bias may effectively be corrected by subtracting from the value of Z the correction: $\frac{p}{2\,(n'-1)}$ »,* in Fisher, 1950, p. 205.

De même, la deuxième erreur systématique se produit si l'on néglige la correction de Sheppard, correction qui vise à débiaiser le calcul d'une variance empirique d'un échantillon dont les valeurs ont été regroupées par classe de même amplitude. Selon Fisher, quand on transforme *r* en Z, il faut prendre la valeur de *r* qui est trouvée sans la correction de Sheppard, car, sinon, cela compliquerait la distribution. Mais en faisant cela, on va introduire une erreur systématique qui va produire l'effet inverse de la première erreur systématique décrite, c'est-à-dire que la valeur moyenne de Z sera plus petite numériquement que ζ. Il est alors possible d'appliquer une correction correspondant à la « moyenne des effets de la correction de Sheppard »[47].

1.3.2 Le biais et la randomisation.

Enfin, Fisher va utiliser le mot « biais » dans le chapitre VIII, intitulé « *Further Applications of the Analysis of Variance* », et plus spécifiquement la section 48 consacrée à la « Technique de l'expérimentation de terrain » (« *Technique of Plot Experiment* »[48]), chapitre qui est à mettre en relation directe avec l'autre ouvrage de Fisher, *The Design of Experiments*[49], notamment son chapitre IV, intitulé : « *An Agricultural Experiment in Randomised Blocks* ». Dans ces chapitres, qui traduisent l'expérience acquise par Fisher comme statisticien à la station expérimentale de Rothamsted, poste qu'il a commencé à occuper en 1919, Fisher va insister sur la nécessité de la randomisation afin de pouvoir disposer d'estimations valides :

> « La première exigence qui gouverne toutes les expérimentations bien planifiées est que l'expérimentation donne non seulement une comparaison des différents engrais, traitements, variétés, etc. mais aussi un moyen de tester la significativité des différences observées » [50]

Le principal moyen consiste d'abord à dupliquer et à répliquer les traitements, de manière à ce qu'on puisse comparer les différences observées entre les différents

[47] « *the average effect of Sheppard's adjustment* », in Fisher, 1950, p. 261.
[48] Fisher, 1950, p. 261.
[49] Fisher, Ronald A., *The Design of Experiments*, Edinburgh, Oliver and Boyd, 1935.
[50] « *The first requirement which governs all well-planned experiments is that the experiment should yield not only a comparison of different manures, treatments, varieties, etc., but also a means of testing the significance of such differences as are observed*. », in Fisher, 1950, p. 261.

traitements avec les réplications qui serviront alors d'étalon[51]. Le problème des expériences en agriculture est en effet qu'il existe une très forte variation quant à la fertilité ou au rendement des différents plants. Or, si l'on veut que « le test statistique de signification soit valide, il faut que les différences observées de fertilité entre les différentes parcelles choisies comme parallèles dans le plan d'expérience soient réellement représentatives des différences entre les parcelles qui ont des traitements différents »[52]. Et le seul moyen de s'assurer que c'est bien le cas est de « répartir les parcelles complètement au hasard[53].

En effet, si l'expérimentateur choisit un « système préarrangé » (« *prearranged system* ») ou une « répartition systématique » (« *systematic arrangement* »), il est fort probable que « nos parcelles aient (…) des traits communs avec la variation systématique de fertilité, et alors notre test de signification est complètement vicié »[54]. La randomisation est ainsi ce qui garantit la validité du test statistique de signification, test qui est fondé sur l'estimation de l'erreur garantie par la réplication des traitements. Cette randomisation permet en effet que « que deux parcelles, quelles qu'elles soient, qui ne sont pas dans une même série, aient la même probabilité d'être traitées de la même manière, et la même probabilité d'être traitée différemment de toutes les façons possibles »[55].

Ainsi, en ne répartissant pas les traitements au hasard, cela risque de produire un biais. Fisher va alors décrire le mécanisme de ce biais dans la section 27 du chapitre IV de son ouvrage *The Design of Experiments*, intitulé explicitement : « *Bias of Systematic Arrangements »*[56]. Selon lui, l'effet principal de ce biais consiste dans une perte d'exactitude (« *accuracy* ») dans l'estimation de l'erreur :

> « Dans n'importe quelle situation particulière, il sera sans doute possible d'assigner à des ensembles de parcelles [« *plots* »] sur un terrain [« *area* »]

[51] « *Consequently all treatments must at least be duplicated, and preferably further replicated, in order that a comparison of replicates may be used as a standard with which to compare the observed differences.* », in Fisher, 1950, p. 261.

[52] « *For our test of significance to be valid the differences in fertility between plots chosen as parallels must be truly representative of the differences between plots with different treatments* », in Fisher, 1950, p. 261.

[53] « *to arrange the plots wholly at random*. », in Fisher, 1950, p. 264.

[54] « *for the systematic arrangement of our plots may have (…) features in common with the systematic variation of fertility, and thus the test of significance is wholly vitiated.* », in Fisher, 1950, p. 261-262.

[55] « *The validity of our estimate of error for this purpose is guaranteed by the provision that any two plots, not in the same block, shall have the same probability of being treated alike, and the same probability of being treated differently in each of the ways in which this is possible* » in Fisher, 1935, p. 71.

[56] Fisher, 1935, p. 71

> donné de nombreux traitements de manière à égaliser leur fertilité plus complètement que si l'on procédait par un arrangement au hasard. (…). L'effet d'une telle procédure sur le test de signification peut être conçu en imaginant qu'il est effectué sur un terrain [« *area* »] soumis à un traitement uniforme, de sorte que les rendements réels ne soient pas affectés par une réallocation des parcelles [« *plots* »]. Dès lors, dans l'analyse de la variance, la somme totale des carrés reste inchangée, tout comme la portion assignable aux blocs [« *blocks* »]. Si, dès lors, l'ingéniosité de l'agronome a réussi à diminuer les différences de fertilité entre les traitements, la diminution de la somme des carrés dans cette ligne du tableau aura été complètement contrebalancée par une augmentation de la somme des carrés sur laquelle l'estimation de l'erreur est fondée. L'effet de ce réarrangement aura été de diminuer les erreurs réelles de l'expérimentation, mais au prix de l'augmentation de l'estimation de l'erreur, de sorte que, alors que la précision [« *precision* »] des comparaisons a été effectivement augmentée, elles apparaitront comme moins exactes [« *accurate* »] qu'avant, et on aura moins confiance dans les résultats ». (Fisher, 1935, p. 71-72).

Ce passage technique peut être illustré par l'expérience décrite par Fisher dans l'exemple 44 qui est donné dans ses *Statistical Methods for Research Workers*. L'expérience consiste à tester cinq traitements différents sur vingt bandes de terres, chacun des traitements étant répliqués quatre fois, et la répartition des traitements se faisant au hasard en assignant à chaque traitement une lettre (A, B, C, D, E) et en mélangeant 20 cartes (5 cartes de A, 5 cartes de B, etc.). Fisher utilise alors les données d'une expérience menée par Mercer et Hall, qui porte sur le poids de racines de blettes obtenu pour chaque traitement. Voici les résultats :

Figure 1-6 : Poids de racines de blettes obtenus pour 5 traitements différents et répartis sur 20 bandes de terre (Fisher)[57]

TABLE 57

B	C	A	C	E	E	E	A	D	A
3504	3430	3376	3334	3253	3314	3287	3361	3404	3366
B	C	B	D	D	B	A	D	C	E
3416	3291	3244	3210	3168	3195	3330	3118	3029	3085

Fisher peut alors calculer la somme des écarts de la moyenne de chaque traitement par rapport à la moyenne générale de tous les traitements. Cela donne :

A = + 290

B = + 216

C = - 59

D = - 243

E = -204

Fisher va alors calculer la variance, c'est-à-dire la moyenne de la somme du carré des déviations, ce qui lui permet de montrer que la somme des carrés correspondant au « traitement » sera le quart de la somme des carrés de ces déviations (puisqu'il y a 4 degrés de libertés, c'est-à-dire 5 observations – 1). De même, la somme des carrés des 20 déviations par rapport à la moyenne générale est de 289 766, ce qui donne le tableau suivant:

[57] Fisher, 1950, p. 262

Figure 1-7 : Analyse de la variance appliquée aux poids des racines de blettes en fonction des 5 traitements (Fisher)[58]

TABLE 58

Variance due to	Degrees of Freedom.	Sum of Squares.	Mean Square.	Standard Déviation.
Treatment . .	4	58,726	14,681	121·1
Experimental error .	15	231,040	15,403	124·1
Total . . .	19	289,766	15,251	123·5

Fisher peut alors conclure que l'écart type d'une seule parcelle est estimée à 124,1, alors que sa vraie valeur est de 123,5 : ce faible écart, et même cette « quasi-concordance » (« *exceedingly close agreement* ») entre la valeur estimée et la valeur réelle de l'écart-type illustre selon lui « la manière par laquelle un arrangement des parcelles effectué purement au hasard assure que l'erreur expérimentale calculée est une estimation non-biaisée des erreurs effectivement présentes »[59].

Si nous revenons à présent sur le biais soulevé par un arrangement systématique, les choses semblent plus claires : dans ce cas-là, Fisher nous dit qu'il y aura moins de déviations par rapport à la moyenne générale et que donc la somme des carrés pour le traitement sera moins élevée. Or, le problème est que la somme totale des carrés reste inchangée, ce qui fait que la somme des carrés pour les erreurs expérimentales sera plus élevée et donc l'écart-type aussi[60]. En conséquence de quoi l'estimation sera plus éloignée de la valeur réelle : l'estimateur sera dit « biaisé ». Par exemple, si nous attribuons arbitrairement la valeur 35 226 à la somme des carrés

[58] Fisher, 1950, p. 263.

[59] « *This is an exceedingly close agreement, and illustrates the manner in which a purely random arrangement of plots ensures that the experimental error calculated shall be an unbiased estimate of the errors actually present.* », in Fisher, 1950, p. 263.

[60] Pour une description plus abstraite du plan de l'expérience et de son analyse statistique, voir la section 23 du chapitre IV, in Fisher, 1935, p. 57-64.

dans la ligne « Traitement », à la place de 58 726, la somme des carrés des erreurs expérimentales devient : 289 766 – 35226 = 254 450. La moyenne des carrés, c'est-à-dire la variance, est donc de : 254 450/ 15 = 16 969,33 (au lieu de 15 403). Et l'écart-type, c'est-à-dire la racine carrée de la variance devient $\sqrt{16969.33}$ =130.3. La vraie valeur étant de 123.5, la valeur estimée de 130.3 est donc très éloignée de cette valeur réelle, ce qui tendrait à prouver que l'estimateur est biaisé : il serait d'ailleurs possible de faire un test statistique pour déterminer si l'estimateur est réellement biaisé.

1.3.3 Le biais est-il un problème d'estimation ou de signification ?

Pourtant, en regardant les différents textes de Fisher où il utilise le mot « biais » il apparait qu'il l'utilise en deux sens différents, que l'on peut catégoriser de la manière suivante :

- Un sens faible qui renvoie à la notion d'estimation, au sens statistique actuel où l'on dit d'un estimateur qu'il est biaisé ou sans biais. Plus précisément, la notion de biais est l'antonyme de celle de convergence (« *consistency* »), qui est un des trois critères en fonction desquels on peut évaluer un estimateur, les deux autres étant l'efficacité (« *efficiency* ») et l'exhaustivité (« *sufficiency* »). Fisher dit ainsi, en 1925, à propos de la convergence : « Une statistique est dite estimation convergente d'un paramètre si, lorsqu'elle est calculée sur un échantillon infiniment grand, elle tend à être égale au paramètre »[61]. En effet, Fisher dit, dans son article sur « La logique de l'inférence inductive »[62], que « la distribution normale n'a que deux caractéristiques, sa moyenne et sa variance. La moyenne détermine le biais de notre estimateur, et la variance détermine sa précision » (Fisher, 1935b, p. 42). Or, de ces deux caractéristiques, seule la variance pose problème[63]. En effet, dans la théorie de Fisher, ce type de biais ne joue « qu'un faible rôle » : « la considération du biais ne nous retiendra guère », car « avec des

[61] Fisher, Ronald A., « Theory of Statistical Estimation », *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 22 / 05, 1925b, p. 700. Cité dans Armatte, Michel, « La construction des notions d'estimation et de vraisemblance chez Ronald A. Fisher », *Journal de la société française de statistique*, vol. 129 / 1-2, 1988, p. 68-95.

[62] Fisher, Ronald A., « The Logic of Inductive Inference », *Journal of the Royal Statistical Society*, vol. 98 / 1, 1935b, p. 39-82

[63] « *is a more serious affair* », in Fisher, 1935b, p. 42.

estimateurs convergents, celui-ci doit tendre vers zéro »[64]. Le biais est donc ici « l'écart espéré entre la vraie valeur du paramètre et son estimateur »[65], il peut être mesuré ou quantifié et il n'est pas nécessaire de l'éliminer. Parfois, un estimateur biaisé peut aussi être plus exact (« *accurate* ») qu'un estimateur non-biaisé.

- Par contre, il existe un deuxième sens du mot « biais » chez Fisher, qu'on pourrait qualifier de sens fort, et qui renvoie non au problème de l'estimation, mais à celui du test statistique. Ce problème apparait très clairement dans le cadre d'une controverse entre Neyman et Fisher qui a lieu en 1935, lorsque Jerzy Neyman lit, devant la Section de la recherche agricole et industrielle de la *Royal Statistical Society,* un article intitulé « *Statistical Problems in Agricultural Experimentation* »[66], séance à laquelle assistent, entre autres, Ronald Fisher et Franck Yates, son ancien assistant à la station expérimentale de Rothamsted (il succède à Fisher quand celui-ci part enseigner à l'*University College* de Londres en 1933), mais aussi Egon S. Pearson, le fils de Karl Pearson[67]. Dans cet article, où le mot biais apparait à cinquante-deux reprises (dix-sept dans l'article, et trente-cinq dans la discussion) Neyman critique en fait la méthode de la randomisation par blocs (« *randomized blocks* ») mais aussi celles du carré latin, méthodes introduites dans l'ouvrage de Fisher, sorti dix ans plus tôt : *Statistical Methods for Research Workers*. Selon Neyman, en effet, le test Z apparait biaisé dans le cas d'un carré latin au sens où il montre « une tendance à découvrir une différenciation là où il n'y en a pas »(Neyman, Iwaszkiewicz, Kolodziejczyk, 1935*,* p. 114) , alors qu'il semble plus valide pour les blocs randomisés. Sans rentrer dans le détail de la controverse, il est intéressant de noter que Fisher lui-même distingue bien deux champs d'application ou

[64] Fisher, 1935b, p. 42. La traduction est de Michel Armatte.
[65] Armatte,1988, p. 83
[66] Neyman, J., Iwaszkiewicz, K. et Kolodziejczyk, St., « Statistical Problems in Agricultural Experimentation », *Supplement to the Journal of the Royal Statistical Society*, vol. 2 / 2, 1935, p. 107.
[67] On peut rappeler, en guise de contextualisation historique, qu'en 1933, le département de statistiques appliquées de l'*University College* de Londres est scindé en deux suite à la retraite de Karl Pearson : Fisher devient le titulaire de la Chaire Galton d'Eugénique et Egon Pearson, le fils de Karl, devient Professeur de Statistiques. La relation entre Karl Pearson est notoirement exécrable (les deux personnages semblent avoir une part égale à cette brouille), et celle de Fisher avec le fils de Karl n'est pas meilleure (ici, Fisher semble être plus responsable qu'Egon Pearson, qui tient ses travaux en haute estime).

deux problèmes liés au concept de biais : le problème de l'estimation (où ce problème est « complètement sans importance »[68]) et celui des tests de signification, où un « biais, s'il était prouvé, serait fatal » (Neyman, Iwaszkiewicz, Kolodziejczyk, 1935, p. 157). Il accuse d'ailleurs Neyman d'avoir confondu ces deux problèmes. Mais pourquoi serait-il fatal dans le cas du test de signification ? La réponse à cette question est formulée par Neyman : le test de signification serait vicié au sens où il détecterait une différence (entre l'effet d'un traitement par exemple) là où il n'y en a pas, ou à l'inverse, de façon logique, au sens où il ne détecterait pas une différence alors qu'il y en a une. En somme c'est la validité du test qui est ici mise en cause : ici, Neyman considère que la distribution Z, et donc le test Z, ne peut pas être appliquée à un arrangement en termes de carré latin.

La dernière question qui subsiste alors est de savoir si et comment ces deux sens différents s'articulent entre eux. En effet, d'un côté, le biais renvoie à la notion d'estimation et de l'autre à la notion de test de signification. N'y a-t-il pas un cercle vicieux au sens où d'un côté, le test statistique, comme le χ^2 ou l'analyse de la variance permet de savoir si l'estimateur est biaisé (comme dans l'exemple du lancer de dé), alors que de l'autre côté, le test statistique ne semble valable que si précisément le plan d'expérience élimine les biais par le moyen de la randomisation ? Autrement dit, l'absence de biais apparait à la fois comme la *condition de possibilité du* test statistique, ou en tout cas la condition de possibilité de sa validité, et le *résultat* du test statistique. Cette contradiction disparait si l'on prend soin de bien distinguer deux étapes dans le plan d'expérience : d'abord il y a l'estimation de l'erreur, qui se fait grâce à la variance qui n'est pas à proprement parler un test statistique mais une mesure statistique, c'est-à-dire un estimateur, estimateur qui a par définition pour fonction d'évaluer un paramètre inconnu d'une population à partir de données obtenues sur un échantillon, (en l'espèce il s'agit ici d'estimer les « causes de variabilité »[69]). Or pour que cette estimation soit valide, il faut que l'échantillon soit choisi au hasard :

[68] Neyman, Iwaszkiewicz, Kolodziejczyk, 1935, p. 157.

[69] Fisher, Sir Ronald A. . «The correlation between relatives on the supposition of Mendelian inheritance. » *Trans. Roy. Soc. Edinb.* 1918; 52:399–433. Cité dans cité dans Bodmer, W., « RA Fisher, statistician and geneticist extraordinary: a personal view », *International Journal of Epidemiology*, vol. 32 / 6, décembre 2003, p. 938-942. La citation est à la page 939.

> « Une façon d'être certain qu'une estimation valide de l'erreur sera obtenue est d'arranger délibérément au hasard les parcelles, de sorte qu'aucune distinction ne pourra s'immiscer entre les paires de parcelles traitées de façon similaire et les paires traitées de façon différente ». [70]

Ainsi, c'est la structure logique du plan d'expérience, via la randomisation, qui permet d'éliminer théoriquement toutes les causes de perturbation, chacune de ces causes affectant ou étant supposée affecter de la même manière les différents échantillons ou membres des échantillons du fait même de leur répartition au hasard, et par là de parvenir à une estimation valide de l'erreur. Et c'est seulement si cette première condition est remplie qu'il est possible de faire un test de signification pour savoir, et tout en quantifiant le risque de se tromper (à 5%), si ces variations sont dues ou non au hasard : la randomisation permet en effet aussi d'éliminer « les causes de perturbation » susceptibles de « corrompre »[71] le test de signification. Dès lors, comme le dit Walter Bodmer, qui fut, selon ses propres termes, « parmi l'un des derniers disciples et étudiants de RA Fisher » (Bodmer, 2003, p. 938), il ne s'agit que d'une seule et même chose, sous deux aspects différents :

> « Ainsi, en effet, son point de vue [celui de Fisher] semble être que le fait que la randomisation fournit d'un côté une estimation valide de l'erreur et, de l'autre, élimine les biais, ne sont en fait que les deux faces d'une même pièce. » (Bodmer, 2003, p. 940).

1.3.4. Continuités et ruptures de la prénotion[72] de biais

Pour conclure sur ce premier chapitre qui vise à faire la préhistoire du concept de biais, il semble que ce concept soit progressivement apparu d'abord dans le cadre d'une discussion sur la théorie de l'évolution par sélection naturelle, et plus particulièrement dans le cadre de la sélection sexuelle chez Galton, discussion où le

[70] Fisher, Sir Ronald Aylmer, « The Arrangement of Field Experiments », *Journal of the Ministry of Agriculture of Great Britain*, 1926, p. 503-513, cité dans Bodmer, 2003, p. 940.

[71] « … *for the full procedure of randomisation, by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have been eliminated* » in Fisher, 1935, cité dans Bodmer, 2003, p. 940.

[72] Nous entendons ici le mot « prénotion » non au sens antique d'une connaissance générale spontanée, mais au sens moderne et durkheimien d'un concept formé spontanément par la pratique et qui n'a pas encore subi l'épreuve de la critique scientifique. Sur ce sens, voir Durkheim, Emile, *Les Règles de la méthode sociologique*, Paris, Alcan, 1901 [1894], p.40.

sens trivial et commun du mot semble effectuer une légère inflexion en direction d'un sens spécifique, bien que non défini par Galton (et sans doute inaperçu), qui renvoie au phénomène d'évolution d'une espèce, ou en termes galtoniens, de modification du « type racial » entendu comme modification du « centre » (c'est-à-dire d'espérance) d'une courbe normale. Puis le concept se déploie au sein de la statistique mathématique qui se développe autour de l'année 1900 avec Weldon et Pearson, là aussi autour d'une problématique liée à l'apparition d'une nouvelle espèce, bien que de façon indirecte à travers l'expérience de Weldon du lancer de dés. Enfin, il prend tout son sens statistique, sans pour autant être défini, avec Fisher et la notion d'estimation et de plan d'expérience, où l'absence de biais apparait comme la condition de possibilité de la validité de l'estimation de l'erreur mais aussi de la validité du test de signification.

Il est donc possible d'apercevoir des ruptures et des continuités dans l'élaboration de cette notion de biais, qui n'est pas, en 1935, tout à fait un concept : l'élément de continuité est à chercher du côté de la notion de variation et donc à celle d'erreur (les deux notions étant étroitement liées dans le calcul des probabilités),au sens où, dès le départ, le biais renvoie à une variation d'un autre genre que la variation classique au sens où en quelque sorte il s'agit non de variations au sein d'une population, mais d'une variation de la population elle-même, qui devient donc une autre population (avec une autre moyenne, une autre variance, etc.) ; non donc une variation mais une spéciation. En un sens il est logique que le mot biais soit utilisé pour désigner une erreur ou une variation plus grande ou d'un autre genre : puisque le calcul des probabilités et la statistique mathématique ont précisément pour tâche essentielle de diminuer, d'étudier ou de quantifier les erreurs (ou les variations ou les déviations), un autre mot est nécessaire pour désigner ce que l'on pourrait appeler une « super-variation » ou une « super-erreur ». Ce n'est d'ailleurs sans doute pas un hasard si Karl Pearson, quand il a voulu étudier les distributions de fréquence anormales ou hétérotypiques, a dû de la même manière recourir à un autre concept, celui de « *skewness* », dont nous avons vu qu'il pouvait être considéré comme un synonyme acceptable du mot « biais », et qui visait précisément à prendre en compte les déviations trop grandes pour être prises en compte par la loi normale et qui, avant Pearson, étaient en général éliminées de l'analyse.

Néanmoins, ce problème apparait comme plus ancien que les analyses de Galton ou Pearson, et semble s'être posé aux fondateurs du calcul des probabilités, et spécifiquement de la théorie des erreurs. Si le mot « biais » n'apparait pas explicitement à cette époque, c'est-à-dire au milieu du XVIIe siècle, l'idée nous semble quant à elle bien présente. L'exemple le plus frappant se situe sans doute dans l'ouvrage de Thomas Simpson et dans sa critique par Thomas Bayes. Pour résumer le propos de Thomas Simpson, celui-ci, s'inscrivant dans le sillage des analyses d'Abraham de Moivre, propose dans une lettre[73] adressée en 1755 au président de la *Royal Society*, Earl Macclesfield, de prendre « la moyenne de nombreuses observations », plutôt que de se fonder sur une seule observation bien faite (« *taken with due care* » in Simpson, 1755, p. 82-83, cité dans Stigler,1986,p. 90) afin d'estimer correctement par exemple la position d'un corps astronomique et de « diminuer les erreurs qui naissent de l'imperfection des instruments, et des organes des sens » (Simpson, 1755, p. 82-83, cité dans Stigler, 1986, p. 90) [74]. Or, Thomas Bayes, dont chacun connait l'influence que ses travaux, et notamment son théorème, ont eu et continuent d'avoir sur la théorie des probabilités, va critiquer la proposition de Simpson dans une lettre adressée au physicien John Canton. Il pose clairement le problème des erreurs, et, selon notre interprétation, du biais :

> « Selon lui [Simpson], en multipliant nos observations et en prenant la moyenne, nous diminuons toujours la probabilité de n'importe quelle erreur donnée, et ce de façon très rapide… Néanmoins, que les erreurs qui naissent de l'imperfection des instruments et des organes des sens soient ainsi réduites à néant ou presque seulement en multipliant le nombre d'observations me semble complètement incroyable. Au contraire, plus vous faites d'observations avec un instrument imparfait, plus il semble certain que l'erreur dans votre conclusion sera proportionnelle à l'imperfection de l'instrument que vous utilisez. (…) Et je pense que c'est manifestement ce qu'il se passe quand nous observons avec

[73] Simpson, Thomas, "A letter to the Right Honorable George Earl of Macclesfield, President of the Royal Societyn on the advantage of taking the mean of a number of observations, in practical astronomy.", *Philosophical Transactions of the Royal Society of London,* 49, p. 82-93. Cité dans Stigler, Stephen M., *The History of Statistics: the Measurement of Uncertainty before 1900*, Cambridge, Mass, Belknap Press of Harvard University Press, 1986, p. 88-98.

[74] Il peut être utile de rappeler ici que la loi des erreurs de Gauss, ou loi normale, centrale dans l'histoire du calcul des probabilités et des statistiques, n'apparaitra qu'en 1809. Stigler montre comment la théorie de Simpson constitue un jalon important dans le développement de la courbe de Gauss-Laplace.

des organes ou des instruments imparfaits ». (Cité dans Stigler, 1986, p. 94-95).

Le problème de Simpson est en effet qu'il suppose que « les chances d'erreurs de même magnitude, en excès ou en défaut, sont en moyenne à peu près égales » (Simpson, 1755, p. 83, cité dans Stigler, 1986, p. 95). Or, cette hypothèse est loin d'être certaine. Mis apparemment au courant de la critique de Bayes, Simpson va proposer une version révisée de sa lettre à Earl Macclesfield et transformer une affirmation en deux suppositions :

> « 1. Qu'il n'y ait rien dans la construction, ou dans la position de l'instrument, qui fasse que les erreurs aient toujours tendance à aller dans la même direction, mais que les chances respectives qu'elles aient lieu par défaut ou par excès soient précisément ou presque identiques.
>
> 2. Qu'il existe certaines limites assignables entre lesquelles ces erreurs sont supposées tomber ; limites qui dépendent de la qualité de l'instrument et de l'habileté de l'observateur. »[75]

Cette description des conditions d'application de la théorie des erreurs par l'un de ses premiers théoriciens nous semble particulièrement pertinente pour définir ce que les statisticiens plus contemporains comme Pearson ou Fisher semblent entendre à travers le concept de biais : le biais est une erreur systématique, et non aléatoire, au sens où elle a tendance à aller toujours dans la même direction, c'est-à-dire soit toujours en excès (au sens où l'estimation est trop haute) , soit toujours en défaut (au sens où elle est trop basse) ; mais aussi au sens où elle excède certaines limites assignables, liées à l'instrument ou l'observateur, auxquels on pourrait ajouter la variation biologique. Un biais est donc une erreur ou une quantité d'erreurs (ou de variations) qui s'accumulent, au lieu de s'annuler mutuellement, ce qui conduit à une estimation faussée, ou bien, dans un contexte galtonien, à une variation telle qu'elle est n'est plus une variation mais une spéciation.

Quant à l'élément de rupture dans la formation progressive de ce concept, elle intervient essentiellement avec Fisher : si ce dernier s'inscrit dans la continuité de cette conception du biais comme « supervariation » (que l'analyse de la variance permet

[75] Simpson, Thomas, *Miscellaneous Tracts on some Curious, and Very Interesting Subjects in Mechanics, Physical-Astronomy and Speculative Mathematics*, Londres: J. Nourse, 1757. Cité dans Stigler, 1986, p. 95.

précisément de faire encore mieux ressortir), il lui assigne une place spécifique dans l'économie conceptuelle qu'il fonde à travers la notion de plan d'expérience. Le biais devient en effet un synonyme de sélection ou d'arrangement systématique, et un antonyme de randomisation, au sens où la randomisation permet d'éviter les biais : en effet, il s'agit via ce procédé de faire que les chances de variation, dans un sens ou dans l'autre, soient à peu près égales. En d'autres termes, il s'agit de rendre les évènements équiprobables, ou comme dirait Laplace les « cas également possibles »: c'est ce que l'on a appelé le « principe de raison insuffisante », renommé par John Maynard Keynes en « principe d'indifférence »[76]. A l'inverse l'arrangement systématique fait que la variation va toujours dans le même sens, et va rendre certains cas plus possibles que d'autres.

La question du biais va alors devenir, au milieu des années 1930, étroitement liée au problème d'échantillonnage, toujours dans le cadre global du plan d'expérience, comme l'attestent la parution de l'article de Neyman en 1934 sur le problème de la méthode représentative[77] ou encore celui de Yates sur différents exemples d' « échantillonnages biaisés »[78]. La portée du plan d'expérience est néanmoins élargie puisqu'il ne s'agit plus ici seulement d'agriculture mais aussi de populations humaines. Trois ans plus tard, un problème similaire de sélection de l'échantillon se posera pour Austin Bradford Hill, le fondateur de l'épidémiologie moderne. C'est vers son œuvre qu'il convient dorénavant de se tourner afin de voir comment la notion de biais va progressivement être introduite dans l'épidémiologie.

[76] Keynes, John Maynard, *A Treatise on Probability.* Londres, Macmillan and Co., 1921. Voir le chapitre IV: « The principle of indifference", p. 41–64.
[77] Neyman, Jerzy, « On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection », *Journal of the Royal Statistical Society*, vol. 97 / 4, 1934, p. 558-625.
[78] Yates, F., « Some examples of biased sampling », *Annals of Eugenics*, vol. 6 / 2, 1935, p. 202–213.

CHAPITRE 2 : L'INTRODUCTION DU CONCEPT DE BIAIS DANS L'EPIDEMIOLOGIE

2.1 A.B. Hill et les *Principles of Medical Statistics*: le biais comme problème de sélection.

2.1.1 Comment disposer d'un échantillon représentatif ? De l'allocation alternée à la randomisation.

L'ouvrage d'Austin Bradford Hill, *Principles of Medical Statistics*, est considéré par beaucoup d'épidémiologistes et d'historiens de l'épidémiologie comme celui qui a le plus influencé l'épidémiologie au vingtième siècle. Publié sous la forme d'articles dans la revue *The Lancet* à partir de 1937, année où il est aussi publié sous la forme d'un « *textbook »,* il ne connaitra pas moins de 9 éditions sous le même titre, la dernière en 1971, jusqu'à sa douzième et dernière édition en 1991, dirigée par le fis d'A.B. Hill, David Hill, qui s'intitule *Bradford Hill's Principles of Medical Statistics*. Farewell et Johnson [1] le considèrent comme le « texte le plus connu sur ce sujet » qui a apporté « une contribution mondiale à la compréhension et à l'enseignement des statistiques médicales durant les 70 dernières années »[2], une influence sans commune mesure selon eux avec les quelques soixante-dix « *textbooks »* parus entre les années 1820 et les années 1930 sur le sujet des statistiques vitales et médicales. Dès lors, il semble logique d'aller chercher si le mot « biais » apparait dans cet ouvrage, et si tel est le cas, le sens que lui donne A.B. Hill. Au total, en se fondant sur la cinquième édition de l'ouvrage, parue en 1950[3], nous dénombrons quinze occurrences du mot « biais » sur les 290 pages de texte de l'ouvrage, ce qui est finalement assez peu. Le mot lui-même n'apparait d'ailleurs pas dans l'index de l'ouvrage, contrairement au mot « sélection »[4], dont il est d'après Hill un synonyme.

[1] Farewell, Vern et Johnson, Anthony, « The origins of Austin Bradford Hill's classic textbook of medical statistics », *Journal of the Royal Society of Medicine*, vol. 105 / 11, 2012, p. 483–489.
[2] Farewell et Johnson, 2012, p. 488.
[3] Hill, A. Bradford, *Principles of Medical Statistics*, 5ème édition, Londres: The Lancet, 1950. Sauf indication contraire explicite, et pour ne pas alourdir inutilement notre texte, les passages de l'œuvre de Hill que nous citons sont traduits par nos soins.
[4] Le mot « sélection » apparaît quant à lui trente fois dans l'ouvrage.

La première occurrence apparait dans le premier chapitre, intitulé : « Le but de la méthode statistique », soit le premier article publié dans le *Lancet* le 2 janvier 1937. Hill commence ainsi son article en s'interrogeant sur l'utilité des statistiques. Puis il constate qu'il y a de plus en plus d'articles dans les revues médicales, dont « l'essence est purement statistique » (Hill, 1950, p.1) et de plus en plus de travailleurs (« *workers* ») qui « cherchent à appliquer la méthode numérique d'analyse à leurs données obtenues dans le champ de la médecine clinique[5] », ainsi que dans « d'autres branches de la médecine », ajoute-t-il lors de la publication de l'ouvrage. Or, d'après Hill, ces articles sont souvent truffés d'erreurs, faute d'une connaissance des principes de base, qui sont « fréquemment oubliés ou ignorés ». C'est pourquoi Hill, souvent sollicité en tant qu'expert statisticien pour valider des études faites par des médecins avant publication, et manifestement lassé de devoir analyser des données erronées et des enquêtes mal planifiées[6], entend apprendre les principes statistiques de base à ceux qui travaillent dans ce domaine. C'est en tout cas le projet tel qu'il le présente dès la préface à la première édition de son ouvrage, où il affirme que le seul moyen d'échapper (« *only one way of escape* », in Hill, 1950, *Preface to first edition*, p. VII) à ces problèmes est que « le travailleur dans le champ des problèmes médicaux, qu'il s'agisse de la clinique ou de la médecine préventive, doit lui-même savoir quelque chose de la technique statistique, aussi bien au niveau des plans expérimentaux que de l'interprétation des figures » (Hill, 1950, *Preface to first edition*, p. VII). C'est pourquoi son ouvrage entend présenter « aussi simplement que possible les méthodes statistiques dont l'expérience [lui] a montré qu'elles étaient les plus utiles pour résoudre les problèmes qui concernent les travailleurs médicaux » (Hill, 1950, *Preface to first edition*, p. VII.).

La première partie de l'article est consacrée à la définition des statistiques. Hill commence par faire une distinction importante entre le travailleur de laboratoire d'un côté et de l'autre le clinicien ou celui qui étudie la société, ou autrement dit entre l'expérimentation proprement dite et l'observation :

> « Alors que le travailleur de laboratoire peut fréquemment exclure les variables qui ne l'intéressent pas et focaliser son attention sur un ou plusieurs facteurs

[5] «*there is an increasing number of workers who endeavour to apply numerical methods of analysis to their records obtained in clinical medicine* », in Hill, 1950, p. 1.

[6] «*I am only persuaded that such is not the case by the recurrence of these mistakes and the neglect of these elementary principles, a feature with which every professional statistician is familiar in the papers submitted to him by their authors for "counsel's opinion.* » in Hill, 1950, p. 2-3.

> contrôlés à la fois, le sociologue ou le clinicien [« the *clinician and social worker* »] sont obligés d'utiliser des données [« *records* »] dont ils savent qu'elles sont susceptibles d'avoir été influencées par des facteurs qu'ils ne peuvent pas contrôler mais qui doivent essentiellement être pris en compte. L'essence de la méthode statistique consiste dans l'élucidation de ces multiples causes. » (Hill, 1950, p. 3).

Après avoir donné l'essence de la méthode statistique, qui consiste donc dans l'« élucidation des effets de multiples causes », il donne ensuite la définition, citant Yule, des statistiques et de la méthode statistique:

> « Par statistiques, ainsi, nous entendons des « données quantitatives affectées dans une large mesure par une multiplicité de causes », et par méthode statistique, des « méthodes spécifiquement adaptées à l'élucidation de données quantitatives affectées par une multiplicité de causes » (Hill, 1950, p. 3).

Il s'agit donc bien, par la méthode statistique, et contrairement à la méthode expérimentale en laboratoire où le chercheur peut choisir quelle variable étudier, de réussir à déterminer, à partir de données quantitatives, quels sont les différents facteurs ou les différentes variables qui influencent ces données, ces facteurs ou variables n'étant précisément pas contrôlables par le clinicien tout en devant être pris en compte.

Pour cela, Hill préconise un principe simple : il faut égaliser (« *equalise* ») les groupes que l'on compare de manière à ce que seule la variable que l'on veut étudier distingue les deux groupes entre eux (Hill prend l'exemple d'un traitement contre la rougeole). Et pour que cela soit efficace il faut le faire dès le début (*ab initio*) de l'expérience, c'est-à-dire au moment de sa planification, ou, si ce n'est pas possible, par la méthode d'analyse elle-même[7]. Néanmoins, Hill avertit ses collègues médecins de l'importance de bien planifier dès le départ l'expérimentation :

> « Se reposer sur la méthode statistique pour éliminer les facteurs perturbateurs au moment de la finalisation du travail constitue une grave erreur. *Aucune* méthode statistique ne peut compenser une expérimentation mal planifiée ». (Hill, 1950, p. 3-4).[8]

[7] « *If we have been unable to equalise the groups ab initio we must equalise them to the utmost extent by the mode of analysis.* », in Hill, 1950, p. 3.

[8] C'est Hill qui souligne.

C'est pourquoi il va consacrer un long développement à la planification et l'interprétation des expérimentations (« *Planning and Interpretation of Experiments* »), où il cite d'ailleurs l'ouvrage de Ronald Fisher, *The Design of Experiments*, paru deux ans avant le premier article de Hill dans le *Lancet*[9]. Selon lui, l'essence du problème dans une expérimentation simple est de s'assurer au préalable qu'autant que possible, « le groupe traité et le groupe contrôle soient identiques sous tous les aspects *pertinents* »[10], ce qu'il résume par une formule, qui revient comme un leitmotiv tout au long de son ouvrage : « *comparing like with the like* », ou autrement dit, il faut comparer ce qui est comparable. Ainsi le problème principal de toute expérimentation est d'abord un problème de comparabilité des groupes : il faut que ces groupes ne diffèrent autant que possible que par le fait d'avoir ou non reçu le traitement étudié, même si l'on ne peut jamais être « *certain* de ne pas avoir négligé un facteur pertinent ou qu'un facteur qui ne peut ni être prévu ni être identifié soit présent »[11]. En effet, il y a toujours selon lui un risque de confusion :

> « Si nous découvrons que le Groupe A diffère du Groupe B en fonction de telle caractéristique, par exemple le taux de mortalité, pouvons-nous être certains que la différence est due au fait que le Groupe A a été inoculé (par exemple) et le Groupe B ne l'a pas été ? Sommes-nous certains que le Groupe A ne diffère pas du Groupe B en fonction d'autres caractéristiques pertinentes pour notre problème lié à la présence ou à l'absence d'inoculation ? » (Hill, 1950, p. 5).

Dès lors, le premier et principal problème de tout statisticien ou médecin qui veut conduire ce genre d'expériences consiste dans la sélection initiale des sujets et groupes de sujets qui vont participer à l'étude :

> « La raison pour laquelle, dans le cadre d'expérimentations sur le traitement d'une maladie, l'allocation alternée des cas aux groupes traités et aux groupes non-traités est souvent satisfaisante est qu'aucun biais conscient ou inconscient

[9] Il est probable que Hill ait eu connaissance des travaux de Fisher sur la randomisation, puisque Hill et Fisher se connaissaient depuis au moins 1929. Néanmoins, les raisons pour lesquelles il a introduit la randomisation ne sont pas des raisons fishériennes. A ce sujet, voir Armitage, 2003, p. 927, et les autres articles du même numéro de l'*International Journal of Epidemiology* de décembre 2003.

[10] « *to ensure beforehand that, as far as is possible, the control and treated groups are the same in all relevant respects.* », in Hill, 1950, p. 4. C'est Hill qui souligne.

[11] « *We can never be certain that we have not overlooked some relevant factor or that some factor is not present which could not be foreseen or identified.* », in Hill, 1950, p. 4. C'est Hill qui souligne

ne risque de s'immiscer dans la sélection des cas, comme cela arrive souvent dans n'importe quelle sélection de cas… »[12]

La première occurrence du mot « biais » dans le texte de Bradford Hill intervient donc dans le contexte de la méthode de sélection des groupes : en adoptant la méthode de l'allocation alternée des cas, répartition qui se fait donc au hasard (« *random allotment »*), cela permet d'éviter que « n'entrent des biais conscients ou inconscients », ce qui arrive dans « n'importe quelle sélection de cas ». Il est donc ici fait clairement référence non à un sens statistique de distorsion systématique, mais bien au sens commun du terme qui est d'ordre psychologique : en effet, celui qui va sélectionner les sujets, c'est-à-dire ici le médecin (puisque c'est à la profession médicale que l'ouvrage de Hill s'adresse) peut être consciemment ou inconsciemment influencé par un certain nombre de préjugés ou d'idées préconçues, préjugés ou préférences qui risquent de le conduire à sélectionner les cas en fonction de critères subjectifs qui risquent de compromettre précisément la comparabilité des groupes. Hill prévient ainsi, dans un paragraphe ajouté[13] par rapport à l'article original de 1937 :

> « Il faut comprendre que la règle de l'allocation alternée nécessite une stricte adhésion – il ne doit pas y avoir de « jonglage » avec l'ordre dans lequel les patients sont inclus dans l'essai afin d'assurer que Me A ou M. B aillent dans un groupe particulier. Tout jonglage de ce genre ruine le caractère hasardeux de la procédure et rend suspecte, voire complètement invalide, la comparaison ultérieure des résultats » (Hill, 1950, p. 5).

Ainsi, si un médecin qui entend faire une expérimentation commence à « jongler » avec la règle d'allocation des patients, pour s'assurer par exemple que telle personne soit dans le groupe traité, cela risque de rendre l'étude non seulement « suspecte » mais aussi « complètement invalide ». En effet, si tel est le cas, toute comparaison est impossible puisque l'expérimentateur ne compare pas ce qui est comparable, au sens où la procédure d'allocation ne s'est pas faite au hasard.

[12] « *The reason why in experiments in the treatment of disease the allocation of alternate cases to the treated and untreated groups is often satisfactory, is because no conscious or unconscious bias can enter in, as it may in any selection of cases…* », in Hill, 1950, p. 5.

[13] Ce paragraphe a été ajouté lors de la cinquième édition, publiée en 1950. La quatrième édition date de 1948.

2.1.2 De la randomisation à la procédure en aveugle : la lutte contre les biais subjectifs.

Dans ce paragraphe, ajouté dans la cinquième édition de 1950, apparait une seconde occurrence du mot « biais », avec un sens similaire. Il préconise ainsi, pour éviter justement tout biais, des tables d'échantillonnage aléatoire des nombres pour garantir que la répartition des sujets inclus dans l'étude se fasse réellement au hasard. Néanmoins, ce biais n'intervient plus au moment de la répartition des sujets après leur inclusion, mais avant même leur inclusion dans l'étude afin d'éviter que le médecin, connaissant la destination de tel patient (groupe traité ou groupe non-traité), voie son jugement modifié ou « biaisé » par cette information, et refuse d'inclure ce patient au motif par exemple qu'il serait dans le groupe non-traité :

> « Une telle méthode serait spécialement importante si le travailleur qui accepte ou refuse d'inclure un patient dans un essai thérapeutique pense que son jugement pourrait être biaisé en sachant dans quel groupe de traitement le patient est susceptible d'entrer – comme il le ferait dans le cas de la méthode alternée » [14]

Cette référence à des tables d'échantillonnage aléatoire renvoie à la méthode utilisée par Hill lors de l'essai sur la streptomycine réalisé en 1947-1948 par le *Medical Research Council* (*MRC*) pour soigner la tuberculose pulmonaire, considéré comme le premier véritable essai clinique randomisé de l'histoire. Dans cet essai, Hill, qui était à la fois l'expert en statistique de l'étude mais aussi le directeur de l'Unité de Recherche Statistique du *MRC*, a ainsi mis en place, ce qui constitue une nouveauté pour l'époque, sa « nouvelle répartition des sujets par l'échantillonnage aléatoire des nombres »[15]. D'après P. D'Arcy Hart, qui fut, en tant que membre de l'équipe scientifique du MRC, le secrétaire du *MRC Streptomycin in Tuberculosis Trials Committee*, cette nouvelle technique remplit une fonction précise [16] :

[14] « *Such a method would be specially important if the worker accepting or refusing a patient for a therapeutic trial thought that his judgment might be biassed through knowing which treatment group the patient was to enter _as he would do with the alternate method.* », in Hill, 1950, p. 6.

[15] « *his novel allocation by random sampling numbers*». Hart, P. D'Arcy, « A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s », *British Medical Journal*, vol. 319 / 7209, 1999, p. 572-573.

[16] Hart, 1999, p. 573

> « Bradford Hill a mûri ces idées d'allocation pendant plusieurs années (avec la randomisation qui vient remplacer la méthode alternée dans le but de mieux cacher la procédure d'allocation) »[17].

Ainsi, l'introduction de la randomisation en médecine, que l'on attribue généralement à Hill, ne remplit pas en fait la même fonction chez Fisher que chez Hill : le but essentiel de Fisher est d'assurer la validité du test statistique, alors que pour Hill il s'agit essentiellement de cacher la procédure aux médecins et à l'enquêteur. Ce raffinement (et même d'après Hart ce « changement scientifique ») de la méthodologie de Hill quant à l'allocation des patients dans les groupes, n'a pas été introduit « pour quelque ésotérique raison statistique », mais pour « contrôler les biais de sélection », comme le souligne son collègue Richard Doll[18]. Il s'agit simplement de « cacher la procédure d'allocation » des patients à celui qui organise l'étude. Hill dit ainsi :

> « Dans une telle position le patient doit d'abord être accepté et l'allocation doit ensuite être déterminée ou pré-déterminée de façon aléatoire, des listes aléatoires peuvent être construites au préalable par une tierce personne et rester inconnues du travailleur qui réalise l'essai. » (Hill, 1950, p. 6.)

Il s'agit bien donc, bien de cacher la procédure de répartition des patients dans les groupes (traités/non traités) à la personne qui organise l'essai, pour ne pas qu'il puisse la manipuler en fonction justement de ses préjugés ou de ses inclinations, donc de ses « biais », conscients ou inconscients. En 1952, Hill se fait d'ailleurs plus explicite sur les raisons qui l'ont conduit à adopter la procédure de randomisation, et qui sont au nombre de trois :

> « Adhérer strictement à cette méthode – et je dois dire avec force que c'est une condition *sine qua non* – garantit trois choses : que ni nos idiosyncrasies personnelles (nos préférences et nos réticences conscientes ou inconscientes) ni notre manque de jugement n'entrent dans la constitution des différents groupes de traitements – l'affectation a été soustraite à notre contrôle et les groupes ne sont donc pas biaisés ; cela écarte le danger, inhérent à une affectation basée sur le jugement personnel, lorsque, croyant que nous pouvons

[17] « *Bradford Hill had formed his allocation ideas over several years (with randomisation replacing alternation in order to better conceal the allocation schedule).* », in Hart, 1999, p. 573.

[18] Doll, Richard, *Presentation at 'Clinical Trials : Into the New Millenium'*. St Anne's College, Oxford, 25 septembre 2000. Cité dans Chalmers, Iain, « Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments », *International Journal of Epidemiology*, vol. 30, 2001, p. 1156-1164.

> être biaisés dans nos jugements, nous essayons de prévenir ces biais pour les exclure et, ce faisant, nous sur-compensons en « tordant le bâton dans l'autre sens », introduisant un déséquilibre contraire ; enfin, si nous avons utilisé une affectation aléatoire, quand vient le moment de la publication, le critique le plus sévère ne peut pas dire qu'il est fort probable que les groupes soient biaisés de différentes façons à cause de nos prédilections ou de notre stupidité. Une fois qu'il a été décidé qu'un patient a le bon type pour être inclus dans l'essai, la méthode d'affectation par randomisation enlève toute responsabilité au clinicien »[19]

Ainsi, selon Hill, la randomisation permet d'éliminer toute trace de subjectivité, et même toute responsabilité, du clinicien ou du chercheur, par son caractère mécanique et procédural : elle garantit aussi l'acceptation ou au moins l'acceptabilité de l'étude au moment de sa publication. Comme le dit Harry Marks : « La mécanisation de la sélection des groupes fut donc organisée pour éliminer les biais introduits par l'activité des cliniciens-chercheurs »[20]. Neutraliser cette subjectivité, en renvoyant à une procédure d'allocation des patients intégralement mécanisée, était la condition *sine qua non* pour garantir la validité des résultats auxquelles elle était parvenue, et par conséquent l'impartialité et l'objectivité de l'étude, donc sa scientificité.

2.1.3 Biais ou sélection ?

Pourtant il est tout aussi manifeste que le concept de biais n'est précisément pas encore un concept mais renvoie à la conception commune du mot de l'époque. D'ailleurs, si l'on se réfère à la deuxième occurrence du mot dans les quatre premières éditions des *Principles of Medical Statistics*[21], qui apparait dans le deuxième chapitre de l'ouvrage de Hill, donc le deuxième article qu'il publie dans *The Lancet*, le mot « biais » est considéré comme un synonyme d'un autre mot : celui de sélection. Or,

[19] Hill, A. Bradford, « La philosophie de l'essai clinique », in Leplège, Alain, Bizouarn, Philippe et Coste, Joël, *De Galton à Rothman: les grands textes de l'épidémiologie au XXe siècle*, Paris, Hermann, 2011, p. 117-129, Trad. Fr. Jean-Paul Amann. L'extrait cité est à la page 124.

[20] Marks, Harry M., « Confiance et méfiance dans le marché : les statistiques et la recherche clinique (1945-1960) », *Sciences sociales et santé*, vol. 18 / 4, 2000, p. 9-27.

[21] Dans la cinquième édition, cette deuxième occurrence devient la sixième occurrence du mot « biais », ce qui dénote une évolution de la conception de Hill, évolution sans nul doute liée à l'évolution de sa carrière qui fait de lui le statisticien en chef de la *MRC*, poste à lourde responsabilité puisqu'il s'agit d'y évaluer, entre autres, l'efficacité des médicaments.

c'est bien ce concept de « sélection » que Hill choisit comme titre de son deuxième article, ce qui manifeste l'importance qu'il attache à ce problème : le premier article étant consacré à la présentation générale de la méthode statistique et de son but, le deuxième article va rentrer dans le détail de la méthode statistique en médecine. Or, le problème central de toute étude statistique est d'abord non pas tant un problème de comparabilité des échantillons entre eux qu'un problème de représentativité de l'échantillon par rapport à la population (ou l'univers) dont il est issu. En effet:

> « *In medical statistics we are nearly always working with samples drawn from large populations. (…) If we wish to argue from this sample to the universe from which it was drawn, to deduce that what is true of our sample of patients is therefore true of the general run of patients, then we must consider very carefully whether in fact our sample is fully representative of all patients, and not in any way biased or "selected". It is important to be clear on the meaning the statistician attaches to the word "selected". By a selected sample he denotes a sample which is not representative of the universe which it is drawn. The selection may have been deliberate, in which case the form of selection is known and the lack of comparability between the sample and the universe is usually perfectly clear. (…) we are clearly not comparing like with the like… (…) More often, however, the "selection" is not deliberate but is quite unforeseen or is unrealised. (…) It may be that with care that selection might have been avoided; often it is unavoidable. Its possible presence cannot be too carefully remembered or taken into account in interpreting statistics.* » (Hill, 1950, p.13).

Le problème qui est ici posé par Hill est celui de l'inférence statistique : comment être sûr que les résultats obtenus en étudiant un échantillon de la population puissent être valables pour la population (ou l'univers) de laquelle est extrait cet échantillon ? Autrement dit, peut-on « déduire que ce qui est vrai pour un échantillon de patients l'est aussi pour n'importe quel patient lambda » (Hill, 1950, p. 13) ? Ou, en d'autres termes, comment être sûr que cet échantillon n'est pas « biaisé ou ''sélectionné'' » ?

L'usage des guillemets par Bradford Hill quand il utilise le mot « sélection » est intéressant pour au moins deux raisons :

- Cela signifie tout d'abord qu'il emploie le terme non en un sens commun mais en un sens manifestement technique ou scientifique. Il ajoute ainsi tout de suite après qu'il « est important de clarifier la signification que le

statisticien accorde au mot ‘’sélectionné’’ », soulignant par là qu'il entend lui donner un sens particulier. Il le précise aussitôt : « par échantillon sélectionné, [le statisticien] dénote un échantillon qui n'est pas représentatif de l'univers duquel il est extrait » (Hill, 1950, p. 13).

- Ceci nous conduit au second point : pourquoi Hill choisit-il le terme sélection plutôt que celui de biais, tout en leur accordant une signification identique (« *biased or ‘’selected’’* ») ? Cela est d'autant plus problématique qu'il semble donner une connotation particulière à ce concept de sélection : en effet, le sens classique et neutre du mot « sélection » en épidémiologie consiste simplement dans le choix des personnes ou des sujets que le médecin ou le statisticien va inclure dans l'étude. Or, ici il nous dit clairement que par sélection, il entend une sélection biaisée, c'est-à-dire qu'elle n'est pas représentative. Notre interprétation est qu'à cette époque, au moins depuis les années 1930, la notion de sélection est largement débattue parmi les statisticiens, au sein desquels Hill est semble-t-il bien introduit[22], et que la sélection est d'emblée suspecte. Ainsi, F. Yates dit dans son article consacré aux échantillonnages biaisés que « malheureusement, la prétention de celui qui choisit l'échantillon à choisir « un échantillon représentatif » à travers son jugement personnel est largement infondée, et sa sélection est en fait sujette à toutes sortes de biais, psychologiques et physiques » (Yates, 1935, p. 202). Dès lors, « afin d'éviter ces biais et de fournir une estimation de la représentativité de l'échantillon, c'est-à-dire de la fluctuation d'échantillonnage [« *sampling error* »] », il est nécessaire de recourir à l'échantillonnage aléatoire (« *random sampling* »), dont Yates montre, comme Neyman dans son article de 1934, la supériorité par rapport à la méthode de la « sélection intentionnelle » (« *purposive selection* »). Dès lors, il est fort possible que Hill considère les notions de biais et de sélection comme strictement équivalentes, toute sélection qui n'aurait pas lieu au hasard, par un procédé mécanique, étant nécessairement biaisée, c'est-à-dire entachée de subjectivité.

[22] Peter Armitage dit ainsi que Hill « avait pénétré dans le cercle restreint des statisticiens des années 30, et était un collègue de J. Oscar Irwin, un fishérien enthousiaste », in Armitage, 2003, p. 927.

Ainsi, pour illustrer la notion de « sélection », Hill donne comme premier exemple de sélection la rubrique des naissances dans le *Times* : il nous dit qu'en compilant les annonces de naissances publiées dans le *Times*, il aboutit à un *sex ratio* de 1089 garçons pour 1000 filles. Or d'après les chiffres officiels, le *sex ratio* en Angleterre et au Pays de Galles ne dépasse pas 1050 garçons pour 1000 filles. C'est donc que l'échantillon en question n'est pas représentatif : ainsi, ceux qui annoncent la naissance de leur enfant dans le *Times* ont peut-être tendance à le faire plus facilement si c'est un garçon qu'une fille, ou que le *sex ratio* diffère en fonction des classes sociales. Il s'agit donc ici d'une sélection différentielle. Hill a-t-il choisi ce terme de sélection, au détriment de celui de biais, pour lui donner une connotation plus scientifique, dans un contexte post-darwinien où Galton ou Pearson mais aussi Fisher ont joué un rôle considérable ? Nous n'avons trouvé nulle part dans l'œuvre de Hill une réponse à cette question. Rétrospectivement, il est bien ici question de ce que les épidémiologistes modernes appellent un biais de sélection, mais pour Hill, cette notion de « biais de sélection » est clairement un pléonasme.

Dans la suite de son ouvrage, c'est bien le sens de « biais » au sens de « sélection », entendu donc comme un défaut de représentativité de l'échantillon par rapport à l'univers ou à la population source d'où cet échantillon est extrait, qui va prédominer. Ainsi, à la page 20, toujours dans ce chapitre sur la notion de sélection, Hill, après avoir donné comme exemple de sélection le *sex ratio* à la naissance, les statistiques hospitalières, le jour du traitement, l'auto-sélection (« *self-selection* »), s'intéresse aux questionnaires. Il dit ainsi:

> « *Inquiries carried out by means of questionnaires are par excellence those in which selection must be suspected.* » (Hill, 1950, p. 20)

En effet, selon lui, le problème des questionnaires est que seule une petite proportion des personnes à qui le questionnaire est envoyé répond. Or,

> « *There can never be the slightest certainty that the individuals who choose to reply are a representative sample of all the individuals approached; indeed very often it is extremely unlikely that they are representative.* » (Hill, 1950, p. 20)

En se fondant sur un questionnaire envoyé par la revue *The Lancet* aux médecins qui se sont enregistrés en 1930 auprès de l'équivalent anglais du Conseil

de l'ordre des Médecins pour « mesurer leur succès »[23] , Hill va alors donner à son lecteur un outil pour savoir s'il y a effectivement un biais mais aussi un moyen arithmétique pour corriger ce biais. D'abord, comment est-il possible de savoir s'il y a un biais, c'est-à-dire de savoir si l'échantillon est ou non représentatif ?

> « *It is possible sometimes, however, to see whether the final sample is or is not biassed in certain known respects. For example, suppose the population to be approached consists of all the persons on the medical register at a given time. For each of these persons we may know such characteristics as sex, age at qualification, degrees or other qualifications obtained, type of medical work upon which the person is engaged-general practice, public health, etc. Only 50 per cent of the total population, let us suppose, answer the questionnaire addressed to them all. In the statistical analysis of the available answers we can at least see, and it is of course essential to do so, whether the sample is representative of the universe in relation to the known characteristics of the latter. If 50 per cent of the men and 50 per cent of the women answered, then the sample obtained is not biassed in its sex ratio; but if 60 per cent of the men and only 25 per cent of the women answered, then a bias has been introduced, for the ratio of males to females in the sample is different from the real ratio in the universe. We must make some allowance for that fact in analysing the results and cannot merely use the sample as it stands. (…) By such means we can then determine whether or no certain classes of persons have tended to answer more or less readily than others, and thus know whether or no our sample is biased in these known respects and, if necessary, make allowances for it. While such a check is highly important, indeed essential, it cannot be entirely conclusive. Even if the sample is representative in the known respects, we cannot be sure that those who chose to answer were in other respects representative of the total. For instance, 50 per cent of men and 50 per cent of women may answer, but in each group those who answer may mainly consist of those who feel more deeply upon the questions addressed to them, or be those, mirabile dictu, who like filling in forms. In other words, the sample is correct in its sex proportions, but for neither sex*

[23] Seules 3 questions sont posées à ces jeunes médecins : Quelle branche de la médecine avez-vous choisie ? Qu'est-ce qui vous a conduit à ce choix ? Quel a été l'année dernière votre revenu professionnel ?

do we know that the sample is such that it will accurately express the views of the total men and women originally approached.

The type of simple correction one can sometimes make for a known bias can be demonstrated arithmetically from the following hypothetical figures[24]*:* »

	Male	*Female*	*Total*
Number of persons in the universe, all of whom were sent questionnaires	*10,000*	*2,000*	*12,000*
Number of persons who answered.	*6,000*	*500*	*6,500*
Mean Income reported by those who answered	*£1,200*	*£800*	*£1,169*

Ainsi, en se fondant uniquement sur les déclarations des personnes qui ont répondu, la moyenne des salaires se situe à 1169£. Or, le problème est que cette estimation du salaire moyen est trop haute. En effet, il y a un décalage entre l'échantillon et l'univers dont il est issu quant à la répartition par sexe : dans l'échantillon (c'est-à-dire ceux qui ont répondu), il y a 12 hommes pour 1 femme, « tandis que dans l'univers le vrai ratio est seulement de 5 hommes pour 1 femme (en raison du fait que 60% des hommes ont répondu, contre 25% seulement pour les femmes) »[25].

« *If we are prepared to believe that for both sexes those who answered were a representative cross-section of the total approached, then the correct estimate of the average income of a person must be obtained by "weighting" the observed mean incomes of the sexes by the correct numbers of persons of each sex. Thus we have (10000 x £1200 + 2000 x £800) / 12000 = £1133. In other words, we are accepting the sample figures as giving a true picture for each sex, but must combine them by using the known true proportion of men to women in the*

[24] Hill,1950, p. 21-23

[25] « *whereas in the universe the real ratio is only 5 men to 1 woman (due to the fact that 60 per cent of the men answered and only 25 per cent of the women).* » in Hill, 1950, p. 23.

population sampled, in place of the untrue proportion given by the sample. » (Hill, 1950, p. 23).

Il est donc possible de parvenir à une estimation correcte du salaire moyen à travers une « simple correction pour un biais connu », c'est-à-dire en pondérant l'échantillon (ici, les personnes qui ont répondu au questionnaire) par l'univers dont il est issu (le nombre de personnes à qui l'on a envoyé le questionnaire), la correction portant ici sur la répartition hommes-femmes. Il est intéressant de noter par ailleurs que nous retrouvons ici, quoique sous une forme simplifiée, la question de l'estimateur biaisé qui était celle de Fisher.

Hill va donner un dernier exemple de biais dans ce chapitre consacré à la notion de sélection, celui de l'échantillonnage aléatoire des maisons tel qu'il a été pratiqué lors d'une enquête sur l'incidence et la mortalité de l'épidémie de grippe espagnole en 1918, menée par le Ministère britannique de la santé à Leicester en 1918-1919. L'échantillonnage consistait à faire du porte-à-porte dans cinq quartiers de la ville, en visitant une maison sur cinq. Le problème était que les maisons où personne ne répondait au moment de la visite devaient être ignorées dans le recensement. Or, nous dit Hill, les enquêteurs ont plus de chances de trouver quelqu'un dans les maisons où il y a de jeunes enfants que des adultes seuls. Ceci doit donc affecter la répartition par âge de l'échantillon, avec une surreprésentation des jeunes et une sous-représentation des adultes. Ce biais dans l'échantillon ne peut pas, d'après Hill, être corrigé.

Enfin, dans le résumé de ce chapitre, Hill va apporter un nouvel élément à la notion de sélection, ou de biais :

« *In taking samples," selection" may occur through the operation of various factors. A selected sample is one which is not representative of the universe, in which one member of the universe sampled has more chance of appearing than another, whether that bias be due to deliberate choice or unconscious selection of the members incorporated in the sample.* » (Hill, 1950, p. 25).

Un échantillon biaisé ou "sélectionné" est donc d'après Hill un échantillon qui n'est pas représentatif de l'univers dont il est issu, car « un membre de cet univers d'où est tiré l'échantillon a plus de chances d'apparaître qu'un autre, que ce biais soit dû à un choix délibéré ou à une sélection inconsciente des membres incorporés dans l'échantillon ». Or, « pour être en position de généraliser, il est nécessaire que

l'échantillon soit représentatif de la population à laquelle il appartient[26] ». C'est pourquoi :

> « *In generalising from a sample, or in making comparisons between one sample and another, the possible presence of selection must always be very closely considered.* »[27]

Ainsi, il y deux problèmes liés mais distincts dans l'esprit de Hill quant au problème du biais ou de la sélection :

- Tout d'abord il y a un problème de comparabilité des échantillons entre eux : si l'on veut par exemple faire ressortir une différence quant à l'exposition à un traitement ou à un facteur de risque, il faut que les deux échantillons soient comparables entre eux et il faut donc choisir les échantillons de telle sorte qu'ils soient autant que possibles égalisés ou appariés en termes de distribution par sexe, par âge, etc., et que seule l'exposition à un traitement ou à un facteur de risque distingue les deux échantillons entre eux. Sans comparabilité des groupes, aucune inférence quant à la causalité d'un traitement ou d'un facteur de risque n'est possible. C'est ce que les épidémiologistes appellent aujourd'hui la validité interne.
- Mais il y a aussi un problème de représentativité de l'échantillon par rapport à l'univers dont il est issu, autrement dit par rapport à la population source. Sans représentativité, aucune inférence statistique n'est possible, au sens où l'estimation qui sera faite à partir de l'échantillon ne peut être étendue à la population cible, car l'échantillon issu de la population et la population source elle-même ne sont pas comparables entre eux et n'ont donc pas les mêmes propriétés (en termes par exemple de distribution par âge ou par sexe, etc.). C'est ce que les épidémiologistes appellent aujourd'hui la validité externe.

Dans les deux cas, il s'agit de rendre l'inférence statistique possible, car si l'échantillon n'est pas représentatif, alors « les valeurs calculées à partir de l'échantillon ne peuvent pas être considérées comme des estimations vraies des valeurs dans la population » (Hill, 1950, p. 25). Cette « élimination des biais », pour reprendre le titre de la première partie du chapitre VII, qui a pour titre « Problèmes

[26] « *to be in a position to generalise, the sample must be representative of the population to which it belongs.* », in Hill, 1950, p. 24-25.
[27] Hill (1950), *Op.cit.* p. 25

d'échantillonnage : les moyennes » (« *Problems of sampling : averages »*), constitue ainsi un préalable à toute enquête statistique, ou plutôt un préalable à l'utilisation des outils statistiques qui visent à rendre compte et à contrôler autant que possible les erreurs aléatoires. En effet, les erreurs dues aux biais ne font pas partie de ces erreurs aléatoires : il s'agit d'un autre type d'erreur et, selon Hill, « aucune technique statistique ne peut rendre compte de ce type d'erreur »[28].

Autrement dit, il est possible de dire que le problème du biais ou de la sélection n'est pas un problème véritablement statistique mais un problème pré-statistique : si, en effet, d'après la définition donnée par Hill, qui cite celle de Yule et Kendall, les statistiques renvoient à « des données quantitatives affectées dans une large mesure par une multiplicité de causes » et que les méthodes statistiques sont les « méthodes spécialement adaptées à l'élucidation des données quantitatives affectées par une multiplicité de causes », le problème du biais, en tant qu'il porte sur les données quantitatives elles-mêmes (qui seraient sujettes à caution, en tant qu'elles ne seraient pas représentatives) , n'est précisément pas susceptible d'être traité par les méthodes statistiques, mais apparait bien plutôt comme la condition de possibilité de l'utilisation de ces méthodes statistiques à ces données quantitatives, et donc la condition de possibilité de toute inférence causale. Ici, Hill semble clairement reprendre le sens proprement statistique du mot « biais », puisqu'il s'agit de disposer d'une bonne estimation, c'est-à-dire non biaisée, afin précisément de garantir la validité du test statistique, ou en d'autres termes, d'établir qu'il y une différence entre deux populations qui sont similaires sous les autres aspects, pour ensuite vérifier si cette différence est significative.

C'est pourquoi Hill insiste autant sur l'importance pour celui qui effectue une enquête de toujours garder à l'esprit la possibilité qu'il existe un biais ou une sélection, notamment quand il interprète les résultats, car la plupart du temps, elle est inévitable (« *unavoidable »*) :

> « *More often, however, the "selection" is not deliberate but is quite unforeseen or is unrealised. (…) It may be that with care that selection might have been avoided; often it is unavoidable. Its possible presence cannot be too carefully remembered or taken into account in interpreting statistics. »* (Hill, 1950, p. 13).

[28] « *no statistical technique can allow for that kind of error*. », in Hill, 1950, p. 82.

2.2 Les premières études cas-témoins sur le lien entre tabagisme et cancer du poumon (1950-1952) : biais de sélection, biais d'information et biais d'enregistrement.

Dans l'ouvrage séminal de Hill sur les statistiques médicales, le mot biais est donc présent mais ne semble pas avoir un sens technique, bien qu'Hill le définisse comme un type d'erreur particulier. Hill lui préfère d'ailleurs le terme de « sélection », qui apparait ainsi comme un véritable concept scientifique, bien plus en tout cas que celui de biais, pour désigner un problème de représentativité de l'échantillon par rapport à la population ou l'univers dont cet échantillon est issu et plus largement un problème de comparabilité soit entre deux échantillons, soit entre un échantillon d'une population et cette population. Pourtant, le sens du mot « biais » n'est pas identique en fonction des auteurs qui vont l'utiliser, et il va de même évoluer chez Hill. Il faut de plus ajouter que tous les épidémiologistes ou statisticiens de l'époque n'emploient pas seulement le mot « biais » en des sens différents mais parfois n'utilisent tout simplement pas le mot « biais ». Ainsi le mot « biais » va être utilisé à de nombreuses reprises, et de plus en plus souvent, à partir du début des années 1950 dans le cadre des études épidémiologiques, notamment celles qui concernent le lien entre le tabagisme et le cancer du poumon. Il semble donc nécessaire, sans prétendre à l'exhaustivité, de recenser les principales occurrences de ce mot dans les articles d'épidémiologie qui paraissent à cette époque.

Tout d'abord, c'est en 1950 que paraissent deux articles qui vont avoir un retentissement important, à la fois parmi les statisticiens et les épidémiologistes, mais aussi dans le grand public[29]. Ces deux articles constituent les comptes rendus de deux études épidémiologiques menées l'une aux Etats-Unis par Ernst Wynder et Evarts Graham qui porte sur 684 cas (et 780 témoins) et est publié dans le *Journal of the American Medical Association*[30], et l'autre en Grande-Bretagne par Bradford Hill et Richard Doll qui porte sur 1732 cas (et 743 témoins) publié dans le *British Medical*

[29] Voir à ce sujet: White, Colin, « Research on smoking and lung cancer: a landmark in the history of chronic disease epidemiology. », *The Yale Journal of Biology and Medicine*, vol. 63 / 1, 1990, p. 29-46

[30] Wynder, Ernest L. et Graham, Evarts A., « Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma », *Journal of the American Medical Association*, vol. 143, mai 1950, p. 329-336

Journal[31]. Toutes deux constituent les premières études qui semblent indiquer un lien entre le tabagisme et le cancer du poumon et sont des études cas-témoin et rétrospectives. Ces deux études sont considérées comme pionnières.

2.2.1 Le biais comme sélection : représentativité et comparabilité des échantillons.

Tout d'abord, dans l'article de Doll et Hill, on compte neuf occurrences du mot « biais », qui peuvent être classées en trois catégories : le biais lié à la sélection différentielle des patients, le biais lié à l'enquêteur, mais aussi le biais d'enregistrement (« *recording bias* »). Les occurrences se répartissent de la façon suivante : quatre occurrences pour le biais de sélection, quatre occurrences pour le biais lié à l'intervieweur, et une occurrence pour le biais d'enregistrement.

La première occurrence apparait dès le début de l'article dans la deuxième partie consacrée à l'enquête (« *Present investigation* »), qui est la partie proprement méthodologique de l'article. Le premier problème rencontré, qui coïncide avec la première occurrence du mot « biais », concerne la notification (ou le signalement) des cas. En effet, Doll et Hill ont demandé à vingt hôpitaux de Londres de coopérer avec eux et de leur signaler « tous les patients hospitalisés avec un cancer du poumon, de l'estomac, du côlon, ou du rectum » (Doll et Hill, 1950, p. 740). Or, le problème est que « la méthode de notification est variable » :

> « Dans certains hôpitaux la notification a été faite par le personnel chargé de l'admission en fonction du diagnostic d'admission, dans d'autres par le médecin de l'établissement une fois qu'un diagnostic clinique suffisamment sûr a été posé, dans d'autres encore par le responsable du registre des cancers ou le service de radiothérapie. » (Doll et Hill, 1950, p. 740).

Il est donc « probable qu'aucune de ces méthodes n'a abouti à la notification de tous les cas, mais rien n'autorise à supposer que les individus qui auraient pu échapper à la notification constituent un groupe particulier – c'est-à-dire un groupe sélectionné

[31] Doll, Richard et Hill, A. Bradford, « Smoking and carcinoma of the lung », British Medical journal, vol. 2 / 4682, 1950, p. 739-748. Cet article a été réédité en 1999 par l'Organisation Mondiale de la Santé, sous la référence suivante : Doll, Richard et Hill, A. Bradford, « Tabagisme et cancer du poumon: rapport préliminaire », Bulletin de l'Organisation mondiale de la Santé, 1999, p. 185-197. Tous les passages cités ici en français proviennent de cette traduction. Néanmoins, pour faciliter la lecture, il a semblé plus simple de ne donner les références qu'au texte original publié en 1950.

de manière à biaiser l'enquête –». Nous retrouvons ici le sens de « biais » comme étant, selon Hill, synonyme de « sélection ». La version originale de l'article est plus explicite car Hill parle non d'un groupe « particulier » mais d'un « *selected group -that is, selected in such a way as to bias the inquiry »* (Doll et Hill, 1950, p. 740). Un groupe « sélectionné », ici les cas qui auraient échappé à la notification, aurait pu en effet conduire à biaiser l'enquête dans la mesure où cet échantillon aurait pu être différent de l'échantillon des cas notifiés, ou plutôt, si l'on s'en tient à la définition donnée par Hill dans ses *Principles of Medical Statistics*, aurait pu être un « échantillon qui n'est pas représentatif de l'univers dont il est issu ».

Pourtant, on peut se demander en quoi cet échantillon aurait pu ne pas être représentatif. Ou, autrement dit, en quoi les cas non signalés auraient pu être substantiellement différents de ceux qui l'ont été ? En réalité, le problème, d'après Hill, ne vient pas tant des cas eux-mêmes que de ceux qui les notifient ou les signalent : il nous dit en effet qu'il « rien n'autorise à supposer que les individus qui auraient pu échapper à la notification constituent un groupe particulier » car « les objectifs de l'investigation étaient soit inconnus des responsables de la notification, soit connus d'eux très vaguement »[32]. Ainsi, le mot « biais » renvoie ici à l'idée d'une sélection différentielle des cas qui aurait été faite intentionnellement par les responsables de la notification de manière à modifier ou à biaiser les résultats, avec donc une volonté manifeste de leur part de tromper ou de plutôt de truquer les résultats.

La deuxième occurrence apparait dans la partie suivante consacrée aux données (« *The data »*). Doll et Hill indiquent tout d'abord qu'entre avril 1948 et octobre 1949, 2370 cas de cancer ont été notifiés, mais qu'il « n'a cependant pas été possible d'interroger tous ces patients ». Doll et Hill avaient décidé, avant de commencer, de ne pas inclure dans l'enquête les patients de plus de 75 ans, « vu qu'il était peu probable d'obtenir des antécédents exacts [« *reliable histories »*] en interrogeant des personnes très âgées. » (Doll et Hill, 1950, p. 740). Ces patients étaient au nombre de 150. Puis Doll et Hill ont éliminé les 80 patients dont le diagnostic était incorrect :

[32] « *The method of notification varied; in some it was made by the admitting clerk on the basis of the admission diagnosis, in others by the house-physician when a reasonably confident clinical diagnosis had been made, and in yet others by the cancer registrar or the radiotherapy department. None of these methods is likely to have resulted in complete notification, but there is no reason to suppose that those who escaped notification were a selected group as the points of interest in the investigation were either not known or known only in broad outline by those responsible for notifying.* », in Doll et Hill, 1950, p. 740.

> « Ces deux groupes enlevés, il restait 2140 patients à interroger. Parmi eux, 408 n'ont pas pu être interrogés pour les raisons suivantes : déjà sortis, 189; trop malades, 116; décédés, 67; trop sourds, 24; incapables de s'exprimer clairement en anglais, 11; pour un patient, l'enquêtrice a abandonné l'entretien, les réponses du patient paraissant totalement dépourvues de vraisemblance. Aucun patient n'a refusé d'être interrogé. La proportion de patients non interrogés est élevée, mais rien n'indique apparemment que les résultats pourraient s'en trouver biaisés »[33].

Ainsi, alors même qu'environ 20% de l'échantillon de cas n'a pas été interrogé, ce qui, de l'aveu même de Hill et Doll, est « élevé », les auteurs considèrent qu'il n'y a aucune raison apparente qui pourrait indiquer que les résultats soient biaisés, sans néanmoins donner de justification à cette assertion. Il est probable que par « raison apparente », ils entendent le fait que ces personnes qui n'ont pas été interrogées ne différaient pas substantiellement des personnes interrogées par leurs caractéristiques et donc que ce groupe n'était pas sélectionné, c'est-à-dire à la fois comparable au groupe des personnes interrogées et représentatif de l'univers dont il était issu.

La troisième occurrence du mot « biais » intervient dans la dixième partie, intitulée « Interprétation des résultats ». Contrairement à ce que l'on pourrait penser de prime abord, il ne s'agit pas d'une partie conclusive, mais d'une partie qui fait office de transition. Il ne s'agit pas de remettre en cause les résultats de l'enquête, qui établissent une association directe entre tabagisme et cancer du poumon, mais de mettre à l'épreuve les résultats et d'envisager d'autres explications possibles à ceux-ci :

> « Si, d'après les tableaux précédents, il ne semble pas douteux qu'il existe une association directe entre le tabagisme et le cancer du poumon, il est nécessaire d'envisager d'autres explications à ces résultats. Pourraient-ils être dus au choix d'un échantillon non représentatif de patients atteints de cancer du poumon ou au choix d'une série témoin qui ne lui est pas réellement comparable ? Pourraient-ils résulter d'une estimation exagérée de leur tabagisme par des patients pensant avoir une maladie susceptible d'être attribuée au tabagisme ?

[33] « *Deducting these two groups leaves 2,140 patients who should have been interviewed. Of these, 408 could not be interviewed for the following reasons: already discharged 189, too ill 116, dead 67, too deaf 24, unable to speak English clearly 11, while in one case the almoner abandoned the interview as the patient's replies appeared wholly unreliable. No patient refused to be interviewed. The proportion not seen is high, but there is no apparent reason why it should bias the results.* » in Doll et Hill, 1950, p. 740.

Pourraient-ils résulter d'un biais dû aux enquêtrices lors du recueil et de l'interprétation des antécédents ? »[34]

Pour notre propos cette citation est particulièrement intéressante car elle est problématique. En effet, Hill et Doll font clairement référence à trois types de biais, au sens où les auteurs les qualifient eux-mêmes de biais :

- Il y tout d'abord le biais de sélection : le problème de la non-représentativité de l'échantillon par rapport à l'univers dont il est issu (ici les patients atteints de cancer du poumon) et le problème de la comparabilité entre le groupe des cas et le groupe des témoins, renvoient clairement à ce que Hill avait défini comme étant un cas de sélection ou de biais dans ces *Principles of Medical Statistics*. Néanmoins, ni le mot « biais » ni le mot « sélection » ne sont utilisés ici.
- Il y a ensuite le biais d'enregistrement, assez peu développé par les auteurs dans cet article, et qui constitue une nouveauté par rapport aux développements de Hill à ce sujet dans son ouvrage séminal. Là encore, le mot « biais » n'est pas utilisé.
- Il y a enfin le biais lié à l'enquêteur (ou à l'intervieweur), explicitement nommé comme tel par Doll et Hill.

Les trois parties suivantes (11, 12 et 13, d'après notre décompte) vont être consacrées à la réponse aux trois questions posées dans cette partie, c'est-à-dire à l'examen des trois biais identifiés comme étant des explications possibles aux résultats auxquels ils sont parvenus dans leur enquête. Elles s'intitulent respectivement : « Sélection des patients interrogés » (« *Selection of patients for Interview* ») « Antécédents tabagiques des patients » (« *Patient's Smoking History* »), « Les enquêtrices » (« *The Interviewers* »).

Tout d'abord se pose donc la question de la sélection des patients : après avoir montré, tableaux statistiques à l'appui (sauf pour le caractère représentatif des cas, car pour les auteurs, « rien ne permet de supposer qu'ils ne constituaient pas un

[34] « *Though from the previous tables there seems to be no doubt that there is a direct association between smoking and carcinoma of the lung it is necessary to consider alternative explanations of the results. Could they be due to an unrepresentative sample of patients with carcinoma of the lung or to a choice of a control series which was not truly comparable? Could they have been produced by an exaggeration of their smoking habits by patients who thought they had an illness which could be attributed to smoking? Could they be produced by bias on the part of the interviewers in taking and interpreting the histories?* »,in Doll et Hill, 1950, p. 740.

échantillon représentatif des patients atteints de cancer du poumon qui s'adressent aux hôpitaux londoniens sélectionnés. »in Doll et Hill, 1950, p. 744) que le groupe des contrôles (ou des témoins) est bien comparable à celui des cas en termes de sexe, d'âge, etc., et que les différences qui existent entre les deux groupes (notamment quant au lieu de résidence) ne sont pas significatives, « il reste la possibilité que les enquêtrices aient choisi d'interroger parmi les patients susceptibles d'être sélectionnés un nombre disproportionné de petits fumeurs », ce qui pourrait expliquer pourquoi il y a plus de cancers du poumon parmi les cas que parmi les témoins. Or, Hill et Doll montrent qu'il n'y a pas de « différence appréciable » si l'on fait un test statistique. Dès lors :

> « on peut par conséquent conclure que rien n'indique la présence d'un biais particulier en faveur des petits fumeurs lors de la sélection des patients de la série témoin. Autrement dit, le groupe de patients interrogés constitue, estimons-nous, une série témoin satisfaisante pour les patients atteints de cancer du poumon, pour ce qui est de la comparaison de leurs habitudes tabagiques »[35]

Ainsi, le terme « biais » renvoie ici clairement à la notion de sélection qui serait ici le fait des enquêtrices : l'on retrouve la définition donnée par Hill dans ses *Principles*, où un échantillon sélectionné est défini comme un échantillon où « un membre de cet univers d'où est tiré l'échantillon a plus de chances d'apparaître qu'un autre ». Dans les termes de l'article, le groupe des témoins serait sélectionné au sens où les enquêtrices auraient « choisi d'interroger parmi les patients susceptibles d'être sélectionnés un nombre disproportionné de petits fumeurs » (Doll et Hill, 1950, p. 745).

2.2.2 Le biais lié au patient et le biais lié à l'enquêteur.

Concernant le biais d'enregistrement, Hill et Doll vont l'aborder, sans le nommer, dans la douzième partie :

[35] « *It can therefore be concluded that there is no evidence of any special bias in favour of light smokers in the selection of the control series of patients. In other words, the group of patients interviewed forms, we believe, a satisfactory control series for the lung-carcinoma patients from the point of view of comparison of smoking habits*.» in Doll et Hill, 1950, p. 745.

> « Une autre hypothèse à envisager est que les patients atteints de cancer du poumon ont eu tendance à exagérer leur consommation de tabac » (Doll et Hill, 1950, p. 745).

En effet, si « la plupart de ces patients ne pouvaient pas savoir qu'ils étaient atteints d'un cancer », néanmoins, « ils ne pouvaient pas ignorer qu'ils avaient des symptômes respiratoires » (Doll et Hill, 1950, p. 745). Ainsi, « une telle connaissance pourrait avoir influé sur les réponses aux questions concernant la quantité de tabac fumée. » (Doll et Hill, 1950, p. 745). Là aussi, même si les patients ont effectivement exagéré leur consommation, cela n'aurait pas changé les résultats :

> « Toutefois, on a déjà montré au Tableau X que les patients atteints d'autres affections respiratoires ne rapportaient pas des antécédents de tabagisme très différents de ceux indiqués par les patients atteints d'affections non respiratoires. Il n'y a donc aucune raison de supposer que l'exagération de leur consommation par les patients atteints de cancer du poumon puisse expliquer les résultats. » (Doll et Hill, 1950, p. 745).

Hill et Doll vont ainsi conclure sur ces deux types de biais dans la partie consacrée à la discussion :

> « Pour résumer, il ne semble pas justifié, à notre avis, d'imputer les résultats à un biais quelconque particulier de sélection ou à un biais d'enregistrement. En d'autres termes, il faut en conclure qu'il existe une association réelle entre le cancer du poumon et le tabagisme.»[36].

Nous retrouvons ici le concept de sélection, qu'Hill identifiait à celui de biais. Néanmoins, Hill et/ou Doll ont ajouté l'épithète « *special* », ce que les traducteurs français de l'article ont traduit, en 1999 par « biais de sélection ».

Enfin, concernant le biais lié à l'enquêtrice, la partie intitulée « Les enquêtrices » lui est consacrée. En effet, Hill et Doll ont dû faire face à un problème :

> « Lorsque l'étude a été planifiée, on pensait que les enquêtrices sauraient uniquement qu'elles interrogeaient des patients atteints de cancer, sans que la localisation exacte (pulmonaire, gastrique, colorectale) soit connue. En pratique, cela n'a malheureusement pas été possible : le siège de la tumeur était indiqué sur la fiche de signalement, ou l'infirmière évoquait le diagnostic

[36] « *To summarize, it is not reasonable, in our view, to attribute the results to any special selection of cases or to bias in recording. In other words, it must be concluded that there is a real association between carcinoma of the lung and smoking »,* in Doll et Hill, 1950, p. 746.

> en désignant le patient, ou elles se sont aperçues que dans un service donné ne se trouvaient que les cas de cancer correspondant à une seule des localisations étudiées. Sur les 1732 patients signalés et interrogés en tant que cas de cancer, le siège de la tumeur était connu de l'enquêtrice au moment de l'entretien, sauf dans 61 cas. » (Doll et Hill, 1950, p. 745).
>
> Dès lors :
>
> « Il faut donc envisager attentivement la possibilité qu'un biais dû à l'enquêtrice ait modifié les résultats (l'enquêtrice ayant eu tendance à surestimer [« *scale up »*] la consommation de tabac des cas de cancer du poumon)[37].

Or, cette hypothèse peut être testée en s'intéressant aux habitudes tabagiques des patients qui ont été classés à tort comme ayant un cancer du poumon :

> « on peut constater que les habitudes tabagiques des patients classés par erreur dans la catégorie cancer du poumon au moment de l'entretien se distinguent nettement des habitudes des patients atteints réellement de cancer du poumon (Tableau XII), mais qu'elles ne diffèrent pas significativement des habitudes des autres patients interrogés (Tableau XIII). Il est donc manifestement impossible d'expliquer les résultats de l'enquête par un biais dû à l'enquêtrice, car, en cas de biais important, les habitudes tabagiques des patients présumés par erreur avoir un cancer du poumon auraient été enregistrées par l'enquêtrice comme celles des sujets réellement atteints de cancer du poumon, et non comme celles des sujets sans cancer du poumon. »[38]

---

[37] « *When the investigation was planned it was hoped that the interviewers would know only that they were interviewing patients with cancer of one of several sites (lung, stomach, or large bowel) but not, at the time, the actual site. This, unfortunately, was impracticable; the site would be written on the notification form, or the nurse would refer to the diagnosis in pointing out the patient, or it would become known that only patients with cancer of one of the sites under investigation would be found in one particular ward. Out of 1,732 patients notified and interviewed as cases of cancer, the site of the growth was known to the interviewer at the time of interview in all but 61. Serious consideration must therefore be given to the possibility of interviewers' bias affecting the results (by the interviewers tending to scale up the smoking habits of the lung-carcinoma cases). »*, in Doll et Hill, 1950, p. 745.

[38] « *The smoking habits of these patients, believed by the interviewers to have carcinoma of the lung, can be compared with the habits of the patients who in fact had carcinoma of the lung and also with the habits of all the other patients. The result of making these comparisons is shown in Tables XII and XIII, and it will be seen that the smoking habits of the patients who were incorrectly thought to have carcinoma of the lung at the time of interview are sharply distinguished from the habits of those patients who did in fact have carcinoma of the lung (Table XII), but they do not differ significantly from the habits of the other patients interviewed (Table XIII). It is therefore clearly not possible to attribute the results of this inquiry to bias on the part of the interviewers, as, had there been any appreciable bias, the smoking habits of the patients thought incorrectly to have carcinoma of the lung would have been recorded as being like those of the true lung-carcinoma subjects and not the same as those without carcinoma of the lung »*, in Doll et Hill, 1950, p. 745.

Les trois types de biais, étant entendus ici comme des explications alternatives aux résultats de l'enquête qui établissent une association entre tabagisme et cancer du poumon, ont donc été éliminés à la fois par un raisonnement *a priori* mais aussi par l'utilisation de tests statistiques, c'est-à-dire en testant les hypothèses alternatives (par exemple, nombre disproportionné de petits fumeurs dans le groupe témoin, ou bien surestimation de la consommation de tabac soit par les malades, soit par les enquêtrices). Or, toutes ces hypothèses, une fois testées, doivent être exclues comme hypothèses alternatives :

> « La possibilité que ces résultats soient dus au choix d'un groupe témoin inapproprié, au choix de patients atteints d'affections respiratoires surévaluant leur consommation de tabac ou à un biais dû à l'enquêtrice a été envisagée. Les raisons qui nous ont permis d'exclure toutes ces éventualités sont exposées, et on conclut de l'étude que le tabagisme est un facteur important dans l'étiologie du cancer du poumon. » [39]

Enfin, dans le second article consacré à cette étude cas-témoins, intitulé « *A study of the Aetiology of Carcinoma of the Lung »,* publié à la fin de l'année 1952[40], Doll et Hill vont utiliser le mot « biais » de façon surprenante : s'il est fait à nouveau référence, dans la partie consacrée aux données (« *Data »*) au problème que 15% des cas de cancers notifiés n'ont pu être interrogés [41], ainsi qu'aux problèmes du biais d'enregistrement et du biais dû à l'enquêteur[42] dans la partie consacrée à la validité des résultats (« *Validity of the Results »*), Hill et Doll vont introduire ce qui s'apparente

[39] « *Consideration has been given to the possibility that the results could have been produced by the selection of an unsuitable group of control patients, by patients with respiratory disease exaggerating their smoking habits, or by bias on the part of the interviewers. Reasons are given for excluding all these possibilities, and it is concluded that smoking is an important factor in the cause of carcinoma of the lung »*, in Doll et Hill, 1950, p. 747.

[40] Doll, Richard et Hill, A. Bradford, « Study of the Aetiology of Carcinoma of the Lung », *British Medical Journal*, vol. 2 / 4797, 1952, p. 1271-1286

[41] « *The reasons why patients were not interviewed were: already discharged from hospital, 213; too ill, 165; dead, 72; too deaf, 33; unable to speak English clearly, 14; while in one case the interview was abandoned because the patient's replies appeared wholly unreliable. No patient refused to be interviewed. With the lung-cancer group alone the proportion not interviewed was also 15 %. We can see no reason why failure to interview all the patients should have biased the results, since it was mainly due to the time that had to elapse between the date of notification' and the date of the almoner's visit »*, *in* Doll et Hill (1952), *Art.cit.*, p. 1272

[42] « *These observations make it unreasonable, we suggest, (a) to attribute the results to exaggeration by the lung carcinoma patients, since patients with other respiratory diseases would presumably be equally inclined to exaggerate their smoking histories; (b) to attribute the results to bias on the part of the interviewers, since patients who were believed by them to have lung carcinoma but who were finally proved not to would have been recorded, had there been bias, as having smoking habits similar to the patients proved to have lung carcinoma »*, in Doll et Hill, 1952, p. 1282

à une nouvelle sorte de biais dans la partie consacrée aux « Autres facteurs étiologiques » (qui précède la partie sur la « Validité des résultats »), et plus spécifiquement aux autres maladies respiratoires[43] :

> « *To avoid bias due to any confusion between an earlier independent respiratory illness (in which our interest lay) and an illness induced by the presence of the tumour, we included in the analysis only such illnesses as had occurred at least five years before the interview. Occasionally illnesses occurring more than five years previously may have been due to a slow-growing tumour, but the number is unlikely, we think, to be important.* » (Doll et Hill, 1952, p. 1280).

Ainsi, la présence d'autres maladies respiratoires précédant la survenue du cancer pourrait être une source de confusion, au sens où ces maladies pourraient être la cause du cancer du poumon en lieu et place du tabagisme. Et c'est cette confusion qui serait alors la cause d'un biais, car la relation causale supposée entre tabagisme et cancer du poumon serait purement artificielle. Mais là encore, les hypothèses alternatives au tabagisme pour expliquer la survenue du cancer du poumon n'apparaissent pas pertinentes. Hill et Doll concluent en disant :

> « *On the present evidence we feel unable to deduce any aetiological relationship between lung carcinoma and previous respiratory illness.* » (Doll et Hill, 1952, p. 1281).

En fait, derrière ce biais de confusion, comme on le qualifierait aujourd'hui, Doll et Hill continuent leur travail argumentatif qui vise à convaincre le lecteur qu'il y a un lien entre tabagisme et cancer du poumon. C'est pourquoi ils avancent une dernière hypothèse alternative qui pourrait expliquer leurs résultats, l'hypothèse d'une maladie respiratoire précédent l'apparition du cancer. En éliminant cette hypothèse ou ce biais, ils renforcent l'hypothèse du tabac.

[43] L'hypothèse des maladies respiratoires comme cause possible, mais rejetée, dans l'étiologie du cancer du poumon intervient dans l'article après celles du l'emploi occupé et de la classe sociale, du lieu de résidence (ville ou campagne), du fait de résider près d'une usine à gaz (*gasworks*), et de l'exposition à différents types de chauffage.

2.2.3 Quelle est la fonction opératoire du concept de biais chez Hill ? Randomisation et aveugle.

Ainsi, comparativement à l'utilisation que Hill faisait de la notion de biais dans son ouvrage, dont la cinquième édition parait pourtant la même année que l'article en question, il est frappant de constater que la notion de biais est moins associée à celle de sélection qu'au biais de l'intervieweur ou de l'enquêteur. Doit-on y voir l'influence de l'article de Wynder et Graham publié quelques mois auparavant et qui est cité dans l'article de Doll et Hill, ce qui pourrait apparaitre comme une tentative d'harmoniser la signification des mots au sein d'une communauté épidémiologique encore embryonnaire ? En effet dans cet article le mot « biais » apparait deux fois, une première fois dans la partie consacrée à la méthode de l'étude (« *Method of Study »*)[44], une seconde fois dans la partie consacrée aux résultats (« *Results »*)[45]: il s'agit exclusivement du biais lié à l'intervieweur, au sens où celui qui conduit l'entretien à l'aide d'un questionnaire ne sait pas si la personne qu'il interroge est ou non atteinte d'un cancer, ou plus précisément, d'un carcinome bronchogénique. Ceux qui conduisent l'entretien sont donc aveugles sur le statut du patient qu'ils interrogent, ce qui permet d'éviter que l'intervieweur, du fait de ses préjugés ou de ses « biais », oriente consciemment ou non les réponses de celui qu'il interroge. Wynder et Graham ont même préféré faire intervenir des enquêteurs qui ne sont pas médecins pour garantir la plus grande neutralité possible. Pour vérifier que cela était bien le cas, une population faisant office de contrôle a été établie afin notamment de « tester la validité des interviews par ceux qui connaissaient à l'avance le diagnostic suspecté dans un cas donné »[46].

---

[44] « *To check all possible bias on the part of the interviewers who saw only patients believed to have bronchiogenic carcinoma, it was deemed advisable to conduct a control study in which a nonmedical investigator would interview every patient admitted to the Chest Service of Barnes Hospital without knowing the diagnosis in advance. Two interviewers were used for this purpose*. », in Wynder et Graham, 1950, p. 331

[45] « *Before the smoking habits of the 605 patients with cancer of the lungs are compared with those of the general hospital population, it might be well to compare the results in the two control studies and the group of 422 patients (study III) interviewed and collected by one of us (E. L. W.) to determine any possible bias in cases in which the suspected diagnosis was known in advance and whether the data are sufficiently similar to warrant their discussion as a group*. », in Wynder et Graham, 1950, p. 332.

[46] « *…to test the validity of the interviews made by those who knew the suspected diagnosis in a given case in advance. »*, in Wynder et Graham, 1950, p. 331.

Si à l'inverse, il ne s'agit pas d'une tentative de se mettre d'accord sur les termes utilisés dans la discipline épidémiologique, doit-on y voir une inflexion de la pensée de Hill à ce sujet, y compris sur la notion de « sélection », lié simplement au fait que les objectifs qu'il poursuit sont différents ? En effet s'il s'agissait dans ses *Principles of Medical Statistics* d'expliquer les bases des statistiques aux médecins et aux chercheurs, il s'agit plutôt dans cet article de convaincre les lecteurs du *British Medical Journal*, qui sont essentiellement des médecins, d'une association entre le tabagisme et le cancer du poumon. Dès lors, il serait plus efficace d'utiliser le sens commun du mot « biais », plutôt qu'un sens technique, afin de pouvoir plus facilement emporter la conviction du lecteur. Une troisième hypothèse serait qu'il aurait intégré les problèmes posés par biais subjectifs lors de l'essai sur la streptomycine réalisé en 1947-1948 par le *Medical Research Council :* il y aurait ainsi une continuité de la conception du « biais » chez Hill au sens où de la même manière que le médecin ou le chercheur peut être influencé par ses préférences dans l'allocation des patients dans le groupe traité ou dans le groupe placebo – d'où la nécessité de lui cacher la procédure d'allocation des patients – de même l'enquêtrice peut être influencée par la connaissance qu'elle a de la maladie du patient (ou plutôt ici de la localisation de la tumeur) et ainsi modifier les réponses des patients, donc les données de l'enquête.

Harry Marks défend une hypothèse dont la portée est beaucoup plus large, et qui permet de concilier les trois hypothèses précédentes. Selon lui, la notion de « biais » va commencer à remplir dans les années 1950 une fonction bien précise : il s'agit pour les médecins réformateurs de faire accepter aux chercheurs et aux cliniciens de l'époque, que « les théories mathématiques n'impressionnent pas »[47], une nouvelle méthodologie (randomisation, double aveugle, etc.) qui vise précisément à faire de la médecine une science objective. Marks dit ainsi :

> « À chaque fois qu'ils expliquaient la nécessité d'un changement de procédure — randomisation, essai en double aveugle, évaluation objective des résultats — les réformateurs invoquaient le mot de « biais ». (…). Les chercheurs étaient régulièrement mis en garde contre les dangers de biais « inconscients » pouvant invalider leurs études.» (Marks, 2000, p. 18).

[47] Kaptchuk, Ted J., « Intentional ignorance: a history of blind assessment and placebo controls in medicine », *Bulletin of the History of Medicine*, vol. 72 / 3, 1998, p. 389-433. La citation est à la page 429.

Kaptchuk, va même aller plus loin dans son article consacré à « l'ignorance intentionnelle », où il montre comment « l'évaluation masquée » (autres noms donnés à la procédure de l'aveugle et du double aveugle), qui était auparavant (au XIXe siècle essentiellement) réservée à l'évaluation du charlatanisme et des médecines alternatives (le mesmérisme, l'homéopathie, etc.), va progressivement s'introduire dans la médecine clinique classique :

> « Auparavant, les accusations infamantes de biais, de préjugé, d'enthousiasme débordant, de crédulité, d'illusion, étaient réservées aux guérisseurs déviants ; à présent, ce qui était autrefois une menace marginale était internalisé. Même les jugements des cliniciens les plus expérimentés concernant l'efficacité des nouvelles thérapeutiques étaient suspectes. Le « biais » hantait la médecine. » (Kaptchuk,1998, p. 430).

Ainsi les années 1950 constituent ce moment où les deux notions de biais, statistique et psychologique, tendent à coïncider, au sens où l'opération permise par la notion de biais est de justifier certaines procédures, dans un essai clinique ou dans une étude épidémiologique, qui visent à éliminer toute trace de subjectivité afin de garantir l'objectivité et la rigueur de l'étude ou de l'essai. Randomisation et évaluation à l'aveugle se renforcent d'ailleurs mutuellement : la mise sous aveugle (« *blinding* ») n'est en effet possible que dans le cadre de la randomisation, et la randomisation ne fonctionne que dans des conditions d'aveugle ou de double aveugle.

Hill a d'ailleurs parfaitement saisi l'avantage de cette technique. Ainsi, dans un passage ajouté à la cinquième édition des *Principles of Medical Statistics*, qui porte sur l'essai clinique sur la streptomycine qu'il vient de réaliser, Hill insiste sur la nécessité de lutter contre les biais par d'autres moyens que la randomisation et introduit la notion d'« aveugle » :

> « *Another aspect of this trial worth noting is that to remove all possibility of bias the radiological progress of the patient was assessed by two radiologists and a clinician independently and each had no knowledge whatever whether the film they saw related to a streptomycin or to a control case. This working "blind" not only guards against all conscious or unconscious bias in the investigator but also – and equally important – against any honest attempt the assessor may make to allow for a possible bias in his attitude. Such a technique in no way questions his intellectual honesty, and it so greatly increases the confidence*

with which the results can be regarded that it is today increasingly welcomed and applied by many workers » (Hill, 1950, p. 8).

Le seul moyen de vérifier que la tuberculose régresse grâce à la streptomycine est en effet de le faire constater par des radiologues. Or, si jamais le radiologue ou le clinicien sait que le patient est ou non sous streptomycine, ou en d'autres termes fait partie du groupe de cas ou du groupe de témoins, cela risque d'influencer son jugement d'une manière ou d'une autre, en fonction par exemple de ses opinions sur l'efficacité de la streptomycine. Dès lors, si les deux radiologues et le clinicien examinent de façon indépendante les radiologies, pour voir s'il y a effectivement un progrès, et s'ils sont « aveugles » sur l'appartenance du patient au groupe des cas ou à celui des témoins, leur opinion ou leur diagnostic doit être neutre, ou autrement dit dépourvu de biais, donc de toute subjectivité, et par là même il doit être objectif, d'autant plus si leurs conclusions concordent. La procédure en aveugle permet ainsi d'« enlever toute possibilité de biais », chez les radiologues et les cliniciens, mais aussi chez l'investigateur (qui peut lui aussi être victime de « biais conscients ou inconscients », comme le désir de prouver par exemple que la streptomycine est efficace) et même chez l'évaluateur de l'essai, qui pourrait tout à fait être poussé par un désir similaire, ou contraire, à celui de l'investigateur. La procédure d'aveuglement leur permet donc, sans remettre en cause leur réputation et leur honnêteté, d'être sûrs qu'ils sont restés objectifs et qu'ils n'ont pas été influencés par quelque préférence ou inclination que ce soit, ce qui a pour conséquence et pour avantage d'« accroître considérablement la confiance dans les résultats » de l'étude.

Hill fait donc ici d'une pierre deux coups : évitant délibérément de parler aux médecins de randomisation de peur de les effrayer[48], et cherchant à les convaincre de la nécessité de cette même randomisation mais aussi de la procédure d'aveugle, le concept de « biais », compris par Hill et par les médecins auxquels ils s'adresse en son sens trivial de préjugé ou de partialité, permet de justifier aux yeux des médecins les procédures qu'il introduit mais aussi de garantir à leurs yeux l'objectivité et donc la scientificité des études épidémiologiques comme des essais cliniques, et par là celles de l'épidémiologie elle-même. En retour, la randomisation et l'évaluation à l'aveugle,

[48] « *I deliberately left out the words "randomization" and "random sampling numbers" at that time, because I was trying to persuade the doctors to come into controlled trials in the very simplest form and I might have scared them off.»,* in Hill, Austin Bradford, « Memories of the British streptomycin trial in tuberculosis: the first randomized clinical trial », *Controlled clinical trials*, vol. 11 / 2, 1990, p. 77–79.

entendues toutes deux comme des procédures visant à éliminer les biais, « pourraient faire de la médecine une science « dure » à part entière » (Kaptchuk,1998, p. 427).

Pour le dire plus clairement, de la même manière que la fonction opératoire du concept de randomisation n'est pas la même chez Fisher et chez Hill, au sens où finalement , comme le soulignent Harry Marks et Iain Chalmers, « Fisher n'a pas grand-chose à voir avec l'introduction de la randomisation en médecine »[49], introduction qui fut l'œuvre essentiellement de Bradford Hill; de la même manière, la fonction opératoire du concept de biais est différente chez Fisher et Hill : chez Fisher il s'agit de disposer d'une estimation valide de l'erreur, et donc de garantir la validité du test statistique, alors que chez Hill il s'agit de mécaniser la procédure afin de la dissimuler à ceux qui y participent pour rendre l'étude impartiale et objective, ce qui justifie, en sus de la randomisation, l'introduction de l'aveugle.

2.3 La conceptualisation du biais par les sciences sociales

2.3.1 Le biais de l'intervieweur

Pourtant, si les thèses de Marks et Kaptchuk sont fondées sur de solides arguments historiques, leur perspective apparait trop centrée sur la médecine. Or, en déportant notre regard sur les sciences humaines et sociales de cette époque, il est possible de renforcer leur thèse, et surtout, du point de vue de notre étude, de mieux comprendre ce que les statisticiens de l'époque entendent par le mot biais. Plus précisément, la définition et la conceptualisation de la notion de biais vont s'opérer à l'orée des années 1950 grâce aux travaux de psychologues et de sociologues. En effet, la discussion sur les biais est assez vive depuis le milieu des années 1930 dans un domaine à mi-chemin entre la psychologie et la sociologie (et de la psychologie sociale) qui est celui des sondages d'opinion, et plus spécifiquement les sondages d'opinion appliqués à la politique, dont la scientificité est encore aujourd'hui encore fort contestée[50]. Cette discussion, d'ordre méthodologique, va se concentrer autour de

[49] « *Fisher had little to do with the introduction of randomization into medicine*. » in Marks, Harry M., « Rigorous uncertainty: why RA Fisher is important », *International Journal of Epidemiology*, vol. 32 / 6, décembre 2003, p. 932-937. La citation est à la page 935.

[50] Sur ce sujet, voir par exemple Bourdieu, « Les doxosophes. », *Minuit*, novembre 1972, p. 26-45 ; Bourdieu, Pierre, « L'opinion publique n'existe pas. », *Les temps modernes*, janvier 1973, p. 1292-1309 ; ou encore Bourdieu, Pierre, « Remarques à propos de la valeur scientifique et des effets

trois problèmes principaux rencontrés par ces études, problèmes qui sont d'ailleurs similaires à ceux rencontrés par les épidémiologistes : le problème de l'échantillonnage, le problème des questionnaires envoyés par courrier et surtout le problème de l'interview et du fameux « biais de l'intervieweur » que l'on retrouve dans les articles de Wynder et Graham, mais aussi de Hill et Doll. Ce que nous proposons ici est donc une explication alternative, qui vise non pas à contredire la thèse de Marks et de Kaptchuk, mais à la compléter par un changement ou un élargissement de la perspective. Notre hypothèse est en effet que les sciences humaines, et notamment la psychologie, disposent en fait déjà d'un concept de biais, ou plutôt d'une notion de biais, synonyme de préjugé, qu'elles vont progressivement constituer comme un objet d'étude, et qui, au fur et à mesure que les sciences humaines vont utiliser les méthodes statistiques, va opérer une forme de jonction avec le concept statistique de biais (dont l'origine est d'ailleurs *in fine* psychologique, que ce soit chez Galton et le choix du partenaire de mariage[51], ou chez Fisher où « l'arrangement systématique » des parcelles, par opposition à leur randomisation, renvoie à « l'ingéniosité de l'agronome »[52]), pour former un concept véritablement scientifique et transdisciplinaire.

La question du biais de l'intervieweur est en fait posée dès 1929 dans un article resté célèbre dans l'histoire de la sociologie et qui ouvre un débat extrêmement fécond au sein des sciences humaines et sociales. Stuart A. Rice, alors professeur de sociologie et de statistiques à l'Université de Pennsylvanie, et qui sera président de l'*American Statistical Association* en 1933, publie un article intitulé : « Un préjugé contagieux dans l'interview : une note méthodologique»[53]. Selon lui en effet l'essentiel de l'information collectée dans la recherche sociale, et notamment par le *United States Census Bureau,* c'est-à-dire le bureau du recensement américain, repose « en dernier ressort sur l'information communiquée par les informateurs aux intervieweurs »[54]. Or, il y a une possibilité, voire une certitude, qu'un « biais », c'est-à-dire « des intérêts

politiques des enquêtes d'opinion. », *Pouvoirs, revue française d'études constitutionnelles et politiques.*, Avril 1985, p. 131-139.

[51] Voir aussi la section 1.1.3 du présent travail pour une explication plus détaillée.

[52] « *the agronomist's ingenuity* », in Fisher, 1935, p. 71. Voir aussi la section 1.3.2 du présent travail pour une explication plus détaillée.

[53] Rice, Stuart A., « Contagious Bias in the Interview: A Methodological Note », *American Journal of Sociology*, vol. 35 / 3, 1929, p. 420–423.

[54] «*Such data as those of the United States Census Bureau, for example, rest ultimately upon information communicated by informants to interviewers*. », in Rice, 1929, p. 420.

sélectifs » (« *selective interests* »[55]) dans l'esprit des deux participants soient présents. En d'autres termes, ni l'intervieweur ni le contenu de l'interview ne sont des médias neutres ou objectifs et l'intervieweur peut « biaiser non seulement la sélection ou l'enregistrement de l'information dans l'esprit de l'interviewé, mais aussi la substance même de l'information », d'où la nécessité de contrôler « rigidement » les « conditions de l'interview » (Rice, 1929, p. 421). Rice prend comme exemple une étude menée en 1914 à New York pour déterminer les « caractéristiques physiques, mentales et sociales » (Rice, 1929, p. 421) de 2 000 personnes sans-abri : douze investigateurs chevronnés (« *skilled* »[56]) effectuent l'interview des sans-abri d'une durée de vingt à trente minutes, chaque sans-abri étant interviewé individuellement par un seul intervieweur. Parmi ces questions, l'une consiste à déterminer la cause qui explique, selon le sans-abri, son dénuement ou sa misère (« *destitution* »[57]), avec deux catégories de réponse : l'alcool (« *liquor* ») ou bien une cause « industrielle » (qui inclut le « licenciement », un « travail saisonnier » ou bien encore la « fermeture de l'usine »). Or, en examinant les données, Rice se rend compte de « certains types de réponses uniformes chez des hommes interviewés par certains enquêteurs ».

> « Ainsi, alors que [l'enquêteur] A voyait les effets mineurs ou majeurs de l'usage d'alcool dans 78% des hommes qu'il interviewait, [l'enquêteur] B voyait ces mêmes effets dans seulement 37% des cas. Alors que B assignait la dépendance à des causes industrielles et non-personnelles dans 73% des cas, A ne discernait ces facteurs que dans 29% des cas »[58]

Or, après enquête, il s'est avéré que l'intervieweur A était « un ardent partisan de la prohibition » tandis que l'intervieweur B « était considéré par ses associés comme un socialiste »(Rice, 1929, p. 422).

Ainsi, comme le souligne Rice :

[55] Rice, 1929, p. 420.
[56] Rice, 1929, p. 421.
[57] Rice, 1929, p. 421.
[58] « *That is, while A saw the major or minor effects of the use of alcohol in 78 per cent of the men before him, B saw these effects in the case of but 37 per cent. While B ascribed dependency to non-personal, industrial causes in the case of 73 per cent, A discerned these factors in the case of but 29 per cent*.» , in Rice, 1929, p. 422.

> « Le préjugé [« *bias* »] dans l'esprit de l'intervieweur a été communiqué par un processus quelconque de suggestion à l'esprit de l'interviewé, et a alors été reproduit dans les réponses de celui-ci »[59]

Le concept psychologique trivial de « *bias* » comme préjugé prend ici une nette coloration scientifique, au sens où il devient un objet d'étude aussi bien pour le psychologue que pour le sociologue et le statisticien : l'idéologie de l'intervieweur déteint sur l'esprit de l'interviewé qui finit par répondre à l'intervieweur ce que celui-ci a envie d'entendre, ce qui compromet la validité des résultats de l'enquête. Ce biais de l'intervieweur (aussi appelé « effet de l'intervieweur ») n'est évidemment pas sans rappeler les problèmes que rencontrent les épidémiologistes dans les études observationnelles, comme nous venons de le voir avec les articles de Doll et Hill ou de Wynder et Graham, ou plus encore dans les essais cliniques randomisés où il y a toujours un risque de ce qu'il est convenu d'appeler l'effet placebo, effet au sein duquel Daniel Schwartz[60] distingue deux types de suggestion : l'autosuggestion (le patient se persuadant consciemment ou non que le traitement est efficace) et l'hétérosuggestion (la « foi du médecin » dans le traitement pouvant « influencer le cours de la maladie »[61]). Ce biais de l'intervieweur va continuer à être étudié de façon soutenue par aussi bien les sociologues que les psychologues, mais aussi par les statisticiens, et va être contrasté notamment avec les enquêtes par courrier, qui se multiplient dans l'entre-deux-guerres, et qui ne sont pas non plus exemptes de biais (de sélection ou de non-réponse par exemple) mais qui sont censées permettre d'éviter ce « biais contagieux ».

2.3.2 Naissance de l'opinion publique : sondages et biais.

En effet, il faut souligner que c'est précisément à cette époque, dans les années 1930, que les premiers instituts de sondage apparaissent aux Etats-Unis d'Amérique comme en France[62] : en 1936, le journaliste G.H. Gallup fonde l'«*American Institute of*

[59] « *the bias in the mind of the interviewer was communicated by some process of suggestion to the mind of the interviewed, and was there reproduced in response to questioning by the latter*», in Rice, 1929, p. 423.

[60] Schwartz, Daniel, « Peut-on évaluer les médecines douces ? », *Sciences sociales et santé*, vol. 4 / 2, 1986, p. 75-88.

[61] Schwartz, 1986, p. 78.

[62] Pour une histoire globale des sondages, voir Blondiaux, Loïc, *La fabrique de l'opinion: une histoire sociale des sondages*, Paris, Seuil, Collection « Science politique », 1998

Public Opinion», tandis que Jean Stoetzel[63] fonde, en 1938 en France, l'IFOP (Institut Français d'Opinion Publique), dont le modèle comme le nom sont directement empruntés à son homologue américain, et réalise la première enquête de l'histoire sur l'opinion publique française[64]. Les sondages préélectoraux existent alors depuis une vingtaine d'années aux Etats-Unis : le *Literary Digest*, hebdomadaire d'actualité américain, a en effet commencé dès 1916 à sonder les américains sur leurs intentions de vote et a d'ailleurs prédit correctement les résultats des différentes élections présidentielles de 1916 jusqu'à 1936, année où il fait une erreur restée célèbre dans l'histoire des enquêtes d'opinion comme le paradigme de ce qu'il ne faut pas faire en la matière : ainsi, il prévoyait que Alfred Landon, le candidat républicain, l'emporterait avec 55% des voix contre seulement 41% pour son adversaire le démocrate Franklin Delano Roosevelt, qui terminait alors son premier mandat. Le résultat final donna 61% des voix à Roosevelt contre seulement 37% à Landon. Cet échec fut analysé par la suite comme étant lié à un biais d'échantillonnage (l'échantillon, pourtant composé d'environ 2,4 millions de personnes, n'était pas représentatif car fondé sur l'annuaire téléphonique, ou la liste des abonnés au magazine, toutes choses que les plus pauvres ne pouvaient pas s'offrir, l'Amérique sortant à peine de la crise de 1929) et à un biais de non-réponse[65] (le magazine avait contacté dix millions de personnes, mais seul un quart d'entre eux avait répondu), même si le mot « biais » n'apparait pas dans les comptes rendus de l'époque[66].

La question de la méthodologie des sondages continue néanmoins à préoccuper les statisticiens, mais aussi les sociologues et psychologues dans l'immédiat après-guerre. Elle les préoccupe d'autant plus que l'élection présidentielle américaine de 1948 marque un échec pour les cinq instituts de sondage nationaux qui avaient tous donné Thomas Dewey gagnant contre Harry Truman (depuis le début de

[63] Sur les liens entre Stoetzel et Gallup via Hadley Cantril, professeur de psychologie sociale à la Columbia University de New York, où Stoetzel séjourne une année (1937-138), voir MARCEL, Jean-Christophe, « Le premier sondage d'opinion », *Revue d'Histoire des Sciences Humaines*, vol. 6 / 1, 2002, p. 145

[64] Stoetzel, Jean, « Une enquête sur l'opinion publique française », *Revue d'Histoire des Sciences Humaines*, vol. 6 / 1, 2002, p. 155.

[65] Sur les raisons de cet échec, voir Squire, Peverill, « Why the 1936 Literary Digest poll failed », *Public Opinion Quarterly*, vol. 52 / 1, 1988, p. 125–133. Il faut noter que l'institut de Gallup a prédit correctement la victoire de Roosevelt, avec un échantillon de 50 000 personnes.

[66] Voir par exemple Crossley, Archibald M., « Straw polls in 1936 », *Public Opinion Quarterly*, vol. 1 / 1, 1937, p. 24–35; ou Cantril, Hadley, « How Accurate Were the Polls? », *Public Opinion Quaterly*, vol. 1 / 1, 1937, p. 97-109. Il est intéressant de constater que le premier numéro de l'histoire de la revue *Public Opinion Quaterly*, émanation de l'*American Association for Public Opinion Research*, soit pour l'essentiel consacré à analyser les raisons de l'échec des sondages de l'élection de 1936.

la campagne jusqu'à une semaine avant l'élection où ils décidèrent, sûrs de leur prédiction, d'arrêter les sondages), qui finit pourtant par l'emporter. C'est pourquoi sont engagés des travaux portant sur les biais des enquêtes, qu'il s'agisse des enquêtes par interview ou par courrier : le *National Opinion Research Center* publie ainsi en 1953 un volumineux rapport[67] de presque 500 pages sur le problème du biais de l'intervieweur qui se veut être, comme l'indique le sous-titre du rapport, une « étude systématique des sources d'erreur dans l'étude empirique des attitudes, des opinions, et d'autres aspects du comportement humain ». Il ne s'agit pas ici d'étudier dans le détail ce rapport, ni d'ailleurs la question du biais de l'intervieweur, rapport qui fait d'ailleurs suite à d'innombrables articles sur la question[68], car cela nous conduirait trop loin de notre sujet. Le fait remarquable du point de vue de notre étude est que c'est à travers ce débat sur le biais de l'intervieweur que va s'opérer, au début des années 1950, une véritable conceptualisation de la notion de biais, qui ne se résume pas d'ailleurs au cadre de la psychologie sociale mais qui, comme on va le voir, le dépasse et concerne finalement toutes les sciences humaines et sociales, mais aussi la médecine et l'épidémiologie. En effet, comme le souligne Hyman, la technique de l'interview est « universelle dans les sciences sociales » (Hyman, 1953, p. 1) : elle est utilisée aussi bien en anthropologie, en sociologie, dans les recensements démographiques[69], mais aussi par les « psychiatres, cliniciens et les psychanalystes » (Hyman, 1953, p. 1). Surtout, la « recherche sur l'opinion publique, ressource commune au politologue, à l'administrateur public, au psychologue social, et l'historien, est fondée sur l'interview »[70].

---

[67] Hyman, Herbert, « Isolation, Measurement, and Control of Interview Effect », *National Opinion Research Center*, University of Chicago, 1953.

[68] Pour un panorama, en français et d'époque, des différents articles et problèmes rencontrés dans les enquêtes d'opinion, voir : Dubost, J. et Durandin, G., « 1° Opinion publique et attitudes », *L'année psychologique*, vol. 52 / 2, 1952, p. 594–613. Le mot « *bias* », quand il est traduit, l'est par le mot « déformation » (voir par exemple la page 598).

[69] Il faut préciser ici, dans un souci de contextualisation historique, que c'est au sortir de la deuxième guerre mondiale que, à l'instar ce qu'il se passe alors dans le domaine de l'épidémiologie, sont lancées des enquêtes sociologiques de grande ampleur, notamment en Grande-Bretagne où la « *Government Social Survey »*, une unité rattachée directement au gouvernement britannique, s'intéresse à des sujets aussi divers que les statistiques de morbidité, le problème du chauffage et de l'éclairage domestiques, ou encore les personnes âgées à charge des familles ; tout cela étant lié à une volonté politique d'améliorer la vie quotidienne des gens, en lien avec l'apparition d'un Etat-providence (« *Welfare State* »), qui fait suite notamment au rapport Beveridge. Ce phénomène est tout aussi visible aux Etats-Unis et en France : l'Institut national d'études démographiques (Ined) est créé en 1945, l'Institut national de la statistique et des études économiques (INSEE) en 1946.

[70] « *Public opinion research, as a common resource of the political scientist, public administrator, social psychologist, and historian is built upon the foundations of interviewing.»,* in Hyman, 1953, p. 1.

En d'autres termes l'instrument essentiel à travers lequel la recherche sur l'opinion publique obtient de l'information et des connaissances, et sur lequel repose sa méthode et donc sa scientificité (ou sa prétention à la scientificité), est essentiellement l'interview. En ce sens, il est parfaitement logique que l'essentiel de la discussion méthodologique ou épistémologique porte sur la validité de cet instrument, c'est-à-dire sa capacité à fournir correctement les mesures qu'il a pour fonction de fournir sur les opinions ou les attitudes des personnes interrogées. A l'inverse, dans le cadre des études épidémiologiques, la question de l'interview n'est qu'un des multiples problèmes rencontrés dans la méthodologie de l'enquête: c'est ce problème de l'interview, conçu comme un cas particulier du biais d'information (comment le patient ou l'intervieweur ou le médecin sont influencés par la connaissance d'un statut pathologique particulier, et comment cette connaissance modifie la façon que le patient a de répondre aux questions, l'intervieweur de les poser et de les noter, ou le médecin de poser ces questions dans le cadre d'un examen clinique ou de poser un diagnostic) qui va justifier le recours à la procédure du « *blinding* » ou de ce que Kaptchuk appelle « l'ignorance intentionnelle ». Néanmoins, il semble assez évident que les sciences humaines et sociales, même si leur objectif est différent, partagent avec l'épidémiologie un certain nombre de problèmes méthodologiques communs liés à l'usage des méthodes et des outils statistiques, méthodes et outils dont la nouveauté est bien souvent problématique.

2.3.3 La première définition du concept de biais :

Ainsi, ce n'est sans doute pas un hasard si l'on trouve, au détour d'un texte consacré au biais de l'intervieweur, une des rares définitions (en fait, la seule à notre connaissance) claires et complètes de la notion de biais à l'époque, ainsi qu'une ébauche de classification des différents biais, ce qui atteste selon nous que cette notion n'a plus rien de trivial mais constitue bien un concept scientifique à part entière. Le texte, publié en 1951, est écrit par le Baron Claus Adolf Moser, un statisticien germano-britannique alors en poste à la *London School of Economics* (où il fera l'essentiel de sa carrière en tant que professeur de statistiques sociales ; il dirige aussi

le *Central Statistical Office* de 1967 à 1978), et a pour titre : « Le biais de l'interview »[71]. Selon lui en effet, ce biais de l'interview constitue l'un des principaux « maillons faibles » des enquêtes qui sont menées à l'époque, notamment celles de la *Government Social Survey* :

> « Un des maillons faibles est l'interview personnelle. La plupart des enquêtes sus-citées utilisent des intervieweurs de terrain, et pourtant on ne sait que très peu de choses sur les erreurs, systématiques ou non, qu'ils pourraient introduire. Dans la mesure où la valeur ultime d'une enquête dépend largement de la fiabilité et de la validité des données obtenues par les intervieweurs, il semble logique que le biais de l'interview soit considéré comme de la plus haute priorité dans la recherche.»[72]

Moser va alors donner ce qui constitue, à notre connaissance, la première définition explicite de la notion statistique de biais dans l'histoire :

> « Par biais, nous voulons dire l'élément d'erreur systématique, non-aléatoire, non-annulable »[73].

Cette définition appelle essentiellement deux commentaires, le premier sur le contexte de publication au sens large, le second sur la définition elle-même. Concernent le contexte, il faut noter d'abord que cet article parait dans la *Revue de l'Institut International de Statistique* : or, cet institut, fondé en 1885 suite à une proposition lancée par Adolphe Quételet lors d'une réunion de la « *Royal Statistical Society* » en 1851 de « fonder une Société internationale de statistique et, en attendant, de réunir périodiquement des congrès internationaux de statisticiens » [74], vise essentiellement à uniformiser les méthodes et les concepts de la statistique au plan international[75]. Dès lors cette définition du concept de biais a nécessairement une portée importante au sein de la communauté des statisticiens et le fait même que cette

---

[71] Moser, C. A., « Interview Bias », *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, vol. 19 / 1, 1951, p. 28-40.

[72] «*One of the weakest links is the personal interview. Most of the inquiries under the above heads use field interviewers, yet very little is known about the errors, both systematic and otherwise, which may be introduced by them. As the ultimate worth of a survey depends largely on the reliability and validity of the data obtained by the interviewers, it seems right that interview bias should now have the highest priority in research* », in Moser, 1951, p. 28.

[73] « *By bias we mean the systematic, non-random, non-cancelling error element* », in Moser, 1951, p. 28.

[74] Chevry, Gabriel, « L'Institut international de statistique (I.I.S.) », *Economie et statistique*, vol. 13 / 1, 1970, p. 63-65. L'extrait cité est à la page 65. Voir aussi sur l'histoire de cet institut : Nixon, James, W., *A History of the International Statistical Institute 1885-1960*, The Hague, International Statistical Institute, 1960.

[75] Voir les statuts de l'Institut votés le 12 avril 1887, notamment l'article 1, reproduits dans Nixon (1960), *Op.cit.*, p. 138

définition soit explicitement formulée prouve qu'une définition du concept de biais était nécessaire pour en fixer le sens. Néanmoins, il est bien difficile de dire si cette définition est connue des épidémiologistes de l'époque. Nous n'avons d'ailleurs trouvé aucune référence à ce texte de Moser dans aucun des textes proprement épidémiologiques parus après cet article. Tout juste peut-on noter que Bradford Hill, dans la cinquième édition de ses *Principles of Medical Statistics* publiée en 1950, apparait comme membre de l'Institut International de Statistique[76], et il est donc probable qu'il ait, en tant que membre si ce n'est lu, au moins reçu le numéro de la revue de l'Institut où apparait l'article de Moser. Le second commentaire porte sur la définition elle-même : qu'est-ce qu'un élément d'erreur « systématique, non-aléatoire, non-annulable » ?

Moser va préciser la définition du biais plus loin dans son article :

> « Le biais renvoie à des erreurs non-aléatoires, systématiques, qui forment un élément d'erreur constant et qui ne diminuent pas nécessairement à mesure que la taille de l'échantillon est augmentée.»[77]

En effet, un des principes fondamentaux de la loi des grands nombres (dont Jacques Bernoulli définit le premier modèle mathématique vers 1690, publié en 1715 dans la quatrième partie de son *Ars conjectandi*), sur laquelle reposent d'ailleurs la plupart des sondages, est que les caractéristiques d'un échantillon aléatoire se rapprochent des caractéristiques statistiques de la population lorsque la taille de l'échantillon augmente. A l'inverse, dans le cas d'un biais, les erreurs ne vont pas s'annuler mais rester constantes : il faut noter que Moser ne fait aucune référence à l'idée que les erreurs s'accumuleraient ou augmenteraient, mais simplement qu'elles restent constantes, donc ne s'annulent pas ou ne diminuent pas. Si l'on reprend la critique de Thomas Bayes adressée à Thomas Simpson que nous avons évoquée dans notre section 1.3.4., et où le mot biais n'apparait pas, il y a dans l'idée de constance de l'erreur, que Moser avance, sensiblement la même idée que Simpson quand il parle d'erreurs qui ont « toujours tendance à aller dans la même direction » (Simpson, 1755, p. 83, cité dans Stigler, 1986, p. 95). Ainsi, les erreurs aléatoires sont des erreurs qui, à la longue, au fur et à mesure de la répétition des épreuves, vont

[76] Hill, 1950, voir la page de titre de l'ouvrage.

[77] « *bias means non-random, systematic errors, which form a constant error element and do not necessarily decrease as the size of sample is increased*. », in Moser, 1951, p. 36.

diminuer et même s'annuler : par exemple, il est parfaitement possible, quoiqu'improbable, de tirer un six dix fois de suite en lançant un dé, mais sur un million de lancers, la probabilité de tirer un six doit tendre vers 1/6. A l'inverse, si le dé est biaisé ou pipé, alors certaines faces du dé auront plus de chance d'apparaitre et la probabilité de tirer la face 6 ne sera pas alors de 1/6. Les erreurs systématiques, à l'inverse, sont donc des erreurs non-aléatoires au sens où elles ne s'annulent pas et éventuellement s'accumulent au fur et à mesure des tirages.

Un autre point important de cet article est que Moser va proposer une classification des biais, une des premières à notre connaissance, en les classant en deux grandes catégories : les « biais dans l'échantillon » (« *Bias in the sample* »), qu'il ne vas pas traiter dans l'article et les biais dans les données collectées (« *Bias in the collected data*»[78]) :

« A. Biais dans l'échantillon :

I. Dus à un échantillonnage à partir d'un cadre [« *frame* »] qui est inadéquat, imprécis ou incomplet.

II. Dus à un échantillonnage à partir d'un cadre qui inclut une forme de périodicité dont on ne tient pas compte.

III. Dû à la réalisation incomplète de l'échantillon sélectionné – en raison de refus, d'absence de contacts etc. (les non-réponses dans un questionnaire envoyé par courrier entrent dans cette catégorie).

IV. Dus à l'utilisation d'une méthode non-aléatoire de sélection. L'échantillonnage intentionnel [« *purposive* »] et l'échantillonnage par quota tombent dans cette catégorie. (…)

B. Biais dans les données collectées :

Complètement différents, et beaucoup plus difficiles, des biais qui apparaissent suite non pas à la composition de l'échantillon final, mais dans la collection des données, posent problème. Dans cette catégorie, nous pouvons distinguer les biais :

I. Dus à l'effet de l'intervieweur sur le répondant, du répondant sur l'intervieweur, ou, de façon plus réaliste, de l'interaction entre celui qui interviewe et celui qui est interviewé.

II. Dus à des facteurs en relation avec le questionnaire.

[78] Moser, 1951, p. 28-29.

III. Dus à des facteurs en relation avec l'organisation et les circonstances de l'interview. »[79]

Ainsi, il est intéressant de noter que même dans une enquête d'opinion, où il s'agit finalement simplement de recueillir l'opinion des personnes interrogées, chaque pièce du dispositif de l'enquête, dispositif qui se résume finalement à la structure suivante : enquêteur-enquête-enquêté, est susceptible d'être biaisée, au sens où précisément l'opinion qui va être exprimée par l'enquêté (celui qui est interrogé) et enregistrée par l'enquêteur (ou l'interviewer) n'est pas valide, car elle ne correspond pas à l'opinion réelle de l'enquêté. Moser oppose d'ailleurs la « validité » et la « fiabilité » des résultats et définit cette dernière comme la « convergence [« *consistency* »] entre les interviews et les intervieweurs » (Moser, 1951, p. 36). Moser liste ainsi les éléments qui pourrait biaiser les réponses, qu'il s'agisse des caractéristiques biologiques, sociologiques ou psychologiques de l'intervieweur[80], des techniques d'interview[81] (le codage des réponses, l'enregistrement incomplet ou incorrect des réponses, etc.), du risque de triche (« *cheating* »[82]) de la part de l'intervieweur, ou encore du questionnaire[83] lui-même (sa durée, l'ordre des questions ou la manière de les formuler…) ou d'autres facteurs[84]. Ainsi, les sources de biais, même dans une enquête d'opinion, sont multiples, et il est donc décisif, si l'on veut que ces résultats soient utiles et utilisables, de détecter les biais, c'est-à-dire selon Moser, de « tester la validité des résultats » (Moser, 1951, p. 36). Comment procéder ?

Moser liste trois difficultés pour détecter les biais :

---

[79] « *- A. Bias in the sample*
I. due to sampling from a frame which is inadequate, inaccurate or incomplete.
II. due to sampling from a frame which contains some form of periodicity which is not taken into account
III. due to incomplete achievement of the selected sample - on account of refusals, non-contacts etc. (Non-response in mail questionnaires falls into this category).
IV. due to the use of a non-random method of selection. Purposive and quota sampling fall into this category. (…)
B. Bias in the collected data
Altogether different, and much more difficult, problems are raised by bias arising, not from the composition of the final sample, but in the collection of the data. Under this category, we may distinguish bias
I. due to the effect of the interviewer on the respondent, the respondent on the inter- viewer or, more realistically, the interaction of interviewer and respondent.
II. due to factors connected with the questionnaire.
III. due to factors connected with the set-up and circumstances of the survey and the interview. », in Moser, 1951, p. 28-29.
[80] Moser, 1951, p. 29-31.
[81] Moser, 1951, p. 31-33.
[82] Moser, 1951, p. 33.
[83] Moser, 1951, p. 34.
[84] Moser, 1951, p. 34-35.

- D'abord, il est difficile de distinguer les erreurs aléatoires et les erreurs systématiques : « ce qui constitue un biais dans un type d'enquête, ou même de question, peut ne pas en constituer un dans un autre type. Dans un cas, les erreurs dues à une source particulière s'accumulent, dans l'autre elles s'annulent »
- Ensuite, le biais peut être absent de la somme totale d'un tableau mais présent dans les sous-totaux. Là aussi « les résultats des différents sous-groupes peuvent chacun être biaisés, mais tous ces biais peuvent s'annuler pour donner des résultats corrects »
- Dès lors, enfin, la « validité est beaucoup plus difficile à vérifier que la fiabilité » et il faut sans doute comparer les résultats avec les « chiffres connus d'autres populations »[85].

Ainsi, selon Moser, le principe méthodologique fondamental dans les enquêtes est de considérer que toutes les erreurs sont des erreurs qui ne s'annulent pas, « jusqu'à preuve du contraire »(Moser, 1951, p. 36), et c'est sur ce type d'erreurs systématiques que doit se concentrer la recherche, y compris par le truchement d'expérimentation, recherche que Moser considère comme « *urgently necessary* » (Moser, 1951, p. 38). Et cette recherche doit inclure aussi bien les statisticiens, que les sociologues et les psychologues.

Ce détour par les sciences humaines et sociales montre à quel point la question du biais, conçu comme erreur systématique et donc comme menace à l'utilisation des statistiques et de son arsenal de tests de signification, apparait décisive aux yeux de ceux – et ils sont nombreux – qui entendent appliquer les outils statistiques à la psychologique, la sociologie, ou encore à la psychologie sociale et aux études sur l'opinion publique. Il permet aussi d'avancer l'hypothèse que le biais est devenu un concept scientifique à travers une rencontre entre son sens psychologique trivial de préjugé et son sens statistique d'erreur systématique, c'est-à-dire non-aléatoire. La

[85] « *Now, three difficulties seem to be inherent in trying to, detect bias: a. Obviously, what constitutes bias on one type of survey, or even question, may not do so on another. In the one case errors due to a particular source accumulate, in the other they cancel out. b. Just as important, but not as often recognised, is the possibility that bias may be absent from all the marginal totals, but present in the sub-totals or cells, of a table. The results of different sub-groups may each be biased, but all the biases may cancel out to give correct total results. If the various breakdowns are to be presented and used separately, this is of great importance. It is not sufficient therefore to demonstrate the accuracy of the over-all distributions. c. Validity is much more difficult to check than reliability. Results must be compared with known population figures and such checks, especially for sub-groups.»,* in Moser, 1951, p. 36.

question qui reste en suspens est de savoir s'il y a un concept spécifiquement épidémiologique de biais. Il convient donc à présent de reprendre le fil de notre histoire de l'épidémiologie moderne, là où nous l'avons laissée, c'est-à-dire au début des années 1950.

CHAPITRE 3 : DU PROBLEME DE L'ECHANTILLONNAGE AU PROBLEME DE LA VALIDITE DE L'INFERENCE STATISTIQUE.

En 1953, un article nécessite d'être analysé plus en détail étant donné le nombre anormalement élevé (pour l'époque) d'occurrences du mot « biais » : celui-ci apparait en effet 47 fois dans l'article écrit par Donald Mainland, publié dans le *American Heart Journal*[1], et qui porte sur le risque de tirer des conclusions fallacieuses sur l'incidence des maladies en se fondant sur les données de l'autopsie.

Tout d'abord, il faut noter que cet article est de nature explicitement méthodologique et ne porte pas sur un problème médical ou épidémiologique spécifique comme ceux de Doll et Hill par exemple, et ce même s'il est publié dans une revue médicale et non statistique : Mainland s'adresse donc, tout comme Hill dans ses articles publiés dans le *Lancet*, aux médecins, et spécifiquement pour Mainland aux médecins nord-américains. Ceci peut expliquer en partie pourquoi le mot « biais » apparait si souvent. De même, il convient de préciser qui est l'auteur de cet article. Donald Mainland (1902-1985) est, à l'époque où l'article est publié, professeur de statistiques médicales à l'Université de New York, après avoir obtenu son diplôme de médecin en Angleterre et être parti au Canada au département d'Anatomie de l'Université Dalhousie à Halifax. En 1953, il a déjà publié deux ouvrages consacrés aux statistiques médicales, le premier en 1938 (celui qui d'après Farewell aurait poussé Hill à accélérer la publication de ses articles dans le *Lancet* sous la forme d'un ouvrage, après qu'Hill en aurait été informé par Ronald Fisher[2]) et le second en 1950. Il a aussi publié un article en 1936 dans le *British Medical Journal*, intitulé « *Problems of Chance in Clinical Work »*[3], où apparait d'ailleurs la notion de biais :

> « *In medical work there is serious neglect of highly important problems of chance and of variation between samples, exemplified in a recent estimate of chances of cure in facial paralysis where odds are wrongly estimated, and too great reliance is*

[1] Mainland, Donald, « The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease », *American Heart Journal*, vol. 45 / 5, Mai 1953, p. 644-654

[2] « *However Fisher's letter also alarmed Hill somewhat by asking 'I do not know if you know Donald Mainland, who has, I understand, a work on tests of significance in medicine now in the hands of the publishers.' Hill replied (12th April 1937) that he did not know Donald Mainland and asked 'is he the man who had an article in the BMJ some time back?' He repeated the question to Pamela Kettle at The Lancet and continued 'whoever he may be his appearance is rather annoying. I leave it to you to use as an argument for speed or to know nothing about.* », in Farewell et Johnson, 2012, p. 485.

[3] Mainland, Donald, « Problems of Chance in Clinical Work », *British Medical Journal*, vol. 2 / 3943, 1936, p. 221-224.

placed on samples. What is required in clinical work is not elaborate mathematical tests, but an understanding of the meaning of chance, and adequate precautions that the samples, however small, are unbiased. » (Mainland, 1936, p. 224).

Nous retrouvons ici le problème de l'échantillon et de sa représentativité vis-à-vis de l'univers dont il est issu. C'est ce problème de la représentativité que Mainland va traiter dans l'article en question. Il va pour cela s'inscrire dans la continuité d'un article, désormais connu de tous les épidémiologistes, écrit par Berkson et publié en 1946, où celui-ci va démontrer ce que l'on appelle depuis le « biais de Berkson » (« *Berkson's bias* ») ou le « sophisme de Berkson » (« *Berkson's fallacy* ») ou encore le « paradoxe de Berkson », bien qu'aucun de ces mots n'apparaisse dans l'article en question. Il est donc nécessaire de faire un détour par cet article si l'on veut comprendre les enjeux liés à cette question.

3.1 Le biais de Berkson.

Joseph Berkson (1899-1982) a la particularité d'être à la fois physicien, médecin et statisticien ; et à l'époque où il publie son article intitulé *Limitations of the Application of Fourfold Table Analysis to Hospital Data*[4], il est le chef de la division « Biométrique et statistiques médicales » de la *Mayo Clinic* de Rochester : il est donc un statisticien réputé, et même, selon Richard Doll, « le plus éminent statisticien médical des Etats-Unis »[5]. Dans cet article, abondamment commenté et discuté depuis sa parution, Berkson montre que deux maladies, en l'occurrence dans l'article la cholécystite et le diabète, qui sont indépendantes dans la population générale peuvent devenir « artificiellement corrélées » (« *spuriously associated* ») dans une étude cas-témoins réalisée sur une population de patients hospitalisés. Cet article est très intéressant du point de vue de notre enquête car il est symptomatique de la confusion importante qui règne autour de la notion de biais et de la difficulté qu'il y a à en saisir le concept.

En effet, si on lit les commentaires récents et moins récents sur cet article, comparativement par exemple à ce que peut en dire Mainland, on s'aperçoit que Berkson et ses commentateurs ne parlent pas de la même chose. Ainsi, pour Vinéis,

[4] Berkson, J. « Limitations of the Application of Fourfold Table Analysis to Hospital Data. » *Biometrics Bulletin* 2 1946, p. 47-53.

[5] Doll, Richard, « Sir Austin Bradford Hill: A personal view of his contribution to epidemiology », *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 1995, p. 155–163. La citation est à la page 161.

Berkson a dans son article « soulevé des doutes sur la validité des recherches épidémiologiques dans un contexte hospitalier »[6] et Vinéis en fait une sous-catégorie du biais de sélection. De même, Snoep, Morabia, Hernandez-Diaz, Hernan, et Vandenbroucke, retracent dans leur article[7] les débats sur le biais de Berkson en opposant ceux qui considèrent ce biais comme s'appliquant uniquement à une relation maladie-maladie, alors que certains pensent qu'il peut aussi s'appliquer à une relation exposition-maladie dans le cadre d'une étude cas-témoins à l'hôpital. Ils considèrent même que « le sophisme de Berkson n'a eu que très peu, voire pas du tout d'influence sur les résultats des études épidémiologiques »[8].

Or, il ne nous semble pas évident que ce que Berkson voulait démontrer était la non-représentativité d'un échantillon de patients hospitalisés par rapport à la population générale, ou plutôt que ce problème de la non-représentativité d'un échantillon de patients hospitalisés par rapport à la population générale n'est qu'un cas particulier d'un problème plus général qui tient à l'utilisation même du tableau 2x2, utilisé, selon les mots mêmes de Berkson, comme « un paradigme de l'analyse statistique »[9].

3.1.1 Méthode statistique et méthode expérimentale.

En effet, Berkson débute son article par une présentation de la méthode utilisée en laboratoire, donc de la méthode expérimentale :

> « *In the biologic laboratory we have a method of procedure for determining the effect of an agent or process that may be considered typical. It consists in dividing a group of animals into two cohorts, one considered the "experimental group," the other the "control." On the experimental group some variable is brought to play; the control is left alone. The results are set up as in table 1-a. If the results show that the ratio a:a + b is different from the ratio c:c + d, it is*

[6] « *He raised doubts about the validity of epidemiological research within hospital settings*», in Vineis, Paolo et Michael, Anthony J., « Bias and confounding in molecular epidemiological studies: special considerations », *Carcinogenesis*, vol. 19 / 12, 1998, p. 2063-2067

[7] Snoep, J. D., Morabia, Alfredo, Hernandez-Diaz, S., Hernan, M. A., et Vandenbroucke, Jan P. « Commentary: A Structural Approach to Berkson's Fallacy and a Guide to a History of Opinions about It ». *International Journal of Epidemiology* 43, 2, 1 avril 2014, p. 515-21.

[8] « *It is likely that Berkson's fallacy has had very limited, if any, impact on the findings of epidemiological studies.* », in Snoep, Morabia, Hernandez-Diaz, Hernan, et Vandenbroucke, 2014, p. 516.

[9] « *it is used as a paradigm of statistical analysis* », in Berkson, 1946, p. 47.

considered demonstrated that the process brought to bear on the experimental group has had a significant effect. »(Berkson, 1946, p. 47).

Puis il met en rapport cette méthode avec la méthode utilisée en statistique, pointant la grande similarité entre les deux méthodes, similarité ou « équivalence apparente » avec la procédure expérimentale qui pourrait expliquer selon lui pourquoi cette méthode en est venue à faire autorité dans le domaine des statistiques :

« A similar method is prevalent in statistical practice, which I venture to think has come into authority because of its apparent equivalence to the experimental procedure. In Biometrika it is referred to as the fourfold table and it is used as a paradigm of statistical analysis. The usual arrangement is that given in table 1-b. The entries, a, b, c and d are manipulated arithmetically to determine whether there is any correlation between A and B. A considerable number of indices have been elaborated to measure this correlation. (…). In essence, however, all these indices measure in different ways whether and how much, in comparison with the variation of random sampling, the ratio a:a + b differs from the ratio c:c + d. If the difference departs significantly from zero, there is said to be correlation, and the correlation is the greater the greater the difference. » (Berkson, 1946, p. 47)[10]

Berkson va même illustrer la grande similarité entre les deux méthodes en mettant côte à côte les deux tableaux à double entrée, celui à gauche (a) étant « typique de la situation expérimentale », celui à droite (b) étant le tableau sous sa « forme statistique » (Berkson, 1946, p. 47) :

[10] C'est Berkson qui souligne.

Figure 3-1 : Comparaison par Berkson entre un tableau à double entrée dans la situation expérimentale et sous sa forme statistique[11]

Table 1. Fourfold Tables

a

Typical of experimental situation

Group	Effect	No Effect	Total
Experimental	a	b	a+b
Control	c	d	c+d
Total	a+c	b+d	a+b+c+d

b

Statistical form

Group	A	Not A	Total
B	a	b	a+b
Not B	c	d	c+d
Total	a+c	b+d	a+b+c+d

La simple vue de ces tableaux indique que les deux méthodes sont identiques mais pour Berkson cette identité n'est qu'apparente. En effet, il y a une « distinction majeure entre la méthode utilisée en laboratoire et celle utilisée dans les statistiques pratiques » :

> « *In the experimental situation, the groups, B and not B, are selected before the subgroupings, A and not A, are effected; that is, we start with a total group of unaffected animals.* »
>
> A l'inverse,
>
> « *In the statistical application, the groupings, B and not B, are made after the subgroupings, A and not A, are already determined; that is, all the effects are already produced before the investigation starts.* » (Berkson, 1946, p. 47-48)[12]

Ainsi, si le résultat sous la forme d'un tableau est identique, il reste que les deux méthodes ne sont identiques qu'en apparence, car dans le cas expérimental, l'effet est

[11] Berkson, 1946, p. 50.
[12] Berkson qui souligne.

produit au cours de l'expérience elle-même, tous les sujets inclus au début de l'expérience étant en quelque sorte identiques, et c'est au cours de l'expérience qu'ils vont se différencier, le groupe contrôle n'étant pas soumis à la variable ou au facteur testé, contrairement au groupe expérimental. A l'inverse, dans le cas statistique, et en l'occurrence ici, pour le dire en termes plus contemporains, dans une enquête cas-témoins rétrospective, tous les effets du facteur ou de la variable testé se sont déjà produits avant le début de l'investigation :

> « *In the end, the tables of the results which are drawn up look alike for the two cases, but they have been arrived at differently.* » (Berkson, 1946, p. 48)[13]

Dès lors, cette différence de méthode dans l'obtention du tableau de résultats implique une interprétation différente des résultats :

> « Corrélativement à cette différence, une interprétation différente pourrait s'appliquer aux résultats. » (Berkson, 1946, p. 48).

C'est donc bien la question de l'échantillon choisi par l'investigateur qui pose problème, dans la mesure où cet échantillon n'est pas neutre, puisque déjà affecté par le facteur dont on cherche à démontrer l'éventuelle causalité, mais aussi affecté par de nombreux autres facteurs possibles, notamment ici le fait que les données qui vont être utilisées dans le tableau et testées statistiquement sont des données issues d'enquêtes faites à l'hôpital. Mais pour Berkson, les données hospitalières (« *Hospital data »*) ne sont qu'un « cas particulier d'un problème plus général » (« *a specific case of a kind »*[14]). Berkson ajoute même en conclusion de son article que « des résultats identiques à ceux de cet article apparaitraient si l'échantillonnage était appliqué à des cartes distribuées au hasard plutôt qu'à des patients » (Berkson, 1946, p. 51).

Quel est donc le problème soulevé par Berkson, c'est-à-dire, pour reprendre les éléments du titre, quelles sont les limites de l'application d'un tableau 2x2 aux données hospitalières ? En d'autres termes, en quoi consiste le biais ou le sophisme de Berkson?

[13] C'est Berkson qui souligne
[14] Berkson, 1946, p. 48.

3.1.2 La première démonstration mathématique d'un biais de sélection.

En fait le problème soulevé par Berkson est un problème purement théorique et proprement statistique, qui est d'ailleurs traité de manière complètement *a priori*, c'est-à-dire ici algébrique : pour Greenland (1987) il s'agit même de « l'analyse algébrique la plus précoce d'un biais de sélection »[15]. Pour dire les choses plus clairement, aucun des chiffres qui sont donnés dans les huit tableaux de l'article ne réfère à une quelconque réalité ou à des données hospitalières réelles : il faudra ainsi attendre l'article de Roberts, Spitzer, Delmore, et Sackett en 1978[16] pour que la démonstration empirique, c'est-à-dire fondée sur des données réelles, soit faite de ce biais. Berkson entend en fait montrer – en partant de l'exemple d'un préjugé répandu chez les médecins des années 1930[17] qui pensaient que la cholécystite était un agent qui causait ou qui aggravait le diabète (ce qui conduisait certains à une ablation de la vésicule biliaire comme traitement préventif du diabète) – non seulement que deux maladies peuvent apparaitre comme étant artificiellement corrélées dans le cadre d'une étude cas-témoins effectuée à l'hôpital, comme le souligneront la plupart des commentateurs, mais plus largement que :

> « les corrélations trompeuses auxquelles il est fait référence [dans cet article] ne découlent pas d'hypothèses portant sur des forces biologiques, ou sur la sélection directe de probabilités corrélées, mais sont simplement le résultat de la combinaison de probabilités indépendantes »[18].

Il est important de noter à ce stade que Berkson emprunte la notion de «*spurious correlation* » à Karl Pearson, qui est le premier à l'utiliser dans son article de 1896[19] : elle peut être définie comme une corrélation qui est due uniquement à la

---

[15] Greenland, Sander. *Evolution of Epidemiologic Ideas: Annotated Readings and Concepts*. Epidemiology Resources, 1987, p. 86.

16 Roberts, Robin S., Spitzer, Walter O., Delmore, Terry, Sackett, David L. , « An empirical demonstration of Berkson's bias », *Journal of Chronic Diseases*, vol. 31 / 2, 1978, p. 119–128.

17 La conférence qui précède l'article en question est prononcée en 1938 devant l'*American Statistical Association*. Berkson explique en 1955 que l'idée de cet article lui est venue suite à une étude de 1929 menée par Raymond Pearl à partir de données d'autopsie, qui suggérait, à tort, un rôle protecteur de la tuberculose vis-à-vis du cancer du poumon. Cet article serait donc l'explication statistique donnée par Berkson à l'établissement d'une corrélation négative trompeuse (« *spurious* ») entre tuberculose et cancer du poumon.

[18] « *the spurious correlations referred to are not a consequence of any assumptions regarding biologic forces, or the direct selection of correlated probabilities, but are the result merely of the ordinary compounding of independent probabilities.* », in Berkson, 1946, p. 51.

[19] Pearson, Karl, « Mathematical Contributions to the Theory of Evolution. — On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs », *Proceedings of the Royal Society of London (1854-1905)*, vol. 60 / 1, janvier 1896, p. 489-498.

« manipulation des observations » ou encore, comme il la définit dans un cours donné en 1913, comme une « corrélation qui est produite par une opération arithmétique [« *a process of arithmetic* »] et non par une quelconque relation organique parmi les quantités traitées »[20]. Il faut ajouter que pour Pearson, la causalité n'est qu'un cas particulier de corrélation, ou plutôt, que la corrélation constitue « la catégorie plus large par laquelle nous devons remplacer la vieille idée de causalité »[21]. Or, c'est exactement ce que Berkson veut montrer ici : comment une simple manipulation des chiffres ou des données peut conduire à montrer une corrélation alors qu'il n'y en a pas.

Dès lors, pour bien comprendre la portée de la démonstration de Berkson, il convient de la retracer dans le détail. Berkson pose le problème de la façon suivante :

> « *The authorities of a hospital wish to know whether their accumulated records of incidence, examined statistically, support this practice [l'ablation de la vésicule biliaire]. On the face of it, it would appear that we have here the typical and elementary problem of the comparison of rates in a fourfold table.* » (Berkson, 1946, p. 48).

Berkson dresse alors un tableau 2x2 qui fait apparaitre une différence significative entre la cholécystite et le diabète (de l'ordre de + 1.58%). Mais il soulève aussitôt l'objection selon laquelle le groupe contrôle (ici ceux qui ne sont pas diabétiques) ne soit pas représentatif, et pour l'éviter il va sélectionner au sein du groupe des non-diabétiques ceux qui souffrent d' « erreurs de réfraction » (autrement dit, de problèmes de vue). Là encore la « différence est positive » (Berkson, 1946, p. 48), de l'ordre de 2.32%. Il ajoute que « bien sûr, dans toute analyse détaillée, nous souhaiterions garder l'âge et le sexe constants, nous renseigner sur la fiabilité du diagnostic, et ainsi de suite »[22], pour aussitôt préciser son intention : « mais le problème dont il s'agit dans cet article n'a rien à voir avec ces questions » et il faut

[20] Voir à ce sujet l'article de Aldrich, John, « Correlations genuine and spurious in Pearson and Yule », *Statistical science*, 1995, p. 364–376, en particulier p. 365-366.

[21] Pearson, Karl, *The Grammar of Science*, 3ème édition, Edinburgh, Black, 1910, p. 157. Cité dans Aldrich, 1995, p. 365.

[22] « *Of course, in any detailed analysis we should wish to keep age and sex constant, inquire into the reliability of the diagnoses, and so forth.* », in Berkson, 1946, p. 48.

donc « dans l'intérêt de notre argument, considérer que tous ces facteurs ont été adéquatement contrôlés »[23]. Pour autant, ajoute-t-il :

> « Même si c'est le cas, est-ce que les résultats permettent de conclure quant à la question de savoir si la cholécystite est biologiquement corrélée avec le diabète ? »[24]

Berkson rentre alors au cœur de son argumentaire, qui porte ici sur la question de la représentativité de la population hospitalière par rapport à la population générale. Il assigne alors à chaque maladie une probabilité d'incidence ($P_{d\ (pour\ diabète)} = 0.01$, $P_{c\ (pour\ cholécystite)} = 0.03$, and $P_{r\ (pour\ réfraction)} = 0.10$) et se fonde sur une population de 10 000 000 de personnes pour distinguer différents cas de figure (ceux qui ont un diabète et rien d'autre, ceux qui ont une cholécystite et rien d'autre, ceux qui ont un diabète et des problèmes de vue, ceux qui ont les trois, ou encore ceux qui n'ont rien). Cela donne le tableau 4 dans son article (Berkson, 1946, p. 51), dans lequel il constitue sa cohorte en fonction de la probabilité différentielle des maladies. Une fois cette population fictive constituée, Berkson fait deux nouveaux tableaux à double entrée Le premier tableau (Berkson, 1946, p. 52) montre clairement qu'il n'y aucune différence entre les trois groupes, que l'on compare le groupe des personnes qui sont diabétiques et qui ont une cholécystite avec le groupe des non-diabétiques ou bien avec le groupe de ceux qui ont des problèmes de vue, puisque la prévalence de la cholécystite est dans tous les cas de 3%, ce qui est normal, puisque selon Berkson, « il n'y a pas de corrélation » (Berkson, 1946, p. 48), alors même que leur taux d'incidence était différent. Mais ce tableau n'est que la prémisse de son argument.

En effet il va ensuite assigner à chacune des trois maladies une probabilité « que leurs victimes soient sélectionnées pour aller à l'hôpital », et considérer que les « probabilités de sélection opèrent de façon indépendante » (Berkson, 1946, p. 48-49). Ceci va lui permettre de montrer plusieurs choses :

- Tout d'abord, en posant les équations de probabilité pour chaque maladie et en distinguant la population générale (N) et la population de l'hôpital (N'), Berkson montre qu'une personne avec plusieurs maladies a beaucoup plus de chances d'aller à l'hôpital qu'une personne avec une seule maladie ; plus

[23] « *But the point referred to in this paper has no relation to such questions, and for the sake of the argument we shall consider that all such factors have been adequately controlled.* », in Berkson, 1946, p. 48.

[24] « *Even so, do the results permit any conclusion as to whether cholecystitis is biologically correlated with diabetes?* », in Berkson, 1946, p. 48.

précisément que la probabilité d'aller à l'hôpital augmente corrélativement avec le nombre de maladies : une personne qui a deux maladies a deux fois plus de chances d'aller à l'hôpital qu'une personne qui n'en a qu'une, une personne qui trois maladies a trois fois plus de chances d'aller à l'hôpital qu'une personne qui n'en a qu'une, et ainsi de suite (Berkson, 1946, p. 49).

- Ensuite Berkson va considérer que le taux de sélection (pour aller à l'hôpital) est égal pour chaque maladie dans la population générale et leur assigner à chacune la probabilité de 0.05, ce qui lui permet de calculer, toujours à partir de la population fictive de 10 000 000 de personnes (N), la population espérée à l'hôpital.

Dans le second tableau (Berkson, 1946, p. 52), ainsi, la corrélation entre la cholécystite et le diabète apparait un peu négative (-0.24%), en raison du fait que si, dans la population générale, l'incidence de la cholécystite était identique chez les personnes diabétiques et chez celles qui avaient des problèmes de vue, ce n'est pas le cas dans la population hospitalière où l'incidence de la cholécystite est inférieure chez les personnes diabétiques par rapport à celles qui ont des problèmes de vue.

Enfin, Berkson va assigner des taux de sélection différents en fonction des maladies : 0.15 pour la cholécystite, 0.05 pour le diabète et 0.20 pour les problèmes de vue. Ceci lui permet de faire un nouveau tableau 2x2 (Berkson, 1946, p. 53) où il apparait que l'incidence de la cholécystite dans le groupe diabétique (8,55%) est presque le double de celle du groupe contrôle (4,72%), ce qui montrerait une corrélation positive entre cholécystite et diabète. Or, nous dit Berkson : « ce ne serait pas représentatif de la population générale et n'aurait aucune signification biologique [25]», précisément parce qu'il s'agit en réalité d'un problème algébrique, dont Berkson donne d'ailleurs la formule (Berkson, 1946, p. 53).

3.1.3 Taux d'admission différentiel et représentativité de l'échantillon.

Ainsi, si l'on assigne à chaque maladie une probabilité sélective indépendante de 0.05, la probabilité d'avoir la cholécystite et des problèmes de vue (P=0.93) est supérieure à celle d'avoir la cholécystite et un diabète (P=0.91), tandis que si l'on

[25] « *It would be quite unrepresentative of the situation in the general population and of no biologic significance*. », in Berkson, 1946, p. 49.

assigne la probabilité de 0.15 à la cholécystite, de 0.05 au diabète, et de 0.2 aux problèmes de vue, on constate que la probabilité d'avoir la cholécystite et des problèmes de vue (P=0.22) est inférieure à celle d'avoir la cholécystite et le diabète (P=0.31). Dès lors, tout dépend effectivement de la probabilité initiale, qui est par ailleurs inconnue et inconnaissable, que la pathologie en question induise ou non une hospitalisation : on peut donc en conclure que toutes les corrélations qui sont mises en évidence dans les différents tableaux ne traduisent pas une corrélation (ou une absence de corrélation) entre une pathologie et une autre, mais sont le reflet d'une probabilité différentielle d'être admis à l'hôpital en fonction de la pathologie ou des pathologies dont la personne souffre, ou, comme le dit Greenland, que « des maladies puissent paraître associées, dans une étude cas-témoins fondée sur une population hospitalisée, uniquement en raison de taux d'admission supérieurs parmi les personnes qui ont plusieurs maladies, même si ces maladies déterminent de façon indépendante les taux d'admission »[26].

L'article se termine par un commentaire de Berkson où il va se justifier et conclure. Tout d'abord il pose que le taux de sélection différentielle appliquée aux différentes maladies est assez conforme à la réalité dans la mesure où un patient a plus de chances d'aller à l'hôpital en fonction de la maladie dont il souffre. Par contre il précise que le fait que ces probabilités soient indépendantes est très simplificateur (« *oversimple »*) car, « en général, on peut supposer que si un patient souffre de deux maladies, leurs symptômes vont mutuellement s'aggraver, et le patient va plus facilement s'en apercevoir », ce qui va mécaniquement surreprésenter les polypathologies à l'hôpital, et donc accroitre la discordance entre la population générale et la population à l'hôpital[27]. Berkson peut ainsi conclure qu'il est « dangereux d'appliquer à une population hospitalisée la méthode de l'analyse du tableau 2x2 dans le cadre d'une enquête sur une corrélation entre différentes maladies [28]», ainsi que

[26] « *Described in modern terminology, Berkson studied the phenomenon that diseases can appear associated in a hospital-based case-control study solely on account of higher admission rates among persons with multiple conditions, even if the conditions affect admission rates independently* », in Greenland, 1987, p. 86.

[27] « *In general we may guess that if a patient is suffering from two diseases, each disease is itself aggravated in its symptoms and more likely to be noted by the patient. So far as this difference of fact from assumption goes, its effect would be to increase relatively the representation of multiple diagnoses in the hospital, and in general to increase the discrepancy between hospital and parent population, even more than if the probabilities were independent.* », in Berkson, 1946, p. 50.

[28] « *It appears from the development that it is hazardous to apply in a hospital population the method of the fourfold table analysis for an inquiry into the correlation of diseases*. », in Berkson, 1946, p. 50.

dans d'autres cas, comme la comparaison entre les ouvriers et les agriculteurs, puisqu'ils ne sont pas représentés dans les mêmes proportions à l'hôpital que dans la population générale.

Par contre, il ajoute que dans certains cas, la « comparaison n'est pas fondamentalement invalide », notamment quand le taux de sélection pour une condition particulière (comme la couleur des yeux, ou le type anthropologique) est égal à zéro et qu'on cherche à savoir si ces caractéristiques particulières sont en corrélation avec une maladie particulière, par exemple les yeux bleus et des problèmes de vue ; ou encore quand chacun des groupes de cas souffre d'une seule maladie, et que le taux de sélection (pour entrer à l'hôpital) est globalement identique. Dans tous les autres cas, la comparaison n'est pas valide et il ne semble y avoir aucun moyen de corriger cette corrélation artificielle qui existe au sein de la population hospitalisée[29]. C'est pourquoi Berkson insiste dès le début de son article sur la distinction entre la méthode du laboratoire et la méthode statistique : dans le premier cas on va de l'exposition à la maladie (de la cause à l'effet), tandis que dans le cas statistique, on va de la maladie à l'exposition (de l'effet à la cause). Or dans la mesure où la maladie influence la sélection des patients dans le groupe ou dans l'échantillon (ici le taux d'admission des patients à l'hôpital), il est impossible de déterminer une corrélation entre l'exposition et la maladie qui ne soit pas le simple effet d'un taux d'admission différentiel qui ferait que telle pathologie serait sous ou sur-représentée dans l'échantillon, d'autant plus si le fait d'avoir plusieurs pathologies augmente d'autant la probabilité d'être admis à l'hôpital.

Dès lors, l'affirmation de Berkson selon laquelle les données hospitalières (« *Hospital data »*) ne sont qu'un « cas particulier d'un problème plus général » (« *a specific case of a kind »*[30]) et que « des résultats identiques à ceux de cet article apparaitraient si l'échantillonnage était appliqué à des cartes distribuées au hasard plutôt qu'à des patients »(Berkson, 1946, p. 51), est logique car il s'agit simplement

[29] « *However, the formulas given indicate some special cases in which comparison is not basically invalid. If the selective rate for any particular condition is zero, the relative incidence of that condition in several disease groups may be validly examined, regardless of the selective rates affecting the other groups. This refers to inquiries in which for instance eye color or anthropologic type is examined in various disease groups to ascertain whether there is correlation between these characters and disease. If each of the disease groups examined consists of only one disease, for example, diabetes or refractive errors but not both, and if the selective rates for these two groups do not differ appreciably then also it is valid to compare the incidence in them of cholecystitis, even though the latter disease is not fairly represented in the hospital* », in Berkson, 1946, p. 50.

[30] Berkson, 1946, p. 48.

du « résultat de la combinaison de probabilités indépendantes » (Berkson, 1946, p. 48). Nous retrouvons en fait ici le problème tel qu'il est soulevé par Bradford Hill dans ses *Principles of Medical Statistics* : la population d'un hôpital est biaisée au sens où c'est un échantillon qui n'est pas représentatif de la population générale, car la probabilité d'être admis à l'hôpital varie en fonction de la maladie en question mais aussi de la présence ou non de plusieurs maladies. Comme le dirait Hill, « un membre de cet univers d'où est tiré l'échantillon a plus de chances d'apparaître qu'un autre»[31] , ce qui empêche toute généralisation qui irait de la population à l'hôpital vers la population générale, donc toute inférence statistique. C'est d'ailleurs ce problème de la représentativité de l'échantillon qui va occuper Donald Mainland dans son article de 1953, qu'il convient à présent d'examiner.

3.2 Mainland et le problème du biais.

3.2.1 Berkson et Mainland : du sophisme au biais.

L'article de Mainland s'inscrit directement dans le prolongement de celui de Berkson, dont il résume les points essentiels dans les trois premiers paragraphes. Berkson a d'ailleurs relu une version préparatoire de l'article de Mainland (Mainland, 1953, p. 654), ce qui permet à Mainland d'affirmer que les « différents aspects du problème ont été correctement et suffisamment traités » (Mainland, 1953, p. 654). Mainland résume ainsi l'argument de Berkson:

> « *The frequency of cholecystitis was found to be higher in the diabetics by an amount that was statistically significant, i.e., greater than investigators are prepared to attribute to chance. Berkson showed, however, that such results could be entirely fallacious under two conditions which must very often exist. In brief these are: (1) that the occurrence of two disorders in the same person gives him an increased probability of admission to a hospital or clinic, and (2) that the persons with the disorders under investigation are not represented in the hospital or clinic population in the same proportions as in the general population.*» (Mainland, 1953, p. 644).

[31] « *to be in a position to generalise, the sample must be representative of the population to which it belongs.* », in Hill, 1950, p. 24-25

Le but de Mainland est en fait de vulgariser la démonstration de Berkson, non pas pour le grand public mais pour la communauté médicale, cette démonstration étant restée selon lui confinée au cercle des statisticiens : le choix de publier son article dans l'*American Heart Journal*, une des principales revues américaines de cardiologie, alors même qu'il ne fait aucune référence dans son article à quoi que ce soit qui pourrait relever de cette spécialité médicale, est sur ce point significatif. De même il prend la peine de préciser que la démonstration de Berkson ne dépend pas de « considérations complexes ou théoriques ». Il dit ainsi :

> « *The importance of the demonstration is well recognized by those who are acquainted with it; but six years after its publication, if one can judge from research papers brought to statisticians for analysis, or sent to statistical referees by editors of medical journals, the demonstration has not yet become widely known, perhaps because it appeared in a statistical journal.(…) It does not, however, depend on complex or theoretical considerations, either statistical or medical* » (Mainland, 1953, p. 644).

Il précise même directement les cibles de son propos, qui sont aussi bien des investigateurs d'une recherche que les cliniciens :

> « *It seems necessary to bring the fallacy directly to the attention of investigators, and that is the purpose of this article. (…) Clinical readers who are unfamiliar with Berkson's work will, however, be able to see from the discussion how the bias can occur in their own fields* » (Mainland, 1953, p. 644).

Avant de rentrer dans le détail de son argumentation, l'utilisation du mot « biais » par Mainland nous semble problématique : en effet, il considère que le sophisme (« *fallacy »*) de Berkson est un biais, ou même un « genre de biais » (« *a kind of bias »* [32]). Or nous savons que Berkson n'a jamais employé ce mot dans son article. Et Mainland fait même dire à Berkson, au style indirect libre, la chose suivante :

> « *More recently he*[2] [Berkson] *expressed the situation in general terms: Although we have long recognized hospital samples as biased samples (not representative of the general population of sick people) we have been slow to recognize the simple corollary, that unbiased conclusions about the*

[32] Mainland, 1953, p. 644.

relationships of two diseases cannot be derived from such biased samples. » (Mainland, 1953, p. 644-645).[33]

Deux choses méritent d'être notées :

- Tout d'abord il nous semble douteux, mais néanmoins possible, que Berkson ait effectivement employé le mot « biais » et ce pour deux raisons : premièrement il n'emploie jamais ce mot au cours de l'article de 1946 ; deuxièmement, Berkson « n'aime pas le mot « biais », car cela paraitrait impliquer une forme de tromperie délibérée »[34], comme il le souligne en commentant l'article de Doll et Hill de 1956 sur la mortalité des médecins. Dans cet article d'ailleurs, quand il emploie le terme « biais », il le met entre guillemets, comme s'il citait Doll et Hill tout en s'en désolidarisant :

  «*The observed associations are "spurious," that is, they have no biologic significance, but are the result of the interplay of various subtle and complicated "biases"* » (Berkson,1958, p. 37).
- Ensuite on peut se demander pourquoi Mainland utilise quant à lui le terme de « biais » (à 47 reprises) dans son article, et ce qu'il entend par là. En effet, il nous semble par exemple que, dans la citation ci-dessus, l'adjectif « *unbiased* » qui qualifie les « conclusions » et l'adjectif « *biased* » qui qualifie les échantillons (« *samples* ») ne soient pas tout à fait des antonymes stricts. En effet, un échantillon biaisé est un échantillon qui a été « sélectionné », au sens où Hill, ainsi d'ailleurs que Berkson dans son article de 1958[35], entendent ce terme, c'est-à-dire que cet échantillon, ici la population de patients hospitalisés, n'est pas « représentatif de la population générale des patients malades », comme le dit Mainland[36]. A l'inverse, une conclusion non-biaisée ne peut pas être une conclusion qui serait représentative de la population, cela n'aurait aucun sens. Un synonyme de

[33] La note 2 dans l'article de Berkson renvoie à une conférence de l'*American Statistical Association*, où Berkson aurait tenu ses propos. La référence exacte est : Berkson, Joseph: *Remarks in Discussion at Meeting of American Statistical Association*, Boston, 1951. Nous n'avons malheureusement pas retrouvé le texte en question, et nous supposons qu'il s'agit d'une discussion qui n'a pas fait l'objet d'une retranscription écrite.

[34] « *I do not like the term "bias," since it may seem to imply conscious deception* », in Berkson, Joseph, « Smoking and Lung Cancer: Some Observations on Two Recent Reports », *Journal of the American Statistical Association*, vol. 53 / 281, mars 1958, p. 28-38.

[35] « *However, it must be pointed out, that the observations constituting the two main variables under study, namely a history of smoking and the cause of death, are subject to considerable and even extreme error, and the samples are selected, possibly highly selected* », in Berkson, 1958, p. 35

[36] Mainland, 1953, p. 645

> « *unbiased »* dans ce contexte renverrait plutôt à l'idée de validité, au sens logique du terme où l'on dit qu'une conclusion est valide en fonction des prémisses qui sont posées et du raisonnement suivi : ici, par exemple, on ne peut, *par principe,* c'est-à-dire indépendamment du résultat du test statistique, tirer aucune conclusion quant à la corrélation entre deux maladies, précisément parce que les prémisses, en l'occurrence ici les échantillons, seraient invalides, et qu'il est donc strictement impossible de faire une inférence quelconque (ici une inférence statistique) à partir de ces prémisses.

Pour clarifier la situation, il convient donc d'examiner plus en détail l'argumentation de Mainland. Celui-ci va en fait appliquer le sophisme de Berkson à une autre situation, tout aussi fictive et à travers un « exemple simplifié ». Mainland prend ainsi trois maladies, nommées A, B et X : il s'agit de savoir si la maladie X est plus fréquente chez les personnes qui ont A ou chez celles qui ont B, en se fondant sur les rapports d'autopsie. Il suppose que la prévalence de X est identique chez les patients qui ont A et chez ceux qui ont B (10%), et qu'elle est identique dans toute la population. Pour simplifier encore plus son exemple et « enlever des biais potentiels » (Mainland, 1953, p. 645), il suppose que tous les morts sont autopsiés, que l'investigateur dispose de tous les rapports d'autopsie, et que les seules causes de décès dans la population sont A, B et X. Enfin la population des personnes qui ont A est identique à celle de ceux qui ont B : 1000 personnes pour chaque. Il y a donc 100 A qui ont X et 900 A qui n'ont pas X, ainsi que 100 B qui ont X, et 900 qui n'ont pas X.

Après avoir égalisé toutes ces conditions, Mainland introduit une différence entre les trois maladies A, B et X, qui porte sur leur taux de létalité respectif : 50% pour A, 20% pour B, et 40% pour X. Il fait ensuite un tableau à double entrée où il apparait que pour les A, 13,46% ont la maladie X tandis que pour les B, 22,41% ont la maladie X, ce qui fait une différence d'environ 9%.

Mainland effectue ensuite un test du $\chi^{2,}$ qui lui donne une valeur de 8.81 avec une probabilité d'aboutir à une telle valeur uniquement au hasard inférieure à 1 chance sur 300. Ceci devrait ainsi nous convaincre que « le hasard ne peut pas être le seul responsable » et, dès lors, « nous devrions conclure, à juste titre, que quelque chose de plus que le hasard était à l'œuvre » (Mainland, 1953, p. 646). Or,

> « nous aurions tort de conclure que ce quelque chose est une association plus forte entre X et B qu'en entre X et A dans la population source. Ce « quelque chose » est en réalité un biais dans l'échantillonnage, qui est dû à une létalité plus basse de B, qui cause une plus forte proportion des B décédés à souffrir de X que de A à souffrir de X. Ce phénomène peut être décrit comme une compétition entre les taux de létalité. »[37]

3.2.2 Mainland et la première définition du biais en épidémiologie : le biais comme « *mislabelling* ».

Un biais, selon Mainland, est donc une entité indéfinie, un « quelque chose » (dans l'exemple, une compétition entre les taux de létalité de différentes maladies) qui va venir « masquer une réelle association ou bien créer une association fallacieuse » (Mainland, 1953, p. 654), et qui va se manifester auprès de l'enquêteur essentiellement à travers ses effets. Ainsi dans la partie suivante, Mainland va procéder à des modifications des valeurs numériques à l'exemple qu'il a donné précédemment, en réduisant notamment la taille de l'échantillon : cette réduction de l'échantillon conduit à la conclusion qu'il n'y pas de différence significative entre la fréquence de X chez les A et la fréquence de X chez les B. Selon lui, cela « souligne la présence d'un biais, car cela conduit à la situation paradoxale où avec un échantillon plus petit, c'est-à-dire avec moins d'information, il semble que nous nous rapprochions de la vérité (aucune différence réelle dans la fréquence de X entre A et B) , qu'avec un échantillon plus large , c'est-à-dire avec plus d'information, obtenu de la même manière» (Mainland, 1953, p. 647).

Dans un autre article publié deux ans plus tard, en 1955, qui est la première partie d'une série d'articles consacrés à « L'emploi des dossiers des malades dans l'étude de la thérapie et d'autres traits d'une maladie chronique » [38], et qui porte plus

---

[37] « *We should rightly conclude, therefore, that something more than chance was operating; but we should be wrong if we thought that this "something" was a closer association in the parent population between X and B than between X and A. The "something" is in reality a bias in the sampling, due to the lower fatality of B, which causes a higher proportion of the dead B's to be afflicted with X than is found among the dead A's. The phenomenon can be described as a competition among fatality rates*. », in Mainland, 1953, p. 646.

[38] Mainland, Donald, « Use of Case Records in the Study of Therapy and Other Features in Chronic Disease I. Planning the Survey », *Annals of the Rheumatic Diseases*, vol. 14 / 4, décembre 1955, p. 337-352

précisément sur la planification ou le « projet » de l'enquête (« *Planning the survey »*), donc la partie proprement méthodologique, Mainland, après avoir énoncé les neuf questions qu'il faut se poser en planifiant l'enquête[39] va proposer une définition un peu plus précise, ou plutôt un peu moins vague, de la notion de biais :

> « *The nine questions are designed to show:*
>
> *(1) How far such pessimism is justified,*
>
> *(2) How forethought can reduce biases, which can be defined here for reference as "things that make a sample different from what it purports to be".* » (Mainland, 1955, p. 338).

Les biais, ce sont donc les choses qui modifient un échantillon de telle sorte qu'il en devient différent de ce qu'il prétend être ou de ce qu'il est censé être. Il reprendra la même définition tout en la précisant à nouveau un peu plus, dans un autre article en trois parties[40], publié trois ans plus tard en 1958, et consacré intégralement à la planification et à l'évaluation des recherches (« *planning and evaluation of research »*). Cette définition apparait dans la troisième partie de la méthode proposée par Mainland, consacrée à la « Subdivision de la population », et plus spécifiquement à une sous-partie intitulée : « Réduction des biais » (« *Reduction of bias »*) :

> « *Bias is anything that makes a sample different from what it purports to be – it is <u>a mislabeling</u>*. »[41]

Un biais est donc une erreur d'étiquetage ou de caractérisation de l'échantillon : en d'autres termes, les propriétés assignées à l'échantillon ne sont pas correctes, ce

---

[39] « (*I) Who ?-Persons responsible for the original observations and for the survey.*
(2) Why ?-The purposes of the survey.
(3) What?-The population sampled.
(4) Where?-Location and environment (e.g. of a clinic and its patients).
(5) When ?-Time factors.
(6) How?-Methods of clinical observation and of survey.
(7) How much ?-Measurement. (In the broad sense of assessing results, this is covered by Question 6.)
(8) How many ?-Enumeration (numbers of patients and observations).
(9) Why ?-Why did this happen ?-Causal relationships. », in Mainland, 1955, p. 338.

[40] Mainland, Donald, « Notes on the planning and evaluation of research, with examples from cardiovascular investigations. Part I », *American Heart Journal*, vol. 55 / 5, 1958, p. 644–655. Nous ferons référence à ce texte sous la forme abrégée: Mainland, 1958a.

Mainland, Donald, « Notes on the planning and evaluation of research, with examples from cardiovascular investigations. Part II », *American Heart Journal*, vol. 55 / 6, 1958, p. 824–837. Nous ferons référence à ce texte sous la forme abrégée: Mainland, 1958b

Mainland, Donald, « Notes on the planning and evaluation of research, with examples from cardiovascular investigations. Part III », *American Heart Journal*, vol. 55 / 6, 1958, p. 838-850. Nous ferons référence à ce texte sous la forme abrégée: Mainland, 1958c.

[41] Mainland, 1958a, p. 647. C'est Mainland qui souligne.

qui risque d'induire des erreurs au niveau des conclusions qu'on va pouvoir tirer à partir de cet échantillon. Mainland donne un exemple pour expliquer ce qu'il veut dire :

> « *For example, if children and adults had been grouped together in the Rheumatic Fever Trial, an excess of children in, say, the aspirin group, as compared with the cortisone group, might have produced a difference in outcome that was not due to any treatment difference at all, or it might have masked a real benefit due to one of the treatments. The true labels of the groups would have been "Aspirin (Many Children)" and "Cortisone (Few Children)." We should always ask: "What may our labels be hiding?"* » (Mainland, 1958a, p. 647-648).

Ici, l'erreur de caractérisation ou d'étiquetage renvoie à l'âge des membres de l'échantillon : les échantillons ne sont pas comparables entre eux car les membres de l'échantillon A et de l'échantillon B ne sont pas comparables en âge. Il s'agit donc ici plutôt du problème de la comparabilité entre les échantillons, que du problème de la représentativité de l'échantillon par rapport à la population générale ou à l'univers dont l'échantillon est issu, comme c'était le cas dans l'article de 1953.

De façon intéressante pour notre propos, Mainland précise implicitement que cette notion de biais est différente de la notion psychologique et populaire du mot « biais », qui renvoie à la notion de préjugé ou de parti-pris. Il semble même faire ironiquement référence à certains chercheurs médicaux qui en seraient restés à cette signification populaire, le verbe « *to linger* » renvoyant en anglais à l'idée de « persister » ou de « s'attarder » :

> « *The popular notion of bias as a psychologic phenomenon still lingers with some medical investigators*. » (Mainland, 1958a, p. 647).

Et pourtant, dans son article de 1955, où apparait la première définition du mot « biais » que donne Mainland, la notion de biais est précisément traduite, par l'éditeur de la revue, en français par « parti-pris » et en espagnol par « fact*or de parcialidad* » (Mainland, 1955, p. 351), ce qui renvoie non pas au problème d'un échantillon qui ne serait pas ce qu'il prétend être mais à la notion populaire et psychologique du biais, avec laquelle Mainland prend pourtant ses distances. Est-ce une simple erreur de traduction de la part de l'éditeur, ce qui soulignerait d'ailleurs l'ambiguïté qui règne autour de cette notion ? Ou bien faut-il comprendre à l'inverse qu'il existe un lien entre les deux notions, ou plutôt entre les deux sens de cette même notion ?

Une clé de compréhension se trouve peut-être dans un autre article de Mainland, publié celui-ci en 1956 et intitulé « The risk of biased selection in forward-going surveys with nonprofessional interviewers »[42]. Cet article, relativement court, est en fait une étude menée par Mainland et sa collègue Lee Herrera, sur un échantillon de leurs propres étudiants en première année de la faculté de médecine de New York, durant le premier cours de l'année universitaire 1955-1956. Ce questionnaire vise à évaluer le surgissement d'un biais dans la sélection des patients dans le cadre d'une enquête prospective et constitue une critique de la méthode adoptée par Hammond et Horn dans leur étude sur la relation entre le tabagisme et les taux de mortalité menée sur 187 766 hommes[43]. En effet ceux-ci, afin de pouvoir interroger autant de personnes ont choisi de faire appel à des volontaires, en l'occurrence à 22 000 membres de diverses divisions locales de l'*American Cancer Society*, volontaires qui avaient notamment pour tâche de « sélectionner les sujets, distribuer les questionnaires, les récupérer, et de revoir les sujets à intervalles annuels » (Mainland et Herrera, 1956, p. 240). Ces volontaires ont reçu des instructions, aussi bien orales qu'écrites, par les statisticiens de l'*American Cancer Society* quant à la marche à suivre. Mais pour Mainland cela ne garantit en rien la validité scientifique de l'investigation car selon lui :

> *« The characteristics of a scientific investigation (objectivity, avoidance of biased sampling, the use of controls) are, however, not very easy to comprehend, for even a complete medical course does not always eradicate unscientific concepts regarding the requirements of valid evidence.* » (Mainland et Herrera, 1956, p. 240).

Or, justement, il est fort probable, selon Mainland, que les volontaires en question aient sélectionné les « bons cas », c'est-à-dire par exemple ceux qui ont à la fois le comportement (ici, fumer) et la maladie en question ou au moins des symptômes en rapport avec la maladie comme la toux du fumeur, non pas tant dans le but de truquer les résultats mais simplement par acquit de conscience :

> *« It is conceivable, therefore, that some nonprofessional interviewers might believe that the best kind of subject in a habit-disease survey was one who*

[42] Mainland, Donald et Herrera, Lee, « The risk of biased selection in forward-going surveys with nonprofessional interviewers », *Journal of Chronic Diseases*, vol. 4 / 3, 1956, p. 240–244.

[43] Hammond, E. Cuyler et Horn, Daniel, « The relationship between human smoking habits and death rates: a follow-up study of 187,766 men. », *Journal of the American Medical Association*, vol. 155 / 15, août 1954, p. 1316-1328

showed both the habit and the disease. This does not imply a desire to distort the evidence, nor does it imply ability to diagnose a disease. It simply implies that some interviewers might feel that they would be doing a better job if they selected subjects who practiced the habit under investigation, and also had some symptoms attributable to the habit. In a smoking survey, the symptoms might be "smoker's cough" or "chest trouble. » (Mainland et Herrera, 1956, p. 240)

Pour tester l'hypothèse selon laquelle les volontaires auraient sélectionné leur échantillon de telle manière qu'il s'en trouve biaisé, c'est-à-dire ici non représentatif de la population dont il est issu, et afin de quantifier le risque que ce soit le cas, Mainland et Herrera ont ainsi créé un bref questionnaire (6 questions) à destination de leurs 129 étudiants afin de voir ce qui pour eux constituerait un bon sujet («*a good case* ») pour ce genre d'études évaluant le lien entre un comportement et une maladie. Voici le questionnaire avec, entre crochets, les résultats :

Figure 3-2 : Questionnaire donné par Mainland et Herrera à leurs étudiants pour évaluer le risque de biais dans le choix des sujets d'une étude[44]

Q. 1. Would it be valuable, if, before starting the field work, you learned something about the diseases (or their symptoms) attributable to drinking?
YES CERTAINLY [37.2] PROBABLY YES [17.1] CERTAINLY NOT [15.5]
PROBABLY NOT [30.2] NO OPINION [0.0]

Q. 2. Which, if either, would be more valuable?
A—A heavy drinker who is healthy.
B—A heavy drinker who is ill with a disease attributable to drinking.
CERTAINLY A [0.8] PROBABLY A [5.4] CERTAINLY B [1.6] PROBABLY B [7.8]
BOTH EQUALLY VALUABLE [80.6] BOTH OF LITTLE USE [1.6]
NO OPINION [2.3]

Q. 3. Which, if either, would be more valuable?
C—A heavy drinker who is healthy.
D—A heavy drinker who is ill with a disease *not* attributable to drinking.
CERTAINLY C [16.3] PROBABLY C [23.3] CERTAINLY D [0.8]
PROBABLY D [6.2] BOTH EQUALLY VALUABLE [45.0]
BOTH OF LITTLE USE [6.2] NO OPINION [2.3]

Q. 4. Which, if either, would be more valuable?
E—A heavy drinker who is ill with a disease attributable to drinking.
F A nondrinker who is ill with a disease attributable to drinking.
CERTAINLY E [4.7] PROBABLY E [8.5] CERTAINLY F [7.0] PROBABLY F [6.2]
BOTH EQUALLY VALUABLE [69.0] BOTH OF LITTLE USE [3.9]
NO OPINION [0.8]

Q. 5. Which, if either, would be more valuable?
G—A heavy drinker who is ill with a disease *not* attributable to drinking.
H—A nondrinker who is ill with a disease *not* attributable to drinking.
CERTAINLY G [7.8] PROBABLY G [20.9] CERTAINLY H [0.8] PROBABLY H [2.3]
BOTH EQUALLY VALUABLE [33.3] BOTH OF LITTLE USE [33.3]
NO OPINION [1.6]

Q. 6. Have you, at any time, and anywhere, had a course in statistics?
YES [17.8] [NO 82.2]

[44] Mainland et Herrera, 1956, p. 242.

Ainsi, si les réponses aux questions 2, 4, 5 (Mainland et Herrera ont exclu la 3 car ils ont considéré qu'elle était mal formulée) sont « Certainement » ou « Probablement » suivies d'une lettre (A, B, E, F, G, H), alors il y un risque de biais. A l'inverse, si la réponse est « D'égale valeur » (« *Both equally valuable* »), il n'y a pas de risque de biais (Mainland et Herrera, 1956, p. 243). Cela donne le tableau 2x2 suivant, que nous reproduisons et traduisons ici :

Figure 3-3 : Tableau à double entrée sur le risque de biais dans la sélection par les étudiants des sujets de l'étude (Mainland et Herrera) [45]

	Nombre d'étudiants	Pourcentage
Risque avéré de biais	71	55.0
Douteux	28	21.7
Pas de preuve de biais	28	21.7
Sans opinion	2	1.6
	129	100.0

Ainsi, plus de la moitié des étudiants auraient sélectionné des sujets qui auraient pu en dernier ressort biaiser l'échantillon et fausser les résultats. Pour Mainland il n'y a d'ailleurs pas de solution à ce problème, si ce n'est la sélection automatique des sujets de l'échantillon. En effet :

> *« Indeed, the basic fault lies not in the interviewer, but in the nonautomatic selection of subjects.* » (Mainland et Herrera, 1956, p. 244).

Il ajoute dans le résumé de son article :

> « *The only safe method is automatic selection.* » (Mainland et Herrera, 1956, p. 244).

Si Mainland n'emploie pas explicitement le mot « biais » pour désigner le parti-pris des volontaires qui sélectionnent les patients dans l'étude de Hammond et Horn, ou ici des étudiants qui sélectionneraient les sujets inclus dans l'étude selon des critères de pertinence qui auraient pour effet de biaiser les résultats, il est néanmoins possible de penser que la signification familière et psychologique du mot « biais », comme synonyme de parti-pris ou de préjugé, rejoint ici la signification proprement

[45] Mainland et Herrera, 1956, p. 242.

statistique de ce mot conçue comme une absence de représentativité de l'échantillon par rapport à l'univers dont il est issu, ou comme un problème de comparabilité entre les différents groupes de l'échantillon, ce qui, dans ces deux cas, empêche de tirer la moindre conclusion valide à partir de l'étude qui aurait été menée.

Plus précisément il semble que Mainland considère que le sens psychologique du mot « biais » est beaucoup trop restrictif et ne semble finalement n'être qu'une sous-catégorie d'un ensemble de phénomènes qui sont susceptibles de fausser les résultats d'une étude. Ainsi dit-il dans la deuxième partie de son article de 1958 :

> *« This kind of procedure[46], now familiar to many clinical investigators, is apparently often thought to be simply a device for avoiding conscious or unconscious psychological bias. It plays a far more fundamental role, which has nothing to do with sealed envelopes*. » (Mainland, 1958c, p. 825).

Quel est donc ce « rôle beaucoup plus fondamental », qui n'a rien à voir avec les enveloppes scellées, que remplit ce « genre de procédures » ? Et quel est précisément ce « genre de procédures » auquel Mainland fait référence ?

3.2.3 Une nouvelle catégorisation : la cause, le hasard, le biais.

Ce genre de procédures renvoie simplement à la randomisation, notion que Mainland hérite directement de Ronald Fisher. En effet, il lui rend un hommage appuyé à ce propos, dans un article qui oscille entre l'autobiographie et la description, relativement pessimiste, de la situation institutionnelle des statistiques médicales de son époque (l'article est publié en 1954)[47] :

> *« He* [Mainland] *then saw that the methods prescribed by Fisher for avoiding bias and allowing for chance, experimental error, biological variation, and sample size, were applicable in all fields of medicine. (…)*
>
> *The second book [The Design of Experiments] emphasized the fact that a significance test has no useful meaning unless an experiment has been properly designed. In terms of a simple experiment, the comparison of two treatments on animals or patients, "properly designed" means designed in such a way that, at*

[46] Mainland fait référence à la procédure des enveloppes scellées, qui permet dans l'exemple qu'il donne de maintenir une allocation au hasard des traitements

[47] Mainland, Donald, « The rise of experimental statistics and the problems of a medical statistician », *The Yale Journal of Biology and Medicine*, vol. 27 / 1, 1954, p. 1-10.

the end of the experiment, one can say: "Chance would so rarely cause such a large difference in outcome that I shall attribute the observed difference to the treatments." There must be only two possibilities: chance and the treatments; and this situation can be reached only by allocating the treatments to the subjects by what Fisher' called "a physical experimental process of randomization," which is now most easily performed by a table of random numbers. This demonstration of the logical necessity of randomization is one important contribution of Fisher's second book on statistics. Another is his demonstration of how to study multiple factors in the same experiment. (…) He further called attention to the way in which different factors interact with each other, and demonstrated designs that could most economically reveal not only the main effects but the interactions. In brief, the principles are: (i) balance, i.e., the subjection of equal numbers of subjects to each factor under test, and (ii) randomization of all the factors, known and unknown, that are not under test.» (Mainland, 1954, p. 1-3).

Cet extrait est à mettre directement en relation avec un passage extrait d'un article de 1950, qui vise à présenter les principes généraux de l'usage des statistiques dans les recherches cliniques [48]. Il dit ainsi, dans la partie consacrée aux conditions requises pour disposer d'un échantillon adéquat (« *The Requirements for an Adequate Sample »*), dans la première sous-partie où il divise les différents facteurs à étudier en trois catégories (facteurs majeurs connaissables, facteurs majeurs inconnus, et facteurs mineurs) :

« Let us consider a very simple type of investigation. We wish to take two samples of patients with the same disease and apply one treatment to one sample and another treatment to the other sample. We avoid some misleading differences by standardizing our methods of observation and our other techniques, but there are many factors still left, apart from the two treatments, that will produce differences in the outcome (…)

Also, we must be sure that, in our present ignorance, we do not vitiate our results by a preponderance of one type of the disease in one of our samples. Likewise,

[48] Mainland, Donald, « Statistics in clinical research: some general principles », *Annals of the New York Academy of Sciences*, vol. 52, 1950, p. 922-930.

there may be hidden environmental factors, and we must avoid the risk of all such hidden bias. » (Mainland, 1950, p.924).

Mainland va ensuite distinguer quatre exigences requises pour disposer d'un échantillon adéquat :

- d'abord, l'échantillonnage doit être raisonné ou dirigé (« *purposive »*) : il s'agit de réduire la variation en catégorisant les individus ou les groupes d'individus selon différents facteurs ou critères (âge, sexe, etc.) afin d'égaliser les facteurs, de sorte que le facteur étudié soit la seule différence entre les groupes
- ensuite, il faut procéder à une randomisation : les traitements doivent être donnés au hasard, notamment à travers l'utilisation d'une table de nombres au hasard (« *Table of random numbers »*)
- enfin vient l'analyse ou l'évaluation (« *assessment »*) des résultats, qui fait intervenir la notion de « signification statistique » qu'il va développer par la suite. C'est là qu'intervient à nouveau la notion de biais :

  « Having made chance operate in the selection of samples, we can, after the experiment, use our knowledge of chance to assess the results, because we know how often various differences in the results would occur by chance, i.e., if there were no difference between the treatments. (…)

  In contrast, let us consider our verdict if we had not randomized. We could then say: "The difference is due either to chance or to something else. Such differences rarely being due to chance, we believe that it is probably due to something else." But this "something else" may be either the difference between the treatments or some bias due to unknown factors, or perhaps treatment plus bias. Because we did not randomize, we have no way of telling. » (Mainland, 1950, p.925-926).

Ces propositions seront reprises quasiment à l'identique dans l'article de 1958, où il distingue les enquêtes épidémiologiques (au sens d'enquêtes observationnelles), des expérimentations :

« In a survey we have no protection against hidden bias. Therefore, our interpretation of results cannot take the simple "either-or" form. We have to consider three possible causes of the observed differences: (1) chance, (2) the factors under test, and (3) hidden bias, not controlled by chance. A statistical test can tell us only that chance was an unlikely cause, perhaps extremely

unlikely, with probability (P) infinitesimally small. » (Mainland, 1958b, p. 826-827).

Il ajoute que ceci est tout aussi valable pour les expérimentations qui seraient conduites sans randomisation préalable :

« Experiments that are conducted without randomization are equivalent to surveys with respect to uncertainty of inference. When an investigator asks for a "statistical analysis" (i.e., a significance test) after such an experiment, he apparently believes that he has left nothing in his experiment except chance and the factors under test – that he has eliminated biases, or has made them so small that they are trivial in comparison with the difference that he is measuring. He may be perfectly right, but a statistical test does nothing to prove him right. As E. B. Wilson, a Harvard professor of chemistry, has written, "fifty pages of higher mathematics will not salvage an experiment with a hidden bias." Elimination of chance as a likely cause of an observed difference is of no importance if we cannot answer the question: "What has really happened to the biases that are always present in any experiment?" » (Mainland, 1958b, p. 827).

Ainsi, un biais n'est pas simplement un problème d'étiquetage des membres de l'échantillon mais une des trois explications possibles d'une différence observée entre deux groupes : soit la différence est due au hasard, soit elle est due au facteur testé (une exposition ou un traitement), soit elle est due à un biais caché. Le test statistique permet d'éliminer la première hypothèse, et laisse donc penser que la différence observée est bien liée au facteur soumis au test. Le problème est que le test statistique n'a aucune valeur si l'étude ou l'expérimentation est entachée de biais, c'est-à-dire si le plan de l'étude a été mal conçu : il est absolument inutile.

Il apparait ainsi clairement que la conception qu'a Mainland de la notion de biais est beaucoup plus proche de celle de Fisher que de celle de Hill, et explicitement posée comme tel dans son article autobiographique de 1954 qui est cité au début de cette partie. Comme le souligne Harry Marks :

« Dans les années 1950 et 1960, les chercheurs en médecine et en statistique ont de la même façon beaucoup insisté sur la capacité de la randomisation à réguler et à contrôler les biais (…). Seuls ceux qui avaient étudié avec Fisher

ou qui l'avaient lu soigneusement ont discuté du rôle crucial de la randomisation pour garantir des estimations valides de l'erreur expérimentale.[54-56] »[49]

Et, pour Marks, Mainland fait clairement partie des disciples de Fisher, comme l'attestent les notes 54 et 56 de l'article, qui renvoient à deux articles de Mainland, celui de 1950 que nous avons étudié et un autre en date de 1960 sur l'usage et le mésusage des statistiques dans les publications médicales[50]. Dès lors, il semble que le même mot ne soit pas le même concept chez Mainland et chez Hill, qu'il s'agisse d'ailleurs du concept de randomisation ou du concept de biais. En effet, là où chez Hill la randomisation a essentiellement pour fonction de dissimuler la procédure, et vise, comme d'ailleurs la procédure de l'aveugle et du double-aveugle, à produire intentionnellement de l'ignorance (ignorance de qui est malade ou non au moment du diagnostic ou de l'interview, ignorance de qui va dans le groupe traité et de qui va dans le groupe non traité au moment de l'allocation des patients) et par là à éviter les biais, ici explicitement entendus en un sens subjectif et psychologique ; chez Mainland, au contraire, la randomisation ne se résume pas à cette fonction de dissimulation mais permet d'obtenir une estimation valide de l'erreur (d'où la notion faible de biais comme propriété de l'échantillon) mais aussi par là de garantir la validité du test statistique (d'où la notion forte de biais comme propriété de l'étude et d'où la tripartition : cause, hasard, biais). Dès lors il est parfaitement logique que Mainland critique la notion psychologique de biais pour lui préférer la notion statistique de biais héritée de Fisher et conçue comme une erreur systématique. Toute la question est de savoir si et comment il est possible de concilier ces deux notions.

[49] « *In the 1950s and 1960s, medical and statistical researchers alike placed great emphasis on randomization's capacity to regulate and control bias (ref. 17, p. 144–47). Only those who had studied with Fisher or read him carefully discussed randomization's crucial role in ensuring valid estimates of experimental error[54-56]* » in Marks, 2003, p. 935.

[50] Mainland, Donald, « The use and misuse of statistics in medical publications », *Clinical Pharmacology and Therapeutics*, vol. 1, août 1960, p. 411-422. La note 55 fait reference à l'article suivant: Greenberg, B. G., « Why Randomize? », *Biometrics*, vol. 7 / 4, décembre 1951, p. 309.

3.3 Les premières études prospectives sur le lien entre tabagisme et cancer du poumon (1954-1956) : Hill contre Berkson

3.3.1 L'étude de Doll et Hill de 1954.

Dans les deux articles[51] [52] qui présentent les premiers résultats de leur enquête prospective sur le lien entre le tabagisme et le cancer du poumon, Hill et Doll vont restreindre l'usage de la notion de biais à deux significations essentielles, en continuité avec leurs articles précédents : tout d'abord le problème de la sélection de l'échantillon (qui porte sur la proportion mais aussi l'état de santé des médecins qui ont répondu au questionnaire), ensuite le problème de l'information (la question de savoir si les médecins n'ont pas été influencés dans leur diagnostic par la connaissance des antécédents tabagiques des patients). Ces deux articles sont spécialement intéressants pour notre propos car le second article est publié deux ans plus tard et contient une réponse (dans une partie dédié aux « Questions relatives aux biais ») à un article de Berkson, publié en 1955[53], article qui va remettre en cause la validité même non seulement des enquêtes rétrospectives, mais aussi des enquêtes prospectives, et spécifiquement celle de Doll et Hill, bien que les critiques de Berkson soient plus spécialement adressées à l'enquête menée par Hammond et Horn aux Etats-Unis, qui porte quant à elle sur 187 766 hommes et organisée grâce à l'*American Cancer Society*[54], celle-là même qui est critiquée par Mainland dans son article de 1956.

Dès l'introduction à leur article, Doll et Hill dressent le bilan des études rétrospectives sur le tabac et le cancer du poumon et justifient leur recours à une nouvelle méthode, qu'ils qualifient de « prospective », définie ici, dans une note de bas de page, par le fait « qu'elle est tournée vers le futur » (« c*haracterized by looking forward into the future* »[55]). En effet, selon eux, toutes les études qui ont été menées

[51] Doll, Richard et Hill, A. Bradford, « The mortality of doctors in relation to their smoking habits », *British Medical Journal*, vol. 1 / 4877, 1954, p. 1451–1455

[52] Doll, Richard et Hill, A. Bradford, « Lung cancer and other causes of death in relation to smoking », *British Medical Journal*, vol. 2 / 5001, 1956, p. 1071-1081

[53] Berkson, Joseph, « The Statistical Study of Association between Smoking and Lung Cancer », *Proceedings of the Staff Meetings. Mayo Clinic*, vol. 30, 1955, p. 319-348.

[54] Hammond, E. Cuyler et Horn, Daniel, « The relationship between human smoking habits and death rates: a follow-up study of 187,766 men. », *Journal of the American Medical Association*, vol. 155 / 15, août 1954, p. 1316-1328

[55] Doll et Hill, 1954, p. 1451

depuis cinq ans (ils en citent plus d'une dizaine) aboutissent à la même conclusion : il y a une association entre le tabagisme et le cancer du poumon. Néanmoins, tous les auteurs de ces études ne sont pas d'accord sur l'interprétation qu'il faut donner à cette association :

> *« Some have considered that the only reasonable explanation is that smoking is a factor in the production of the disease; others have not been prepared to deduce causation and have left the association unexplained.* » (Doll et Hill, 1954, p. 1451).

Dès lors, pour Doll et Hill, qui font partie de ceux qui concluent à une relation de causalité entre tabagisme et cancer du poumon, il ne sert à rien de faire de nouvelles études rétrospectives car cela ne permettrait pas d'apprendre quelque chose de nouveau, et donc de convaincre les sceptiques. Plus important encore, un nouveau type d'enquête permettrait peut-être de détecter un « défaut inaperçu » dans les études rétrospectives :

> *« If, too, there were any undetected flaw in the evidence that such studies have produced, it would be exposed only by some entirely new approach.* » (Doll et Hill, 1954, p. 1451).

Il s'agit donc d'inverser la perspective : au lieu de partir de la maladie (le cancer du poumon) pour aller vers l'exposition (le tabac), Doll et Hill vont partir de l'exposition pour aller vers la maladie :

> *« It should determine the frequency with which the disease appeared, in the future, among groups of persons whose smoking habits were already known.* » (Doll et Hill, 1954, p. 1451).

L'enquête porte à l'origine sur 59 600 médecins, hommes et femmes, inscrits sur le Registre des médecins du Royaume-Uni, à qui Hill et Doll ont envoyé en octobre 1951 un questionnaire. Sur les 59 600 questionnaires envoyés, Doll et Hill ont reçu 41 024 réponses, dont 40 654 étaient suffisamment complètes pour être utilisées. Ils ont encore retranché de ces 40 564 réponses 10 017 hommes de moins de 35 ans, et 6 158 femmes de tout âge, car le cancer du poumon est « peu commun » chez les femmes et « rare » (Doll et Hill, 1954, p. 1452) chez les hommes de moins de 35 ans. Ainsi, pendant 29 mois (de novembre 1951 à mars 1954 inclus), dès qu'il y avait un mort parmi ces 24 389 hommes de plus de 35 ans, le chef de l'état civil du Royaume-Uni (*the Registrars-General of the U.K.)* envoyait le certificat de décès à Doll et Hill.

Pendant ces 29 mois, il y eut 789 décès parmi cette population, dont 35 ont été certifiés comme ayant pour cause de décès le cancer du poumon.

Après avoir exposé les résultats préliminaires de leur enquête, qui montrent qu'il y a bien une association entre le tabagisme et le cancer du poumon (le signe le plus intéressant, qu'ils qualifient même de « découverte biologique » étant la relation positive entre la quantité de tabac fumée et l'augmentation du nombres de morts par cancer du poumon, symbolisé par un graphique qui montre que le ratio morts observées/ morts attendues augmente en fonction de la quantité de tabac fumée) , Doll et Hill vont mettre en relation les résultats de cette étude avec ceux de leur précédente enquête rétrospective dans la partie intitulée « *Comparison Between the Results of the Retrospective and Prospective Inquiries »*. C'est là qu'intervient un questionnement sur la présence d'un biais éventuel. Le problème, ou plutôt « l'incompatibilité » (« *incompatibility »*[56]) des résultats entre les deux études, vient du faible taux de mortalité des médecins par rapport à la population du Grand Londres (le taux standardisé de mortalité est de 1.97 pour la population du Grand Londres, contre 0.73 pour les médecins). Pour les auteurs, cela renvoie à un problème de sélection, au sens où les médecins les plus malades n'auraient pas répondu à leur questionnaire. En effet :

> *« One important reason– and one which applies to all causes of death and not only to lung cancer– is, we believe, that doctors who were already ill of a disease likely to prove fatal within a short space of time would have been disinclined, or indeed unable, to answer our inquiries.* » (Doll et Hill, 1954, p. 1454).

Pour Doll et Hill, cela ne pose néanmoins pas un véritable problème car ce biais va progressivement disparaitre avec le temps, du fait que les plus malades sont déjà décédés, ce qui est selon eux déjà le cas :

> *« If persons sick of a fatal illness were unwilling to reply, or, indeed, never saw our communication, that bias would tend to wear off with the passage of time – as it shows signs of doing.* » (Doll et Hill, 1954, p. 1454).

Une question plus épineuse est celle de savoir si ce biais de sélection va affecter différemment la mortalité du groupe des fumeurs par rapport à celui des non-fumeurs, c'est-à-dire que les gros fumeurs qui savent déjà qu'ils sont atteints d'un

[56] Doll et Hill, 1954, p. 1454.

cancer auraient plus répondu que les autres, c'est-à-dire les non-fumeurs et les petits fumeurs :

> *« The question is whether such a bias would differentially affect the mortality of the smoking group. Could it artificially produce the gradient that we have observed with cancer of the lung, and probably with coronary thrombosis, whilst not, producing any gradient with other causes of death? For such an effect we should have to suppose that the heavier smokers who already knew that they had cancer of the lung tended to reply more often than non-smokers, or lighter smokers, in a similar situation.* » (Doll et Hill, 1954, p. 1454).
>
> Cette hypothèse leur semble hautement improbable pour deux raisons :
>
> *«That would not seem probable to us. As evidence to the contrary we would also add (a) that, although the numbers of deaths are admittedly very small, we have not seen any obvious change in the lung cancer gradient over the 29 months of the inquiry, and (b) that it would be surprising if a gradient produced in this way so closely resembled the gradient we obtained in our retrospective inquiry.* » (Doll et Hill, 1954, p. 1454).

Le second biais possible, enfin, est abordé dans la courte partie consacrée aux diagnostics (« *The Diagnoses »*), et renvoie à ce qu'on appellerait aujourd'hui un biais d'information, plus spécifiquement un biais de détection :

> « *It might perhaps be argued that physicians in reaching a diagnosis of cancer of the lung have been biased by the patient's smoking history.* » (Doll et Hill, 1954, p. 1454).

Ainsi, les médecins auraient plus promptement diagnostiqué un cancer du poumon car ils connaissaient les antécédents tabagiques du patient. Or, pour Doll et Hill cet argument ne tient pas car, si tel était le cas, « les décès dus aux autres causes devraient être proportionnellement moins élevés dans le groupe des gros fumeurs »[57], ce qui n'est pas le cas. Un tel biais est donc très peu probable (« *very unlikely »*). Dès lors :

> « *The association between smoking and the disease is real and not due to some such bias as we have discussed* »(Doll et Hill, 1954, p. 1454).

[57] « *Deaths from other causes would have to be proportionately less in the groups of heavier smokers*. », in Doll et Hill, 1954, p. 1454

3.3.2 La critique logique de Berkson : le biais comme sophisme.

La critique de Berkson parait un an plus tard, en juillet 1955, et va porter simultanément sur l'article de Hammond et Horn et sur celui de Doll et Hill. Berkson commence son article par une référence à l'étude de Pearl, publiée en 1929, qui établissait que la tuberculose constituait un facteur protecteur contre le cancer du poumon :

> *« When I encountered the first of the series of statistical studies on the association between smoking and lung cancer which have recently appeared, it immediately recalled to me a prior investigation on the association between tuberculosis and cancer* » (Berkson, 1955, p. 319).

Sans rentrer dans les détails de cette étude cas-témoins, qui montrait que la tuberculose était trouvée dans seulement 6.6% des cas des 816 personnes autopsiées qui avaient une tumeur maligne, alors qu'elle était trouvée chez 16.3% des 816 personnes autopsiées qui n'avaient pas de tumeur maligne, et concluait donc que la tuberculose protégeait contre le cancer du poumon, il est important de noter que pour Berkson, cette étude, menée qui plus est par un statisticien reconnu, était parfaite d'une point de vue méthodologique, quand bien même la conclusion était fausse (et due au fait que les personnes qui avaient un cancer n'avaient pas le temps de développer la tuberculose, car ils mouraient avant). Il dit ainsi :

> *« Although retrospectively I agree that the conclusion reached by Pearl in his first investigation is not correct (…), still, on the basis of very generally accepted principles of statistical procedure it seems to me that he was invulnerably right. If in two "cohorts" of a population, differentiated in respect of only one relevant characteristic x, the finding of a unquestionable difference between the two, in the relative frequencies of a character y, establishes association between x and y, irrespective of the character of the population itself, then Pearl's investigation did establish negative association between cancer and tuberculosis and in fact it was an impeccable example of such a demonstration* » (Berkson, 1955, p. 322).

Le problème selon Berkson est que les études sur le tabac et le cancer du poumon reposent sur la même méthodologie, et leurs conclusions sur la « validité de ce principe général » (Berkson, 1955, p. 322). Dès lors, il semble possible de penser que ces études sont victimes de la même erreur ou du même sophisme dans leurs

conclusions, du fait que l'échantillon a été sélectionné, comme cela était le cas pour l'enquête de Pearl. En effet, pour Berkson, l'étude de Pearl était, du point de vue de sa construction logique (« *the logical development on which it was planned* ») convaincante (« *cogent* ») mais fausse, du fait que la population étudiée était une population de morts (« *a population of the dead* »[58]). Berkson conclut ainsi :

> « *I resolved never to study association of diseases in a dead population* » (Berkson, 1955, p. 324)

C'est suite à cette expérience que Berkson se rend compte, après qu'on lui a demandé de travailler comme consultant sur une étude qui voulait montrer si l'ulcère duodénal protégeait du cancer de l'estomac, que le même genre de « conclusion fallacieuse » (« *fallacious character of the conclusion* »[59]) pouvait être tirée si l'on prenait pour base la population, morte ou vivante, d'un hôpital. Mais Berkson va aller encore plus loin, sur la suggestion, d'ailleurs, de Mainland[60] : ce genre de sophisme peut non seulement s'appliquer aux études rétrospectives, comme la littérature de l'époque l'a, d'après lui, bien compris, mais aussi aux études prospectives, et donc à celles de Doll et Hill, ou de Hammond et Horn. Il dit ainsi :

> *« A simple mechanism may be operating which will produce spurious association in the selection population similar to that referred to in the study of association of diseases in a hospital population* » (Berkson, 1955, p. 325).
>
> Et en effet, dans l'étude de Doll et Hill:
>
> *« The population actually composing the material of comparison is only a certain portion of the physicians registered at the time of inception of the study and these are only a portion of the general population*. » (Berkson, 1955, p. 325).

Il y a donc bien selon Berkson un risque de sélection, entendue ici comme un problème de représentativité de l'échantillon, équivalent à celui qu'on retrouve dans une étude rétrospective. Pour lui c'est ce phénomène de sélection qui explique les résultats aussi bien des études rétrospectives que prospectives, phénomène qui est lié à la procédure statistique elle-même. Berkson, comme dans son article de 1946, utilise un exemple fictif[61] pour illustrer son argument et distingue deux groupes pour la comparaison : le premier groupe est constitué de personnes qui sont sérieusement

[58] Berkson, 1955, p. 324.
[59] Berkson, 1955, p. 324.
[60] *« It was Pr Donald Mainland who suggested me rather pointedly what I had thought of only vaguely», in* Berkson, 1955, p. 325.
[61] Berkson, 1955, p. 326.

malades, et représente 3% de la population de référence (une cohorte de 100 000 personnes), avec un taux de mortalité de 99% , et dont 50% est recruté pour l'enquête. Le second groupe est constitué de personnes qui ne sont pas malades, avec un taux de mortalité de 0.03%, et dans ce groupe, 99% de la population de non-fumeurs répond à l'enquête et est inclus dans l'investigation tandis que 65% des fumeurs répondent et sont inclus dans l'enquête. Pour résumer, le critère d'inclusion dans l'enquête pour le groupe 1 dépend de l'état de santé, tandis que le critère d'inclusion pour le groupe 2 dépend du fait de fumer ou non. Quant à la population de référence, on considère que 80% de cette population fume, et que le taux de mortalité annuelle est de 3%. Berkson montre que, alors qu'il n'y a pas d'association entre le tabagisme et les taux de mortalité dans la population de référence, une association positive notable apparait dans la population sélectionnée, puisque le taux de mortalité pour les fumeurs est de 2.6% alors qu'il n'est que de 1.6% pour les non-fumeurs. Berkson précise que cette différence n'est pas liée au fait que des personnes s'éliminent de l'enquête parce qu'ils sont malades avec une plus grande probabilité s'ils sont fumeurs que s'ils ne le sont pas, mais qu'il s'agit ici de « l'opération simultanée selon des intensités différentes de la sélection à la fois sur le tabagisme et sur les morts » (Berkson, 1955, p. 328). S'il ne s'agit selon lui que d'un « modèle statistique » (Berkson, 1955, p. 328) qui ne prétend pas déterminer exactement comment le mécanisme de sélection opère, il reste que cette possibilité de ce qui s'apparente à un biais (même s'il refuse d'employer ce terme, et lui préfère celui de *« fallacy »*) permet de mettre en doute la méthode prospective, dont on pense qu'elle n'est pas sujette, comme l'étude rétrospective, à ce genre de sophisme[62]. Cela renvoie en fait au problème plus général, étudié d'après Berkson, par Neyman et Fix, des « risques compétitifs » (« *competitive risks »*). C'est d'ailleurs ce qui explique selon lui pourquoi toutes les études portant sur le rôle du tabac dans le cancer du poumon produisent les mêmes résultats, cette convergence des résultats étant selon lui non la preuve d'une association, mais le « signe d'une corrélation statistique fallacieuse » (« *the hallmark of spurious statistical correlation »*[63]). Le défaut se situe en effet dans la procédure statistique elle-même :

[62] *« It is only a "statistical model", and has been presented because wide acceptance of the prospective studies as probative appears to be based on the idea that with this method of investigation no fallacy is possible.* », in Berkson, 1955, p. 328.
[63] Berkson, 1955, p. 332.

> *« If correlation is produced by some elements of the statistical procedure itself, it is almost inevitable that the correlation will appear whenever the statistical procedure is used* » (Berkson, 1955, p. 332).

Il dira même un peu plus loin dans son article, dans la partie intitulée « Considérations biologiques » que le problème principal est précisément que la corrélation entre tabagisme et cancer du poumon est « si exclusivement statistique » (« *so exclusively statistical »*[64]). Berkson va même plus loin car il considère que les études qu'on appelle prospectives sont en réalité elles aussi des études rétrospectives :

> *« In a crucially important sense, however, both of these types of study [rétrospective et prospective] are retrospective. If in the prospective type of study – exemplified in those of Doll and Hill and of Hammond and Horn – the designation of the cases with and without cancer is "prospective", the designation of each individual as to whether he is a smoker or a non-smoker is not prospective, for this is already accomplished at the initiation of the study. It is just this which opens the way to "selection*". » (Berkson, 1955, p. 341).

On retrouve ici le problème que Berkson avait soulevé dans son article de 1946 : les deux échantillons comparés sont différents avants que commence l'étude, contrairement à ce qu'il se passe dans une expérimentation. Il dit ainsi :

> *«The type of study which is genuinely prospective is the experimental study, for here one begins with neither variable predetermined, but instead with the entire group of individuals undifferentiated in respect of either of the two variables, the association of which is under investigation. Separation of individuals is then made in respect of one variable – the putatively "causal" variable – at the will of the experimenter, and according to well-defined statistical principles of randomization.* » (Berkson, 1955, p. 341)[65]

Le problème est donc que dans une étude observationnelle, rétrospective ou prospective, il est impossible, pour des raisons pratiques, d'effectuer une randomisation véritable, contrairement par exemple aux essais cliniques, dont Berkson loue d'ailleurs la méthodologie élaborée par Hill. Or, sans randomisation, il est impossible d'effectuer « une inférence statistique valide » (« *a valid statistical*

[64] Berkson, 1955, p. 339.
[65] C'est Berkson qui souligne.

inference »[66]). Dès lors le seul moyen d'établir une relation causale entre le tabagisme et le cancer du poumon est de faire une expérimentation, car « dans la science, il n'y pas de substitut aux expérimentations » (Berkson, 1955, p. 342). C'est d'ailleurs le titre de la dernière partie de son article : « *Wanted - An adequate program of experimental verification. » (*Berkson, 1955, p. 341). Selon lui :

> « si une association biologique importante doit être établie comme une conclusion scientifique définitive, c'est-à-dire si elle doit être considérée comme « prouvée », la population ne doit pas être autre chose qu'une population expérimentale » (Berkson, 1955, p. 323).

Ainsi seule l'expérimentation peut permettre de prouver une association démontrée sur une base purement statistique. La critique de Berkson peut donc être qualifiée de « logique » ou de « logiciste » au sens où elle ne porte pas sur les données ou sur les résultats de l'étude, mais sur l'étude elle-même, c'est-à-dire sur son plan ou sa structure logique : étant donné qu'il n'y pas eu et qu'il n'est pas possible d'avoir, ni dans les études cas-témoins ni dans les études de cohorte, de randomisation, alors il n'est pas possible de tirer la moindre conclusion ou de faire une inférence qui soit valide, c'est-à-dire qui soit *formellement* vraie, et a fortiori *matériellement* vraie.

3.3.3 La réponse empirique de Doll et Hill : le biais comme explication possible et l'inférence à la meilleure explication.

Dans le deuxième article consacré à leur étude prospective sur le lien entre tabagisme et cancer du poumon, publié en 1956[67], Doll et Hill vont consacrer spécifiquement la quatrième partie de leur article à la question des biais (« *Questions of Bias »*[68]), qu'ils vont diviser en deux sous-parties : « Le diagnostic des causes de décès » (« *Diagnosis of Cause of Deaths »*), et « La population à risque » (« *The Population at Risk »*). La première sous-partie n'est pas concernée par la critique de Berkson car elle porte sur la possibilité que les médecins qui ont signé les certificats de décès aient été influencés dans leurs diagnostics par la connaissance des antécédents tabagiques du patient :

[66] Berkson, 1955, p. 342.
[67] Doll, Richard et Hill, A. Bradford, « Lung cancer and other causes of death in relation to smoking », *British Medical Journal*, vol. 2 / 5001, 1956, p. 1071-1081
[68] Doll et Hill, 1956, p. 1076. Nous avons dénombré 14 occurrences du mot « biais » dans cet article.

> « *It might perhaps be argued that doctors have more readily diagnosed lung cancer in heavy smokers than in light smokers or in non-smokers, and have thus produced the gradient of mortality recorded here*. » (Doll et Hill, 1956, p. 1076).

Le premier test pour évaluer si les médecins avaient réellement été influencés a consisté pour Doll et Hill à écrire directement aux médecins pour leur demander : sur les 47 médecins impliqués, 40 avaient une connaissance préalable des antécédents des patients, et sur ces 40, 36 ont affirmé que cette connaissance n'avait pas influencé leur jugement. Ce test reste néanmoins peu convaincant, seul un médecin ayant avoué que cela l'a influencé, et un deuxième que cela aurait pu l'influencer inconsciemment (« *it might have done so subconsciously »*[69]*)*. Doll et Hill proposent donc un second test :

> *« A second, and perhaps more convincing, test of this possible bias can be made by comparing the mortality gradient with smoking for those cases in which the diagnosis was firmly established (category I in Table IV) with that greater element of doubt (categories II and III in Table IV).* » (Doll et Hill, 1956, p. 1076).

Doll et Hill vont alors résumer ces informations dans le tableau X de l'article (Doll et Hill, 1956, p. 1076) où il apparait que pour les cas fermement établis, c'est-à-dire par des preuves histologiques, et donc de façon objective, c'est-à-dire indépendamment du jugement du médecin qui aurait pu être influencé, la tendance est aussi forte, voire plus forte, que pour les cas plus douteux. Doll et Hill peuvent alors conclure :

> *«In view of these results it seems to us most improbable that the relationship we have observed between smoking and lung cancer can be attributed merely to a biased attitude among the medical profession.* » *(*Doll et Hill, 1956, p. 1076).

Vient maintenant la question beaucoup plus critique d'un éventuel biais de sélection tel qu'il est soulevé par Berkson dans sa critique. Doll et Hill avaient, comme nous l'avons vu, anticipé ce problème dans leur article de 1954, mais ils vont accorder à la discussion de ce problème beaucoup plus de place que dans le premier article, afin de répondre précisément au problème soulevé par Berkson. Doll et Hill commencent par reprendre, dans les mêmes termes ou presque, le problème qu'ils avaient soulevé dans le premier article, à savoir le faible taux de mortalité des

[69] Doll et Hill, 1956, p. 1076.

médecins par rapport à la population générale. Ils vont ajouter les résultats de l'étude de Hammond et Horn, qui exhibent la même différence entre les taux de mortalité, pour toutes les causes de mortalité, des sujets de leur enquête par rapport à la population générale. Ils vont alors résumer, de façon limpide, l'argument de Berskon :

> *« These results led Berkson (1955) to suggest that not only is the total population in these studies biased, by the absence of the seriously ill at the time of initial inquiry into smoking habits, but that the component smoking and non-smoking groups may be differentially biased to the advantage of the latter in the subsequent mortality experience. He points out that this would be the effect if non-smokers in good health came more readily into the study than smokers in good health – for example, because answering the questionary is a simpler task for the non-smoker – whereas the chances of inclusion in the study were low for men seriously ill and unrelated to the smoking habits. In such circumstances the already seriously ill component would be artificially low, but still representative of the parent population; the component in good health would be large but unrepresentative. It would contain proportionately too many healthy non-smokers. It follows that the total mortality would be lower than that anticipated from general population rates, and the mortality among non-smokers would be less than that amongst smokers – and for all causes of death.* » (Doll et Hill, 1956, p. 1076). [70]

Il y a donc deux problèmes liés mais distincts : un problème de représentativité de l'échantillon par rapport à la population générale, du fait que ceux qui sont sérieusement malades seraient artificiellement moins nombreux, quoique représentatifs, tandis que ceux qui sont en bonne santé seraient nombreux mais non-représentatifs. Mais il y a aussi un problème de comparabilité des deux groupes au sens où le groupe des non-fumeurs serait avantagé en termes de mortalité, au sens où leur mortalité serait moins élevée, du fait qu'il y aurait plus de personnes en bonne santé dans le groupe des non-fumeurs, ce qui ferait qu'au final la mortalité des non-fumeurs serait moins élevée que celle des fumeurs, non parce qu'ils ne fument pas mais parce qu'il y a plus de personnes en bonne santé dans ce groupe.

Pour Doll et Hill, le test final de la thèse de Berkson réside dans le passage du temps :

[70] Ce sont Doll et Hill qui soulignent.

« The final test of Berkson's thesis lies with the passage of time. For as time passes it becomes progressively less likely that the shadow of death could have been foreseen at the start of the inquiry, less likely that such pre-knowledge could have influenced response to our questionary » (Doll et Hill, 1956, p. 1076).

Berkson avait d'ailleurs souligné ce fait, qui disait :

« A check will be available when prospective studies have been continued long enough. Experience of insurance companies indicates that selection effected by initial medical examination, and by self-selection is « worn off » in about 3 to 5 years » (Berkson, 1955, p. 347).

Doll et Hill vont alors proposer deux arguments :

- Tout d'abord, celui du temps : les taux de mortalité continuent d'augmenter durant la troisième et la quatrième année, par rapport aux deux premières années.
- Ensuite, celui du gradient : le gradient de mortalité en fonction de la quantité de tabac fumée est « remarquablement constant » au cours des quatre années de l'étude.

Si donc la population de départ était biaisée, c'est-à-dire sélectionnée, ce qui expliquerait le faible taux de mortalité de l'échantillon par rapport à la population générale, et que l'effet de ce biais diminuerait en fonction du temps passé, le gradient devrait logiquement baisser. Or,

« *it certainly has not become any less pronounced with more representative death rates.* » (Doll et Hill, 1956, p. 1077).

Dès lors,

«the observations do not seem to us to support Berkson's thesis. » (Doll et Hill, 1956, p. 1077).

Doll et Hill vont alors essayer de vérifier si le taux de mortalité de leur échantillon de docteurs est plus faible que celui des docteurs en général, afin de vérifier si leur échantillon est ou non représentatif. La technique utilisée par les auteurs a consisté à constituer un échantillon de 10% (tiré au hasard) des médecins qui n'avaient pas répondu à leur enquête, même si cela n'a pu être fait que plusieurs mois après le début de l'enquête, ce qui les a empêchés de « reconstruire la population totale » mais aussi de « mesurer la mortalité durant la première année de ceux qui n'avaient pas répondu » (Doll et Hill, 1956, p. 1077). Néanmoins, il est possible de faire une

comparaison pour les années suivantes et d'obtenir une estimation du taux de mortalité de la population totale des médecins en agrégeant les données de ceux et de ceux qui n'ont pas répondu, puis de comparer, en faisant un ratio entre le taux de mortalité de la population totale et celui de l'échantillon : ce ratio est de 72% pour la deuxième année (14.7‰ pour ceux qui ont répondu, contre 20.4‰ pour la population totale), 87% pour la troisième année, et 92% pour la quatrième. Doll et Hill concluent :

> *« We see, therefore, that though the effect of self-selection initially present may still not have entirely worn off, it is certainly no longer large.* » (Doll et Hill, 1956, p. 1077).

Dès lors, il semble que leur échantillon de médecins soit représentatif de la population générale des médecins. En effet comparer le taux de mortalité de la population totale des médecins avec celui des médecins de leur échantillon est « le critère correct de comparaison pour révéler à quel point notre groupe est représentatif du total » (« *the proper standard of comparison, to reveal how far our group is representative of the total »*[71]*)*

Mais Doll et Hill vont montrer que même si l'échantillon est en meilleure santé et a un taux de mortalité moindre que celui de la population de médecins, et serait en ce sens pas ou pas assez représentatif, différence qui pourrait bien « ne jamais complètement disparaitre » [72], cela ne change rien car ce qui est important est le contraste entre les deux groupes comparés :

> *« In other words, we should <u>always</u> have a population which – in total – has a relatively favourable mortality experience. But it does not follow that its components (smokers and non-smokers) cannot be validly contrasted. That very marked contrast is not, as we have shown above, diminishing with the passage of years* » (Doll et Hill, 1956, p. 1077)*.* [73]

Ainsi, Doll et Hill ont démontré que la suspicion de biais, due à leur méthode d'investigation, qui produirait artificiellement une relation entre le tabagisme et le cancer du poumon, telle qu'elle était soulevée par Berkson, ne tenait pas face à l'examen détaillé des différentes données de leur investigation. Ainsi disent-ils, dans la partie consacrée au résumé et aux conclusions :

---

[71] Doll et Hill, 1956, p. 1077*.*
[72] *« the difference may never wholly vanish »*[72], In Doll et Hill, 1956, p. 1077.
[73] Ce sont Doll et Hill qui soulignent.

> *«The relationship cannot therefore be attributed to a biased attitude in the medical profession in certifying cancer of the lung as the cause of death. 8. On these grounds we do not believe that the gradient of mortality with smoking can be regarded as merely an artifact due to bias in those who chose to reply to the questionary*. » (Doll et Hill, 1956, p. 1080)*.*

La différence entre le traitement de la question par Berkson et son traitement par Hill et Doll est assez saisissante : là où Berkson se fonde sur un argument logique (absence de randomisation) et des données mathématiques fictives pour justifier sa critique des études prospectives, Doll et Hill font quant à eux preuve de pragmatisme et se fondent uniquement sur les données dont il dispose, et qui sont réelles, voire vont chercher d'autres données (en allant par exemple chercher des données sur les médecins qui n'ont pas participé à leur enquête), afin de justifier de la validité de leurs conclusions et de montrer que leurs résultats ne sont pas artificiels. Le concept de biais joue ici le rôle d'explication alternative possible aux résultats, alternative dont Doll et Hill montrent qu'elles ne sont pas fondées. En ce sens, l'explication du tabagisme comme cause du cancer apparait comme la meilleure explication, les autres explications possibles (les biais) ayant toutes été tour à tour éliminées. Bien évidemment, il est douteux que ces arguments aient réellement convaincu Berkson, car ils ne répondent pas véritablement à sa critique purement logique et ne peuvent pas y répondre puisque la critique de Berkson remet en cause la validité même des études et leur légitimité à démontrer ce qu'elles prétendent démontrer. Les deux positions semblent ainsi irréconciliables. C'est la thèse que soutient Mark Parascandola[74] en distinguant deux approches différentes de la nature de la preuve apportée par les études épidémiologiques qui se font jour à la faveur de la polémique sur le rôle du tabac dans le cancer du poumon :

> « *The differing opinions on the evidence reflected two different models of etiological research – controlled experiment as the crucial, objective test of a causal hypothesis versus inferential judgment based on a diverse body of evidence.* » (Parascandola, 2004*,* p. 81).

[74] Parascandola, Mark, « Two approaches to etiology: the debate over smoking and lung cancer in the 1950s », *Endeavour*, vol. 28 / 2, juin 2004, p. 81-86. Voir aussi, sur le même sujet, Parascandola, Mark, « Epidemiology in Transition : Tobacco and Lung Cancer in the 1950s », in Jorland, Gérard, Opinel, Annick et Weisz , George (éds.) *Body counts: medical quantification in historical and sociological perspective / La quantification médicale, perspectives historiques et sociologiques*, Montréal ; Ithaca, McGill-Queen's University Press, 2005, p. 226-248.

Parascandola considère d'ailleurs Berkson, mais aussi Fisher, comme un partisan du test objectif de l'hypothèse causale via l'expérimentation contrôlée, tandis que Hill serait plutôt partisan du jugement inférentiel fondé sur un ensemble de preuves diverses. Mais cette polémique entre Berkson d'un côté et Hill et Doll de l'autre n'est qu'une des multiples controverses qui fleurissent à cette époque dans un contexte plus large de discussion sur la méthodologie des études épidémiologiques et sur la nature de la preuve apportée par les études épidémiologiques. Ce sont ces controverses, au cœur desquelles la notion de biais joue un rôle important, qu'il s'agit à présent d'étudier.

PARTIE 2 : DU CONCEPT EPIDEMIOLOGIQUE AU CONCEPT MEDICAL DE BIAIS, ET RETOUR

CHAPITRE 4 : PROBLEMES EPISTEMOLOGIQUES DE LA NOTION EPIDEMIOLOGIQUE DE BIAIS (1956-1965)

La controverse sur le rôle du tabagisme dans le cancer du poumon, qui apparait dans les années 1950 pour se prolonger jusqu'au milieu des années 1960 constitue plus largement, comme le montre Parascandola, « un test crucial pour la discipline de l'épidémiologie » (Parascandola, 2005, p. 226) et aboutit à un questionnement épistémologique sur la scientificité même de l'épidémiologie, de ses méthodes, et du genre de connaissances qu'elle est en mesure de fournir. Le concept de biais va progressivement remplir plusieurs fonctions relatives à ces diverses interrogations épistémologiques sur l'épidémiologie. Ces interrogations peuvent être selon nous réduites à trois problèmes centraux, entremêlés dans la pratique mais qu'il nous faut distinguer :

- Le premier problème concerne directement la théorie épidémiologique elle-même : en effet, si l'épidémiologie des maladies chroniques commence à produire des résultats au cours des années 1950, c'est-à-dire commence à identifier un certain nombre de facteurs de risque de maladie, notamment dans le cadre de la controverse autour du rôle causal du tabagisme dans le cancer du poumon, son statut scientifique se voit contesté par des statisticiens ou même des épidémiologistes qui soulignent l'absence de théorie cohérente de cette discipline et qui remettent en question la possibilité même d'une quelconque preuve épidémiologique. Le concept de biais fournit alors un angle d'attaque pour démontrer que les études épidémiologiques sont incapables de démontrer ce qu'elles prétendent démontrer, et que la notion de preuve en épidémiologie est éminemment problématique. En effet, la difficulté ou la critique principale à laquelle se sont heurtées notamment les études sur le lien entre tabagisme et cancer du poumon est de ne pas avoir apporté de preuve de la causalité, ou d'avoir apporté des preuves qui étaient seulement statistiques au sens où elles ne permettaient pas d'exhiber un mécanisme biologique qui expliquerait comment le tabac ou ses constituants causaient le cancer, au sens aussi où elles s'appuyaient uniquement sur l'observation et non sur l'expérimentation. Pour répondre à ces objections, les épidémiologistes vont alors procéder à

ce que nous pouvons appeler un « renversement de la charge de la preuve » en forgeant un nouveau concept, celui de risque relatif : en montrant que la mortalité des fumeurs par cancer du poumon est neuf fois supérieure à celle des non-fumeurs, les partisans d'un lien causal entre tabagisme et cancer du poumon objectent aux partisans de l'hypothèse constitutionnelle ou génétique qu'un facteur génétique ou constitutionnel ne saurait expliquer une telle différence. Progressivement c'est une véritable théorie épidémiologique qui va alors se mettre en place au tournant des années 1950-1960, théorie au sein de laquelle le concept de biais va progressivement prendre sa place comme un concept important.

- Le deuxième problème est relatif à la question de la causalité : en effet, toute étude épidémiologique de nature analytique a pour fonction idéale d'établir une relation de causalité entre un facteur de risque et une maladie. Cette discussion sur la notion de causalité nous conduira à étudier les réponses d'un certain nombre d'épidémiologistes, comme Hill ou Cornfield, aux critiques faites notamment par des statisticiens de renom comme Berkson ou Fisher sur la méthodologie et sur les résultats des études épidémiologiques observationnelles. La question centrale est en somme celle de la nature de l'inférence qui est permise par les études épidémiologiques. Or, la question de l'inférence, en particulier de l'inférence causale va nécessiter une redéfinition même de la notion de causalité, bientôt conçu sous la forme d'un réseau ou d'une toile, afin de pouvoir saisir la complexité des relations causales qui existent dans le cas des maladies chroniques. Ceci conduira les épidémiologistes à distinguer des critères qui permettent d'affirmer un lien de causalité, en particulier Bradford Hill dont les neuf « critères » de causalité passeront à la postérité. Surtout, cette discussion sur la causalité, sa nature et la manière de la prouver, conduira les épidémiologistes à affirmer l*a priori*té de l'action sur la connaissance, donc de la pratique sur la théorie, ce qui constitue aussi une manière de montrer qu'une présomption de causalité suffit souvent à déclencher une action de santé publique, la décision devant être ultimement posée non en termes strictement scientifiques, mais en fonction d'impératifs politiques, économiques et sanitaires.

- Le troisième et dernier problème concerne la question du plan d'expérience et est étroitement lié au à celui de la causalité. Cette question du plan d'expérience dépasse le simple cadre de l'épidémiologie et se pose de la même manière à la psychologie sociale ou à la recherche sur l'éducation. Les questions qui se posent au sein de ces disciplines sont néanmoins différentes de celles de Fisher : il s'agit en effet, en épidémiologie comme dans les sciences sociales[1], non d'expériences randomisées où l'expérimentateur peut contrôler les différentes variables et qui portent sur des plantes ou des légumes, mais d'études observationnelles qui portent sur des êtres humains où il est très difficile, voire impossible, autant de randomiser que de contrôler la plupart des variables à l'étude. La question centrale porte notamment, comme nous l'avons vu, sur la question de la sélection de l'échantillon, et donc sur sa représentativité, mais aussi sur la question de la possibilité de généraliser les résultats de l'échantillon à la population générale : le concept de biais étant étroitement lié à celui de sélection et d'échantillon, il est logique qu'il occupe un rôle central dans ce questionnement méthodologique. Le concept de biais va ainsi être redéfini, notamment chez Stanley et Campbell, donc en dehors du cadre de l'épidémiologie, comme une menace à la validité interne du plan d'expérience, et même comme une menace à sa validité tout court. Cette question de la validité interne mérite une analyse plus détaillée qui nous permettra de faire retour au concept de biais dans l'épidémiologie, sur le terrain de la logique de l'étude épidémiologique comme procédure de recueil et d'analyses des données. Nous retrouverons alors la question de la preuve, mais dans une perspective différente, qu'on pourrait qualifier de formelle, en tant que cette interrogation sur les plans d'expérience, et la hiérarchisation qui en découle entre les différents plans, va permettre de distinguer différents niveaux de preuve : autrement dit, plus un plan est exempt de biais, plus la preuve qu'il apporte est solide.

[1] Le terme « sciences sociales » est évidemment anachronique pour l'époque : par là nous désignons la psychologie, la sociologie, la psychologie sociale mais aussi les sciences de l'éducation (terme lui aussi anachronique) car c'est dans ces disciplines que la question du biais mais aussi celle du plan d'expérience (et celle du biais en relation avec le plan d'expérience) apparait problématique.

4.1 L'épidémiologie en quête de théorie :

4.1.1 Epidémiologie de la « boîte noire » et sous-détermination de la théorie épidémiologique

A la fin du chapitre précédent, nous nous sommes arrêtés sur une opposition entre deux approches irréconciliables de la causalité et de la preuve de la causalité, entre d'un côté des statisticiens comme Berkson et Fisher, partisans d'une expérience cruciale, et de l'autre Hill et Doll considérés comme les défenseurs d'une approche plus empirique ou plus pragmatique de l'établissement d'une relation causale fondée sur un ensemble de preuves concordantes et sur la nécessité du jugement de l'épidémiologiste pour asserter ce lien de causalité. Le problème principal auquel les épidémiologistes sont en effet confrontés dans le cadre de cette polémique autour d'un lien de causalité entre le tabagisme et le cancer du poumon est qu'ils établissent une relation causale sans en exhiber le mécanisme biologique ou la pathogénèse: c'est ce que Mervyn et son fils Ezra Susser[2] ont appelé le « paradigme de la boîte noire » caractéristique de l'ère de l'épidémiologie des maladies chroniques, une boîte noire étant une boîte où l'on connait l'entrée (l'input), la sortie (l'output), mais où l'on ne sait pas ce qui se passe dans la boîte ou, comme la définissent les Susser « la métaphore commune pour une unité autonome dont les processus internes sont cachés de celui qui la regarde »[3]. En termes épistémologiques, et dans le sillage de l'analyse de W.V.O. Quine[4], nous avons donc à faire à un cas de sous-détermination des théories par l'expérience ou par les faits : en d'autres termes, un même ensemble de données observables peut s'interpréter de plusieurs manières ou peut être expliqué par plusieurs théories scientifiques concurrentes voire contradictoires entre elles.

Nous avons déjà étudié une grande partie de ce débat dans le chapitre précédent à travers le prisme de la controverse entre Berkson et Hill en nous arrêtant

2 Susser, Mervyn et Susser, Ezra, « Choosing a future for epidemiology: I. Eras and paradigms. », *American Journal of Public Health,* vol. 86 / 5, 1996, p. 668–673.

3 « *the general metaphor for a self-contained unit whose inner processes are hidden from the viewer*. » in Susser et Susser, 1996, p. 670.

4 Voir à ce sujet: Quine, W. V., « On the Reasons for Indeterminacy of Translation », *The Journal of Philosophy*, vol. 67 / 6, mars 1970, p. 178.

à l'année 1956. Pourtant cette controverse va s'étendre au moins jusqu'en 1965[5], quand l'article célèbre d'Austin Bradford Hill[6] sur ce que les épidémiologistes ont appelé par la suite les critères de causalité de Bradford Hill (bien que Hill n'utilise jamais ce terme) va fournir un outil décisif pour établir une relation de causalité. Néanmoins, il ne s'agit pas pour nous d'étudier cette controverse en tant que telle, et ce pour deux raisons : la première est qu'elle a été déjà beaucoup étudiée, et l'on peut d'ailleurs dire sans se tromper que c'est l'épisode le plus étudié de l'histoire de l'épidémiologie ; la seconde est que cela nous conduirait hors du champ de notre sujet. C'est pourquoi nous allons l'étudier sous l'angle spécifique de notre étude : nous soutenons que c'est à l'occasion de cette controverse que le concept de biais va acquérir son rôle spécifique au sein de l'épidémiologie moderne ou, autrement dit, que c'est à cette occasion que son usage va se fixer, non sans doute de façon définitive mais de façon suffisamment ferme pour que cela détermine la suite de son évolution. En effet, un des aspects intéressants de cette polémique réside dans sa structure à deux étages: ce que nous voulons dire par là est que la critique de la relation causale entre le tabac et le cancer du poumon porte d'abord (d'abord étant entendu ici au sens logique et non chronologique) sur la réalité de l'association statistique entre tabac et cancer du poumon, c'est-à-dire sur la question de savoir si l'association statistique est réelle ou bien artificielle et fallacieuse (« *spurious* ») ; et ensuite sur la question de savoir quelle interprétation donner à cette association statistique, à supposer qu'elle soit réelle, et donc la question de savoir si elle renvoie à une véritable relation causale donc pas simplement formelle ou logique (statistique) mais matérielle ou ontologique. La fonction et l'importance du concept de biais varient bien évidemment en fonction de la question qui est à l'étude et apparaissent comme bien plus importantes dans la controverse statistique que dans la controverse proprement causale.

Dans le chapitre précédent, nous avons arrêté notre analyse en 1956 avec l'article de Doll et Hill qui constitue, entre autres, une réponse à la critique de Berskon qu'il a émise dans son article de 1955. Bien évidemment la polémique est loin d'être terminée et elle ne saurait se réduire à l'opposition entre Doll et Hill d'une côté et

5 La controverse est en fait encore vive en 1978 comme l'atteste l'article suivant et la discussion qui le suit: Burch, Paul R.J., « Smoking and Lung Cancer: The Problem of Inferring Cause », *Journal of the Royal Statistical Society. Series A (General*), vol. 141 / 4, 1978, p. 437-477.

6 Hill, Austin Bradford, « The environment and disease: association or causation? », *Proceedings of the Royal Society of Medicine*, vol. 58 / 5, 1965, p. 295-300.

Berkson de l'autre. En fait, elle mobilise une bonne partie des épidémiologistes de l'époque, ainsi que certains des statisticiens les plus connus. Néanmoins, nous pouvons partir de Berkson, qui propose trois arguments contre le lien causal entre tabac et cancer du poumon, arguments qui résument l'essentiel des griefs dirigés contre l'association tabac-cancer du poumon, avant de nous intéresser aux arguments donnés par d'autres épidémiologistes ou statisticiens. Ces trois arguments sont exposés dans son article de 1958[7], et répétés dans son article de 1959[8], et qui est le texte d'une conférence faite par Berkson le 10 septembre 1958 durant le congrès de la Fédération internationale pharmaceutique réuni à Bruxelles :

1) Le premier argument consiste à dire que l'association statistique observée est fausse ou fallacieuse (« *spurious* »[9]) : en d'autres termes, cette association n'aurait pas de « signification biologique » (Berkson,1958, p. 37), et ne serait en fait qu'un « phénomène statistique artificiel» (Berkson,1959, p. 215), dû « à l'interaction de divers « biais » subtils et compliqués »[10] . Selon Berkson, ces « biais » (il retirera finalement les guillemets au mot biais dans son article de 1959) sont de deux sortes :
   - il s'agit tout d'abord du problème « des caprices des données » (« *the vagaries of the data* »[11]) concernant en premier lieu les antécédents du tabagisme et en second lieu les causes de décès, données qui sont toutes deux sujettes à des « erreurs » ou à des « variations » considérables et arbitraires.
   - Le second biais renvoie à la question des échantillons qui n'ont pas été « obtenus selon des méthodes scientifiques d'échantillonnage » (Berkson, 1958, p. 37) et qui sont donc « sélectionnés », c'est-à-dire non représentatifs de la population dont ils sont issus. En d'autres termes, il n'y a pas eu de randomisation.

   Ainsi, selon Berkson, dans la mesure où la méthodologie des enquêtes, prospectives comme rétrospectives, est problématique puisqu'elles s'appuient

[7] Berkson, Joseph, « Smoking and Lung Cancer: Some Observations on Two Recent Reports », Journal of the American Statistical Association, vol. 53 / 281, mars 1958, p. 28-38.
[8] Berkson, Joseph, « The statistical investigation of smoking and cancer of the lung. », *Proceedings of the Staff Meetings. Mayo Clinic*, vol. 34 / 8, avril 1959, p. 206-224.
[9] Berkson, 1958, p. 37 et Berkson,1959, p. 215
[10]« *the result of the interplay of various subtle and complicated "biases."* », in Berkson, 1958, p. 37
[11] Berkson,1959, p. 215

sur des données qui ne sont pas fiables (« *unreliable data*»[12]), ces enquêtes sont susceptibles de produire des résultats erronés (« *erroneous results* »[13]).

2) Le deuxième argument renvoie à l'hypothèse de la constitution, et il est défendu aussi par Fisher (dans une version plus génétique) ou Jacob Yerushalmy[14]. La substance de cet argument est que le fait de fumer et le fait de développer un cancer auraient une cause commune, une sorte de prédisposition qui ferait que ces personnes auraient plus tendance à développer un cancer du poumon et à fumer. En d'autres termes, comme le souligne Anne Fagot-Largeault, « la fumée du tabac, plutôt que la cause du cancer (le facteur qui induit l'augmentation du risque), [peut] être un simple marqueur (un témoin de l'augmentation du risque) »[15]
3) Le troisième argument renvoie à la notion que Berkson emprunte à Raymond Pearl et qui est celle du « *rate of living* » et qu'il explique ainsi : « les fumeurs, à un âge donné sont biologiquement plus âgés que leur âge chronologique » (Berkson, 1958, p. 37). Cet argument réapparait dans l'article de 1959 mais disparait par la suite.

Dans un article de 1962[16] en apparence consacré au lien entre mariage et mortalité mais qui constitue en réalité un ensemble de « réflexions sur la déduction de l'étiologie à partir des statistiques », Berkson va faire des trois arguments un seul et même argument en donnant finalement un nom à ce sophisme qu'il avait soulevé dès son article de 1942. Selon lui en effet :

> « En essayant de trouver, à partir de corrélations statistiques, une signification étiologique à des facteurs environnementaux dans la causalité d'une maladie, il est facile de tomber dans la pratique qui consiste à sélectionner certaines maladies particulières associées à certains facteurs qui semblent plausibles, tout en ignorant des associations similaires de ces mêmes maladies avec d'autres facteurs, ou de ces mêmes facteurs avec d'autres maladies. Je

[12] Berkson,1959, p. 216.
[13] Berkson,1959, p. 216.
[14] Yerushalmy, Jacob et Palmer, Carroll E., « On the methodology of investigations of etiologic factors in chronic diseases », *Journal of Chronic Diseases*, vol. 10 / 1, 1959, p. 27–40. L'hypothèse constitutionnelle est avancée à la page 36.
[15] Fagot-Largeault A., « Épidémiologie et causalité », in Valleron, A.-J. (Ed.) *L'Épidémiologie humaine: conditions de son développement en France, et rôle des mathématiques*, Académie des sciences, Rapport sur la science et la technologie n° 23, Paris: EDP sciences, 2006, chap. 7, 237-245.
[16] Berkson, Joseph, « Mortality and marital status. Reflections on the derivation of etiology from statistics », *American Journal of Public Health and the Nations Health*, vol. 52 / 8, 1962, p. 1318–1329.

considère cette attitude comme un exemple de « sophisme de la concrétude mal placée »[17]

Ce « sophisme de la concrétude mal placée » est explicitement emprunté à Alfred North Whitehead qui le définit comme le fait de considérer une abstraction (une croyance, une opinion, une idée) comme un objet concret : autrement dit, il s'agit d'une réification et l'on pourrait plus justement parler de « sophisme de la réification ». Dans le contexte de la controverse tabagisme-cancer du poumon, la thèse de Berkson consiste ainsi à soutenir que la relation statistique qui apparait entre le tabagisme et le cancer du poumon est uniquement statistique (« *so exclusively statistical* »[18],comme il le soulignait en 1955) et donc en quelque sorte purement nominale : en ce sens elle pourrait n'être que le résultat d'un ensemble de biais à la fois de sélection (de l'échantillon) et d'information (antécédents de tabagisme et diagnostic des causes de décès), comme il l'explique dans son article de 1958 (Berkson,1958, p. 37). Ceci va conduire Berkson à adopter une position radicale qui consiste à rejeter purement et simplement l'idée même d'une preuve statistique : selon lui en effet, il n'est pas possible de savoir si les études épidémiologiques sont valides ou non, c'est-à-dire si elles sont entachées de biais ou non. En ce sens, la relation de causalité qui semble établie par la corrélation statistique est purement artificielle ou comme il le dit, abstraite : rien ne permet de savoir si la relation est réelle ou non, s'il y a donc une causalité entre tabac et cancer du poumon, et conclure à la réalité de l'association sur le seul fondement de l'association statistique reviendrait à commettre un raisonnement fallacieux, c'est-à-dire un sophisme, celui de la réification. Selon lui, la seule réalité, et donc la seule explication scientifique possible, est biologique, et il n'y a que la biologie et les biologistes qui puissent trancher la question de la causalité du tabagisme dans le cancer du poumon, les statistiques n'ayant qu'un rôle subalterne, :

> « Le cancer est un problème biologique, et non statistique. Les statistiques peuvent parfaitement jouer un rôle auxiliaire dans son élucidation. Mais si les

[17] « *In seeking to find, from statistical correlations, etiologic significance of environmental factors in causation of disease, it is easy to fall into the practice of selecting particular diseases associated with particular factors that appear plausible, while ignoring similar associations of the same diseases with other factors. and of the same factors with other diseases. I have referred to this as an example of the "fallacy of misplaced concreteness."* », in Berkson,1962, p. 1327-1328.

[18] Berkson,1955, p. 339.

biologistes autorisent les statisticiens à devenir les arbitres des questions biologiques, le désastre scientifique est inévitable »[19]

De même, la seule méthode à même d'établir la preuve scientifique, c'est-à-dire biologique, d'un lien entre tabac et cancer du poumon est la méthode de l'expérimentation et non de l'observation. Aussi appelle-t-il à la mise en place d'une « épidémiologie expérimentale des maladies non-infectieuses »(Berkson,1958, p. 37)[20], qui pourrait être réalisée sur des animaux. Selon lui, en effet, seule la preuve expérimentale est à même de clore le débat.

Ce scepticisme radical de Berkson vis-à-vis des études épidémiologiques observationnelles n'est pas selon nous lié à des facteurs extrascientifiques[21], mais au contraire parfaitement justifié d'un point de vue épistémologique. En réalité, à l'époque, il y a une impossibilité de savoir si l'étude est valide ou non qui tient au fait que l'on ne dispose pas encore de critères pour évaluer la méthodologie d'une étude. C'est d'ailleurs ce que soulignent Yerushalmy et Palmer dans leur article qui porte sur la « méthodologie des investigations des facteurs étiologiques dans les maladies chroniques » :

---

[19] « *Cancer is a biologic, not a statistical, problem. Statistics can soundly play an ancillary role in its elucidation. But if biologists permit statisticians to become the arbiters of biologic questions, scientific disaster is inevitable*. », in Berkson,1958, p. 32, note 4.

[20] Selon lui cette épidémiologie expérimentale permettrait de prouver non que l'inhalation de la fumée du tabac cause spécifiquement le cancer du poumon chez les animaux, mais qu'elle augmente le « *rate of living* ».

[21] Berkson, bien que non-fumeur, a en effet souvent été accusé, comme Fisher, de travailler pour l'industrie du tabac et d'être un relais de son argumentaire (Voir à ce sujet Berlivet, Luc, « "Association or Causation? «The Debate on the Scientific Status of Risk Factor Epidemiology, 1947–c. 1965 », *Clio Medica/The Wellcome Series in the History of Medicine*, vol. 75 / 1, 2005, p. 39–74, notamment la page 53). Pourtant à la lecture de certains documents déclassifiés provenant des représentants du lobby du tabac, le tableau apparait bien plus nuancé. Ce qui ressort plutôt est que Berkson semble convaincu que ses arguments sont pertinents mais restent largement incompris de la part des épidémiologistes, qu'il semble par ailleurs mépriser. A ce sujet, il est intéressant de lire ces deux lettres, l'une de 1957 (https://www.industrydocumentslibrary.ucsf.edu/docs/#id=tkny0014) où un représentant de *The American Tobacco Company* fait un rapport à sa direction sur Berkson, et le décrit comme « solitaire », « aigri », « consumé par la haine pour Dorn, Cornfield» ou encore « incontrôlable », et conseille de ne pas l'embaucher comme conseiller scientifique (« *Scientific advisor* ») au sein de la société ; l'autre de 1962 (https://www.industrydocumentslibrary.ucsf.edu/docs/#id=njhy0059) où un membre du *Tobacco Industry Research Comittee* (le faux-nez « scientifique » de l'industrie du tabac) présente Berkson comme un homme d'une « intégrité intellectuelle extraordinaire » qui est réellement convaincu qu'il n'y a pas de causalité entre tabagisme et cancer du poumon. Nous considérons donc que ce n'est pas parce qu'il travaille pour l'industrie du tabac qu'il nie la causalité entre tabagisme et cancer du poumon, mais parce qu'il nie cette causalité qu'il s'est rapproché de l'industrie du tabac, tout en critiquant d'ailleurs le principe même du *Tobacco Industry Research Committee*, qui selon lui crée un problème là où il n'y en a pas et donne une cible facile aux anti-tabac, tout en jetant le doute sur l'objectivité des personnes qui sont liées à ce comité. Nous pensons avec Berkson que son argument global a en effet été mal compris car il a été réduit au problème de la représentativité d'un échantillon pris à l'hôpital alors que l'argument porte non sur un cas particulier de sélection mais sur la procédure elle-même de l'étude cas-témoin et de l'étude prospective.

> « La faiblesse majeure des observations faites sur les êtres humains provient du fait qu'elles ne possèdent pas la caractéristique de la comparabilité des groupes, une condition pourtant essentielle qui est accomplie, par un effort conscient, dans une expérimentation à travers la randomisation. Dès lors, il existe toujours une possibilité que les associations qui sont observées soient, à un degré plus ou moins grand, dues à d'autres facteurs que ceux qui sont étudiés. Ainsi, que l'investigation soit fondée sur un seul ou plusieurs types d'observations, il est nécessaire d'évaluer, pour chacun de ces types, sa justesse et sa validité.
>
> Malheureusement, *la méthodologie et les critères d'évaluation n'ont pas encore été adéquatement développés*. Il n'est dès lors peut-être pas surprenant que des données issues d'observations non-contrôlées soient acceptées comme concluantes par certains investigateurs et rejetées par d'autres investigateurs tout aussi compétents »[22]

Ainsi, l'outil statistique qui est utilisé par les épidémiologistes ne peut pas, en l'absence de randomisation, être testé ou évalué : les résultats qu'il produit peuvent être interprétés différemment, et même de façon contradictoire, et donc ne sont pas fiables, dans la mesure où la méthode qui a conduit à ces résultats n'est pas elle-même fiable, puisqu'elle n'a pas été développée ni évaluée. Comme le dit Luc Berlivet[23], s'appuyant sur les travaux de Harry Collins[24], il y a une sorte de « régression

[22] « *The major weakness of observations on humans stems from the fact that they often do not possess the characteristic of group comparability, a basic requirement which in experimentation is accomplished by conscious effort through randomization. The possibility always exists, therefore, that such associations as are observed may, to a greater or lesser degree, be due to factors other than those under study. Thus, whether the investigation is based on a single type or a multiplicity of types of observation, it is necessary to evaluate each for its soundness and validity. Unfortunately, the methodology and criteria for evaluation have not yet been adequately developed. It is, therefore, perhaps not surprising that some data derived from uncontrolled observations are accepted as conclusive by some investigators and rejected by other equally competent observers.* », in Yerushalmy et Palmer, 1959, p. 28. Nous soulignons

[23] Berlivet, Luc, « "Association or Causation? «The Debate on the Scientific Status of Risk Factor Epidemiology, 1947–c. 1965 », *Clio Medica/The Wellcome Series in the History of Medicine*, vol. 75 / 1, 2005, p. 39–74.

[24] Collins, H. M., *Changing order: replication and induction in scientific practice*, London ; Beverly Hills, Sage Publications, 1985. Cette régression de l'expérimentateur est une régression logique à l'infini qu'il expose par exemple à la page 85 de son livre à propos du détecteur d'ondes gravitationnelles : « Ce qu'est le bon résultat dépend s'il y a ou non des ondes gravitationnelles qui frappent la Terre à travers des flux détectables. Pour déterminer si c'est le cas nous devons construire un bon détecteur d'ondes gravitationnelles et regarder. Mais nous ne saurons pas si nous avons construit un bon détecteur tant que nous ne l'aurons pas essayé et obtenu le résultat correct ! Mais nous ne savons pas quel est le résultat correct jusqu'à… et ainsi de suite à l'infini. L'existence de ce cercle, que j'ai appelé « la régression de l'expérimentation », constitue l'argument central de ce livre. » En somme, nous savons qu'un résultat expérimental est correct quand il est obtenu à travers un bon appareil expérimental, et

de l'observation » au sens où l'innovation scientifique, ici l'innovation méthodologique que constituent les études cas-témoins et les études de cohorte (grâce à Bradford Hill), engendre un redoutable problème qui est :

> « l'absence de critères formel indépendants des résultats de l'expérience elle-même pour permettre une évaluation de la qualité des méthodes, précisément parce qu'il n'y a pas d'antécédents de résultats solides qui auraient été établis à travers ces méthodes » (Berlivet, 2005, p. 52-53)

4.1.2 La question de la preuve épidémiologique :

Le problème semble donc insoluble sans prendre des positions épistémologiques radicales : soit, comme Berkson rejeter complètement l'outil épidémiologique (les études cas-témoins ou les études de cohorte) comme incapable de produire des connaissances qui soient vraies et justifiées, notamment du fait de la présence de biais, soit démontrer la légitimité de la preuve statistique ou de la preuve épidémiologique, soit en tant que telle, soit au côté d'autres types de preuves.

Cette question de la preuve statistique fondée sur la seule observation, et de son opposition supposée à la véritable preuve qui serait la preuve expérimentale, est en réalité au cœur des débats depuis le début des années 1950 et la publication de la première étude rétrospective de Doll et Hill sur le lien entre tabagisme et cancer du poumon. Deux textes méritent à cet égard d'être étudiés : le premier est l'œuvre de Bradford Hill et est intitulé « Observation et expérimentation », en date de 1953[25] ; le second, bien qu'il soit attribué à Ernest Rubin, est en fait rédigé par Jérome Cornfield, et date de la fin de l'année 1954[26]. Si le thème est identique dans les deux textes[27], la

nous savons que notre appareil expérimental est bon quand il donne un résultat expérimental correct : il y a là un cercle vicieux ou une régression à l'infini. Selon Collins, qui appartient au courant de la sociologie des sciences, cela empêche la réplication des expérimentations, et le critère qui permet de juger si une expérimentation est bonne ou mauvaise doit être extérieure à l'expérimentation elle-même, et doit être construit socialement notamment par la communauté scientifique.

[25] Hill, A. Bradford, « Observation and Experiment », *New England Journal of Medicine*, vol. 248 / 24, juin 1953, p. 995-1001. Ce texte est en fait une conférence faite à la *Harvard School of Public Health* le 25 mars 1953, dans le cadre de la *Cutter Lecture on Preventive Medicine*.

[26] Rubin, Ernest, « Questions and Answers », *The American Statistician*, vol. 8 / 5, décembre 1954, p. 19-21.

[27] Cornfield, dans une note, souligne qu'après avoir rédigé sa réponse, son « attention a été attirée » sur la conférence de Hill « dans laquelle la plupart des questions que nous avons traités ont été examinées – d'une façon néanmoins beaucoup plus complète, lucide (et raisonnable) », in Rubin, 1954, note 4, p. 21.

question qui est posée par chacun est différente. Dans le premier, la problématique de Hill est la suivante :

> « En d'autres termes, est- ce que les observations peuvent être faites de sorte qu'elles puissent satisfaire autant que possible les exigences expérimentales ? »[28]

Dans le texte de Cornfield, la problématique est différente en tant qu'elle est d'emblée orientée vers la question de la causalité et la nature de la preuve :

> « Y a-t-il des exemples dans l'histoire de la médecine où une association statistique a prouvé la causalité ou bien est-ce que la preuve d'une relation causale en médecine a toujours dépendu de l'expérimentation directe ? [29]»

Autrement dit, dans le premier texte, Hill va justifier la méthode observationnelle comme capable de produire des connaissances et de fournir des preuves d'une relation causale, tandis que Cornfield va plutôt chercher à justifier la notion de preuve statistique, et par là il entend la preuve que les études épidémiologiques peuvent tirer de l'observation. Selon Hill, tout d'abord, l'étude expérimentale fonctionne comme un étalon ou un standard duquel il faut se rapprocher, et duquel Doll et Hill ont cherché à se rapprocher dans l'étude cas-témoins sur le tabagisme et le cancer du poumon, en contrôlant autant que possible ce qu'ils ont appelé dans leur article le biais de sélection et le biais de l'intervieweur :

> « Notre but était de faire que les observations de terrain reflètent autant que possible le plan expérimental. Pour chaque cas de patient souffrant d'un cancer du poumon, nous avons cherché un patient « contrôle » qui souffrait d'une autre maladie – un patient du même sexe, de la même tranche d'âge, dans le même hôpital ou au même moment, ou sinon choisi au hasard. En d'autres termes, nous avons cherché, comme dans une expérimentation, à limiter les variables. Nous les avons aussi limitées, pas seulement de cette façon, mais aussi en employant, dans l'interrogation sur les antécédents, seulement un petit nombre

[28] « *In other words, can observations be made in such a way as to fulfil, as far as possible, experimental requirements*? », in Hill, 1953, p. 995.

[29] « *Are there instances in the history of medicine in which a statistical association proved causation or has the proof of a causal relationship in medicine always depended on direct experimentation*? », in Rubin,1954, p. 19.

> d'interviewers qualifiés, chacun armé d'un ensemble déterminé de questions »[30]

Selon Hill d'ailleurs, il n'y a pas de différence de nature entre une étude observationnelle et une étude expérimentale, mais seulement une différence de degré. En effet, l'approche observationnelle comporte une « faiblesse » dont est exempte l'approche expérimentale :

> « Avec la première, nous pouvons déterminer l'explication la plus probable de notre contraste dans les données ; sous réserve que nous ayons pris suffisamment soin d'éliminer les causes perturbatrices, cette probabilité peut être très haute. Mais avec une expérimentation bien construite il doit être possible d'éliminer (ou de tenir compte de) presque toutes les causes perturbatrices et ainsi de rendre l'interprétation du contraste encore plus certaine. »[31]

Ainsi, selon Hill, si l'expérimentation permet d'apporter une certitude plus grande que l'observation, cette certitude n'est pas pour autant absolue. C'est pourquoi une expérimentation sur l'animal, par exemple en exposant des souris à la fumée de tabac, si elle peut « renforcer la preuve» (« *evidence* »), devra être confirmée par une étude sur l'homme pour avoir « la preuve des preuves finale » (« *final proof of proofs* »[32]), ce qui est malheureusement impossible pour des raisons éthiques. Observation et expérimentation sont donc étroitement liées et doivent se renforcer mutuellement en matière de santé publique ou de médecine préventive[33]. L'observateur doit simplement être « plus patient que l'expérimentateur » mais aussi plus « imaginatif », en « sentant les corrélations qui gisent sous la surface de ses observations », plus « logique et moins dogmatique, en évitant comme la peste, le sophisme *post hoc ergo propter hoc*, qui consiste à confondre corrélation et

---

[30] « *Our aim was to make the field observations mirror an experimental design as nearly as possible. For each patient with cancer of the lung we sought a "control" patient with some other disease - a patient of the same sex, of the same age group, in the same hospital at or about the same time, but otherwise chosen at random. In other words, we sought, as in an experiment, to limit the variables. We limited them, too, not only in this way but also by employing, in history taking, only a few skilled interviewers, each armed with a prescribed set of questions*. », in Hill, 1953, p. 998.

[31] «*With the former we can determine the most probable explanation of a contrast in our data; given the provision that we have taken sufficient care to remove disturbing causes, that probability can be very high. But with a well defined experiment it should be possible to eliminate (or allow for) nearly all disturbing causes and thus to render the interpretation of the contrast even more certain.* », in Hill, 1953, p. 998.

[32] Hill, 1953, p. 998.

[33] « *In the world of public health and preventive medicine each will _ or should _ constantly react beneficially upon the other.* », in Hill, 1953, p. 1000.

causalité »(Hill, 1953, p. 1000). En somme, comme le souligne S. Greenland, dans la présentation qu'il fait de ce texte, les biais d'information ou de sélection propres aux études observationnelles « font partie des problèmes que le chercheur a pour tâche de résoudre » et ne constituent pas des « défauts insurmontables inhérents à toute recherche non-expérimentale »[34], comme le pense Berkson. Nous retrouvons donc ici la thèse de Parascandola sur l'importance accordée non à une quelconque expérience cruciale qui permettrait de trancher entre des hypothèses mais au jugement de l'épidémiologiste qui doit faire appel à son imagination théorique mais aussi à la déduction logique afin de déterminer, à partir d'une ensemble d'indices et de preuves, et en examinant avec soin la présence et l'effet d'éventuels biais, si une association statistique entre deux variables reflète ou non une connexion causale entre des entités du monde physique.

Cornfield, dans son article, aborde la question sous un angle similaire : selon lui, en effet, la différence entre observation et expérimentation repose sur le fait que dans le second cas, « nous sentons que les effets d'autres variables importantes sont contrôlés et ne peuvent pas expliquer l'association » (Rubin,1954, p. 19). Pourtant, Cornfield montre que ce contrôle est tout à fait possible dans le cas de l'observation en se fondant sur l'exemple de John Snow et de sa découverte que le choléra était transmis par l'eau. De même, il soutient ensuite qu'on ne peut jamais être sûr, à l'inverse, que toutes les « variables exogènes [« *extraneous variables*»] aient été réellement contrôlées par l'expérimentation directe » (Rubin,1954, p. 19), y compris à travers la randomisation. Ces deux arguments lui permettent d'affirmer sa thèse :« Il n'y a pas de différence de nature entre les deux types de preuves [observationnelles et expérimentales] », mais il y a « une grande différence de degré » (Rubin,1954, p. 20), parce qu'il est beaucoup « plus difficile de contrôler les variables » dans les études observationnelles que dans les études expérimentales. En d'autres termes :

> « Il n'existe pas de catégories telles que des preuves de première classe et des preuves de seconde classe »[35]

Il ajoute :

[34] Greenland, Sander, *Evolution of Epidemiologic Ideas: Annotated Readings on Concepts and Methods*, Chestnut Hill, Mass., Epidemiology Resources, 1987; p. 2.

[35] «*There are no such categories as first-class evidence and second-class evidence*», in Rubin,1954, p. 20.

> « Distinguer entre, d'un côté, des associations statistiques et, de l'autre, des relations qui sont établies par l'expérimentation, sans aucune référence à des variables alternatives qui seraient présentes dans un cas mais pas dans l'autre, ne nous semble pas être de la bonne statistique, ni de la bonne science, ni de la bonne philosophie – bien que cela puisse constituer une bonne diversion »[36]

Dès lors, il semble que ce que Hill comme Cornfield entendent montrer est que l'étude observationnelle comme l'étude expérimentale peuvent être affectées de différents biais, même si l'étude observationnelle y est plus sujette, notamment l'étude cas-témoin. En ce sens, une étude expérimentale, à la condition qu'elle soit correctement planifiée et exécutée, apporte une ou des preuves plus solides qu'une étude observationnelle, mais cela ne signifie pas qu'une étude observationnelle n'apporte aucune preuve. C'est pourquoi l'une des stratégies adoptées par Cornfield et nombre de ses collègues épidémiologistes va consister à collationner les différentes preuves du lien entre tabagisme et cancer du poumon, notamment dans un article resté célèbre dans l'histoire de l'épidémiologie, publié en 1959[37], et qui réunit « les leaders de l'épidémiologie américaine des années 1950 »[38] : Jérôme Cornfield bien sûr, mais aussi William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin et Ernst L. Wynder.

L'article de Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin et Wynder, qui est présenté comme un « rapport » (« *report* ») et dont la bibliographie compte pas moins de 85 références, poursuit essentiellement deux objectifs :

> « passer en revue les découvertes épidémiologiques et expérimentales parmi les plus récentes quant à la relation entre le tabagisme et le cancer du poumon, et discuter certaines des critiques dirigées contre la conclusion que fumer du tabac, particulièrement des cigarettes, joue un rôle causal dans l'augmentation du cancer broncho-pulmonaire » (Cornfield, Haenszel, Hammond *et al.,* 1959, p. 173).

---

[36] «*To distinguish between statistical association on the one hand and relationships that are established by experimentation on the other, without any reference to alternative variables that are present in one case but not the other, seems to us to be neither good statistics, good science, nor good philosophy- though it may be good red herring*. » in Rubin,1954, p. 20.

[37] Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, Abraham, Shimkin, Michael B. et Wynder, Ernst L., « Smoking and lung cancer: recent evidence and a discussion of some questions », *Journal of the National Cancer Institute*, vol. 22 / 1, janvier 1959, p. 173-203.

[38] Comme le souligne Alfredo Morabia sur son site internet en réponse à une question sur la première occurrence du terme « risque relatif » : http://www.epidemiology.ch/history/questions2.htm

Deux arguments sont mobilisés à cet effet, tous deux visant à affaiblir l'hypothèse d'un troisième facteur, c'est-à-dire par exemple l'hypothèse constitutionnelle avancée par Berkson (son article de 1955 et celui de 1958 sont cités et vivement critiqués) ou bien par Fisher[39]:

- Le premier argument stipule que « la magnitude de l'excès de risque de cancer du poumon parmi les fumeurs de cigarette est tellement grande que les résultats ne peuvent pas être interprétés comme provenant d'une association indirecte avec un autre agent ou une autre caractéristique ». En effet, « cet agent hypothétique devrait être au moins aussi fortement associé à l'usage de cigarette ». Or, « aucun agent de cette sorte n'a été trouvé ni suggéré »[40].
- Le deuxième argument est celui de la convergence (« *consistency* ») entre « toutes les preuves épidémiologiques et expérimentales ».

D'un point de vue historique, il est intéressant de noter que l'opposition entre deux conceptions de la notion de biais qui parcourt les années 1950 (opposition qui recouvre à notre sens celle posée par Parascandola entre les partisans de l'expérience cruciale, où la notion forte de biais, comme chez Mainland, invalide par principe l'étude; et les épidémiologistes qui font appel au jugement et à l'ensemble des preuves disponibles, où la notion faible de biais n'invalide pas par principe l'étude, mais nécessite un véritable travail de jugement de l'épidémiologiste pour évaluer la présence et la magnitude des différents biais) va en réalité se transformer à la fin des années 1950, notamment sous l'effet de l'article de Mantel et Haenszel, publié en 1959[41], qui constitue, avec l'article de Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin et Wynder, un progrès méthodologique considérable en épidémiologie. En effet, deux notions décisives pour l'épidémiologie moderne apparaissent précisément

[39] Fisher, apparemment choqué par la « propagande terrifiante » contre le tabac commence à critiquer les résultats des études sur le tabac et le cancer du poumon à partir de 1957 où il envoie ses premières « Lettres à l'éditeur » du *British Medical Journal* (une en juillet 1957 et l'autre en août 1957), puis de la *Centennial Review* (1958) et à l'éditeur de de la prestigieuse revue *Nature* (la première en juillet 1958 et la seconde en août 1958). Il compilera ces textes dans un pamphlet publié en 1959 : Fisher, Ronald A., *Smoking: the cancer controversy: some attempts to assess the evidence*, Oliver and Boyd Edinburgh, 1959.

[40] « *The magnitude of the excess lung-cancer risk among cigarette smokers is so great that the results can not be interpreted as arising from an indirect association of cigarette smoking with some other agent or characteristic, since this hypothetical agent would have to be at least as strongly associated with lung cancer as cigarette use; no such agent has been found or suggested*. », in Cornfield, Haenszel, Hammond, *et al.,* 1959, p. 173.

[41] Mantel, Nathan et Haenszel, William, « Statistical aspects of the analysis of data from retrospective studies of disease », *Journal of the National Cancer Institute*, vol. 22 / 4, 1959, p. 719–748.

en 1959 : il s'agit de la notion de « risque relatif », définie et exprimée mathématiquement par Jerome Cornfield, mais aussi celle d' « odds ratio » (qu'ils expriment sous la forme *R*= AD/BC, dans le cadre d'un tableau de contingence à double entrée[42]), qui est formulée mathématiquement par Mantel et Haenszel, même si le mot n'apparait pas sous leur plume. Ceci constitue sans nul doute un événement important dans l'histoire de l'épidémiologie, puisque celle-ci dispose d'un instrument de mesure qui lui est propre et qui va lui permettre non seulement de quantifier plus précisément la force de l'association mais aussi par là même, dans le cadre de la controverse sur le lien entre tabagisme et cancer du poumon, de réfuter l'hypothèse d'un tiers facteur, chère à Berkson et à Fisher qu'ils expriment sous la forme de l'hypothèse constitutionnelle ou génétique. Tout ceci n'est pas sans conséquences sur la notion de biais qui, comme celle de risque, va être relativisée.

4.1.3 Relativisation du risque, relativisation du biais

Cornfield et ses collègues, en plus de la formulation mathématique du risque relatif, vont aussi justifier son usage en donnant trois arguments qui montrent que l'usage de mesures relatives est plus utile que l'usage de mesures absolues

Tout d'abord, ils entendent entendent réfuter l'argument de Berkson qui soutient selon eux qu'une « mesure relative est inappropriée dans une enquête sur le tabagisme et la mortalité » (Cornfield, Haenszel, Hammond *et al,* 1959, p. 193)[43]. Au contraire, selon eux :

> « quand un agent a un effet apparent sur plusieurs maladies, le classement des maladies par la magnitude de l'effet va dépendre de l'utilisation d'une mesure absolue ou d'une mesure relative »[44]

[42] Mantel et Haenszel, 1959, p. 731.

[43] Cornfield, Haenszel, Hammond, *et al.,* 1959, p. 193. La citation exacte de Berkson est la suivante: « *Doll and Hill measured the differences in death rates among the different classes of smokers relatively, sometimes in relation to the death rate among all men, sometimes in relation to the death rate among nonsmokers. It seems almost instinctive to express differences relatively in some such manner, but in a case like the present application, it may be misleading. There are, of course, situations in which such relative measures are appropriate, but it does not appear to me that this is such a situation*. », in Berkson, 1958, p. 29-30. L'appendice A à l'article de Cornfield, Haenszel, Hammond, *et al.,* est explicitement une réponse et une « justification rationnelle » au qualificatif d'instinctif utilisé ici par Berkson. Voir Cornfield, Haenszel, Hammond, *et al.,* 1959, Appendix A, p. 198, où le mot « *instinctive* » est cité entre guillemets mais sans référence précise.

[44] « *When an agent has an apparent effect on several diseases, the ranking of the diseases by the magnitude of the effect will depend on whether an absolute or a relative measure is used.* », in Cornfield, Haenszel, Hammond, *et al.,* 1959, p. 193.

Ils se fondent alors sur l'étude de Dorn sur les vétérans américains pour démontrer l'intérêt des mesures relatives : ainsi, il y a eu 187 morts par cancer du poumon, contre 20 morts attendues (en se fondant sur les taux de mortalité chez les non fumeurs), soit 167 morts 'en excès. Pour les maladies cardio-vasculaires, il y a eu 1 780 morts contre 1 165 attendues, soit un excès de 615 morts. En termes absolus, il apparaitrait alors que le tabac tue bien plus par maladies cardio-vasculaires que par cancer du poumon. Or, si l'on prend maintenant une mesure relative, c'est-à-dire si l'on divise le nombre de morts par cancer du poumon par le nombre de morts attendues par cancer du poumon, on obtient 187/20=9,35. Si l'on effectue la même opération pour les maladies cardio-vasculaires, on obtient : 1780/1165=1,53. Ainsi :

> « Relativement, les cigarettes ont un effet beaucoup plus grand sur le cancer du poumon que sur les maladies cardio-vasculaires, alors que l'inverse est vrai si on utilise une mesure absolue ».

Cornfield et ses collègues n'invalident pas pour autant les mesures absolues : chacune des mesures a en effet son utilité. Les mesures absolues permettraient par exemple d'évaluer « l'importance en terme de santé publique d'un effet connu pour être causal »[45]. Mais ce sont surtout les mesures relatives dont l'utilisation va être précisément justifiée par les auteurs, ce qui montre à quel point ces mesures sont importantes pour les épidémiologistes, et importantes dans l'histoire de l'épidémiologie. La première justification est que les mesures relatives permettent « l'évaluation de la possible nature non causale d'un agent qui a un effet apparent »[46]:

> « Si un agent, A, qui n'a aucun effet causal sur le risque d'une maladie, mais qui néanmoins, en raison d'une corrélation positive avec un autre agent causal indéterminé, montre un risque apparent r, pour ceux exposés à A, alors la prévalence de B, parmi ceux exposés à A, relativement à la prévalence parmi ceux qui n'y sont pas exposés, doit être plus grande que r »[47]

[45] « *The absolute measure would be important in appraising the public health significance of an effect known to be causal*. », in Cornfield, Haenszel, Hammond *et al.,* 1959, p. 194.

[46] « *appraising the possible noncausal nature of an agent having an apparent effect* », in Cornfield, Haenszel, Hammond *et al.,* 1959, p. 194.

[47] « *If an agent, A, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, B, shows an apparent risk, r, for those exposed to A, relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r*. », in Cornfield, Haenszel, Hammond *et al.,* 1959, p. 194.

En d'autres termes, une explication alternative, en termes de tiers facteur ou de facteur de confusion, au rôle causal du tabagisme pour le cancer du poumon (Cornfield et ses collègues donnent l'exemple d'une hormone, mais la cible est évidemment l'hypothèse constitutionnelle de Berkson[48] ou génétique de Fisher) doit prouver que le facteur en question, l'hormone X, a un effet aussi grand que le tabagisme, c'est-à-dire montrer que « la proportion de producteurs d'hormone X parmi les fumeurs de cigarette doit être au moins 9 fois plus grande que celle des non-fumeurs. » Si ce n'est pas le cas, « alors l'hormone X ne peut rendre compte de la magnitude de l'effet apparent »[49]. Les auteurs renvoient alors à la démonstration mathématique qui est faite dans l'appendice A.

La seconde justification aux mesures relatives est la suivante :

> « Si deux agents non corrélés, A et B, augmentent tous deux le risque d'une maladie, et si le risque de la maladie en l'absence de l'un ou de l'autre agent est petit (en un sens à définir), alors le risque relatif apparent pour A, r, est moindre que le risque pour A en l'absence de B. »[50]

L'argument, formulé mathématiquement dans l'appendice B, est ici plus subtil[51] : le risque relatif apparent de la population (ou du sous-groupe) exposée à A sans être exposée à B va être plus élevé que le même risque relatif pour la population plus large qui est exposée à la fois à A et à B. En s'appuyant sur un autre article écrit par Cornfield et Haenszel et publié en 1960[52], où ils reformulent les arguments

---

[48] Il est intéressant de noter sur ce point que dans l'article écrit par Cornfield et Haenszel en 1960, ceux-ci vont reprendre l'étude de Raymond Pearl sur le lien entre cancer et tuberculose, celle-là même sur laquelle s'appuyait l'argumentation de Berkson pour refuser toute validité aux études rétrospectives, puis aux études prospectives.

[49] « *Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke in a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone-X-producers among cigarette smokers must be at least 9 times greater than that of nonsmokers. If the relative prevalence of hormone-X-producers is considerably less than ninefold, then hormone X cannot account for the magnitude of the apparent effect*. », in Cornfield, Haenszel, Hammond *et al.,* 1959, p. 194.

[50] « *If two uncorrelated agents, A and B, each increase the risk of a disease, and if the risk of the disease in the absence of either agent is small (in a sense to be defined), then the apparent relative risk for A, r, is less than the risk for A in the absence of B.* », in Cornfield, Haenszel, Hammond *et al.,* 1959, p. 194.

[51] Selon Morabia et Szklo, il s'agit d'une approche novatrice au sens où elle permet de traiter les « groupes à petit risque » (« *low-risk group* »). Voir Morabia, A. et Szklo, M., « Letters to the Editor. Re: Smoking and lung cancer: recent evidence and a discussion of some questions. », *International Journal of Epidemiology*, vol. 39 / 6, décembre 2010, p. 1676. Selon Charles Poole, cet argument, ainsi que la démonstration algébrique de l'appendice B est purement circulaire. Poole, Charles, « On the Origin of Risk Relativism », *Epidemiology*, vol. 21 / 1, janvier 2010, p. 3-9.

[52] Cornfield, Jerome et Haenszel, William, « Some Aspects of Retrospective Studies », *Journal of Chronic Diseases*, vol. 11 / 5, Mai 1960, p. 523–534.

énoncés dans l'article de 1959, il apparait que le problème traité est toujours celui du facteur de confusion : ici le risque est effectivement d'attribuer à tort un pouvoir causal à une caractéristique ou à un facteur. Le risque relatif va alors permettre de suspecter l'influence causale d'autres caractéristiques. Cornfield et Haenszel disent ainsi :

> « Le ratio d'incidence fournit une indication de l'importance d'autres caractéristiques que celles soumises à l'étude. Si la caractéristique étudiée est seulement l'une des nombreuses caractéristiques indépendantes associées à la maladie, le ratio sera plus proche de l'unité que si ce n'est pas le cas. »[53]

L'avantage du risque relatif est donc ici évident : si ce risque relatif est égal à 1 ou proche de 1, alors, il est fort possible que la caractéristique étudiée ne soit qu'une cause parmi d'autres de la maladie : il faut donc « faire preuve de la plus grande prudence » (« *exercising great caution* »[54]) si l'on veut attribuer un rôle causal à ce facteur. Les auteurs de l'article montrent alors que cela est tout aussi valable s'il y deux maladies en question et non une seule :

> « Si, par exemple, le risque relatif de développer soit la maladie I soit la maladie II suite à l'exposition à A est le même en l'absence d'autres causes, et si la maladie I, mais pas la maladie II, est liée à la présence de l'agent B, alors le risque relatif apparent de développer la maladie I suite à l'exposition à A sera moindre que le risque relatif pour la maladie II »[55]

En d'autres termes :

> « Si un seul agent est associé à deux maladies, nous pouvons dire que l'association avec la maladie qui a le plus grand risque relatif est moins susceptible d'être expliquée par une troisième cause commune. Dès lors, l'augmentation de 70% du risque de maladie cardio-vasculaire parmi les fumeurs de cigarette (…) pourrait être expliquée comme le résultat possible d'une troisième caractéristique commune dont la prévalence chez les fumeurs de cigarette est deux fois plus élevée que chez les non-fumeurs. Ce serait

[53] « *the incidence ratio provides some indication of the importance of characteristics other than the one being studied. If the characteristic under study is only one of many independent characteristics associated with the disease, the ratio will be closer to unity than if this is not the case.* », in Cornfield et Haenszel, 1960, p.531.

[54] Cornfield et Haenszel, 1960, p.531.

[55] « *If, for example, the relative risk of developing either disease I or disease II on exposure to A is the same in the absence of other causes, and if disease I, but not disease II, also has agent B present, then the apparent relative risk of developing disease I on exposure to A will be less than that for disease II* », in Cornfield, Haenszel, Hammond *et al.,* 1959, p. 194.

> néanmoins arithmétiquement impossible que la caractéristique en question explique que la différence pour le cancer du poumon soit neuf fois plus élevée »[56]

Il ne reste plus alors à Cornfield et ses collègues qu'à énoncer la dernière justification à l'usage d'une mesure relative, qui est particulièrement intéressante:

> « Si un agent causal A augmente le risque de la maladie I et n'a pas d'effet sur le risque de la maladie II, alors le risque relatif de développer la maladie I est à lui tout seul plus grand que le risque relatif de développer la maladie I et la maladie II combinées, tandis que la mesure absolue n'est pas affectée. »[57]

Cela permet selon eux un « affinement de la classification »[58] dans un double sens : en effet, non seulement cela va améliorer la « classification par caractéristique » au niveau de l'exposition (par exemple, la distinction chez les fumeurs entre les fumeurs de cigarettes, de cigares et de pipes, et ceux qui ne fument que des cigarettes), mais aussi la classification au niveau de la maladie : par exemple, le « cancer du poumon peut ainsi être défini de manière à inclure tous les types histologiques ou bien être restreint au cancer épidermoïde ». L'usage du risque relatif permet ainsi de montrer que le tabagisme est beaucoup plus associé au cancer épidermoïde ou au cancer indifférencié qu'à l'adénocarcinome. Ainsi, « on peut dire que la meilleure classification à la fois quant à l'axe de la caractéristique et à celui de la maladie est celle qui conduit au risque relatif le plus grand »[59].L'invention de ce nouvel instrument de mesure qu'est le risque relatif constitue donc une avancée considérable dans l'histoire de l'épidémiologie, et ce pour trois raisons :d'abord il permet d'évaluer « la possible nature non-causale d'un agent qui a un effet

---

[56] « *Similarly, if a single agent is associated with two diseases, we may say that the association with the disease having the higher relative risk is less likely to be explained by a common third cause. Thus, the 70 per cent elevation in risk from coronary heart disease among cigarette smokers that has been reported could possibly be explained as the result of a common third characteristic whose prevalence among cigarette smokers is twice that among nonsmokers. It would be arithmetically impossible, however, for this same characteristic to explain the ninefold difference in lung cancer*. », in Cornfield et Haenszel, 1960, p.531.

[57] « *If a causal agent A increases the risk for disease I and has no effect on the risk for disease II, then the relative risk of developing disease I, alone, is greater than the relative risk of developing disease I and II combined, while the absolute measure is unaffected.* », in Cornfield, Haenszel, Hammond *et al.,* 1959, p. 194.

[58] « *A refinement of classification* », in Cornfield et Haenszel, 1960, p.531.

[59] « *A priori considerations will not often indicate exactly what the classification by characteristic should be; neither will they indicate exactly how the disease should be defined. Thus, smokers may be defined as those who smoke either cigarettes, cigars, or pipes or as those who smoke only cigarettes. Lung cancer may be defined to include all histologic types or restricted to epidermoid carcinoma of the lung*.», in Cornfield et Haenszel, 1960, p.531.

apparent » ; ensuite il permet d'évaluer « l'importance d''un agent au regard des autres agents qui induisent le même effet » ; enfin il « reflète correctement les effets d'une mauvaise classification des maladies ou d'une classification plus précise »[60]

Surtout, du point de vue de notre étude, la relativisation du concept de risque va de pair avec une relativisation du concept de biais. Cette relativisation est particulièrement prégnante dans l'article de Mantel et Haenszel de cette même année 1959 qui est consacré aux « aspects statistiques de l'analyse des données dans les études rétrospectives des maladies». L'article commence en effet par une affirmation pour le moins surprenante pour qui a lu la littérature épidémiologique depuis les années 1940, où l'étude cas-témoins n'a cessé d'être critiquée par de nombreux statisticiens comme étant très peu fiable quant à ses résultats :

> « La méthode rétrospective peut être considérée, selon la théorie statistique établie, comme la méthode d'étude de prédilection. (…) L'étude rétrospective peut être regardée comme une extension naturelle de la pratique des médecins depuis l'époque d'Hippocrate qui utilisaient comme aide au diagnostic les histoires de cas » (Mantel et Haenszel, 1959, p. 720).

Ils ajoutent tout de même que pour que ce soit le cas, il ne doit pas y avoir de « biais importants dans l'étude »[61]. Selon eux, deux types de biais principaux menacent la validité des études rétrospectives : le biais dans la sélection de l'échantillon (qu'il s'agisse des cas ou des témoins) et le biais de l'intervieweur[62]. Pourtant la présence éventuelle de biais n'entame pas nécessairement la crédibilité ou la validité de l'étude et c'est à l'épidémiologiste d'évaluer la situation :

> « Cependant, un simple catalogue des biais résultant du caractère possiblement non représentatif des échantillons des cas et des témoins ne doit pas *ipso facto* invalider les découvertes d'une étude, quelles qu'elles soient.

[60] « *The relative measure is helpful in 1) appraising the possible noncausal nature of an agent having an apparent effect; 2) appraising the importance of an agent with respect to other possible agents inducing the same effect; and 3) properly reflecting the effects of disease misclassification or further refinement of classification.* », in Cornfield, Haenszel, Hammond *et al.,* 1959, p. 194.

[61] « *In the absence of important biases in the study setting, the retrospective method could be regarded, according to sound statistical theory, as the study method of choice. (…) The retrospective study might be looked upon as a natural extension of the practice of physicians since the time of Hippocrates, to take case histories as an aid to diagnosis*. », in Mantel et Haenszel, 1959, p. 720.

[62] Nous avons dénombré six occurrences de cet « *interviewer bias* » de la page 727 à la page 728. La catégorisation est identique dans Cornfield, Haenszel, Hammond *et al.,* à la nuance près qu'ils ajoutent au biais de l'interviewer le biais lié au patient et la question de ses antécédents tabagiques. Ils forment tous deux le problème de l'exactitude de l'information (« *accuracy of information* », in Cornfield, Haenszel, Hammond *et al.,* 1959, p. 181).

> C'est un problème substantiel qui doit être résolu sur le fond pour chaque investigation »[63]

De plus, il est toujours possible de faire appel à des « preuves collatérales » (« *collateral evidence* ») pour fournir des informations quant à « la magnitude potentielle du biais et à la taille de l'association fallacieuse qui pourrait en résulter »[64]. Surtout la grande différence est que les épidémiologistes disposent désormais d'outils et de procédures statistiques qui vont leur permettre de quantifier aussi bien le risque que le biais, et même de l'éliminer :

> « Un problème majeur dans toute étude épidémiologique est d'éviter les associations fallacieuses. Il a été remarqué que là où le risque de maladie change avec l'âge, une association apparente de la maladie avec d'autres facteurs liés à l'âge peut apparaitre. Cependant, il existe des procédures statistiques appropriées pour contrôler ces facteurs connus ou suspectés d'être liés à l'occurrence de la maladie. Ils sont utiles non seulement pour éliminer les biais de l'enquête mais ils peuvent aussi améliorer sa précision ».[65]

Cette procédure, à laquelle Mantel et Haenszel laisseront leur nom, est une méthode d'ajustement, que l'on peut effectuer *a posteriori*, qui, en scindant les variables en classes (par exemple en tranches d'âge de 10 ans, selon Mantel et Haenszel[66]), et en estimant le risque relatif ou l'odds ratio par exemple par tranche d'âge, permet d'obtenir une mesure plus précise, un risque ajusté, pour l'ensemble des strates en ajoutant les différentes mesures par strates. Il constitue un outil essentiel pour éliminer les facteurs de confusion dans une étude épidémiologique. Surtout, l'épidémiologie dispose à présent d'estimateurs qui lui sont propres et la notion de biais va alors pouvoir s'appliquer, comme dans la statistique mathématique de Fisher ou de Pearson, pour ne citer qu'eux, à ces mêmes estimateurs : le biais va alors devenir un problème essentiel de mesure, en l'espèce la mesure du risque. Un

---

[63] « *However, a mere catalogue of biases arising from the possibly unrepresentative nature of a sample of cases and controls should not ipso facto invalidate any study findings. This is a substantive issue to be resolved on its merits for a specific investigation*. », in Mantel et Haenszel, 1959, p. 725.

[64] « *Collateral evidence may provide information on the potential magnitude of bias and the size of spurious associations which could result*», in Mantel et Haenszel, 1959, p. 725.

[65] « *A major problem in any epidemiological study is the avoidance of spurious associations. It has been remarked that where the risk of disease changes with age, apparent association of the disease with other age-related factors can result. However, there are appropriate statistical procedures for controlling those factors known or suspected to be related to disease occurrence. They serve not only to remove bias from the investigation but, in addition, can add to its precision.* », in Mantel et Haenszel, 1959, p. 733.

[66] Mantel et Haenszel, 1959, p. 733.

passage de l'ouvrage de Brian MacMahon, Thomas Pugh et Johannes Ipsen[67], considéré comme « le premier manuel formel d'épidémiologie jamais publié aux Etats-Unis[68] et signe d'une « nouvelle phase de professionnalisation de l'épidémiologie » (Krieger, 1994, p. 889) illustre cet aspect métrologique du concept de biais, en relation avec les concepts de risque relatif et de risque attribuable[69] :

> « Les deux arguments ci-dessus reposent sur l'hypothèse que l'augmentation du taux de mortalité associé à un tabagisme important est le résultat d'une association causale. Une explication alternative serait que cette augmentation résulte d'un biais systématique quelconque inhérent à la méthode d'étude, ou que l'association statistique existe réellement mais résulte d'une association entre d'un côté un tabagisme important et les types spécifiques de mortalité, et de l'autre côté une troisième variable non-identifiée. La taille du risque relatif est habituellement un meilleur indice de la probabilité d'une relation causale que la taille du risque attribuable. Ainsi une différence de 10% (risque attribuable) que l'on remarque entre deux groupes serait moins susceptible d'être une erreur de mesure si elle survenait entre des taux de 1 et de 11 qu'entre des taux de 110 et 120. Cela nécessiterait un biais systématique moins important pour augmenter un taux de 110 à 120 que de 1 à 11, tout comme il est plus facile de se tromper d'un centimètre en mesurant un kilomètre qu'en mesurant un mètre. »[70]

---

[67] MacMahon, Brian, Pugh, Thomas F. et Ipsen, Johannes, *Epidemiologic Methods*, 1ère, Boston, MA, Little, Brown & Co, 1960.

[68] « *the first formal epidemiologic textbook ever published in the United States* », in Krieger, Nancy, « Epidemiology and the web of causation: has anyone seen the spider? », *Social Science and Medicine*, vol. 39 / 7, octobre 1994, p. 887-903. Nous avons choisi de traduire littéralement le mot « *formal* » par son équivalent français « formel » : « officiel » ou « scolaire » (ici, plutôt « universitaire ») constituent néanmoins d'autres traductions possibles étant donné le contexte.

[69] Le concept de risque attribuable a été proposé par Morton Levin en 1953, dans Levin, M. L., « The occurrence of lung cancer in man », *Acta - Unio Internationalis Contra Cancrum*, vol. 9 / 3, 1953, p. 531-541

[70] « *Both the above arguments depend on the assumption that the increase in death rate associated with heavy smoking is the result of a causal association. Alternative explanation would be that the increase resulted from some systematic bias inherent in the study method, or that a statistical association existed in fact but resulted from association of both heavy smoking and the specified types of mortality with some unidentified third variable. The size of the relative risk usually is a better index of the probability of causal relationship than the size of the attributable risk. Thus a difference of 10 per thousand (attributable risk) noted between two groups would be less likely to be an error of measurement if it occurred between rates of 1 and 11 than between rates of 110 and 120. It would take less systematic bias to raise a rate from 110 to 120 than to raise it from 1 to 11, just as it is easier to make an error of one inch in measuring a mile than in measuring a foot*.» in MacMahon, Pugh et Ipsen,1960, p. 231-232.

Ainsi, le ou les biais dans le plan de l'étude pourraient modifier la valeur du risque relatif ou du risque attribuable et conduire à une erreur de mesure de ce risque. En ce sens, plus le risque relatif est grand, plus le risque de biais est faible. En d'autres termes, plus l'association entre deux variables est forte, plus il y a de chances qu'elle soit réelle, et non artificielle en raison d'un ou de plusieurs biais. De même, selon Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin et Wynder, mais aussi Mantel et Haenszel, plus le risque relatif est élevé, plus le risque que l'association soit liée à un tiers facteur s'amenuise, car il faudrait que ce tiers facteur ou cette « troisième variable non-identifiée », selon les mots de MacMahon et Pugh, aient un effet de même magnitude. Le concept de biais a ainsi été relativisé : avec les nouveaux outils à leur disposition, les épidémiologistes peuvent détecter les biais, les mesurer, quantifier leur force ou leur effet, mais aussi réduire cet effet et éventuellement les éliminer de l'étude par des procédures statistiques. La notion forte de biais (qu'on pourrait aussi qualifier d'absolue) a été remplacée par la notion faible (ou relative) de biais.

4.2 Biais et causalité

4.2.1 Les postulats de Koch et les critères de la causalité en épidémiologie, ou la recherche d'un nouveau paradigme.

Avec l'apparition des notions de risque relatif, d'odds ratio, et leur formulation mathématique, ce sont les concepts centraux mais aussi les outils d'analyse essentiels et actuels de l'épidémiologie moderne qui se mettent en place à la fin des années 1950. De plus, dans le cadre de la controverse sur le rôle du tabagisme dans le cancer du poumon, les textes de Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin et Wynder, ainsi que celui de Mantel et Haenszel, permettent sans nul doute un renversement de la charge de la preuve par rapport aux critiques de Berkson ou de Fisher : étant donné l'importance du risque relatif de développer un cancer du poumon en étant fumeur (de l'ordre de 9 pour les fumeurs par rapport aux non-fumeurs, et jusqu'à 24 pour les gros fumeurs, d'après les données de l'enquête de Doll et Hill de 1956 recalculées par MacMahon et Pugh[71]) c'est à Fisher et Berkson de démontrer désormais que leur hypothèse d'un facteur constitutionnel et génétique peut expliquer

[71] MacMahon, Pugh et Ipsen,1960, p. 231.

une telle force de l'association. En ce sens, si cette épidémiologie se situe certes toujours dans le paradigme de la boîte noire, c'est bien à la constitution d'une théorie épidémiologique que nous assistons entre la fin des années 1950 et le début des années 1960, ou, à tout le moins, à la définition d'une méthodologie spécifique pour l'étude des maladies chroniques. Le problème central qui anime ces débats méthodologiques propres à l'épidémiologie est sans nul doute celui de la causalité, et de la possibilité de faire une inférence causale à partir des études observationnelles : il culminera et se clora, au moins provisoirement, avec la parution de l'article de Bradford Hill en 1965 et intitulé : « Environnement et maladie : association ou causalité ? »[72], où sont énoncés les fameux neuf critères de la causalité de Hill (bien que le mot « critère » n'apparaisse pas dans l'article). Les épidémiologistes sont en effet depuis les années 1950 en quête de règles pour effectuer un jugement causal : ils ont donc besoin de critères (qui, pour rappel, vient du grec « κ ρ ι΄ ν ε ι ν » qui signifie « juger ») ou de « fils directeurs » (« *guideposts* »[73]) ou encore de « lignes directrices » (« *guide lines* »[74]). Pour ce faire, les épidémiologistes vont s'inspirer du paradigme dominant à l'époque dans l'épidémiologie et qui concerne les maladies infectieuses, à savoir le paradigme bactériologique des postulats de Henlé-Koch. Un débat a alors lieu dans la section consacrée à la méthodologie de la revue *Journal of Chronic Diseases*, en 1959, dans une série de trois articles qui portent sur « la méthodologie des investigations sur les facteurs étiologiques dans les maladies

[72] Hill, Austin Bradford, « The environment and disease: association or causation? », *Proceedings of the Royal Society of Medicine*, vol. 58 / 5, 1965, p. 295-300.

[73] Yerushalmy et Palmer, 1959, p. 28.

[74] Lilienfeld, Abraham M., « "On the methodology of investigations of etiologic factors in chronic diseases"—some comments », *Journal of Chronic Diseases*, vol. 10 / 1, 1959, p. 41–46. Les mots « *criteria* » et « *guide lines* » apparaissent à la page 41.

chroniques », débat qui va impliquer, dans l'ordre de publication, Jacob Yerushalmy et Carrol Palmer, Abraham Lilienfeld[75], et Philip Sartwell[76].

L'article de Yerushalmy et Palmer commence par une critique traditionnelle des études épidémiologiques (le problème du tiers facteur, l'absence de randomisation) et déplore l'absence de « méthodologie et de critères pour l'évaluation »(Yerushalmy et Palmer, 1959, p. 28) des études épidémiologiques, absence qui rend possible une divergence dans l'interprétation des données. Dès lors, dans la mesure où « la formulation de ces postulats par Koch et leur utilisation dans le champ de la bactériologie a grandement contribué à l'identification méthodique et systématique des organismes causaux dans les maladies chroniques » (Yerushalmy et Palmer, 1959, p. 28), l'idée de Yerushalmy et Palmer est de « comparer » ou de « développer un parallélisme élémentaire entre les enquêtes sur les facteurs étiologiques de nombreuses maladies chroniques et ceux des maladies bactériennes » (Yerushalmy et Palmer, 1959, p. 28). L'objectif ultime des études épidémiologiques consiste en effet selon eux à identifier « une entité unique, définie et finale comme un agent causal » (Yerushalmy et Palmer, 1959, p. 28). Or, concernant les maladies chroniques, les études épidémiologiques ne permettent selon eux que d'étudier les « conditions,

[75] Pour le contexte historique, on peut noter qu'Abraham Lilienfeld (1920-1984) est le beau-frère de Jacob Yerushalmy (1904-1973), ce dernier étant le beau-frère de la femme de Lilienfeld, Lorraine. Ils vivent même ensemble, avec leur femme respective, pendant les années 1940 à Bethesda, dans le Maryland. Carroll Palmer (1903-1972) est quant à lui l'ami et le chef de Yerushalmy : Yerushalmy, après son doctorat de mathématiques, rencontre Palmer, démographe de profession, à la faculté de biostatistiques de l'Université Johns Hopkins, au début des années 1930. Ils ne se quitteront pas du début des années 1940 jusqu'à la fin des années 1960. Dans les années 1930, Yerushalmy se lie aussi d'amitié avec le Docteur Morton Levin (1904-1995), à qui il présentera Lilienfeld, et qui sponsorisera la première étude cas-témoins aux Etats-Unis sur le tabagisme et le cancer du poumon à la fin des années 1940. Ensuite, en 1938, Yerushalmy est embauché au NIH (« *National Institute of Health* ») où il rencontre Harold Dorn (1906-1963), lui aussi démographe qui dirige la première enquête nationale sur le cancer aux États-Unis. Pendant la seconde guerre mondiale, Yerushalmy a de nombreuses discussions avec Joseph Berkson, sur la possible utilisation des statistiques vitales pour expliquer la mortalité néo-natale. En 1947, suite à l'essai clinique randomisé sur la streptomycine réalisé sous la direction de Bradford Hill, Palmer envoie Yerushalmy en Angleterre pour rencontrer Hill. En rentrant, il dit à Lilienfeld et Levin que Hill est un statisticien compétent qui sait bien analyser les données mais qui ne se préoccupe pas assez de la qualité de ces données. Il avoue même : « je ne croirais rien de ce que Hill publiera à propos de ces données ». Quant à Lilienfeld, c'est un médecin diplômé de la faculté de médecine de l'Université du Maryland qui, se désintéressant progressivement de la clinique, rejoint le service de santé publique (« *Public Health Service* ») en 1946, et se voit assigné au Bureau de la tuberculose, dirigé alors par Yerushalmy. En 1948, sur les conseils de Levin, il commence un Master de santé publique à la *Johns Hopkins School of Hygiene and Public Health*, où il aura notamment comme professeur William Cochran (1909-1980) et Philip Sartwell (1908-1999). Toutes ces informations sont extraites de: Lilienfeld, David E., « Abe and Yak: The Interactions of Abraham M. Lilienfeld and Jacob Yerushalmy in the Development of Modern Epidemiology (1945-1973) », *Epidemiology*, vol. 18 / 4, juillet 2007, p. 507-514.

[76] Sartwell, Philip E., « "On the methodology of investigations of etiologic factors in chronic diseases"—Further comments », *Journal of chronic diseases*, vol. 11 / 1, 1960, p. 61–63.

souvent environnementales, qui pourraient être impliquées dans la causalité d'une maladie donnée »[77], en somme, les « vecteurs » (« *vectors* ») ou les « véhicules » (« *vehicles*[78]) de la cause, non la cause elle-même. Selon eux, d'ailleurs, le tabac ne serait qu'un vecteur de la vraie cause, tout comme l'eau était le vecteur de la fièvre typhoïde, la vraie cause étant ici le bacille de la typhoïde que l'eau véhicule. En fait, le premier problème dans l'étude des maladies chroniques est que les épidémiologistes sont confrontés à une « causalité multiple » (« *multiple causation* »[79]). Au fur et à mesure des études néanmoins, ce problème va se réduire et passer progressivement des vecteurs à l'agent causal spécifique. Mais la principale difficulté réside dans le fait que l'agent causal n'est ni nécessaire (par exemple, il y a des cancers du poumon chez des personnes qui n'ont jamais fumé), ni suffisant (par exemple, il y a des fumeurs qui ne développent pas de cancer du poumon), et si le deuxième cas se produit dans le cadre des maladies infectieuses (les porteurs sains), le premier ne se produit jamais d'après Yerushalmy et Palmer. Dès lors, ceux-ci vont reformuler les trois postulats de Koch pour les faire correspondre à la particularité de l'étude des maladies chroniques. Les trois postulats de Koch, tels qu'ils sont formulés par Yerushalmy et Palmer sont les suivants :

> « I. L'organisme doit être trouvé dans tous les cas de la maladie en question.
>
> II. L'organisme doit être isolé des patients et cultivé *in vitro*.
>
> III. Quand l'organisme cultivé est inoculé chez des animaux ou des êtres humains sains, il doit reproduire la maladie. »[80]
>
> Selon les auteurs, ces trois postulats impliquent « deux types de preuves (…) :
>
> A. La présence simultanée de l'organisme et de la maladie, et leur apparition dans la séquence temporelle correcte, et
> B. La spécificité de l'effet de l'organisme sur le développement de la maladie. »[81]

---

[77] « *Rather we are concerned with the investigation of conditions, often environmental, which may be involved in the causation of a given disease*.», in Yerushalmy et Palmer, 1959, p. 28.

[78] Yerushalmy et Palmer, 1959, p. 28.

[79] Yerushalmy et Palmer, 1959, p. 28.

[80] « *I. The organism must be found in all cases of the disease in question. II. It must be isolated from patients and grown in pure culture. III. When the pure culture is inoculated into susceptible animals or man, it must reproduce the disease.* », in Yerushalmy et Palmer, 1959, p. 30.

[81] « *A. The simultaneous presence of organism and disease and their appearance in the correct sequence, and*
B. The specificity of effect of the organism on the development of the disease. », in Yerushalmy et Palmer, 1959, p. 31.

Yerushalmy et Palmer vont alors reformuler le premier postulat de Koch dans le vocabulaire de l'épidémiologie des maladies chroniques, sous deux formes différentes qui correspondent aux deux types d'étude épidémiologique :

> « 1. La caractéristique suspecte doit être trouvée plus souvent chez les personnes qui ont la maladie que chez les personnes qui ne l'ont pas.
>
> 2. Les personnes qui ont la caractéristique doivent développer la maladie plus souvent que les personnes qui n'ont pas la caractéristique. »[82]

La première formulation correspond aux études rétrospectives, la seconde aux études prospectives. Les deux formulations visent à satisfaire le premier critère ou le premier type de preuve, qui renvoie à la présence simultanée de l'organisme (ici, la caractéristique) et de la maladie, mais aussi à la séquence temporelle. La première formulation peut poser problème quant à la séquence temporelle (les auteurs font une référence implicite au biais d'information de l'enquêteur ou bien du patient), mais pas la seconde. Les deux posent évidemment le problème de la sélection du groupe contrôle, afin de pouvoir « jauger l'augmentation de la fréquence »[83] de la maladie ou de l'exposition.

Concernant à présent le second type de preuves, c'est-à-dire la spécificité de l'effet de l'organisme sur le développement de la maladie, et qui correspond aux deuxième et troisième postulats de Koch (Yerushalmy et Palmer, 1959, p. 34), sa fonction est d' « éliminer les cas où un organisme suspect n'est pas une cause réelle, mais ne fait qu'accompagner cette cause »[84], donc d'éliminer les facteurs de confusion : ainsi fumer pourrait être considéré comme « l'analogue d'un « satellite », l'organisme qui accompagne d'autres situations de maladie »[85], la vraie cause étant peut-être constitutionnelle ou liée au mode de vie. Le problème en effet est qu'« établir qu'une caractéristique a un effet spécifique sur une maladie chronique implique des

[82] « *1. The suspected characteristic must be found more frequently in persons with the disease in question than in persons without the disease, or*
2. Persons possessing the characteristic must develop the disease more frequently than do persons not possessing the characteristic.», in Yerushalmy et Palmer, 1959, p. 32.

[83] « *selection of controls against whom an increase in frequency may be gauged*. », in Yerushalmy et Palmer, 1959, p. 32.

[84] «*The requirements of Koch's second and third postulates (…), serve the purpose of ruling out cases in which a suspected organism is not the real cause, but merely accompanies it.* », in Yerushalmy et Palmer, 1959, p. 34.

[85] «*Smoking by itself, as an etiologic factor in lung cancer, may be the analogue of the "satellite," the accompanying organism of other disease situations*. », in Yerushalmy et Palmer, 1959, p. 36.

considérations extrêmement complexes »[86]. Se pose la question de l'autorecrutement (« *self-selection* »), du tiers facteur (le mode de vie, des différences de personnalité, la constitution, etc.) ... Pourtant, selon Yerushalmy et Palmer, c'est bien le critère de la spécificité qui est le plus « crucial » :

> « La démonstration de fréquences relatives élevées dans le groupe d'étude n'est ainsi que la première étape dans le processus de recherche des facteurs étiologiques. L'investigation doit procéder à la seconde et plus cruciale étape (que nous avons appelée, faute d'un meilleur terme, la spécificité de l'effet), c'est-à-dire la démonstration que la différence entre les fréquences relatives reflète une relation spécifique et significative entre la caractéristique suspecte et la maladie considérée »[87]

Dès lors, Yerushalmy et Palmer vont formuler le critère de la spécificité d'une manière extrêmement précautionneuse :

> « Une association observée entre une caractéristique et une maladie doit être testée quant à sa validité en recherchant une relation entre la caractéristique et d'autres maladies et, si possible, en recherchant une relation entre des caractéristiques similaires et reliées entre elles à la maladie en question. La caractéristique suspecte peut être considérée comme spécifiquement liée à la maladie en question quand les résultats d'une telle investigation indiquent que des relations similaires n'existent pas avec d'autres entités morbides quand de telles relations ne sont pas prédictibles sur des bases physiologiques, pathologiques, expérimentales ou épidémiologiques. En général, plus la fréquence de ce genre d'associations est faible, plus la spécificité de l'observation originellement observée est forte, et plus forte est la validité de l'inférence causale. »[88]

---

[86] «*establishing that a characteristic has a specific effect on the development of a chronic disease involves extremely complex considerations*. », in Yerushalmy et Palmer, 1959, p. 34.

[87] « *The demonstration of high relative frequencies in the study group is thus only a first step in the process of searching for etiologic factors. The investigation must proceed to the second and more crucial consideration (which, for want of a better term, is denoted here as that of specificity of effect), i.e., to the demonstration that the difference in relative frequencies reflects a specific and meaningful relationship between the characteristic under suspicion and the disease under consideration*.» , in Yerushalmy et Palmer, 1959, p. 36.

[88] « *An observed association between a characteristic and a disease must be tested for validity by investigating the relationship between the characteristic and other diseases and, if possible, the relationship of similar or related characteristics to the disease in question. The suspected characteristic can be said to be specifically related to the disease in question when the results of such investigation indicate that similar relationships do not exist with a variety of characteristics and with many disease entities when such relationships are not predictable on physiologic, pathologic, experimental, or*

En faisant du critère de la spécificité le critère essentiel de l'inférence causale, Yerushalmy et Palmer retrouvent un argument proposé notamment par Berkson, en 1958, à l'occasion de sa critique de l'étude prospective de Doll et Hill qui montrait que fumer augmentait non seulement le risque de mourir par cancer du poumon mais aussi de nombreuses autres maladies (autres cancers, maladies cardio-vasculaires, etc.) :

> « Cependant, quand on met en place une enquête pour tester la théorie, suggérée par des preuves précédemment obtenues, que fumer cause le cancer du poumon, et que cette enquête montre finalement que fumer cause ou provoque tout un éventail de maladies, cela provoque inévitablement le soupçon qu'on a raté quelque chose »[89]

Il reformule cet argument en 1959, disant que les résultats de ces études (celle de Hammond et Horn, et celle de Doll et Hill) « vont tellement plus loin que les résultats escomptés que l'on pourrait dire qu'ils la réfutent [l'hypothèse que fumer cause le cancer du poumon], car ils la « prouvent trop »[90]. En 1955, Berkson, en étudiant les premiers résultats de l'enquête de Hammond et Horn, avait déjà résumé cette idée sous la forme d'une boutade :

> « En effet, la question soulevée par les découvertes de l'étude menée par l'*American Cancer Society* sur les taux de mortalité plus élevés parmi les fumeurs de cigarette est moins : « Est-ce que fumer des cigarettes cause le cancer du poumon ? », que : « Quelle maladie fumer des cigarettes ne cause pas ? »[91]

Cependant, par-delà la boutade, c'est un argument épistémologique sérieux que Berkson adresse aux études épidémiologiques, qu'elles soient prospectives ou rétrospectives. Plus précisément, c'est la scientificité même de l'épidémiologique qu'il met sérieusement en doute précisément dans la capacité que l'épidémiologie a de prouver une relation de causalité entre une caractéristique et une maladie. En fait,

---

epidemiologic grounds. In general, the lower the frequency of these other associations, the higher is the specificity of the original observed association and the higher the validity of the causal inference. », in Yerushalmy et Palmer, 1959, p. 39.

[89] « *However, when an investigation set up to test the theory suggested by evidence previously obtained, that smoking causes lung cancer, turns out to indicate that smoking causes or provokes a whole gamut of diseases, inevitably it raises the suspicion that something is amiss*.», in Berkson, (1958), *Art.cit.*, p. 34

[90] « *They do not contradict that theory, but they go so far beyond the expected findings that one might say they almost disprove it, because they « prove too much*», in Berkson, 1959, p. 10.

[91] « *Indeed the question raised by the findings in the American Cancer Society study of higher death rates among cigaret smokers is not, "Does cigaret smoking cause cancer of the lungs?"so much as it is, "What disease does cigaret smoking does not cause?"»,* in Berkson,1955, p. 335.

Berkson se situe explicitement dans l'idéal newtonien de la « *vera causa* », c'est-à-dire de la vraie cause, et se réfère aussi bien au principe méthodologique du rasoir d'Ockham (ne pas multiplier les entités sans nécessité) qu'aux règles de raisonnement en science exposées dans le livre III des *Principes de la philosophie naturelle* d'Isaac Newton. Il dit ainsi, après avoir montré que fumer cause de nombreuses maladies :

> « Nous devons trouver une explication unifiée pour expliquer l'association générale observée. En d'autres termes, nous ne devrions pas donner une explication hypothétique pour l'association entre le tabagisme et le cancer du poumon, une autre pour les maladies cardio-vasculaires, une autre pour le cancer de la vessie, une autre pour le cancer de l'estomac (si on peut en trouver une), une autre pour la cirrhose du foie et ainsi de suite (…). Faire cela revient à violer la règle fondamentale, parfois appelée « rasoir d'Ockham », et incarnée dans les célèbres « règles de raisonnement » de Newton : 1. « Il ne faut admettre de causes que celles qui sont nécessaires et suffisantes pour expliquer les faits observés. 2. Dès lors, aux mêmes effets on doit, autant que possible, assigner les mêmes causes. »[92]

Cet appel aux canons de la science moderne et à l'autorité de Newton fait écho à l'appel lancé aux postulats de Koch par Yerushalmy et Palmer dans leur article qui parait la même année que celui de Berkson. Ces deux appels sont symptomatiques selon nous, à travers la question cruciale de la causalité, que ces trois auteurs estiment devoir être spécifique (c'est-à-dire dans une relation de type : une cause-une effet), d'une interrogation ou d'une tentative de fonder l'épidémiologie des maladies chroniques comme une science à part entière. En termes kuhniens, on pourrait y voir les symptômes d'une crise de l'épidémiologie, le paradigme bactériologique n'étant plus adapté à l'étude des maladies chroniques. Cette discussion sur les critères de causalité n'en est encore pourtant qu'à ses prémices, et l'épidémiologie ne pourra faire l'économie d'une redéfinition de la notion même de causalité pour résoudre cette crise.

[92] « *We should seek a unified explanation for the general association observed. That is, we ought not give one hypothetical explanation for the association of smoking with cancer of the lung, another for the vascular diseases, another for cancer of the bladder, another for cancer of the stomach (if one can be found), another for cirrhosis of liver and so forth (…). To do this would violate the fundamental rule, sometimes referred to as "Occam's razor", and epitomized in Newton's famous "rules of reasoning": A. We are not to assume more causes than are sufficient and necessary for the explanation of the observed facts. 2. Hence as far as possible similar effects must be assigned to same causes*», in Berkson, 1959, p. 10.

4.2.2 La critique de la spécificité et la multiplication des critères de causalité.

L'article de Yerushalmy et Palmer est suivi par deux articles qui en sont le commentaire. Le premier est l'œuvre de Abraham Lilienfeld. Il débute par le constat partagé d'un « besoin pour une formulation générale de méthodes procédurales et inférentielles qui serviraient de lignes directrices aux enquêteurs »[93] dans les études épidémiologiques. Sa critique va porter essentiellement sur deux points : la distinction faite par Yerushalmy et Palmer entre vecteur et agent de la maladie, et la notion de spécificité de l'effet. Selon lui, la différence entre vecteur et agent tient essentiellement au « cadre de référence dans lequel l'enquêteur opère, qui en retour reflète des différences dans les niveaux de développement du savoir de ces deux types de maladie [infectieuses et non infectieuses], plutôt qu'une différence dans la logique de la situation»[94]. En somme, selon lui, que l'on dise que c'est l'eau polluée qui cause la fièvre typhoïde (le vecteur), ou bien le bacille de la typhoïde (l'agent), la distinction est finalement purement logique ou nominale et tient au niveau d'explication souhaité ou adopté: on pourrait en effet tout à fait considérer que « d'un point de vue moléculaire, le bacille de la typhoïde peut aussi être considéré comme le vecteur d'un agent chimique spécifique qui est la cause « réelle » de la maladie »[95]. Dès lors, selon les propres critères de Yerushalmy et Palmer, le bacille de la typhoïde ne pourrait pas être considéré comme la cause de la typhoïde. Après avoir pris Yerushalmy et Palmer à leur propre piège, Lilienfeld conclut en montrant que la distinction vecteur-agent n'est pas plus pertinente du point de vue théorique que du point de vue pratique : même en ayant « une connaissance complète des agents causaux à un niveau moléculaire, il serait quand même nécessaire, dans de nombreuses circonstances, de disposer d'informations à propos des vecteurs afin d'instituer de manière effective les mesures requises pour la prévention et le contrôle de la maladie »[96]. En d'autres termes, que

[93] « *there is a need for a general formulation of procedural and inferential methods to serve as a guide line to investigators*.», in Lilienfeld,1959, p. 41.

[94] «*this difference merely reflects differences in the frame of reference within which an investigator operates, which in turn reflects differences in levels of development of knowledge of these two types of disease, rather than differences in the logic of the situation*. », in Lilienfeld,1959, p. 41.

[95] « *From a molecular viewpoint, the typhoid bacillus can also be considered a vector of a specific chemical agent which is the "real" cause of the disease*. », in Lilienfeld,1959, p. 42.

[96] « *At this point it is rather pertinent to indicate that even if we had complete knowledge of causative agents at a molecular level, it would still be necessary, in many instances, to have information concerning vectors in order to institute effectively the measures required for the prevention and control of the disease*. », in Lilienfeld,1959, p. 42.

fumer soit un vecteur ou la cause du cancer du poumon, cela ne change rien d'un point de vue scientifique ni du point de vue de la santé publique.

Concernant à présent le critère de la spécificité, Lilienfeld ne discute pas de la pertinence du critère mais entend par contre en « spécifier les conditions d'application »[97] : en effet, selon lui, « la spécificité de l'effet doit être interprétée en termes de degré d'association de la caractéristique avec la maladie »[98], ou, autrement dit, en termes de fréquence relatives. Ainsi en prenant le ratio de mortalité entre fumeurs et non-fumeurs ou fumeurs occasionnels, donc en prenant comme Cornfield le risque relatif comme mesure de l'association, ce ratio est de 9,35 en ce qui concerne le cancer du poumon, 2,76 pour les maladies respiratoires, et seulement 1,58 pour les maladies coronariennes, et 1,30 pour les cancers autres que celui du poumon.

Dès lors, « on pourrait dire que cette association est « spécifique » si on prendre en compte l'aspect quantitatif de cette association »[99] : le « poids » (« *weight* »[100]) de l'association doit être pris en compte dans l'évaluation. Cette quantification du poids de l'association, ou cette « force de l'association »[101], permet d'ailleurs de détecter des associations qui ne seraient peut-être pas réelles, car, selon Lilienfeld, et contrairement à ce que disent Yerushalmy et Palmer (nous pourrions ajouter Berkson):

> « En second lieu, il ne parait pas improbable qu'il y ait des raisons différentes à ces diverses associations. Par exemple, mon opinion personnelle est que l'association entre tabagisme et cancer du poumon est indicative d'une relation causale, tandis que d'autres associations peuvent être le résultat d'une auto-sélection telle qu'elle est discutée par Yerushalmy et Palmer, et d'autres encore pourraient être artificielles »[102].

[97] « *Generally speaking, it is difficult to quarrel with such a position, although there is need to qualify the application of this criterion.* », in Lilienfeld,1959, p. 42.

[98] « *Specificity of effect must be interpreted in terms of the degree of association of the characteristic with the disease.* », in Lilienfeld,1959, p. 42.

[99] « *In fact, one might state that this association is "specific" when the quantitative aspect of the association is taken into account*. », in Lilienfeld,1959, p. 43.

[100] Lilienfeld,1959, p. 43.

[101] Cette notion de « *strength of association* » n'apparait pas dans l'article de Lilienfeld, mais dans celui de Sartwell (Sartwell, 1959, p. 61). Hill en fera son premier critère de la causalité, hiérarchiquement parlant, dans son article de 1965 (Hill,1965, p. 295).

[102] « *Second, it does not seem improbable that there may be different reasons for the various associations. For example, my personal opinion is that the association of smoking and lung cancer is indicative of a causal relationship, whereas some other associations may be a result of self-selection as discussed by Yerushalmy and Palmer, and still others may be spurious*.», in Lilienfeld,1959, p. 43.

C'est alors que Lilienfeld fait intervenir un autre critère intéressant : celui de la « plausibilité biologique » (« *biologic plausibility* »[103]). Ce critère est cependant problématique car il est relatif à l'état présent des connaissances biologiques et donc limité par celui-ci. De plus, « la découverte d'une association biologiquement invraisemblable pourrait constituer la première piste à cette extension de notre savoir »[104]. Il faut donc toujours garder à l'esprit ce critère, qui permet notamment d'éliminer certaines associations comme non-causales : on ne considérerait pas en effet une « une association entre le tabagisme et les ongles incarnés » (Lilienfeld,1959, p. 43) comme une relation causale.

Enfin, Lilienfeld va avancer un dernier critère pour déterminer si une association est causale, ou plutôt pour « augmenter la vraisemblance que l'hypothèse causale soit vraie »[105] : ce critère est celui de la « convergence» (« *consistency* »[106]). Par ce terme, Lilienfeld désigne le fait que la distribution de la maladie répond en quelque sorte à la distribution du possible facteur étiologique dans la population, ou, en d'autres termes, que les deux distributions convergent :

> « Par exemple, si on trouve une relation entre le tabagisme et le cancer du poumon, et si on trouve aussi que le cancer du poumon apparait plus fréquemment dans un certain segment de la population, il serait de la plus grande importance de déterminer si la raison de l'augmentation de la fréquence du cancer du poumon dans ce segment est le résultat de l'augmentation de la fréquence du tabagisme dans ce segment. La découverte d'une telle convergence augmente la confiance dans la réalité de l'association »[107]

---

[103] Lilienfeld,1959, p. 43.

[104] « *In fact, the finding of a biologically implausible association may be the first lead to this extension of knowledge*.» , in Lilienfeld,1959, p. 43.

[105] « *The finding of consistency increases the likelihood of the causal hypothesis being true*,» , in Lilienfeld,1959, p. 46.

[106] La traduction du mot « *consistency* » par celui de « convergence » serait d'ailleurs plus conforme à l'usage qui est fait de ce mot dans la théorie des probabilités, où l'on dit d'un estimateur qu'il est « convergent » c'est-à-dire « *consistent* ». Nous avons donc choisi de traduire systématiquement « *consistency* » par « convergence » plutôt que par cohérence. La présence des deux notions de « *consistency* » et de « *coherence* » dans les critères présentés dans le rapport du *Surgeon General* de 1964 et dans l'article de Bradford Hill de 1965 (voir *infra*, Section 4.2.3.) renforce selon nous ce choix en permettant d'éviter une confusion ou une redondance entre ces deux critères.

[107] « *For example, if one finds a relationship between smoking and lung cancer, and if one also finds that lung cancer occurs more frequently in a certain segment of the population, it would be of considerable importance to determine whether the reason for the increased frequency of lung cancer in that segment is a result of the increased frequency of smoking in that segment. The finding of such consistency increases the confidence with which one regards the association.»* , in Lilienfeld,1959, p. 44.

Lilienfeld va même plus loin puisqu'il considère que de « telles observations peuvent être considérées comme étant similaires au processus de réplication dans les expérimentations. C'est comme si la nature nous avait fourni des séries d'expérimentations naturelles»[108]. Cette référence à la notion de réplication dans l'expérimentation n'est pas anodine et peut être interprétée, avec les précautions d'usage comme une référence au sens que cette notion a chez Fisher[109]. En effet, chez Fisher, la réplication permet de garantir une estimation correcte de l'erreur, et apparait comme une condition de possibilité du test statistique de signification, test dont la validité est garantie par la randomisation. Or, Lilienfeld ne dit pas autre chose lorsqu'il affirme que « la détermination d'une telle convergence peut être accomplie uniquement par le moyen d'un échantillonnage aléatoire du groupe contrôle »[110], pour éviter justement des échantillons biaisés ou auto-sélectionnés. Si cela peut s'avérer « plus difficile et plus cher » (Lilienfeld,1959, p. 44) d'obtenir un échantillon aléatoire, Lilienfeld propose de réserver cette éventualité à une deuxième phase de l'enquête épidémiologique lorsqu'il s'agit non plus de « d'explorer une série d'hypothèses pour développer une piste » mais de rechercher « un type spécifique de relation (…) de façon plus définitive » ; notamment dans le cadre d'une étude à l'échelle d'une communauté[111].

Enfin, l'article de Philip Sartwell, beaucoup plus court (trois pages), reprend pour l'essentiel, mais de façon plus explicite, les critères avancés par Lilienfeld, tout en critiquant de façon plus virulente le critère de spécificité de Yerushalmy et Palmer. Selon Sartwell, cinq critères (bien qu'il n'emploie pas le mot) permettent de donner un « poids » à la « preuve épidémiologique de l'hypothèse qu'un facteur environnemental (E) est un agent étiologique (ou un vecteur de l'agent étiologique) d'une maladie chronique »[112] :

---

[108] «*Such observations can be regarded as being similar to the process of replication in experimentation. It is as though nature has provided us with a series of natural experiments*. », in Lilienfeld,1959, p. 44.

[109] Voir la section 1.3.2. du présent travail.

[110] «*The determination of such consistency can be accomplished only by means of probability sample controls* », in Lilienfeld,1959, p. 44.

[111] « *If one is exploring a series of hypotheses in order to develop a lead, he will perhaps limit himself to a hospital population and he may utilize various types of controls. On the other hand, if a specific type of relationship is being investigated in a more definitive manner, and a community-wide population study is being contemplated, it would appear more desirable to select probability samples*.» in Lilienfeld,1959, p. 44.

[112] Sartwell,1959, p. 61.

1. La « force de l'association » (« *The Strengh of the Association* »[113]) : cette force peut se constater selon deux modalités différentes :
    a. A quelle fréquence trouve-t-on E chez les personnes malades ? Plus il y a de cas de maladies sans E, plus l'hypothèse de la causalité s'affaiblit, sans pour autant l'invalider
    b. A quelle fréquence apparait E quand la maladie est absente ? Une fréquence considérable de E dans la population générale (donc non-malade) n'affaiblit pas nécessairement l'hypothèse de la causalité.
2. La confirmation de l'association par la réplication (« *Confirmation of the Association by Replication* »[114]) : il s'agit ici des résultats de différentes études, faites par différents enquêteurs dans différentes populations
3. La relation quantitative entre la quantité de E et la fréquence de la maladie (« *The Quantitative Relationship Between the Amount of E and Frequency of Disease* »[115]) : cette relation peut se mesurer en termes de durée d'exposition ou en intensité d'exposition, ou mieux en intégrant les deux.
4. Des relations chronologiques (« *Chronologic Relationships* »[116]) : E doit précéder la maladie, et il doit être possible parfois de « démontrer une courbe de périodes latentes ».

Sartwell précise alors que les points 1 à 4 « ne permettent pas de répondre à l'argument logique selon lequel E pourrait ne pas être l'agent étiologique ou son vecteur, mais un facteur indirectement associé au vrai agent étiologique, encore non-identifié ». Cet argument classique du marqueur ou du tiers facteur va être réfuté par Sartwell en montrant que « plus la correspondance approche la perfection, à différentes époques, en différents lieux, et dans différents groupes de populations », plus la présence d'un tiers facteur ou d'un facteur de confusion devient « improbable »[117].

---

[113] Sartwell,1959, p. 61.
[114] Sartwell,1959, p. 61.
[115] Sartwell,1959, p. 61.
[116] Sartwell,1959, p. 62.
[117] « *It may be charged that points 1 through 4 provide **no** answer to the logical argument that E may not be the etiologic agent or its vector, but merely a factor indirectly associated with the real, unidentified etiologic agent. But it appears quite logical that the more completely points 1 to 4 are satisfied, the greater the likelihood that we are indeed dealing with agent or vector, because otherwise we should have to postulate an association between E and the true unidentified agent or vector which approaches nearer and nearer to perfect correspondence, in different times, at different places, and in different population groups. This would seem to become more improbable as the association between E and the disease becomes closer*.», in Sartwell,1959, p. 62.

5. Le caractère biologiquement raisonnable de l'association (« *The Biologic Reasonableness of the Association* »[118]) : comme Lilienfeld, Sartwell montre qu'il faut être prudent avec ce critère, notre connaissance biologique étant forcement limitée.

Enfin, Sartwell termine son article en critiquant, de façon plus virulente que Lilienfeld, le critère de la spécificité proposé par Yerushalmy et Palmer, et considère que soutenir qu'une caractéristique particulière peut causer une seule maladie est un « sophisme » (« *fallacy* »[119]) et montre que dans le cadre des maladies infectieuses, un même agent peut causer plusieurs maladies, comme le streptocoque qui est associé au « mal de gorge, à la scarlatine, à l'otite moyenne, à l'érésipèle, à la fièvre puerpérale, à l'ostéomyélite, et d'autres conditions moins communes »(Sartwell,1959, p. 63). Selon lui, il est donc clair que le critère de spécificité doit être rejeté.

Ainsi, ce débat organisé par la revue *Journal of Chronic Diseases* en 1959 autour de la méthodologie des enquêtes sur les facteurs étiologiques des maladies chroniques permet l'émergence d'une discussion autour d'un certain nombre de critères ou de lignes directrices, ou autrement dit de preuves, pour déterminer si une association peut être ou non considérée comme causale. Il n'est pas inutile d'en dresser la liste provisoire, et les correspondances entre les différents critères avancés par les différents auteurs :

1. Plus grande co-fréquence de la caractéristique et de la maladie (Yerushalmy et Palmer) / Degré d'association entre la caractéristique et la maladie (Lilienfeld)/ Force de l'association (Sartwell)
2. Séquence temporelle correcte (Yerushalmy et Palmer)/ Relation chronologique (Sartwell)
3. Spécificité de l'effet (Yerushalmy et Palmer)
4. Plausibilité biologique (Lilienfeld / Sartwell)
5. Convergence (« *Consistency* ») des résultats de différentes études / Réplication (Lilienfeld / Sartwell)
6. Relation quantitative entre la quantité de E et la fréquence de la maladie (Sartwell)

[118] Sartwell,1959, p. 62.
[119] Sartwell,1959, p. 63.

Le critère 6 semble assez proche du critère 1, mais dans la mesure où Sartwell distingue ce critère de celui de la force de l'association, il est nécessaire de ne pas les confondre. En réalité, il ne serait pas illogique d'assimiler ce critère à celui du « gradient biologique » chez Hill (Hill,1965, p. 298), c'est-à-dire l'idée d'une relation dose-réponse entre la caractéristique et la maladie : l'idée est que plus on est exposé (soit sur la durée, soit en intensité, soit les deux), plus la maladie a des chances de se développer. A l'inverse le critère 1 renvoie plutôt à l'idée qu'à la fréquence de l'exposition dans la population doit correspondre la fréquence de la maladie dans cette même population.

4.2.3 Les critères de la causalité et la redéfinition de la notion de causalité : de la connaissance à l'action.

En 1959, nous sommes donc en présence de six critères ou lignes directrices qui permettent de déterminer si une association peut être considérée comme causale ou non. Il est temps à présent de clore cette période afin de déterminer ce que cette discussion sur les critères de la causalité en épidémiologie a changé pour l'épidémiologie, ses méthodes, ses buts et les concepts qui forment ce que l'on peut appeler la théorie épidémiologique, et notamment le concept de biais qui brille surtout par son absence dans le cadre de cette discussion sur les critères de causalité. Avant d'aborder la formulation des critères de causalité par Austin Bradford Hill, il est nécessaire de mentionner les cinq critères qui sont proposés dans le rapport fait par le Comité consultatif (où siège notamment le statisticien William G. Cochran) sur les liens entre tabagisme et santé, au *Surgeon General*[120] des Etats-Unis d'Amérique et publié en 1964[121]. Ces cinq critères ont en fait été élaborés par Reuel Stallones et présentés (dans un ordre différent) dans un mémorandum interne de 1963 au *Surgeon General*, qui n'a été publié qu'en 2015[122], et qui porte sur le lien entre le tabagisme et

[120] Le *Surgeon General* est à l'époque le médecin Luther Terry, qui a, entre autres, enseigné à la *Johns Hopkins University* entre 1944 et 1961, et qui a été nommé à ce poste par John Fitzgerald Kennedy en 1961, poste qu'il quittera en 1965

[121] « Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service », *US Department of Health, Education and Welfare, Public Health Service Publication*, Publication n° 1103, 1964. [En ligne : https://profiles.nlm.nih.gov/NN/B/B/M/Q/_/nnbbmq.ocr].

[122] Stallones, Reuel A, « The association between tobacco smoking and coronary heart disease », *International Journal of Epidemiology*, vol. 44 / 3, juin 2015, p. 735-743.

les maladies cardio-vasculaires. Les voici tels qu'ils sont formulés dans le rapport au *Surgeon General* :

1. « La convergence (« *consistency* ») de l'association
2. La force (« *strength* ») de l'association
3. La spécificité (« *specificity* ») de l'association
4. La relation temporelle (« *temporal relationship* ») de l'association
5. La cohérence (« *coherence* [123]») de l'association »[124]

Dans la mesure où nous avons déjà commenté ces critères, il est inutile de le faire à nouveau, et temps de s'intéresser aux fameux neuf critères avancés par A.B. Hill. Le point intéressant dans l'article de Hill, dans une perspective historique, ne tient pas tant d'ailleurs aux critères eux-mêmes, car les plus importants d'entre eux sont déjà présents et discutés depuis au moins 1959, soit six ans avant la parution de l'article en question, que dans le fait que, pour la première fois à notre connaissance, ces critères sont ordonnés selon une hiérarchie, c'est-à-dire du plus important au moins important. C'est là à notre sens la principale nouveauté de l'article de Hill, ainsi que quelques distinctions et clarifications quant aux critères eux-mêmes. Un autre point intéressant est la question telle qu'elle est formulée par Hill et à laquelle les critères (que Hill ravale au rang d'« aspects », et plus tard de « points de vue »[125]) ont pour fonction de répondre :

> « Nos observations révèlent une association entre deux variables, parfaitement claire et au-delà de ce que nous pourrions attribuer au hasard. Quels aspects de l'association devons-nous spécialement considérer avant de décider que l'interprétation la plus probable est celle de la causalité ? [126]»

[123] La notion de « *coherence* » est assez proche de celle de plausibilité biologique de l'association, sans être totalement identique. Stallones parle ainsi d'une cohérence avec « les faits biologiques connus sur cette maladie ») : il y a donc l'idée que l'association ne contredise pas ce qu'on sait déjà sur l'aspect biologique de la maladie. A l'inverse, la notion de « *consistency* » renvoie au fait que les résultats de différentes études réalisées dans différents pays avec différentes populations sont similaires ou convergents. Voir Stallones,2015, p. 735. Hill quant à lui séparera la plausibilité biologique et la cohérence avec les faits biologiques.

[124] « Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service », *US Department of Health, Education and Welfare, Public Health Service Publication,* Publication n° 1103, 1964, p. 20.

[125] «*Here then are nine different viewpoints from all of which we should study association before we cry causation.* », in Hill,1965, p. 299.

[126] «*Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?* », in Hill, 1965, p. 299.

Voici donc les neuf critères de Hill, classés par ordre d'importance, et leur justification respective :

1. La force (« *strength* »[127]) : Hill donne comme exemple le « cancer du ramoneur » découvert par le chirurgien Percival Pott en 1775 et s'appuie sur une étude de Doll qui montre en 1964 que la mortalité des ramoneurs par cancer du crotum est 200 fois celle des travailleurs non-exposés à des goudrons ou des huiles minérales. Cette « augmentation *énorme* »[128] a permis à Pott de faire une conclusion correcte. Ceci vaut aussi pour le tabagisme et le cancer du poumon puisque le taux de mortalité des fumeurs est neuf fois supérieur à celui des non-fumeurs, et même vingt à trente fois supérieur pour les gros fumeurs.
2. La convergence (« *consistency* »[129]) : il cite alors le rapport de 1964 du *Surgeon General* qui montre que « 29 études rétrospectives et 7 études prospectives » ont trouvé une association entre tabagisme et cancer du poumon. Dès lors, « nous pouvons légitimement inférer que l'association n'est pas due à une erreur ou à un sophisme constant qui imprégnerait toutes les études. »[130]
3. La spécificité (« *specificity* »[131]) : si cette troisième caractéristique, vivement critiquée par Lilienfeld et surtout Sartwell, doit être « invariablement considérée », il ne faut pas pour autant en « exagérer l'importance » (« *over-emphasize the importance* »[132]). Ainsi, ce n'est pas parce que le taux de mortalité est plus élevé chez les fumeurs que chez les non-fumeurs pour de nombreuses causes de maladie (Hill fait ici implicitement mais clairement référence à la critique de Berkson), il reste qu'il y une « spécificité dans la magnitude de l'association », donc dans la force de l'association(« *here surely one must return to my first characteristic, the strengh of the association* »[133]), puisque le cancer du

[127] Hill, 1965, p. 295.
[128] « *the enormous increase* », in Hill, 1965, p. 295. C'est Hill qui souligne.
[129] Hill, 1965, p. 296.
[130] « *In other words we can justifiably infer that the association is not due to some constant error or fallacy that permeates every inquiry.»,* in Hill, 1965, p. 296. Si le mot « biais » n'est pas employé, il est difficile de ne pas y voir une référence ici.
[131] Hill, 1965, p. 297.
[132] Hill, 1965, p. 297.
[133] Hill, 1965, p. 297.

poumon augmente de 900 à 1000% avec le tabagisme, tandis que les autres causes de décès n'augmentent que de 10, 20 ou 50%. En d'autres termes, « si la spécificité existe on peut tirer des conclusions sans hésitation ; si elle n'est pas apparente, nous ne sommes pas pour autant nécessairement laissés irrésolus comme l'âne de Buridan »[134]

4. La temporalité (« *temporality*»[135]) : ce problème se pose surtout pour les maladies qui ont un long temps de latence, comme certaines maladies professionnelles, et c'est un problème qui ne se rencontre pas souvent.
5. Le gradient biologique (« *biological gradient* »[136]), aussi appelé la « courbe dose-réponse » (« *dose-response curve* »[137]) : cette courbe renforce la preuve (« *adds a very great deal* ») d'une relation causale, et son absence l'affaiblit mais sans la détruire. Ainsi le fait que le taux de mortalité par cancer du poumon augmente de façon linéaire avec la quantité de cigarettes fumées « admet une explication simple et met sans aucun doute clairement en lumière le fait »[138].
6. La plausibilité (« *plausibility* »[139]) : il s'agit de la plausibilité biologique, qui dépend du savoir de l'époque. Hill la considère comme utile (« *helpful* ») mais il est convaincu que c'est une « caractéristique (…) qu'on ne peut exiger »[140].
7. La cohérence (« *coherence* »[141]) : il s'agit ici de montrer que « l'interprétation en termes de cause à effet de nos données ne doit pas entrer sérieusement en conflit avec ce que nous savons de l'histoire naturelle et de la biologie de la maladie »[142]. Hill ajoute que si les preuves histopathologiques ou les preuves de laboratoire renforcent la preuve en

[134] « *In short, if specificity exists we may be able to draw conclusions without hesitation; if it is not apparent, we are not thereby necessarily left sitting irresolutely on the fence*» in Hill, 1965, p. 297.
[135] Hill, 1965, p. 297.
[136] Hill, 1965, p. 298.
[137] Hill, 1965, p. 298.
[138] « *The clear dose-response curve admits of a simple explanation and obviously puts the case in a clearer light*.», in Hill, 1965, p. 298.
[139] Hill, 1965, p. 298.
[140] « *It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand*.» in Hill, 1965, p. 298.
[141] Hill, 1965, p. 298. Il emprunte explicitement le concept de « *coherence* » au rapport américain du *Surgeon General* de 1964.
[142] « *On the other hand the cause-and-effect interpretation of our data should not seriously conflict with the generally known facts of the natural history and biology of the disease* » in Hill, 1965, p. 298.

faveur d'un lien de causalité, leur absence n'invalide (« *nullify* »[143]) pas pour autant les observations épidémiologiques faites sur les hommes.

8. L'expérimentation (« *experiment* »[144]) : contrairement à ce que l'on pourrait croire il ne s'agit pas ici de preuves établies en laboratoire mais plutôt d'expériences grandeur nature. Hill cite comme exemple le fait de savoir ce qui se passe quand par exemple les gens arrêtent de fumer : si la fréquence du cancer du poumon diminue, alors « la preuve la plus forte de l'hypothèse de la causalité pourrait être révélée »[145].
9. L'analogie (« *analogy* »[146]) : ceci n'est valable qu'en certaines circonstances. Par exemple, Hill dit qu'après avoir les effets de le thalidomide[147] ou de la rubéole chez les femmes enceintes et surtout sur le fœtus, « on serait sûrement plus enclin à accepter une preuve plus faible mais similaire avec un autre médicament ou une autre maladie virale durant la grossesse. »[148].

Ces neuf critères de la causalité ont été et sont toujours abondamment commentés depuis la parution de cet article, aussi bien par les épidémiologistes que par les épistémologues[149] et constituent encore de nos jours un élément essentiel de la boite à outils des épidémiologistes. Pour notre part, nous nous contenterons de quelques commentaires essentiels pour notre étude.

Le premier point important tient au fait que cet article, tout comme le rapport du *Surgeon General*, soutient la thèse selon laquelle affirmer un lien de causalité entre deux variables relève essentiellement, et même exclusivement, d'un jugement et même d'une décision (Hill dit ainsi que ses neuf points de vue ont pour fonction de « nous aider à nous décider »[150] mais aussi que passer de l'association à la causalité

[143] Hill, 1965, p. 298.
[144] Hill, 1965, p. 298.
[145] «*The dust in the workshop is reduced, lubricating oils are changed, persons stop smoking cigarettes. Is the frequency of the associated events affected? Here the strongest support for the causation hypothesis may be revealed*. » in Hill, 1965, p. 298-299.
[146] Hill, 1965, p. 299.
[147] Le thalidomide est un médicament utilisé durant les années 1950 et 1960 comme sédatif et anti-nauséeux, notamment chez les femmes enceintes, qui provoque de graves malformations congénitales.
[148] « *With the effects of thalidomide and rubella before us we would surely be ready to accept slighter but similar evidence with another drug or another viral disease in pregnancy*.», in Hill, 1965, p. 299.
[149] Pour une discussion proprement épistémologique sur la fonction des critères de Hill, voir par exemple Bird, Alexander, « The epistemological function of Hill's criteria », *Preventive Medicine*, vol. 53 / 4-5, octobre 2011, p. 242-245.
[150] « *to help us to make up our minds*», in Hill, 1965, p. 299.

relève d'une « décision », dont il faut examiner les conséquences[151]), jugement qui va bien au-delà d'une quelconque expérience cruciale et même des statistiques. Le rapport du *Surgeon General* de 1964 est explicite sur ce point :

> « Les méthodes statistiques ne peuvent pas établir la preuve d'une relation causale dans une association. La signification causale d'une association est une affaire de jugement qui va au-delà de tout énoncé de probabilité statistique »[152]

Hill ne dit pas autre chose quand il soutient qu'aucun « test formel de signification ne permet de répondre à ces questions »[153]. Il souligne aussi que les tests de signification ont pris une importance considérable à son époque, notamment aux Etats-Unis :

> « Cependant, je soupçonne que, trop souvent, nous perdons beaucoup de temps, nous attrapons l'ombre et nous perdons la substance, nous affaiblissons notre capacité à interpréter les données et à prendre des décisions raisonnables, quelle que soit la valeur de P. (…). Comme le feu, le test du χ^2 est un excellent serviteur mais un mauvais maître »[154].

Ceci nous conduit au deuxième point important qui ressort de ce problème de la causalité en épidémiologie et qui concerne directement le statut épistémologique de l'épidémiologie, et ce, selon deux modalités différentes. La première modalité renvoie au but même de l'épidémiologie ou à sa finalité : a-t-elle pour but de produire des connaissances pour elles-mêmes ou bien de produire des connaissances pour l'action, en l'espèce une action de santé publique ? La réponse de nombreux épidémiologistes de l'époque est claire : il s'agit avant tout d'agir. Ainsi, la dernière partie de l'article de Hill est intitulée : « *The Case for Action* »[155], que l'on pourrait traduire par : « La nécessité d'agir ». Selon lui en effet :

---

[151] « *Finally, in passing from association to causation I believe in 'real life' we shall have to consider what flows from that decision*.» , in Hill, 1965, p. 300.

[152] « *Statistical methods cannot establish proof of a causal relationship in an association. The causal significance of an association is a matter of judgment which goes beyond any statement of statistical probability*.», in « Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service », *US Department of Health, Education and Welfare, Public Health Service Publication*, Publication n° 1103, 1964, p. 20.

[153] « *No formal tests of significance can answer those questions.* », in Hill, 1965, p. 299.

[154] « *Yet too often I suspect we waste a deal of time, we grasp the shadow andlose the substance, we weaken our capacity to interpret data and to take reasonable decisions whatever the value of P.(…). Like fire, the X2 test is an excellent servant and a bad master.*» , in Hill, 1965, p. 299.

[155] Hill, 1965, p. 300.

> « Finalement, en passant de l'association à la causalité, je crois que dans la « vraie vie » nous devons réfléchir à ce qui découle de notre décision. Sur des fondements scientifiques, nous ne devrions pas procéder de la sorte. »[156]

En effet, si la question de la causalité est bien évidemment une question scientifique, comme Berkson ne cessera de l'affirmer, il reste que quand il s'agit de risques ou de dangers environnementaux ou comportementaux, quand il s'agit de la santé et de la vie de milliers ou de millions de personnes, la question de l'action, la question pratique, prime sur la question scientifique ou théorique. Il ne faut donc pas juger des preuves en elles-mêmes, indépendamment de ce qu'elles impliquent et de façon objective comme dans la science, mais il faut adapter le standard de la preuve en fonction de la situation donnée, en fonction des conséquences que cela implique. Sartwell soutient une idée similaire en disant qu'il ne faut pas être « trop frileux [« *overcautious* »] en rejetant la preuve épidémiologique » (Sartwell, 1960, p. 63) car en ce cas, quand Snow a découvert que l'eau polluée était responsable de l'épidémie de choléra, les autorités sanitaires n'auraient dû prendre aucune mesure de prévention sous prétexte que cette découverte n'était pas corroborée par « une preuve valide de laboratoire » ou que l'hypothèse que l'eau contenait le poison du choléra n'était pas « plausible biologiquement » (Sartwell, 1960, p. 63) à l'époque.

Stallones est encore plus explicite sur la distinction entre l'aspect scientifique ou objectif, l'aspect décisionnel ou subjectif, et l'aspect social, c'est-à-dire les conséquences économiques et sociales que cela impliquerait si par exemple on considérait que le tabagisme cause des maladies cardio-vasculaires. Ainsi, face à la question de savoir si fumer est ou non un facteur causal de la maladie coronarienne, notre réponse « peut varier en fonction de la sphère d'activité qui est impliquée :

> « 1. La réponse scientifique est évidente. A partir du moment où les données sont inadéquates pour résoudre la question, nous devons chercher plus de données (…).
>
> 2. La réponse personnelle est le choix de chaque individu. Elle sera déterminée par des préjugés personnels, la personnalité, les émotions, ainsi que par l'évaluation intellectuelle de la preuve. Pour différentes raisons, je crois que la

[156] « *Finally, in passing from association to causation I believe in 'real life' we shall have to consider what flows from that decision. On scientific grounds we should do no such thing.»* , in Hill, 1965, p. 300.

seconde hypothèse [fumer n'est pas un agent causal de la maladie coronarienne] est correcte(…).

3. La réponse sociale est une question embarrassante, car elle doit déterminer les attitudes et les actions officielles. Je crois que la réponse sociale doit être d'accepter la première alternative [fumer est un agent causal de la maladie coronarienne] comme une hypothèse de travail et agir en conséquence. Que cela n'est pas justifiable scientifiquement doit être admis, de même que cela ne doit pas porter atteinte au droit de chaque individu à choisir son propre destin. (...). C'est un pari, qui, s'il est pris correctement, met dans la balance un gain social énorme contre une perte économique appréciable, si les actions réussissent. »[157]

On retrouve enfin une position similaire dans le rapport remis au *Surgeon General*, dans la section consacrée au « jugement du Comité » qui apparait, en gras, sous la forme suivante, et qui conduira justement à apposer des avertissements sanitaires sur les paquets de tabac vendus aux Etats-Unis d'Amérique à partir de 1965 :

« **Fumer des cigarettes constitue un danger pour la santé suffisamment important aux Etats-Unis pour justifier des mesures correctives appropriées** »[158]

Ainsi, comme on peut le constater, le jugement de causalité tel qu'il est pratiqué par les épidémiologistes n'est pas tant un jugement scientifique au sens strict, et par là impartial et objectif, indépendamment des conséquences qui pourraient découler de ce jugement, mais un jugement qui induit une décision et une action (ou une non-

[157] « *1. The scientific response is obvious. Since the data are inadequate to resolve the question, we must seek more data. (…)*

2. The personal response is the choice of each individual. This will be determined by personal prejudice, personality and emotion as well as intellectual assessment of the evidence. For various reasons, I believe the second hypothesis is correct (…)

3. The social response is the troublesome one, because this must determine official attitudes and actions. I believe the social response must be to accept the first alternative as a working hypothesis and act accordingly. That this is not scientifically justifiable must be granted, nor should official action impair the right of each individual to determine his own fate. However, the social response that is called for is a massive endeavour to discourage smoking using fully the not inconsiderable resources of the Federal and State Governments. This is a gamble, which if correctly taken balances enormous social gain as against an appreciable economic loss, if the actions are successful. » in Stallones,2015, p. 742

[158] «***Cigarette smoking is a health hazard of sufficient importance in the United States to warrant appropriate remedial action*** », in « Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service », *US Department of Health, Education and Welfare, Public Health Service Publication*, Publication n° 1103, 1964, p. 33.

action), et dont l'action qui en découle pourrait d'ailleurs conduire à confirmer ou à infirmer le jugement, comme le montre Hill dans le critère consacré à l'expérimentation. Plus précisément, il s'agit d'un jugement où les conséquences (l'action ou la décision) font partie du jugement lui-même : il s'agit donc bien d'un raisonnement qui doit tenir ensemble les causes supposées de la maladie et les conséquences économiques et sociales que le fait de supprimer ou de réduire ces causes pourrait engendrer. Ce raisonnement est d'ailleurs sous sa forme logique assez proche de celui du médecin qui doit tenir ensemble diagnostic, pronostic et thérapeutique lorsqu'il prend soin d'un patient, et où des aspects économiques, sociaux ou éthiques sont aussi susceptibles d'entrer en ligne de compte. Il ne serait d'ailleurs sans doute pas inintéressant d'effectuer un rapprochement épistémologique plus soutenu entre la médecine et l'épidémiologie, tant cette dernière apparait, à l'instar de la médecine, autant comme une science que comme une technique.

En ce sens, et c'est le troisième point que nous souhaiterions aborder, l'épidémiologie ne saurait se réduire à une méthode comme le soutient par exemple Alex Broadbent dans un des rares ouvrages contemporains consacré spécifiquement à l'épistémologie de l'épidémiologie[159]. Selon lui, en effet, l'épidémiologie se caractérise par l'absence de théorie, ou plutôt par « l'absence de domaine propre à la théorie épidémiologique », au sens où par exemple l'énoncé : « le tabac cause le cancer du poumon » ne peut être dit vrai ou faux que par d'autres « branches des sciences biomédicales », non par l'épidémiologie elle-même[160]. Ainsi, d'après Broadbent, une des particularités de la discipline épidémiologique serait d'être une science qui ne prétend pas dire la vérité sur la réalité, et qui sous-traiterait en quelque sorte la justification des énoncés qu'elle produit à d'autres sciences. En somme, l'épidémiologie serait essentiellement une méthode, ou un ensemble de méthodes : « l'expertise d'un épidémiologiste est méthodologique »[161].

Or, il semble bien que ce à quoi nous assistons au tournant des années 1950-1960 soit la constitution d'une théorie épidémiologique avec un domaine propre de recherche (les maladies chroniques), un objet d'étude (les facteurs de risque des maladies chroniques), des buts (décrire la prévalence et l'incidence des maladies,

[159] Broadbent, Alex, *Philosophy of epidemiology*, Basingstoke, Hampshire ; New York, Palgrave Macmillan, 2013.
[160] Broadbent, 2013, p 3-6.
[161] Broadbent, 2013, p. 5.

expliquer cette prévalence et cette incidence en cherchant à en déterminer les causes, prendre des mesures de santé publique), une méthode spécifique (la méthode statistique appliquée aux études observationnelles, prospectives et rétrospectives), des instruments de mesure (le risque relatif, l'odds ratio), mais aussi un ensemble de concepts qui ne lui sont pas forcément propres mais qu'elle va redéfinir en fonction de ses objets d'étude et de ses buts. Ainsi, le concept de causalité se voit par exemple redéfini précisément à cette époque avec l'idée d'une « causalité multiple », notion que Nancy Krieger considère même comme le « canon de l'épidémiologie contemporaine » (Krieger, 1994, p. 887) : cette multiplicité des causes va ainsi être représentée par exemple par Lilienfeld sous la forme d'une « chaîne de relations causales »[162], où chacun des facteurs « agit indépendamment »[163] au niveau cellulaire, sans qu'aucun d'entre eux ne soit ni nécessaire, ni suffisant.

Cette transformation ou cette redéfinition du concept de causalité afin de pouvoir rendre compte des facteurs qui sont susceptibles d'induire des maladies chroniques, dont la discussion sur les critères de causalité est le prolongement, culmine d'ailleurs dans la notion de « *web of causation* » (littéralement, « toile de la causalité », qu'on peut traduire aussi par « réseau de causalité) qui apparait pour la première fois dans l'ouvrage de MacMahon et Pugh en 1960. Elle est conçue, d'après Krieger, comme une critique de l'image de la chaîne causale que MacMahon et Pugh jugeaient incapable de prendre en compte « la « généalogie complexe » des antécédents de chaque composant de la chaine mais aussi comment ces généalogies des divers facteurs pouvaient s'imbriquer et créer une variété d'associations directes et indirectes »[164]. En tout cas, ces discussions denses sur la notion de causalité et sur ses éventuels critères témoignent de la volonté des épidémiologistes de l'époque de donner une consistance scientifique à leur discipline. Sur ce point, la discussion

[162] Lilienfeld, Abraham M., « Epidemiological methods and inferences in studies of noninfectious diseases », *Public Health Reports*, vol. 72 / 1, 1957, p. 51-60

[163] Lilienfeld, 1957, p. 56.

[164] « *It was in this context [celui de la guerre froide et du Maccarthysme, qui n'incitait pas les épidémiologistes à chercher les causes sociales des maladies] that MacMahon et al. introduced the concept of the 'web of causation' . They did so in reaction to the then prevalent notion of 'chains of causation,' which they argued failed to take into account: (1) the 'complex genealogy of antecedents' of each 'component' in the 'chain', and (2) how the genealogies of diverse factors or outcomes might overlap, creating a variety of indirect as well as direct associations.*» in Krieger, 1994, p. 880. Voir aussi MacMahon, Pugh et Ipsen, 1960, p. 18.

proprement épistémologique de Cornfield sur les « principes de la recherche »[165] et sa distinction entre les « champs de recherche » (« *fields of research* ») qui ont un haut « degré d'articulation[166] » comme la physique ou les mathématiques et ceux qui ont un « bas degré d'articulation » comme la géographie ou l'épidémiologie, témoignent selon nous d'une véritable réflexion sur la scientificité propre à l'épidémiologie.

Enfin, le dernier point important à propos de ce problème de la causalité nous reconduit au concept de biais. En effet, ce concept semble quasiment absent de la discussion sur la causalité en épidémiologie. Or, cette absence permet selon nous de manifester *a contrario* la place et la fonction que ce concept occupe dans l'esprit des épidémiologistes de l'époque : selon eux, le concept de biais remplit une fonction précise non quand il s'agit de juger de la causalité entre un ou plusieurs facteurs et une maladie, mais quand il s'agit d'affirmer que l'association entre deux variables est bien réelle et non artificielle ou fallacieuse. Le rapport du *Surgeon General* distingue clairement ces deux étapes de l'inférence :

> « En menant des études où l'on utilise la méthode épidémiologique, de nombreux facteurs, variables, et résultats d'investigations doivent être considérés comme déterminant d'abord si une association entre un agent ou un attribut et une maladie existe réellement. Le jugement sur ce point est fondé sur des mesures directes ou indirectes de l'association suggérée. S'il est montré qu'une association existe, alors la question est posée : « Est-ce que l'association a une signification causale ? »[167]

Dès lors, il apparait que le concept de biais relève de la notion de plan d'étude, l'étude épidémiologique ayant précisément pour fonction d'établir si une association

---

[165] Cornfield, Jerome, « Principles of Research », *Statistics in Medicine*, vol. 31 / 24, octobre 2012, p. 2760-2768. L'article a été originellement publié dans *American Journal on Mental Deficiency,* Volume 64, 2, 1959, p. 240–252

[166] Cornfield mesure et définit le degré d'articulation d'un champ (« *degree of articulation of a field*») ainsi : « l'ampleur avec laquelle les phénomènes par lesquels le champ est concerné sont potentiellement capables d'être expliqués et prédits en termes d'un petit nombre de constantes et de concepts fondamentaux » (« *the extent to which the phenomena with which the field is concerned are potentially capable of being explained and predicted in terms of a small number of fundamental concepts and constants.* »), in Cornfield, 2012, p. 2760.

[167] « *In carrying out studies through the use of this epidemiologic method, many factors, variables, and results of investigations must be considered to determine first whether an association actually exists between an attribute or agent and a disease. Judgment on this point is based upon indirect and direct measures of the suggested association. If it be shown that an association exists, then the question is asked: "Does the association have a causal significance?"»* in « Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service », *US Department of Health, Education and Welfare, Public Health Service Publication*, Publication n° 1103, 1964, p. 20.

entre deux variables existe, et de mesurer justement la force de cette association. En ce sens, le concept épidémiologique de biais est étroitement lié au concept statistique de biais qui apparait chez Fisher en lien avec le plan d'expérience et la notion d'estimation, alors même que nous avons vu que le concept de biais, comme celui d'ailleurs de randomisation, n'avait pas le même sens chez Fisher et chez Hill par exemple. C'est donc le lien entre le concept de biais et celui de plan d'expérience qu'il s'agit d'étudier à présent plus en détail.

4.3 Biais et plan d'expérience :

4.3.1 La notion de plan d'expérience.

Si la théorisation de la méthode expérimentale peut être attribuée à l'œuvre de Francis Bacon, le *Novum Organum*, au XVIe siècle, la notion de plan d'expérience (en anglais : « *design of experiment* ») est quant à elle relativement récente dans l'histoire des sciences et encore assez peu étudiée par les épistémologues. En ce sens, l'origine de la notion de plan expérimental, comme étroitement lié à l'analyse statistique de la variance qui l'accompagne et la possibilité d'étudier plusieurs facteurs ou variables à la fois dans un plan factoriel et non plus, comme dans le protocole expérimental, d'étudier un facteur à la fois[168] (en anglais, on parle de la méthode *OFAT* pour « *one factor at a time* »), est à rattacher au fameux livre éponyme de Ronald Fisher, *The Design of Experiments*, publié en 1935[169]. La question qui se pose immédiatement est de savoir ce qu'est véritablement un plan d'expérience, mais aussi quelle est sa fonction et son but, ou encore ce qu'il permet de faire. La réponse la plus précise de Fisher sur ce point se situe sans doute dans l'article qu'il dédie explicitement à cette

[168] Dans une citation restée célèbre, Fisher souligne, dès 1926, que l'opinion selon laquelle il faut poser une seule question à la nature à la fois est complètement erronée : « *No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed.*», in Fisher, Sir Ronald Aylmer, « The Arrangement of Field Experiments », *Journal of the Ministry of Agriculture of Great Britain*, vol. 33, 1926, p. 503-513.La citation est à la page 511. Cet article est en partie une réponse à un article de Sir John Russell, directeur de la station de Rothamsted, qui défend une approche en termes d'un facteur à la fois : « Field Experiments : How They are Made and What They Are. », *Journal of the Ministry of Agriculture of Great Britain*, vol. 32, 1926, p. 989-1001

[169] Fisher, Ronald A., *The Design of Experiments*, Edinburgh, Oliver and Boyd, 1935.

question[170] et qui a pour titre : « La place du plan d'expérience dans la logique de l'inférence scientifique ». Ce texte est issu d'une conférence faite à Paris en 1961, dans le cadre d'un colloque international organisé par le Centre National de la Recherche Scientifique et intitulé : « Le plan d'expérience ». Selon lui, la planification des expérimentations est devenue au cours des années 1950 une « nouvelle discipline » au sein des « sciences statistiques » (Fisher, 1965, p. 33), avec la publication de « nombreux livres » (Fisher, 1965, p. 33) sur ce sujet. Fisher va même plus loin puisqu'il considère que « le plan d'expérience n'est pas, comme on pouvait le penser il y a encore quelques années une extension accidentelle des études statistiques, mais est central à tout le processus des sciences naturelles »[171] . En effet, Fisher remarque que l'étude et l'amélioration des plans d'expérience permet d'obtenir une plus grande précision dans les résultats (« deux fois plus, ou cinq fois plus, ou davantage ») alors même que l'amélioration des outils statistiques ne permet d'augmenter la précision de quelques pourcents.

Le plan d'expérience est ainsi selon Fisher un outil, une technique et son approche est essentiellement « technologique »[172] : le terme anglais « *design* » montre d'ailleurs bien cette idée de conception ou de dessein qui préside à cette conception, comme on conçoit par exemple un outil ou une machine, plus que sa traduction française par « plan ». La fonction de ces instruments que sont les différents plans d'expérience, et la supériorité de certains sur d'autres, consiste ainsi dans la capacité qu'ils ont « d'extraire des données plus d'« information » à propos du sujet soumis à l'étude, et ainsi de mener à des estimations d'une plus grande précision, et à des tests de signification d'une plus grande sensibilité »[173]. Ainsi, en étudiant spécifiquement et délibérément les plans d'expérience, il est possible d'obtenir « des gains technologiques quantitativement importants », à condition que « le plan d'expérimentation et d'observation soit logiquement cohérent avec les buts de l'expérience, ou, en d'autres termes, avec le genre d'inférence à propos du monde

[170] Fisher, Ronald A., « The place of the design of experiments in the logic of scientific inference », *Sankhyā: The Indian Journal of Statistics, Series A*, 1965, p. 33–38.

[171] « *For, in truth, the Design of Experiments is not, as it might have been thought but a few years ago, a casual extension of statistical studies, but is central to the whole process of the Natural Sciences***.**», in Fisher, 1965, p. 35.

[172] « *My approach at that time was frankly a technological one***.**», in Fisher, 1965, p. 33.

[173] « *Technically, I could see that some methods were superior to others in the concrete sense of extracting from the data more "information" on the subjects under enquiry, and therefore of leading to estimates of higher precision, and to tests of significance of greater sensitivity.»,* in Fisher, 1965, p. 33.

réel »[174]. Ainsi, pour Fisher, les « données » constituent la « fondation logique pour inférer des énoncés exacts de probabilité mathématique »[175] et la « production de *bonnes* données de cette sorte constitue un accomplissement technologique capable d'étayer la forme perfectionnée d'inférence que de telles données rendent possible »[176].

En ce sens, comme le souligne le psychologue français Maurice Reuchlin en 1953[177], le plan de l'expérience, c'est-à-dire sa structure logique, « préside *à la fois* à l'organisation des opérations matérielles dont l'ensemble constitue l'expérience et à l'organisation des procédés statistiques permettant d'interpréter les résultats » (Reuchlin, 1953, p. 59-60)[178]. C'est en effet ce que souligne Fisher dès la première partie de l'introduction de son ouvrage *The Design of Experiments*, où il explique qu'il y a deux manières d'attaquer « une preuve expérimentale » qui est supposée prouver une « conclusion scientifique »[179] : la première est d'affirmer que « l'*interprétation* de l'expérimentation est fautive », ce qui relève du « domaine des *statistiques* », et qui d'ailleurs en général émise par de « prétendus statisticiens »[180] ; la seconde est de soutenir que « l'expérience elle-même a été mal planifiée, ou, bien sûr, mal exécutée », ce qui « renvoie à la question du *plan*, ou de la *structure logique* de l'expérimentation », critique qui est en général émise par « une *autorité* de poids »[181].

---

[174] « *It was thus clear at an early stage that there were quantitatively large technological gains to be obtained through the deliberate study of Experimental Design, and that these gains were to be harvested by making the plan of experimentation and observation logically coherent with the aims of the experiment, or, in other words with the kind of inference about the real world* », in Fisher, 1965, p. 33.

[175] « *To ensure that the data shall provide a logical foundation for inferring exact statements of mathematical probability is one of the tasks to be considered in experimental design***.** », in Fisher, 1965, p. 35.

[176] « *And that the thoroughness with which the logic and the mathematics have been explored with respect to the Gaussian sample, should not blind us to the fact that the production of good data of this kind is a technological accomplishment worthy to sustain the perfected form of inference which such data make possible.* **»**, in Fisher, 1965, p. 38.

[177] Reuchlin, Maurice, « Utilisation en psychologie de certains plans d'expérience », *L'année psychologique*, vol. 53 / 1, 1953, p. 59-81.

[178] C'est Reuchlin qui souligne.

[179] « *When any scientific conclusion is supposed to be proved on experimental evidence, critics who still refuse to accept the conclusion are accustomed to take one of two lines of attack*», in Fisher, 1935, p. 1.

[180] « *They may claim that the interpretation of the experiment is faulty (…). Such criticisms of interpretation are usually treated as falling within the domain of statistics. They are often made by professed statisticians* (…).», in Fisher, 1935, p. 1. C'est Fisher qui souligne.

[181] « *The other type of criticism to which experimental results are exposed is that the experiment itself was ill designed, or, of course, badly executed. (….) Both of these points come down to the questions of the design, or the logical structure of the experiment. This type of criticism is usually made by what I might call a heavyweight authority.* » , in Fisher, 1935, p. 2. C'est Fisher qui souligne.

Pourtant, Fisher précise immédiatement que ces deux sortes de critique reviennent d'un point de vue logique à la même chose :

> « La procédure statistique et le plan expérimental sont seulement deux aspects différents d'un même tout, et ce tout inclut l'ensemble des exigences logiques du processus général par lequel les connaissances naturelles sont accrues par l'expérimentation »[182]

Par « ajouter quelque chose à notre connaissance naturelle », Fisher fait ici référence à la notion d'induction qu'il va développer dans la deuxième partie de son introduction, intitulée « *The Mathematical Attitude towards Induction* » où il défend précisément la possibilité même de l'induction :

> « J'ai présumé, comme l'expérimentateur le présume toujours, qu'il *est* possible de tirer des inférences valides à partir des résultats d'une expérimentation ; qu'il est possible de raisonner des conséquences aux causes, des observations aux hypothèses ; ou, comme dirait un statisticien, d'un échantillon à la population d'où cet échantillon est extrait, ou, comme le dirait un logicien, du particulier au général. »[183]

Cela est possible selon Fisher non pas au sens où l'inférence serait certaine comme elle peut l'être dans la méthode hypothético-déductive, mais au sens où l'on peut exprimer rigoureusement l'incertitude aussi bien quant à la « nature » de cette incertitude que quant à son « degré »[184], grâce à la théorie des probabilités.

Dès lors, il nous semble possible de définir le plan d'expérience comme une procédure ou comme un instrument de recueil et d'analyse de données, et d'extraction d'information à partir de ces données, qui permet, à travers l'utilisation d'outils statistiques, de faire une inférence qui concerne le monde extérieur : son but est de tester une hypothèse donnée, l'hypothèse nulle, afin de la réfuter, en spécifiant à partir de l'expérience elle-même, c'est-à-dire *a posteriori* (contrairement à Neyman et Egon Pearson), le niveau de signification (entre 1 et 5%) donc le risque de se tromper (c'est-

---

[182] « *Statistical procedure and experimental design are only two different aspects of the same whole, and that whole is the logical requirements of the complete process if adding to natural knowledge by experimentation*», in Fisher, 1935, p. 3.

[183] « *I have assumed, as the experimenter always does assume, that it is possible to draw valid inferences from the results of experimentation; that it is possible to argue from consequences to causes, from observation to hypotheses; as a statistician would say, from a sample to the population from which the sample was drawn, or, as a logician might put it, from the particular to the general* », in Fisher, 1935, p. 4. C'est Fisher qui souligne.

[184] « …*for the nature and degree of the uncertainty may itself be capable of rigourous expression*», in Fisher, 1935, p. 4.

à-dire de rejeter l'hypothèse nulle alors qu'elle est correcte)[185] . Les données matérielles, à condition qu'elles soient correctes ou fiables, constituent ainsi les prémisses logiques de l'inférence causale. Néanmoins, la validité de l'inférence logique, et la valeur de l'énoncé portant sur le monde réel qu'elle permet de faire, reposent entièrement sur la structure même du plan expérimental dans son ensemble : certains plans étant meilleurs que d'autres (en termes de quantité d'information extraite, en termes de précision de l'estimation ou en termes de sensibilité du test statistique), certaines inférences sont aussi meilleures que d'autres, c'est-à-dire plus justifiées.

Néanmoins, comme nous avons pu le constater tout au long de notre étude, il est évident que les études épidémiologiques, sauf peut-être l'essai clinique randomisé, ne sont pas comparables à des expérimentations au sens strict, notamment du fait de l'impossibilité de la randomisation. Ce n'est sans doute pas un hasard si Berkson (comme Fisher plus tard), dans sa critique des études observationnelles inaugurée par son article de 1946, a d'abord attaqué le plan de l'étude et montré que la grande différence entre une expérience de laboratoire et une étude observationnelle consiste justement dans l'impossibilité constitutive de cette dernière de manipuler les variables au gré de l'expérimentateur, mais aussi les données des différentes études épidémiologiques. Ce n'est qu'ensuite qu'il a critiqué l'interprétation des résultats des différentes études qui montraient un lien entre le tabagisme et le cancer du poumon. Selon lui, en effet, il n'est pas possible de tirer une inférence valide, d'un point de vue logique, à partir de telles études, et a fortiori, impossible de tirer une inférence causale, c'est-à-dire concernant le « monde réel », de ces mêmes études. Les épidémiologistes de l'époque, comme Hill ou Lilienfeld, en ont d'ailleurs, comme nous l'avons montré précédemment[186], parfaitement conscience en montrant que les preuves obtenues par l'observation sont moins fortes que celles obtenues à travers l'expérimentation. C'est sans doute pourquoi aussi est née cette réflexion sur les critères de causalité qui a conduit les épidémiologistes à la fois à croiser les preuves obtenues par différentes méthodes d'observation, elles-mêmes effectuées par différentes personnes dans

[185] Pour une présentation des « Principes des études expérimentales dans les sciences de la vie », voir le texte éponyme d'Alain Leplège, Philippe Bizouarn, Nicolas Lechopier, Sabine Plaud, Etienne Brun-Rovet, Jean-Louis Auget, Laure Cartron, Jean-Jacques Szczeciniarz, in Leplège, Alain, Bizouarn, Philippe et Coste, Joël, *De Galton à Rothman: les grands textes de l'épidémiologie au XXe siècle*, Paris, Hermann, 2011, p. 25-32

[186] Voir *supra*, Section 4.1.2.

différents endroits à différentes époques, avec des données non-épidémiologiques comme par exemple la plausibilité biologique ou la cohérence avec les connaissances biologiques ou cliniques de la maladie.

Pourtant, si le concept de biais est dès son origine chez Fisher étroitement lié au concept de plan d'expérience, notamment dans la section 27 de *The Design of Experiments*, intitulée « *Bias of Systematic Arrangements* »[187], la conceptualisation proprement dite du biais en relation avec le plan d'expérience, à travers les notions de validité interne et de validité externe, et la conceptualisation de la notion de hiérarchie des plans d'expérience en relation avec leur susceptibilité aux biais, va provenir non de l'épidémiologie elle-même, mais des sciences sociales en général, notamment la psychologie, la sociologie ou encore les sciences de l'éducation.

4.3.2 Plans expérimentaux et plans observationnels en sciences sociales et en épidémiologie.

Le texte de Donald T. Campbell (1916-1996) et Julian C. Stanley (1918-2005), intitulé *Plans expérimentaux et quasi-expérimentaux pour la recherche*[188], bien qu'assez peu connu en France, « fait partie des textes les plus cités dans le domaine des sciences sociales » (Leplège, Bizouarn et Coste, 2011, p. 31) et a été réédité de nombreuses fois. Le sujet qui nous intéresse dans cet ouvrage est bien évidemment la place du concept de biais et sa relation avec la notion de plan d'expérience. Or, le premier point important dans ce texte est que les auteurs précisent d'emblée que « les plans d'expérience dans la tradition de Fisher ne sont pas l'objet de ce chapitre » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 73). En effet, selon eux, les plans fishériens posent deux problèmes : le premier est que dans ces plans, « l'expérimentateur peut, parce qu'il a la maîtrise complète, programmer les traitements et les mesures en vue d'une efficacité statistique optimale » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 73), alors que dans les plans

[187] Fisher,1935, p. 71-72. Voir aussi la section 1.3.2 du présent travail.

[188] Campbell, Donald Thomas et Stanley, Julian Cecil, *Experimental and quasi-experimental designs for research*, Boston, Houghton Mifflin Comp, 1967. A l'origine, ce texte constitue un chapitre de Gage, N.L (éd.), *Handbook of Research on Teaching*, Chicago, Rand McNally & Company, 1963.Nous nous appuierons, sauf indication contraire, sur la traduction française effectuée par Nicolas Lechopier et Laure Cartron dans Leplège, Bizouarn et Coste 2011, p. 73-100. En conséquence nous nous référerons, sauf indication contraire, à la pagination de la traduction.

étudiés par Campbell et Stanley, la raison de leur complexité est à chercher dans « l'intransigeance de l'environnement », autrement dit dans « l'absence pour l'expérimentateur d'une complète maîtrise » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 73) des variables. Le second problème renvoie à la volonté de Campbell et Stanley de renverser la perspective fishérienne : selon eux la tradition fishérienne insiste plutôt sur l'amélioration de l'analyse statistique des données expérimentales, tandis que celle dont ils s'inspirent[189] se concentre plutôt sur « les méthodes permettant d'obtenir avec certitude de bonnes données – des données adéquates – auxquelles appliquer ces procédures statistiques »[190]. Le rapport entre ces deux critiques des plans fishériens est donc simple : l'attention portée aux données, plutôt qu'à leur manipulation statistique, est directement liée à l'objectif poursuivi par Campbell et Stanley qui est de présenter des plans qui ne sont pas nécessairement expérimentaux (où l'expérimentateur a un contrôle total et où il peut utiliser la randomisation), mais qui sont, comme ils les appellent, « quasi-expérimentaux », comme par exemple « les expériences non randomisées de rotation »(Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 75), ou encore « pré-expérimentaux[191] ».

La principale avancée méthodologique, et donc le principal intérêt historique, consiste ainsi selon nous dans cette relativisation des plans d'expérience qui est explicitement assumée par Campbell et Stanley, dans leur présentation de 16 plans d'expérience (dont 3 sont pré-expérimentaux, 10 quasi-expérimentaux, et 3 réellement expérimentaux), et de la « *check list* » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 71) des facteurs de validité qui accompagne chaque plan, et qui permet d'établir une hiérarchie parmi les plans, non pas au sens il s'agirait de « substituer, au dogme du seul ou des deux plans expérimentaux acceptables, un autre dogme des treize plans acceptables », mais au sens où il s'agit « d'encourager l'ouverture d'esprit et une attitude prospective pour de nouvelles organisations de recueils de données » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 96). Plus précisément , il nous semble que c'est dans la notion de validité du plan d'expérience, et plus encore dans la distinction entre validité interne et validité

[189] Campbell et Stanley s'appuient sur l'ouvrage de Willam A. McCall, *How to experiment in education*, publié en 1923.
[190] La citation est de McCall et est extraite de la préface de son livre.
[191] Le préfixe « pré » semble ici être employé au sens logique et non chronologique du terme.

externe du plan d'expérience que se situe le progrès méthodologique crucial effectué par Campbell et Stanley[192] : quels sont alors ces facteurs qui menacent ou mettent en péril (« *jeopardizing* »[193]) la validité des plans ? Et que signifie cette notion de validité ? Enfin, comment distinguer la validité interne de la validité externe ?

Selon Campbell, la notion de validité du plan d'expérience est à mettre en rapport avec la notion de « variables exogènes »[194] : en effet, la fonction principale d'une expérimentation est d'exclure (« *rule out* ») les variables exogènes qui pourraient nous conduire à une confusion dans notre interprétation. En ce sens, plus le plan d'expérience est proche du plan expérimental, plus les variables exogènes de confusion sont contrôlées ou plutôt éliminées dans l'interprétation. A l'inverse, plus on s'éloigne de l'expérimentation, plus le risque de confusion est élevé. Stanley et Campbell dénombrent en tout « douze facteurs qui remettent en cause la validité de différents plans expérimentaux » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 76), c'est-à-dire douze catégories de variables exogènes. Ainsi, la validité d'un plan d'expérience peut être évaluée selon deux critères : le critère de la validité interne, et celui de la validité externe. La validité interne est selon Campbell et Stanley « le minimum exigé pour qu'une expérience soit interprétable : les traitements expérimentaux font-ils une différence dans ce cas expérimental spécifique ? ». Quant à la validité externe, elle « pose la question de la généralisabilité : à quels populations, paramètres, variables dépendantes, mesures, l'effet observé peut-il être généralisé ? ». En d'autres termes, il s'agit de savoir si l'échantillon étudié est ou non représentatif de l'univers dont il issu (les auteurs assimilent plus loin validité externe et « *representativeness* »[195]). Campbell et Stanley vont alors distinguer huit menaces à la validité interne, et quatre menaces à la validité externe, menaces étant ici synonymes de variables exogènes non contrôlées dans le plan d'expérience qui « risquent de produire des effets qui seront confondus avec les effets du stimulus

[192] Pour être tout à fait exact, et comme cela est explicitement précisé par les auteurs (voir la note 3 dans Campbell et Stanley,1967, p. 5), c'est Campbell qui a défini ces facteurs. Voir Campbell, D. T., « Factors relevant to the validity of experiments in social settings », *Psychological Bulletin*, vol. 54 / 4, juillet 1957, p. 297-312.

[193] Campbell et Stanley, 1967, p. 5.

[194] « *extraneous variables* », in Campbell,1957, p. 297.

[195] Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 77.

expérimental » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 76). Les voici sous forme d'un tableau[196] :

Figure 4-1 : Tableau récapitulatif des principales menaces à la validité du plan d'expérience selon Campbell et Stanley

MENACES A LA VALIDITE INTERNE	MENACES A LA VALIDITE EXTERNE
Histoire	Effet réactif ou effet d'interaction du test
Maturation	Interactions entre biais de sélection et variable expérimentale
Test	Effet réactifs du dispositif expérimental
Instrumentation	Interférence entre des traitements multiples dies à des effets de traitement antérieurs
Régression vers la moyenne	
Biais de sélection	
Mortalité expérimentale	
Interaction de la maturation et de la sélection	

A première vue, on pourrait croire que la notion de biais se résume au biais de sélection, c'est-à-dire au phénomène selon lequel « les membres des groupes de comparaison sont sélectionnés de façon différentielle » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 76). Pourtant, la notion de biais est assimilée dans un autre article de Campbell[197] à celle d'artefact (« *artifact* »), les deux constituant une menace à la « validité de l'inférence » (Campbell, 1969, p. 269). Selon nous, cela signifie que le biais est étroitement lié à la question de la validité interne : en effet, les « variables exogènes » que le plan d'expérience a pour fonction de

[196] Pour une présentation plus détaillée, voir Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 76-77.

[197] Campbell, Donald. T., « Prospective: Artifact and Control», in Rosenthal, Robert et Rosnow, Ralph L., *Artifacts in Behavioral Research*, New York, Oxford University Press, 2009, p. 264-286. La première édition date de 1969 : nous abrègerons donc cet article en Campbell,1969.

contrôler risquent de produire une association artificielle entre les variables qui sont étudiées, donc de produire un artefact ou un biais. Cela rejoint d'ailleurs, et confirme, la distinction que nous avons établie auparavant entre la question de l'association et la question de la causalité : selon nous, la question de l'association renvoie à la question de la validité interne, tandis que celle de la validité externe renvoie à la question de la causalité. De même, comme le soulignent Campbell et Stanley, la validité interne « est une condition *sine qua non* » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 76) à la validité externe donc à la généralisation, tout comme la validité de l'association établie par une ou plusieurs études épidémiologiques (cas-témoin ou cohorte) précède nécessairement dans l'ordre logique l'assertion d'un lien de causalité. Ainsi, quand Stanley et Campbell soulignent que « les facteurs menaçant la validité interne sont ceux qui affectent directement la mesure *O* » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 78)[198], cela n'est pas sans rappeler les mesures de risque, absolu et relatif, que les critiques d'une relation causale entre le tabagisme et le cancer du poumon estimaient être biaisées, c'est-à-dire produites artificiellement. Dès lors, il est possible de reformuler les deux principales critiques adressées à cette relation causale entre tabagisme et cancer du poumon, en les reformulant dans le langage de Campbell et Stanley :

- La première critique porte sur la validité interne : l'association entre tabagisme et cancer du poumon est le produit artificiel d'un ou de plusieurs biais, comme par exemple un biais de sélection différentielle à l'admission (le biais de Berkson concernant la clientèle hospitalière par exemple), qui montrerait une association alors qu'il n'y en a pas, cette association étant le produit de l'étude ou du plan d'expérience et donc un artifice statistique (d'où le sophisme de la réification énoncé par Berkson). En d'autres termes, l'association que semble montrer l'expérience ou l'étude épidémiologique est fausse, au sens où l'expérience n'est pas interprétable, au sens où l'on ne peut pas savoir si cette étude-là montre ce qu'elle prétend montrer sur cette population-là (c'est-à-dire par exemple le groupe des cas et le groupe des témoins). Dès lors, à partir du moment où il n'y pas de validité interne, il

[198]. Dans la codification des plans d'expérience effectuée par Campbell et Stanley, « un X représente l'exposition d'un groupe à une variable expérimentale ou un évènement, dont on veut mesurer les effets », tandis qu'un « O désigne un processus de mesure ou d'observation », in Leplège, Bizouarn et Coste, 2011, p. 77.

ne peut y avoir de validité externe ; ou, autrement dit, à partir du moment où l'association est fallacieuse (« *spurious* », selon l'adjectif consacré), aucun lien de causalité ne peut être établi.

- La seconde critique porte sur la validité externe : en effet, l'argument constitutionnel, avancé par exemple par Berkson ou Fisher, consiste finalement à soutenir que l'association exhibée dans les études épidémiologiques n'est pas généralisable. Ainsi, dire qu'il y a une certaine constitution physique, génétique, ou psychologique qui prédispose à la fois au tabagisme et au cancer revient finalement à dire que la relation entre tabagisme et cancer du poumon ne vaut que pour cette population qui a cette constitution précise. En d'autres termes, si l'hypothèse constitutionnelle est vraie, ceux qui ne sont pas membres de cette population qui a cette constitution peuvent fumer (car après tout, fumer reste un choix, contrairement au fait d'avoir un cancer) sans prendre le risque de développer un cancer. De même, ceux qui ont cette constitution peuvent ne pas fumer et développer un cancer quoi qu'il arrive.

Si cette distinction est appelée à devenir fondamentale en épidémiologie, comme l'atteste sa présence dans les ouvrages classiques de l'épidémiologie moderne qui paraitront à partir des années 1980[199], elle joue aussi un rôle fondamental dans le développement d'une nouvelle épistémologie qui est en train de se mettre en place à cette époque, sous l'égide notamment de Donald Campbell, et qui concerne essentiellement les sciences sociales mais aussi l'épidémiologie dans la mesure où ces sciences utilisent des plans expérimentaux, ou quasi ou pré-expérimentaux, qui ont pour fonction d'étudier des populations humaines et dont la méthodologie est fondée essentiellement sur la comparaison et le contraste entre deux groupes, comme dans une étude cas-témoin. Cette nouvelle épistémologie, explicitement qualifiée comme une « épistémologie évolutionniste de la connaissance »[200], nous apparait

[199] Voir par exemple la préface de l'ouvrage de Kleinbaum, Kupper et Morgenstern, où les auteurs soulignent que « la validité doit être le but premier d'une étude épidémiologique, même si cela signifie sacrifier la généralisabilité », in Kleinbaum, David G., Kupper, Lawrence L. et Morgenstern, Hal, *Epidemiologic research. Principles and quantitative methods*, Belmont, CA, Lifetime Learning Publications., 1982, p. XVII. Voir aussi la critique qu'en font Rothman et Greenland dans Rothman, Kenneth J. et Greenland, Sander, *Modern Epidemiology*, 2ème, Philadelphia, PA, Lippincott Williams & Wilkins, 199, spécifiquement le chapitre 8, intitulé « Précision et validité dans les études épidémiologiques », p. 115-134.

[200] Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 76.

comme étant le double produit de la révolution darwinienne et de la révolution statistique. Or, dans cette épistémologie nouvelle, le concept de biais est amené à jouer un rôle important, et c'est sans conteste Campbell qui va le plus approfondir cette notion et en faire un concept scientifique important, notamment dans son rapport avec le problème de l'induction, ce qui ne sera pas sans répercussions sur l'épidémiologie et la médecine. Plus précisément, nous dirons que Campbell va expliciter une épistémologie qui est ou qui était sous-jacente à l'épidémiologie, épistémologie qui est propre à une discipline essentiellement inférentielle, et rarement, voire jamais, hypothético-déductive, et qui expliquerait pourquoi il a été si difficile pour de nombreux épidémiologistes de faire reconnaitre le lien entre tabagisme et cancer du poumon, et par là même de justifier le caractère scientifique de leur discipline. Cette épistémologie est à rattacher à celle de Popper (comme Campbell le fait d'ailleurs explicitement[201] en soulignant que son épistémologie est compatible avec l'épistémologie réfutationniste ou faillibiliste de Popper), épistémologie poppérienne qui semble constituer aujourd'hui le cadre philosophique de l'épidémiologie[202].

4.3.3 Vers une épistémologie du concept de biais ?

Le principe même de l'épistémologie évolutionniste de Donald Campbell consiste à considérer l'expérimentation comme un prolongement de l'évolution. En effet, selon lui :

> « Les applications pratiques et la connaissance scientifique sont considérées comme le résultat de l'accumulation d'une série d'essais laissant de côté de multiples autres tentatives écartées par l'expérience » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 75).

En ce sens, la tradition, c'est-à-dire les pratiques ou les croyances que nous ont léguées nos ancêtres, constitue sans doute « parmi toutes les pratiques possibles, un sous-ensemble valide et bien établi ». Le problème des pratiques traditionnelles, qu'il s'agisse des pratiques pédagogiques dont parlent Campbell et Stanley ou des

[201] Campbell et Stanley, 1967, p. 35.

[202] A notre connaissance, le premier article écrit par un épidémiologiste pour convaincre ses collèges de la pertinence de l'épistémologie de Popper date de 1975 et a pour auteur Carol Buck : Buck, Carol, « Popper's philosophy for epidemiologists », *International Journal of Epidemiology*, vol. 4 / 3, 1975, p. 159–168.

pratiques médicales, est qu'elles n'ont pas été évaluées correctement, que la sélection des pratiques s'est sûrement fait « par pur hasard ». Dès lors,

> « L'expérimentation arrive à ce point pour affuter la pertinence du processus sélectif, celle des tests et des essais » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 75).

Plus précisément, l'expérimentation s'inscrit dans le processus même de la science :

> « La science, tout comme les autres processus cognitifs, implique la proposition de théories, d'hypothèses, de modèles, etc. dont l'acceptation ou le rejet se fait sur la base d'un critère externe. L'expérimentation appartient à cette seconde phase, celle de l'élagage, du rejet ou de la correction » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 84).

En ce sens, d'après les auteurs, il y a une analogie entre le processus de la cognition au plan individuel, le processus de développement de la science au plan collectif, mais aussi le processus de l'expérimentation, qui semble constituer un moyen terme entre le plan individuel et le plan collectif de la connaissance. Ainsi une théorie scientifique n'est pas confirmée au sens où des faits viendraient corroborer l'hypothèse de départ dans une sorte d'accumulation de preuves, mais à l'inverse au fur et à mesure que « le nombre d'*hypothèses rivales plausibles* disponibles rendant raison des données » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 85[203]) est progressivement réduit :

> « Moins il reste d'hypothèse rivales, plus grand est le degré de « confirmation » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 85).

De la même manière, la fonction épistémologique d'une expérience ou d'un plan d'expérience vise essentiellement à infirmer, plutôt qu'à confirmer l'hypothèse de départ, c'est-à-dire l'hypothèse nulle :

> « *La tâche consistant à réunir des données pour tester des théories est par conséquent principalement celle du rejet des hypothèses inadéquates* » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 84[204]).

[203] Ce sont Stanley et Campbell qui soulignent.
[204] Ce sont Stanley et Campbell qui soulignent.

Dès lors, c'est la notion même de preuve, en son sens classique ou mathématique, qu'il faut rejeter, comme le préconisaient d'ailleurs des épidémiologistes comme Hill ou Cornfield[205] :

> « Les résultats expérimentaux ne confirment jamais ni ne prouvent une telle théorie – en fait, la bonne théorie est soumise à des tests et échappe à l'infirmation. Le mot « prouver » a pris pour notre génération une connotation qui ne correspond ni à ses anciens usages ni à des applications à des procédures inductives telles que l'expérimentation. Les résultats d'une expérimentation renforcent mais ne prouvent pas une théorie. Une hypothèse adéquate est une hypothèse qui a survécu à de nombreuses mises à l'épreuve – mais qui peut toujours être récusée lors d'une nouvelle épreuve. » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 84).

La notion de biais acquiert alors un statut particulier au sein de cette épistémologie qu'on pourrait qualifier de réfutationniste. En effet, si la notion de biais est assimilée à celle de menaces à la validité d'un plan d'expérience, et si ces menaces sont en réalité les variables exogènes que le plan d'expérience a pour fonction de contrôler, alors la notion de biais peut être rendue équivalente à celle de variables exogènes. En ce sens, éliminer les hypothèses rivales revient à éliminer les biais :

> « Dans cette perspective, la liste des sources d'invalidité que les plans expérimentaux cherchent à contrôler peut être vue comme une liste d'hypothèses plausibles fréquentes rivales de l'hypothèse selon laquelle c'est la variable expérimentale qui a effectivement produit tel effet. En « contrôlant » un ou plusieurs de ces facteurs, un plan expérimental rend tout simplement cette hypothèse rivale non plausible, même si elle pourrait, à travers un ensemble de coïncidences complexes, continuer à opérer et produire le résultat observé. (…) Dans une quasi-expérimentation, là où les contrôles manquent, lorsqu'on interprète les résultats, il faut évaluer précisément la vraisemblance de facteurs non-contrôlés qui pourraient rendre compte également de ces résultats. Plus l'absence de ces facteurs est plausible, plus l'expérimentation est « valide ». » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 86).

[205] Voir supra la section 4.1.2 consacrée à la question de la preuve en épidémiologie.

Ainsi, dans les plans d'expérience comme dans la science en général, il s'agit toujours de ce qui est appelé par les épistémologues contemporains une « inférence à la meilleure explication », selon la terminologie employée par Gilbert Harman[206] à la même époque, et qui correspond selon lui « approximativement à ce que d'autres ont appelé « abduction », « la méthode des hypothèses », « l'inférence hypothétique », « la méthode d'élimination », « l'induction éliminatoire » et « l'inférence théorique »[207] et qu'il explique de la façon suivante :

> « En faisant cette inférence, on infère, en partant du fait qu'une certaine hypothèse expliquerait la preuve [« *the evidence* »], à la vérité de cette hypothèse. En général, il y aura de nombreuses hypothèses qui pourraient expliquer la preuve, on doit alors être capable de rejeter de telles hypothèses alternatives avant de pouvoir faire légitimement cette inférence »[208]

Comme on peut le constater, c'est bien une nouvelle épistémologie qui semble se mettre en place dans les années 1960, épistémologie qui dépasse le simple cadre de l'épidémiologie mais qui pourrait parfaitement s'y appliquer tant les débats méthodologiques et proprement épistémologiques qui ont lieu parmi les épidémiologistes à la fin des années 1950 et au début des années 1960 rejoignent finalement des débats épistémologiques qui traversent toutes les sciences, et en particulier celles qui font appel à des plans d'expérience qui ne sont pas expérimentaux et qui portent sur des êtres humains. Dès lors, il est logique que dans ces sciences observationnelles, la question du plan de l'étude ou du plan d'expérience soit cruciale car c'est précisément ce plan de l'étude qui permet d'éliminer progressivement les hypothèses alternatives comme par exemple l'hypothèse constitutionnelle dans le cas de la controverse sur le lien entre tabagisme et cancer du poumon. En effet comme le soulignent Campbell et Stanley :

> « La « validité » d'une expérimentation devient une question de crédibilité relative des théories rivales, d'un côté la théorie selon laquelle X a un effet, de

[206] Harman, Gilbert H., « The Inference to the Best Explanation », *The Philosophical Review*, vol. 74 / 1, janvier 1965, p. 88.

[207] « *"The inference to the best explanation" corresponds approximately to what others have called "abduction," the method of hypothesis," "hypothetic inference," "the method of elimination," "eliminative induction," and "theoretical inference."»,* in Harman, 1965, p. 88-89.

[208] « *In making this inference one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there will be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making the inference*», in Harman, 1965, p. 89.

> l'autre la théorie selon laquelle cette causalité implique des facteurs non contrôlés. L'existence de l'effet de X est d'autant plus défendable quand plusieurs ensembles de différences peuvent tous être expliqués par l'hypothèse selon laquelle X a un effet, et qu'en même temps, on suppose différentes variables non contrôlées pour rendre compte de chaque différence. » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 86).

C'est précisément grâce à ce genre de considérations, en montrant que l'hypothèse la plus plausible et la plus simple (Stanley et Campbell évoquent le principe de parcimonie comme « une supposition générale à propos de la nature du monde » qui « sous-tend presque toute théorie en science »[209]) pour expliquer l'augmentation du cancer du poumon est l'augmentation de la consommation de tabac, que les épidémiologistes vont finalement convaincre les autorités scientifiques et sanitaires qu'il y a une relation de causalité entre le tabagisme et le cancer du poumon. De même, pour reprendre la distinction entre validité interne et validité externe, les épidémiologistes ont pu s'appuyer sur la multiplicité des enquêtes menées dans différents pays, à différentes époques, selon différents plans d'expérience (à titre d'illustration, le Comité qui a rendu son rapport sur le tabagisme et la santé au *Surgeon General* a étudié plus de 7000 articles consacrés à ce sujet, et un des critères de sélection des dix membres qui composaient ce Comité était de ne pas avoir pris position sur le sujet auparavant, pour éviter les biais subjectifs et donc garantir l'objectivité de leur avis scientifique) pour montrer que le lien entre le tabagisme et le cancer du poumon (ainsi que de nombreuses autres maladies comme les maladies cardio-vasculaires) n'était pas propre à telle ou telle étude, et donc potentiellement artificiel, mais était parfaitement généralisable, d'autant plus généralisable que les études de cohorte menées sur des dizaines de milliers de personnes démontraient la même relation de causalité. En effet, comme le soulignent Campbell et Stanley :

> « Alors que les questions de validité interne peuvent être résolues à l'intérieur du cadre des statistiques probabilistes, les problèmes de validité externe ne peuvent pas être logiquement résolus de façon nette et définitive. (…) Du point de vue logique, nous ne pouvons pas généraliser au-delà de ces limites. Autrement dit, nous ne pouvons faire aucune généralisation. En réalité, cependant, nous tentons des généralisations en présupposant certaines lois, et

[209] Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 86.

> en vérifiant que ces généralisations sont valables dans d'autres conditions, spécifiques elles aussi mais différentes. » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 80).

En somme, la répétition de la mise à l'épreuve d'une même hypothèse selon des plans d'expérience et dans des contextes différents, permet justement de dépasser, sans la négliger, la question de la représentativité de l'échantillon (et nous avons vu à quel point les épidémiologistes se sont montrés dès le départ soucieux d'éviter le biais de sélection, et aussi à quel point cette question de la représentativité de l'échantillon a constitué un point central de la critique de Berkson ou d'autres statisticiens et épidémiologistes) pour aller vers une sorte de réplicabilité des expériences. Campbell et Stanley maintiennent d'ailleurs, comme Hill ou Cornfield, la supériorité épistémologique de l'expérimentation proprement dite sur la quasi-expérimentation ou sur l'observation :

> « Les sciences qui sont couronnées de « succès », telles que la physique ou la chimie, ont réussi sans porter la moindre attention à la représentativité (mais en insistant sur la répétabilité des expériences par des chercheurs indépendants). Une science de laboratoire, telle une tour d'ivoire, est un bel aboutissement même si elle n'est pas représentative, pourrait bien souvent être essentielle pour analyser les différentes variables – ce qui est fondamental pour obtenir des résultats dans de nombreux domaines. » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 81).

Néanmoins, cela ne doit pas conduire à « étouffer la volonté de construire des plans expérimentaux » car « toute expérience est imparfaite du point de vue de son interprétation finale et de la tentative de l'inscrire dans le développement scientifique » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 83). En ce sens,

> « Le rôle d'une *check-list* de critères de validité est de rendre l'expérimentateur plus attentif aux imperfections résiduelles au sein de son plan expérimental, afin qu'il soit conscient des points sur lesquels d'autres interprétations de ses données pourraient être réalisées » (Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 83).

Ainsi, et pour conclure sur ce quatrième chapitre, il apparait que le concept de biais soit étroitement lié à la notion de plan d'expérience, et plus spécifiquement à la validité de ce plan d'expérience. En ce sens, le biais n'est pas uniquement une

propriété de la mesure comme par exemple du risque relatif, cette mesure étant entendue comme le résultat de l'expérience ; ou une propriété de l'échantillon (la question de la représentativité) ou de la comparabilité des échantillons entre eux comme dans une étude cas-témoins ; ou une propriété des individus qui participent à cette étude ou l'organisent (qu'il s'agisse de l'enquêteur, de l'intervieweur, de l'interviewé, du médecin qui établit par exemple un diagnostic à partir d'une radiographie, ou du patient lui-même, tous étant sujet à des processus de suggestion qui pourraient modifier leur réponse, ou leur diagnostic ou leur interprétation) ; il est une propriété du plan d'expérience dans son ensemble. Ou plutôt, tous ces biais peuvent s'additionner ou/et interagir pour former un biais qui constitue finalement l'ensemble des explications alternatives à l'explication en termes de causalité, que cet agent causal soit une exposition à un risque ou bien un traitement. En ce sens, Donald Mainland est sans doute celui qui a le mieux perçu les enjeux de ce concept à travers la tripartition que nous avons vue à la fin du chapitre précédent : cause, hasard, biais.

La fonction opératoire du concept de biais apparait ainsi à ce stade de notre enquête beaucoup plus claire, et est étroitement liée à l'introduction des méthodes statistiques et du calcul des probabilités en médecine, comme d'ailleurs dans d'autres disciplines, et spécifiquement liée à l'apparition, suite aux travaux de Fisher auxquels il faudrait ajouter selon Stanley et Campbell, McCall et d'autres, de la notion et même de cette nouvelle discipline à l'intérieur des statistiques que sont les plans d'expérience. Le concept de biais, conçu comme erreur systématique, ne peut en effet apparaitre qu'à partir du moment où est apparue la notion d'erreur aléatoire, c'est-à-dire l'idée, centrale dans les probabilités à travers la courbe des erreurs et la loi des grands nombres, que les erreurs ou les variations, sur le long terme, s'annulent et convergent vers un point central qui constitue leur espérance. Or, certaines erreurs, comme l'avait aperçu Thomas Bayes, ne s'annulent pas mais s'accumulent et font varier la valeur mesurée vers une certaine direction qui s'écarte de la vraie valeur, en raison d'une sorte d'artifice statistique produit par le plan d'expérience qui conduit à une interprétation ou à une inférence erronée, la validité de l'inférence reposant en dernier ressort sur la validité du plan d'expérience, cette même validité étant liée à l'absence ou à la minimisation des biais.

La spécificité du concept de biais dans l'épidémiologie tient alors essentiellement à son objet d'étude qui inclut au final beaucoup plus de possibilités de biais qu'une

étude psychologique ou sociologique, mais aussi à la manière d'aborder cet objet qui peut se faire à travers un plan d'expérience très proche du plan expérimental comme peut l'être l'essai clinique randomisé, ou beaucoup plus proche du plan observationnel comme le sont les études cas-témoins, les études de cohorte étant à mi-chemin entre les deux, puisqu'il s'agit non pas de remonter des effets aux causes, mais de partir des causes (ou en tout cas des facteurs de risque, entendus comme des causes probables) pour aller vers les effets, comme dans un cadre expérimental, mais sans la possibilité d'un contrôle total sur les variables. En ce sens, toute la difficulté des épidémiologistes, et de leur outil essentiel qu'est le plan d'expérience, consiste à pouvoir éliminer ou au moins à réduire autant que possible les hypothèses rivales qui permettraient d'expliquer l'apparition d'une maladie, c'est-à-dire au final d'éliminer ou de réduire tous les biais. C'est pourquoi Kaptchuk a raison de dire que le « biais hantait la médecine »[210] à partir des années 1950, à condition de considérer que ce fantôme qu'est le biais ne renvoie pas seulement au problème de la suggestion, de l'effet placebo, et plus globalement de la subjectivité (du médecin, du patient, de l'enquêteur, etc.), donc pas seulement au biais subjectif mais aussi au biais objectif qui est lui étroitement lié au plan d'expérience, donc à la méthodologie mêmes des études épidémiologiques, et qui consiste à commettre l'erreur d'attribuer un pouvoir causal à une variable exogène donnée en raison d'un artifice produit par cet instrument qu'est le plan d'expérience. En d'autres termes, le biais pose la question de la validité des études épidémiologiques, et donc de la scientificité de l'épidémiologie. Et c'est précisément en intégrant ce risque de biais dans leur théorie que les épidémiologistes ont fait progresser scientifiquement leur discipline, tout en définissant ou redéfinissant leurs concepts-clés comme ceux de risque relatif, de facteurs de risque, ou de causalité, mais aussi une épistémologie propre à leur discipline, où le concept de biais joue désormais un rôle important, du fait même, c'est du moins notre thèse, que c'est au sein de l'épidémiologie que ces deux notions de biais, statistique et psychologique, objective et subjective, fusionnent pour créer un concept proprement scientifique au même titre que celui de cause ou celui de hasard.

[210] « *Bias now haunted medicine* », in Kaptchuk, 1998, p. 430.

CHAPITRE 5 : LA REDEFINITION DU CONCEPT DE BIAIS COMME ECART PAR RAPPORT A LA VERITE (1966- 1979).

Pour qui s'intéresse à l'histoire du concept de biais au sein de l'épidémiologie, il est un article qui apparait comme une référence incontournable et dont on peut dire que la définition du concept de biais qui y apparait fait date dans l'histoire, au sens où il y un avant et un après. Ainsi, lorsque le *Dictionary of Epidemiology,* édité pour l' « *International Epidemiological Association* » par John Last, définit le concept de biais comme une « déviation des résultats ou des inférences par rapport à la vérité, ou les processus conduisant à une telle déviation »[1], c'est à cet article qu'il renvoie pour avoir une description des différentes variétés de biais. L'article en question est celui de David Sackett (1934-2015), intitulé « Bias in Analytic Research », publié en 1979[2], cité depuis sa parution plus de 2100 fois, d'après *Google Scholar*[3], où il dresse un « catalogue » de trente-cinq biais. Ce qui est moins connu[4], ou en tout cas moins étudié, est que la définition de la notion de biais présentée par Sackett est empruntée (et modifiée) à un autre ouvrage écrit par Edmond A. Murphy (1925-2009) en 1976 et intitulé *The Logic of Medicine*[5]. Or, selon nous, la définition du concept de biais telle qu'elle apparait sous la plume de Murphy, et la place que celui-ci lui accorde dans ce qu'il appelle lui-même une « théorie de la médecine » (Murphy,1976, p. 6), non seulement dépasse largement, d'un point de vue intensionnel et surtout extensionnel, la définition qui a cours jusque-là au sein de l'épidémiologie et des sciences en général, mais va aussi profondément modifier le devenir de son concept au sein de l'épidémiologie en l'inscrivant plus globalement dans le champ de la médecine. Ce

[1] « *Deviation of results or inferences from the truth, or processes leading to such deviation*. », in Last, John M., *A dictionary of epidemiology*, 2ème Edition, New York, Oxford University Press, 1988, p. 13.

[2] Sackett, David L., « Bias in Analytic Research », *Journal of Chronic Diseases*, vol. 32 / 1-2, 1979, p. 51-63.

[3] https://scholar.google.com/scholar_lookup?title=Bias+in+Analytic+Research.&author=Sackett&publication_year=1979. Accédé le 12 août 2017.

[4] Paolo Vineis, dans son article consacré à l'histoire de la notion de biais, souligne que les années 1970 voient apparaitre les premières taxinomies du biais chez Murphy en 1976 (la taxinomie de Murphy telle qu'elle est donnée par Vineis ne correspond d'ailleurs pas à celle donnée par Murphy lui-même : ainsi les quatre premières catégories données par Vineis ne sont pas des catégories mais des exemples d'une même catégorie qui est celle du « *bias of design* ». Il oublie aussi le « *bias of observation* ») puis chez Sackett en 1979. Mais Vineis ne parle pas des définitions données par Murphy ou Sackett. Voir Vineis, Paolo, « History of bias », *Sozial-und Präventivmedizin*, vol. 47 / 3, 2002, p. 156–161; repris dans Morabia, Alfredo, *A History of Epidemiologic Methods and Concepts*, Boston, MA, Birkhaüser Verlag, 2004, p. 327-336.

[5] Murphy, Edmond M., *The Logic of Medicine*, Baltimore, MD, Johns Hopkins University Press, 1976. Tous les extraits de cet ouvrage cites dans ce travail sont traduits par nos soins.

changement de perspective est concomitant à la naissance de l'épidémiologie clinique, dont Sackett, comme Alvan Feinstein (1925-2001), peuvent être considérés comme les fondateurs, et qui consiste, selon la définition donnée par Sackett en 1969, dans « l'application, par un médecin qui fournit des soins directs à un patient, de méthodes épidémiologiques et biométriques à l'étude des processus thérapeutiques et diagnostiques dans le but de produire une amélioration de la santé du patient »[6].

Dès lors, la perspective se déplace de l'épidémiologiste vers le médecin clinicien et de la population vers l'individu : il s'agit moins de savoir si l'étude épidémiologique est valide quant à son plan ou quant à ses résultats, ou si l'échantillon est suffisamment représentatif de la population pour pouvoir généraliser à la population entière, que de savoir si ces résultats des études épidémiologiques peuvent s'appliquer au patient individuel que le médecin doit diagnostiquer et traiter. Ainsi, on pourrait dire que la perspective ou le point de vue qui est adopté ici est celui du médecin qui se trouve face à un ou plusieurs articles, relatant des études épidémiologiques observationnelles ou des essais cliniques, et qui doit être capable d'en faire une analyse critique. Il est alors difficile de ne pas penser à l'*Evidence-Based Medicine*, terme inventé par Gordon Guyatt dans les années 1990 et dont Dave Sackett est considéré là aussi comme un des fondateurs, conçue d'abord comme une « nouvelle approche de l'enseignement de la pratique médicale »[7], et qui fait suite, selon Sackett, à ce qu'il appelle « l'évaluation critique » (« *critical appraisal* ») de la littérature médicale qu'il a commencé à enseigner aux étudiants en médecine de l'Université McMaster, au Canada, à partir de 1978[8]. Il est d'ailleurs intéressant de noter à ce sujet que le pénultième chapitre de l'ouvrage de Murphy est consacré à « un exercice de critique qualitative » qui consiste à faire une évaluation critique d'un article fictif. En somme, il s'agit dans un premier temps d'étudier comment le concept de biais, dont on a vu l'importance dans le devenir de l'épidémiologie comme discipline

[6] « *the application, by a physician who provides direct patient care, of epidemiologic and biometric methods to the study of diagnostic and therapeutic process in order to effect an improvement in health*. », in Sackett, David. L., « Clinical epidemiology », *American Journal of Epidemiology*, vol. 89 / 2, février 1969, p. 125-128.

[7] Selon le titre même de l'article fondateur de l'*Evidence-Based Medicine*: Guyatt, Gordon, Cairns, John, Churchill, David [*et al.*], « Evidence-based medicine: a new approach to teaching the practice of medicine », *Journal of the American Medical Association,* vol. 268 / 17, 1992, p. 2420–2425

[8] Voir l'interview qu'il a donnée peu avant sa mort, et qu'il a quasi intégralement rédigée : Haynes, R. Brian et Sackett, David L., « An interview with David Sackett, 2014–2015 », 2015, p. 0-104. Voir la page 41 pour le passage du l'évaluation critique.

scientifique à part entière, va acquérir une place importante dans la tentative de fonder la médecine comme science.

Pour autant, il ne saurait être question de quitter le domaine de l'épidémiologie : si l'ouvrage de Murphy n'est pas consacré à l'épidémiologie mais à la médecine, et se veut théorique, l'article de Sackett est quant à lui extrait d'un numéro spécial de la revue *Journal of Chronic Diseases* (le volume 32), édité ensuite sous la forme d'un livre, et qui a pour titre : « L'étude cas-témoin : consensus et controverse »[9]. Il constitue les actes d'un symposium consacré à l'étude cas-témoin, son histoire et ses problèmes méthodologiques, qui a eu lieu en avril 1978 aux Bermudes, symposium où sont présents (pour ne citer que ceux dont nous avons déjà parlé auparavant) Jerome Cornfield, Abraham Lilienfeld, Philip Sartwell, mais aussi Olli Miettinen, David Sackett et Alvan Feinstein, ainsi que d'autres épidémiologistes, statisticiens ou médecins importants des Etats-Unis et d'Europe. Il s'agit là d'un document historique absolument décisif pour notre étude et ce pour quatre raisons :

- d'abord le concept de biais y est omniprésent[10] et est au centre de la controverse sur les études cas-témoins. Les auteurs de l'article conclusif, Michael A. Ibrahim et Walter O. Spitzer, soulignent ainsi que « la susceptibilité de la méthode [de l'étude cas-témoin] aux différents biais a été au centre [de leurs] préoccupations »[11].
- Ensuite, la définition du biais qui y est donnée, par Sackett notamment, ainsi que la relation de ce concept avec ceux voisins de validité, de représentativité et de généralisabilité, continuent à faire autorité aujourd'hui.
- De même, la nécessité d'un catalogue des différents biais (proposition qui fait consensus parmi les participants), et la première formulation d'une taxinomie de ces biais constitue un projet toujours en cours au sein de l'épidémiologie
- Enfin la classification, désormais canonique, des biais en trois grandes catégories qui sont le biais d'information, le biais de sélection et le biais de confusion, alors même que la notion de confusion avait toujours été distinguée de celle de biais par les épidémiologistes, certes parfois de façon

[9] Ibrahim, Michel A., *The Case-Control Study Consensus and Controversy*, New York, Pergamon, 1979.
[10] Nous dénombrons 490 occurrences du mot « *bias* » sur les 151 pages du document (en incluant la page de titre et l'index)
[11] « *The susceptibility of the method to numerous bias has also been the focus of concern* », in Ibrahim, 1979, p. 144.

assez confuse, date très exactement de ce symposium et apparait dans l'article d'Ibrahim et de Spitzer.

Ainsi, selon nous, ces quatre points montrent que la notion de biais devient définitivement, en 1979, un concept scientifique à part entière au sein de l'épidémiologie. C'est cet évènement dont il s'agit à présent de retracer la genèse.

5.1 La notion de biais et le projet d'une médecine scientifique :

5.1.1 La clinique peut-elle être une science ?

Dans son *Essai sur quelques problèmes concernant le normal et le pathologique*, titre donné à sa thèse de doctorat en médecine, publiée en 1943, Georges Canguilhem pose en ces termes la relation entre la clinique et la science :

> « Quand on parle de pathologie objective, quand on pense que l'observation anatomique et histologique, que le test physiologique, que l'examen bactériologique sont des méthodes qui permettent de porter scientifiquement, et certains pensent même en l'absence de tout interrogatoire et exploration clinique, le diagnostic de la maladie, on est victime selon nous de la confusion philosophiquement la plus grave, et thérapeutiquement parfois la plus dangereuse. Un microscope, un thermomètre, un bouillon de culture ne savent pas une médecine que le médecin ignorerait. Ils donnent un résultat. Ce résultat n'a en soi aucune valeur diagnostique. Pour porter un diagnostic, il faut observer le comportement du malade. (…)
>
> En matière de pathologie, le premier mot, historiquement parlant, et le dernier mot, logiquement parlant, revient à la clinique. Or *la clinique n'est pas une science et ne sera jamais une science*, alors même qu'elle usera de moyens à efficacité toujours plus scientifiquement garantie. »[12]

Pour Canguilhem, en effet, la conception quantitative et objective de la santé et de la maladie héritée de Claude Bernard, c'est-à-dire l'identité de l'état normal et de l'état pathologique aux variations quantitatives près, fait perdre de vue aux médecins le point de vue subjectif et qualitatif du patient sur sa maladie, qui donne seul un sens

[12] Canguilhem, Georges, *Le normal et le pathologique*, 7ème éd., Paris, Quadrige/PUF, 1998 [1ère éd. : 1943], p. 152-153. Nous soulignons.

aux mesures objectives des examens cliniques et paracliniques. Les médecins considèrent ainsi souvent la médecine comme une science, à l'instar des autres sciences qu'elle utilise au quotidien, comme la physiologie, l'histologie ou la bactériologie, qui ont connu des progrès sans précédent au cours du XIXe et surtout du XXe siècle, disciplines auxquelles il faudrait ajouter la biologie moléculaire[13], qui apparait justement dans les années 1940. En somme, Canguilhem considère que la médecine est devenue trop scientifique et critique le fait que l'on puisse considérer la clinique elle-même comme une science, confondant ainsi les résultats de laboratoire avec le diagnostic, en oubliant la plainte du patient et sa subjectivité souffrante. Or, selon lui, la médecine doit être considérée comme « une technique ou un art au carrefour de plusieurs sciences, plutôt que comme une science proprement dite »[14].

Pourtant, c'est bien à une tentative de fonder la médecine comme science que vont s'atteler un certain nombre de médecins américains au cours des années 1960, qu'il s'agisse de la médecine en général, ou bien de la médecine clinique en particulier, c'est-à-dire la médecine qui se fait au lit du malade (conformément à l'étymologie latine « *clinice* » : « médecine exercée près du lit du malade », empruntée elle-même au grec « κ λ ι ν ι κ η »). Ainsi, Alvan Feinstein, alors « *Assistant Professor of Medicine*» à la prestigieuse *Yale University School of Medicine* publie en 1963 un éditorial dans le *Journal of Chronic Diseases*, intitulé : « The Basic Elements of Clinical Science »[15], qu'on peut traduire par « Les principes fondamentaux de la science clinique ». Selon lui en effet, les médecins sont victimes d'un préjugé qui leur fait croire que « la recherche digne d'intérêt peut seulement être faite au laboratoire » (Feinstein, 1963a, p. 1125) et qui implique que le « médecin universitaire doit agir soit comme un scientifique non-clinicien, soit comme un clinicien non-scientifique » (Feinstein, 1963a, p. 1125). Or, il est parfaitement possible d'être à la fois un bon scientifique et un bon clinicien selon Feinstein, à condition qu'il y ait « les mêmes normes de recueil et d'organisation des données au lit du patient qu'au laboratoire »[16]. En effet, ce sont ces normes de recueil et d'organisation des données, que Feinstein juge négligées par les cliniciens, qui constituent « les exigences scientifiques pour la reproductibilité

[13] Sur ce sujet, voir Morange, Michel, *Histoire de la biologie moléculaire*, Paris, Découverte/Poche, 2003.
[14] Canguilhem, 1998, p. 7.
[15] Feinstein, Alvan R., « The basic elements of clinical science », *Journal of Chronic Diseases*, vol. 16 / 11, 1963a, p. 1125–1133.
[16] « … *to insist on the same standards for collection and organization of data at the bedside as in the laboratory*. » in Feinstein, 1963a, p. 1125.

des résultats cliniques »[17]. Ce critère de la reproductibilité est directement emprunté aux expérimentations effectuées en laboratoire, et « la reproductibilité du travail scientifique » repose intégralement sur la « description correcte des méthodes utilisées pour acquérir et classer les données primaires ». Ce recueil et cette organisation des données selon une procédure correcte, c'est-à-dire reproductible, constituent ainsi la condition *sine qua non* pour procéder à une analyse statistique correcte. En effet :

> « L'analyse statistique est une procédure secondaire, imposée aux données principales à des fins d'interprétation, de réduction ou d'élaboration : elle ne corrige pas les erreurs fondamentales, elle n'insère pas des faits omis, ni ne valide un raisonnement défectueux »[18]

Or, si de grands progrès ont été faits en ce qui concerne le plan et l'analyste statistique, ce n'est pas le cas pour les méthodes de recueil des données primaires, qui restent encore largement défectueuses. De même, si les expériences réalisées en laboratoire ont permis de comprendre, par l'analyse et la réduction des problèmes complexes à des « mécanismes biologiques fondamentaux » (Feinstein, 1963a, p. 1125) et d'illuminer de « nombreux territoires de la maladie humaine »[19], elles sont néanmoins insuffisantes car il n'est pas toujours possible d'appliquer par exemple à l'homme des données issues d'expériences faites sur les animaux, ou bien de comprendre « un problème clinique complexe » à partir « d'investigations isolées de ses parties individuelles », c'est-à-dire sans prendre en compte « les interactions de ces parties dans un organisme intégré »[20].

Mais quelles sont alors ces données cliniques primaires ? Et comment en améliorer le recueil et l'organisation afin d'accroitre « le potentiel scientifique de la médecine au lit du malade » ? Ces « données basiques des maladies humaines » sont selon lui au nombre de trois : « les symptômes, les signes, et les patterns des

---

[17] « *Absence of such standards has led to neglect of basic scientific requirements for reproducibility of clinical results*. » in Feinstein, 1963a, p. 1125.

[18] « *The reproducibility of scientific work depends upon how well it is described by the methods used to acquire and classify the primary data. Statistical analysis is a secondary procedure, imposed on the main data for interpretation, reduction, or elaboration; it does not correct fundamental errors, insert omitted facts, or validate faulty reasoning*. » in Feinstein, 1963a, p. 1126.

[19] « *This type of fundamental scientific investigation in the laboratory is the best approach to many biologic problems, and the advances resulting from it have illuminated many areas in human disease*. », in Feinstein, 1963a, p. 1125.

[20] « *A complex clinical problem can rarely be solved merely by isolated investigations of its individual parts; its solution also requires separate study of the interaction of those parts in the integrated organism*. », in Feinstein, 1963a, p. 1125.

maladies »[21]. Parmi ces trois données, la notion de « pattern » retient particulièrement l'attention de Feinstein. En effet, selon lui, si la médecine anatomo-clinique a permis de créer un « excellent » système de « classification des maladies » sur une base histopathologique, elle s'en tient à donner des « *noms* » aux maladies, c'est-à-dire à décrire « ce qu'*est* une maladie » et non « ce qu'elle *fait* », c'est-à-dire son « *comportement* clinique »[22], ou en d'autres termes son « pattern », que Feinstein va diviser en « pattern de manifestations » (« *Pattern of manifestations*») et en « pattern de découverte » (« *Pattern of discovery*»[23]). C'est en effet en classant ces patterns de maladie, c'est-à-dire en adoptant une classification comportementale des maladies, que l'on va pouvoir rendre la médecine clinique plus scientifique. Mais après le problème de la classification des données vient celui de « l'exactitude des données » : en effet, le problème est que si certaines variables peuvent être mesurées par des tests de laboratoire ou par une biopsie, d'autres variables, ne le peuvent pas, notamment le recueil des antécédents médicaux (« *history-taking* ») et l'examen physique (« *physical examination* »[24]). En effet, ces données ne peuvent être recueillies et mesurées que par « un appareil nommé le médecin »[25]. Or, cet instrument ne subit pas d'inspections ou de processus de « calibration » comme les instruments de laboratoire :

> « L'absence de meilleurs méthodes pour garantir l'objectivité et l'exactitude [des données recueillies] au lit du malade constitue un défaut critique de la science clinique. »[26]

Dès lors, cet appareil qu'est le médecin est susceptible de faire des erreurs, aussi bien dans le recueil des antécédents médicaux que dans l'examen physique,

---

[21] «*To increase the scientific potential of the bedside, investigators must also improve the methods for obtaining and organizing the basic data of human illness: symptoms, signs, and patterns of disease.*», in Feinstein, 1963a, p. 1125. Nous avons choisi de garder le mot « *pattern* » tel quel dans la mesure où il est considéré comme un mot français à part entière par le CNRTL. Voir http://www.cnrtl.fr/definition/pattern (Accédé le 19 août 2017).

[22] « *Histopathologic terminology is inadequate for the total clinical spectra of disease, and must be supplemented by additional classifications that describe not merely the names of diseases, but their clinical behavior. The new classifications, augmenting those that tell what a disease is, should add what it does, as determined from two major considerations*. », in Feinstein, 1963a, p. 1126. C'est Feinstein qui souligne.

[23] Feinstein, 1963a, p. 1126.

[24] Feinstein, 1963a, p. 1129.

[25] « *The primary data in such diseases must come directly from the patient and are collected by an apparatus called a physician*. », in Feinstein, 1963a, p. 1129.

[26] « *The absence of better methods for objectivity and accuracy at the bedside is a critical defect of clinical science*. », in Feinstein, 1963a, p. 1129.

que Feinstein décompose en trois étapes : la description (« une zone rouge blanchissante est observée sur la peau, avec un centre d'un rouge plus foncé et non-blanchissant »), l'interprétation (« il s'agit d'une pétéchie entourée d'un érythème »), le diagnostic (« cela est lié à une méningococcémie »)[27]. Ainsi, dans le recueil des antécédents médicaux, « les médecins peuvent laisser leur attitude, leur manière de poser des questions, et leur interprétation des réponses, être influencées par leur propres préconceptions conscientes ou inconscientes »[28]. De même, dans le cadre d'un examen physique, « les interprétations peuvent être biaisés par la connaissance d'autres données cliniques »[29].

Dès lors, pour éviter les erreurs, il est nécessaire selon Feinstein de « calibrer » les médecins cliniciens, par exemple par une auscultation à l'aveugle (« *'blind' auscultation* »[30]), et cette calibration peut se faire aussi bien au niveau intra-individuel (le médecin examine son patient sans regarder au préalable le dossier, puis compare ses impressions cliniques avec le dossier du patient) ou au niveau inter-individuel (les médecins doivent vérifier et comparer les résultats obtenus chez un même patient), au risque sinon de « produire des déviations inexplicables dans les données »[31]. Cette calibration des observateurs permet alors d'assurer l'exactitude des données cliniques et par là de garantir leur reproductibilité, qui est la condition de possibilité d'une médecine scientifique. Ainsi,

> « A moins que les principes fondamentaux de la médecine clinique – symptômes, signes, patterns – soient étudiés et utilisés scientifiquement, l'application continue dans la clinique d'avancées isolées faites au laboratoire ne pourra qu'apporter la dignité de la science sans sa clarté. Sans science au lit du malade, la recherche médicale moderne pourrait ne produire, sous couvert

[27] « *By description, a blanching red area is observed on the skin, with a nonblanching darker red center. By interpretation, it is called a petechia surrounded by erythema. By diagnosis, it is attributed to meningococcemia.* », in Feinstein, 1963a, p. 1130.

[28] « *In addition to such errors of omission, errors of commission occur when physicians allow their attitude, manner of questioning, and interpretation of replies to be influenced by their own conscious or subconscious preconceptions*. », in Feinstein, 1963a, p. 1129.

[29] « *The interpretations may be biased by knowledge of other clinical data*. », in Feinstein, 1963a, p. 1130.

[30] Feinstein, 1963a, p. 1131.

[31] « *Examiners must check and compare results in the same patients in order to be certain that the actual examinations are not performed with subtle individual differences that may create otherwise unexplainable deviations in data. »,* in Feinstein, 1963a, p. 1131.

> d'un plan d'expérience sophistiqué, contrôlé, en double-aveugle, et très cher à mettre en œuvre, qu'un chaos statistiquement significatif. »[32]

Cette phrase, qui clôt l'article de Feinstein, ne doit pas être interprétée comme une attaque contre les études épidémiologiques, notamment contre l'essai clinique randomisé, mais bien comme la tentative à la fois de mettre la clinique, et ses données, au niveau de scientificité qui est celui du laboratoire, mais aussi de montrer qu'en matière d'études épidémiologiques, les données cliniques sont premières dans l'ordre logique : il est donc nécessaire, pour faire une inférence valide, que ces données cliniques soient exactes, et donc, dans l'esprit de Feinstein, reproductibles. Selon lui en effet, il est tout à fait possible par exemple que les patients classés comme souffrant de la même maladie dans une étude cas-témoins, ou même dans un essai clinique randomisé, ne souffrent pas en fait de la même maladie, mais de ce qu'il appelle, dans un article[33] publié la même année et qui entend appliquer, dans le cadre de la théorie des ensembles, l'algèbre booléen[34] et les diagrammes de Venn à la taxinomie clinique, « le spectre d'une maladie » (Feinstein,1963b, p. 930). Ainsi, Feinstein, après avoir dit que « les applications et les implications de ces diagrammes (…) devraient être particulièrement apparentes aux épidémiologistes cliniques et aux statisticiens » (Feinstein,1963b, p. 936) souligne la chose suivante:

> « Parce que presque toute maladie possède un spectre complexe tel qu'il est illustré ici, *aucun* ensemble de ceux qui souffrent de cette maladie ne peut véritablement constituer un échantillon au hasard de la maladie dans son intégralité. (…). Pour ces raisons, deux ensembles de patients qui souffrent de la même maladie peuvent avoir de profondes différences qui restent non détectées malgré la similarité superficielle des collections. Ces variations cachées peuvent avoir des effets très importants quand ces deux ensembles doivent être comparés comme groupe expérimental et comme groupe contrôle. La division en groupes similaires est un principe fondamental du plan

[32] « *Unless the basic elements of clinical medicine – symptoms, signs and patterns— are studied and used scientifically, the continued clinical application of isolated advances from the laboratory may bring only scientific dignity without scientific clarity. Without science at the bedside, modem medical research may yield an intricately designed, expensively produced, doubly-blind controlled, statistically significant chaos.* », in Feinstein, 1963a, p. 1133.

[33] Feinstein, Alvan R., « Boolean Algebra and Clinical Taxonomy: Analytic Synthesis of the General Spectrum of a Human Disease », *New England Journal of Medicine*, vol. 269 / 18, 1963b, p. 929-938.

[34] Il faut préciser qu'Alvan Feinstein est titulaire d'un Master en mathématiques, obtenu en 1948 à l'Université de Chicago, soit quatre ans avant son doctorat en médecine.

> expérimental scientifique, mais la raison principale qui justifie cette division est violée si le nombre de patients dans les deux groupes est identique alors que les types de patients inclus dans les deux groupes sont différents. Les variations observées par la suite pourraient donc être l'effet non des procédures thérapeutiques ou expérimentales mais celui des variations dans les populations de base utilisées dans les deux groupes. »[35]

Ainsi, le point central de l'argumentation de Feinstein est que même la randomisation, conçue justement pour éviter le problème du biais de sélection, ne permet pas nécessairement de garantir la représentativité de l'échantillon, puisqu'il est possible que les patients souffrant par exemple d'un cancer du poumon ou d'un rhumatisme articulaire aigu (pour reprendre les exemples étudiés par Feinstein) ne souffrent pas tous de la même maladie, conçue non plus comme une entité aux caractéristiques cliniques bien définies mais comme un spectre constitué de multiples ensembles et sous-ensembles de symptômes et de signes qui peuvent être mutuellement exclusifs, ou se chevaucher, ou bien être dans une relation de subordination l'un par rapport à l'autre[36]. Dès lors, cette nouvelle « technique de classification », qu'il appellera bientôt la science de la « clinimétrie », qui traite des « méthodes quantitatives dans le recueil et l'analyse de données cliniques comparatives, et particulièrement de l'amélioration de la « mesure » des phénomènes spécifiquement cliniques et personnels dans la prise en charge du patient »[37], est nécessaire chaque fois que « des ensembles de patients doivent être comparés dans

---

[35] « *Because almost any disease has the complex spectrum shown here, no single collection of its hosts can truly be a valid random sampling of the entire disease. (…) For these reasons two collections of patients with the same disease may have profound differences that remain undetected despite superficial similarity of the collections. These hidden variables can have a far reaching effects when the two collections are to be compared as experimental and control groups. Division into similar groups is a basic principle of scientific experimental design, but the main reason for making the division is violated if the number of patients in the two groups is the same but the types of patient contained in the two groups are disparate. The variation observed thereafter may arise not from any experimental or therapeutic procedures but from variations in the base populations used in the two groups.* », in Feinstein, 1963b, p. 936. C'est Feinstein qui souligne.

[36] Voir Feinstein,1963b, p. 932, pour une représentation graphique, sous la forme de diagrammes de Venn, de ces relations entre ensembles.

[37] « *The domain of clinimetrics is concerned with quantitative methods in the collection and analysis of comparative clinical data, and particularly with improved "measurement" of the distinctively clinical and personal phenomena of patient care*», in Feinstein, Alvan R., « An Additional Basic Science for Clinical Medicine: IV. The Development of Clinimetrics », *Annals of Internal Medicine*, vol. 99 / 6, décembre 1983, p. 843-848.

le cadre d'investigations épidémiologiques, pronostiques, thérapeutiques ou cliniques. »[38].

En somme, Feinstein entend faire de la clinique une science en l'objectivant à travers un processus de quantification qui permet d'homogénéiser ou de standardiser les données cliniques et ainsi de les rendre reproductibles. Il est intéressant de noter, sans que nous puissions développer plus cette question, la similarité du projet de Feinstein avec ce qui s'est passé dans l'astronomie du XIXe siècle : dans son article[39] sur l'objectivation de l'observation (ou selon l'expression qu'il emploie, « l'observation sans sujet observant ») Zeno G. Switjtink montre ainsi que le développement et l'application de méthodes non-déductives d'inférence, comme la méthode des moindres carrés, « a été rendue possible par une objectivation de la mesure scientifique, qui a elle-même mené à une plus grande objectivation de la pratique scientifique » (Switjtink, 1987, p. 261). En effet, selon lui :

> « La mesure scientifique a été rendue objective, c'est-à-dire indépendante du jugement personnel des observateurs individuels, à travers l'utilisation d'instruments de mesure de précision qui ne permettaient pas des lectures équivoques, ou, quand cela n'était pas possible, en considérant l'observateur humain lui-même comme faisant partie de l'instrument de mesure. »[40]

La lutte contre les biais chez Feinstein est donc d'abord, comme d'ailleurs chez Bradford Hill dans le cadre des études épidémiologiques, une lutte contre la subjectivité, en l'espèce la subjectivité des médecins cliniciens et leurs idiosyncrasies qui empêchent une classification des maladies à travers des critères objectifs (signes, symptômes, patterns), et donc qui empêchent la médecine clinique de devenir une science à part entière. Ce sont en effet les données cliniques, et leur organisation en une taxinomie « efficace » (« *effective* ») et « exacte » (« *accurate* »[41]), qui vont fournir une nouvelle logique à la médecine, inspirée de l'algèbre de Boole, c'est-à-dire

---

[38] « *The new classification technique (…) is necessary, however, whenever collections of diagnosed patients are to be compared in epidemiologic, prognostic, therapeutic or other clinical investigative studies.* », in Feinstein, 1963b, p. 938.

[39] Switjtink, Zeno G., « The Objectification of Observation: Measurement and Statistical Methods in the Nineteenth Century», in Krüger, Lorenz, Daston, Lorraine et Heidelberger, Michael (eds), *The Probabilistic Revolution*, 1: Ideas in History, Cambridge, Mass., MIT Press, 1987, p. 261-285.

[40] « *Scientific measurement has been made objective, that is, independent of the personal judgment of individual observers, through the use of precision measuring instruments that allow for unequivocal readings, or, in cases where this was not possible, by considering the human observer himself as part of the measuring apparatus*. », in Switjtink, 1987, p. 261.

[41] Feinstein, 1963a, p. 1133.

une nouvelle manière de raisonner. La logique de la médecine, c'est précisément le titre d'un ouvrage d'Edmond Murphy, qui parait en 1976, et qu'il convient à présent d'étudier.

5.1.2 Edmond Murphy et le projet d'une logique de la médecine :

Tout d'abord, il convient de souligner que le titre qu'Edmond Murphy[42] donne à son ouvrage, The *Logic of Medicine* (en français : *La logique de la médecine*), doit être pris en son sens le plus littéral, qui est celui d'une théorie de la preuve en médecine. En effet, l'objectif de Murphy est énoncé dès son introduction (intitulée explicitement : « L'objectif ») : « cet ouvrage porte sur les idées [« *this book is about ideas* »], et plus précisément sur « les idées derrière l'évaluation des sources de données et de leur interprétation, à partir desquelles le corpus du savoir médical doit être dérivé » (Murphy, 1976, p. 3). Ces données, ou ces « faits », sont supposées être vraies, mais cette vérité, selon Murphy, n'est que provisoire[43] et c'est pourquoi il est plus important de s'intéresser aux « règles de la preuve » (« *rules of evidence* », in Murphy, 1976, p. 3) qui sont « plus durables » (Murphy, 1976, p. 3), et dont la négligence, aussi bien dans le cursus médical que par les médecins eux-mêmes, est « irresponsable » (Murphy, 1976, p. ix). C'est pourquoi Murphy entend dans son ouvrage, destiné aux « étudiants en première année de médecine »[44] leur permettre de « cultiver un sens commun alerte » dans le domaine de « l'inférence abstraite » (Murphy, 1976, p. x). Murphy part en effet d'un constat similaire à celui de Feinstein concernant la

[42] La source la plus complète d'informations biographiques et bibliographiques sur Edmond Murphy est le site internet qui lui est consacré : https://sites.google.com/site/edmondantonymurphy/home (accédé le 20 septembre 2017). Pour notre sujet, il est important de noter que Murphy est d'abord un médecin, né au Pays de Galles, formé à l'Université de Belfast en Irlande du Nord, où il obtient son doctorat en médecine en 1952, avant de s'exiler aux Etats-Unis en 1956, plus précisément à la *Johns Hopkins University School of Medicine*. C'est à l'université et à l'hôpital de Johns Hopkins qu'il passe l'essentiel de sa carrière d'enseignant et de médecin, notamment dans la Division de génétique médicale, et à la *Johns Hopkins School of Public Health* qu'il obtient son doctorat en biostatistiques en 1964. En plus de son intérêt pour les aspects épistémologiques et éthiques de la médecine, Murphy s'est beaucoup intéressé au conseil génétique, qu'il a pratiqué et dont il a défini les principes dans un ouvrage, écrit en collaboration avec un de ses anciens étudiants, Gary Chase, paru en 1975 et intitulé : *Principles of Genetic Counseling*. Il a été notamment un des premiers, si ce n'est le premier, à intégrer des approches bayésiennes dans l'estimation du risque dans le domaine du conseil génétique. (Sur ce point, voir Harper, Peter S., *A Short History of Medical Genetics*, Oxford ; New York, Oxford University Press, 2008, 557 p., (« Oxford monographs on medical genetics », 57), et plus précisément le chapitre 11.

[43] « *Biological mechanisms are complex and what purports to be a fact has a « short half-life »* (Murphy, 1976, p. 3). Il ajoute, plus loin : « Ce serait une erreur pour le lecteur de considérer que ce qui compte ici ce sont les faits. » (Murphy, 1976, p. 4).

[44] « *medical students in their first year* », in Murphy, 1976, p. x.

scientificité de la médecine. Il considère même que la médecine de son époque affronte une « crise des soins médicaux » (Murphy, 1976, p. 8). Selon lui, la médecine « est insuffisamment académique » (Murphy, 1976, p. 5), l'académisme renvoyant selon lui à trois critères essentiels :

1) D'abord l'exactitude (« *exactitude* ») : il s'agit selon Murphy d'une « valeur ultime qu'aucune autre considération ne doit mettre en danger »[45]. Cette notion d'exactitude renvoie immédiatement à celle de vérité, et plus précisément aux moyens de l'atteindre, c'est-à-dire là aussi aux règles d'inférence. Ainsi, « si le médecin pense qu'il doit satisfaire un sens de la responsabilité dans le fait de connaitre la vérité, alors il doit accepter la responsabilité des moyens par lesquels il connait cette vérité » (Murphy, 1976, p. 5).
2) Ensuite, un « besoin urgent de systématisation » : en effet, selon Murphy, « la vérité transcende les faits » et est plus qu'un « simple agrégat de petits morceaux d'information » (Murphy, 1976, p. 6). Il faut donc systématiser ou généraliser les faits, c'est-à-dire être capable de les relier entre eux, pour disposer d'un savoir véritablement académique. Evidemment, cette systématisation est susceptible d'entrer en conflit avec l'exactitude, conflit qui peut se révéler insoluble (« *incurable conflict* »), une représentation exacte de la réalité nécessitant en effet la totalité des faits.
3) Enfin, la troisième caractéristique de l'académisme consiste à discipliner rigoureusement ses émotions. Cela renvoie à la fois aux aspects éthiques de la médecine mais aussi à ses aspects épistémologiques : le clinicien doit en effet prendre des décisions en situation d'incertitude, et sans pouvoir toujours les justifer d'un point de vue rationnel, c'est-à-dire en s'appuyant sur des preuves ou des règles d'inférence. En ce cas, il ne doit pas hésiter à « laisser l'évaluation de la preuve à ceux qui sont en position de le faire de façon dépassionnée » (Murphy, 1976, p. 6).

C'est pour toutes ces raisons concernant le manque d'académisme de la médecine que Murphy considère qu'une véritable théorie de la médecine est nécessaire, et que la clinique doit développer « sa propre méthode et sa propre

[45] « *First, exactitude is an ultimate value which no other consideration must endanger* », in Murphy, 1976, p. 5.

épistémologie » (p. 7). Ainsi , une théorie de la médecine devrait comprendre aussi bien le « processus diagnostique » (Murphy, 1976, p.7) , dont l'efficience doit être optimisée ; le problème de la classification nosologique ; la relation entre les événements, et donc la question de la causalité, où l'analogie peut permettre au clinicien de prendre de bonnes décisions ; ou encore la protection du clinicien contre les fausses doctrines, qui peut être assurée en donnant à l'étudiant en médecine un « équipement général pour évaluer la preuve » (Murphy, 1976, p. 9). Cette insistance sur la notion de preuve (« *evidence* ») et sur celle d'inférence, tout comme le constat d'une crise de la médecine, et notamment de la médecine clinique, montrent qu'il s'agit bien pour Murphy de fonder une nouvelle logique médicale c'est-à-dire aussi bien une nouvelle manière d'évaluer les preuves que les inférences qui permettent de passer des preuves à la vérité. La situation du concept de biais dans l'ouvrage, qui apparait dans la partie consacrée aux preuves (« *the evidence* ») après les chapitres consacrés à la « superstition », à la notion de « preuve » (« *proof* »), à celle de « *bimodalité* », puis à celle de « cause », et juste avant celui consacré à la « confusion » (« *confounding* ») est sur ce point intéressante : au cœur du biais gît aussi bien le problème de l'inférence (notamment de l'inférence causale) que celui de la vérité.

En effet, le problème du biais, auquel Murphy consacre le chapitre 15 de son ouvrage, renvoie selon lui directement à la question de la vérité et de la causalité. Ainsi le sens courant du mot « biais » est lié à la question de la représentation correcte de la réalité :

> « L'opinion d'un homme est considérée comme biaisée si elle ne représente pas fidèlement [« *fairly* »] les faits » (Murphy, 1976, p. 239)

La seule différence entre le sens commun et l'usage du mot « biais » dans « le contexte épistémologique », où il est employé « de façon analogue » (« *analogously* ») au sens courant, se situe en fait dans l'absence de « connotations morales » (« *moral overtones* ») : dans l'acception commune du mot, dire à quelqu'un que son opinion est biaisée fait ainsi figure de « reproche », ce qui n'est pas le cas dans son sens scientifique. Murphy donne alors sa définition du mot « biais » au sens scientifique, définition qui deviendra canonique en épidémiologie et en médecine :

« *Un biais est un processus qui, à n'importe quelle étape de l'inférence, tend à produire des résultats qui s'écartent systématiquement des vraies valeurs* ».[46]

Cette définition appelle plusieurs commentaires : d'abord, Murphy reprend la distinction entre l'erreur aléatoire et l'erreur systématique, qu'il semble d'ailleurs considérer comme allant de soi. Dès 1969, dans un article[47] en trois parties consacré aux « données médicales et à l'éthique appliquée », et spécifiquement dans la deuxième partie, il avait déjà défini le biais comme erreur systématique, par opposition à l'erreur aléatoire, et insisté sur le fait que le biais renvoyait à un problème de représentativité de l'échantillon. Pourtant, une dizaine d'années plus tard, la définition du concept de biais telle qu'elle est donnée par le même Murphy est beaucoup plus large au sens où la question de la représentativité de l'échantillon n'est plus qu'un exemple parmi d'autres de biais.

En effet, le point le plus original dans le commentaire que fait Murphy de la définition du biais dans son ouvrage de 1976, est que, immédiatement après la distinction, proprement statistique, entre erreur aléatoire et erreur systématique, Murphy fait une référence à Sigmund Freud pour souligner que « le scientifique est conscient des mécanismes profonds et sournois (…) qui permettent de rendre compte des divergences de vue entre deux observateurs aguerris qui regardent un même phénomène » (Murphy, 1976, p. 239), avant d'ajouter que les « sentiments de l'observateur ne sont pas pertinents pour la vérité d'une proposition scientifique » (Murphy, 1976, p. 239), tout comme l'autorité de l'observateur, liée à sa réputation ou à la position scientifique qu'il occuperait, ne constitue pas un argument, les faits et les « régles de la preuve » (« *rules of evidence* ») s'appliquant à tous de la même façon. Il s'agit bien ainsi du problème de la subjectivité, et même de l'inconscient, qui apparait ici, et donc le sens trivial ou psychologique qui existe dans la langue anglaise ou la langue américaine. Relativement à l'histoire du concept de biais que nous venons de retracer, il est difficile de ne pas reconnaitre, en schématisant quelque peu, l'aspect fishérien du concept de biais liée à la notion d'erreur systématique et qui renvoie au

[46] « *A bias is a process at any stage of inference tending to produce results that depart systematically from the truth* », in Murphy, 1976, p. 239. C'est Murphy qui souligne.

[47] Murphy, Edmond A., « Medical Data and Applied Ethics: Part I », *The Linacre Quarterly*, vol. 36 / 3, 1969, p. 158-164.

Murphy, Edmond A., « Medical Data and Applied Ethics: Part II: The Sources of Data », *The Linacre Quarterly*, vol. 36 / 4, 1969, p. 229-235.

Murphy, Edmond A., « Medical Data and Applied Ethics: Part III: The Interpretation of Evidence », *The Linacre Quarterly*, vol. 36 / 3, 1969, p. 236-241

problème de l'estimation, et l'aspect plutôt propre à A. B. Hill[48] de la notion psychologique de biais qui renvoie à la question de la subjectivité et de l'objectivité. En réalité – et nous entendons montrer que c'est là la véritable innovation de Murphy relativement à ce concept, innovation conceptuelle qui fait *définitivement* du biais un concept médical et épidémiologique à part entière – la particularité et l'originalité de la définition que donne Murphy du concept de biais consistent en quelque sorte à réunir et même à fusionner les différentes significations disponibles à l'époque du mot biais, de la plus triviale (son sens psychologique) à la plus scientifique (son sens statistique), à balayer le spectre de ses nuances sémantiques pour aboutir à une épuration de ce qui fait l'essence même du concept de biais.

Ainsi le biais est essentiellement un « processus » (« *process* »), terme suffisamment vague pour s'appliquer aussi bien à de nombreuses classes de phénomènes qu'à de nombreuses disciplines scientifiques : ainsi existe-t-il des phénomènes physiques, biologiques, psychologiques, sociologiques, statistiques, etc. De même, en le définissant comme un processus, Murphy insiste aussi bien sur le caractère dynamique que temporel du concept de biais. Ce processus qu'est le biais s'insère lui-même au sein d'un autre processus, au sens où il comprend de multiples étapes (« *stages* »), qui est celui de l'inférence : il s'agit donc ici d'un processus logique qui consiste à tirer une conclusion ou une conséquence d'un évènement, d'un fait ou d'une donnée ou d'un ensemble d'évènements, de fait ou de données. Enfin, ce processus qu'est le biais se manifeste essentiellement par ses effets : il « tend à produire des résultats qui s'écartent systématiquement des vraies valeurs ». Ainsi, en parlent de « résultats » et de « valeurs », Murphy semble faire référence, au moins implicitement, à des données numériques comme les résultats d'un test sanguin ou d'une étude épidémiologique, ce qui aurait pour effet de limiter sérieusement son propos ou même de friser la contradiction : en effet, si l'on considère les résultats ou les données comme le fondement de l'inférence, c'est-à-dire sa ou ses prémisses, et le biais comme une déviation systématique dans ces données, alors le biais devrait intervenir uniquement à la première étape de l'inférence, qui est celle des données.

[48] Murphy cite d'ailleurs l'ouvrage de Hill, *Principles of Medical Statistics*, et en recommande la lecture dans son introduction, au motif qu'il « encourage un sens de la responsabilité envers ce qu'il y a derrière les données », et souligne que le propos de Hill est « moins général que le sien » au sens où il porte essentiellement sur le problème d'effectuer « des inférences à partir des données épidémiologiques » (Murphy, 1976, p. 10). Ainsi, il s'agit bien pour Murphy de la question de l'inférence à partir de toutes les données, et pas seulement les données épidémiologiques.

Or, Murphy nous précise bien que le biais peut intervenir à « n'importe quelle étape » (« *any stage* ») de l'inférence, et pas seulement à celle des données. Comment alors concilier ces deux aspects ? N'y a-t-il pas ici une contradiction, ou à tout le moins une ambiguïté, au sein même de la définition du concept de biais donnée par Murphy ?

En un sens, il y a effectivement une certaine ambigüité dans la définition du concept de biais telle qu'elle est donnée par Murphy en 1976, qui reflète sans doute soit une certaine indécision, soit une volonté de la part de Murphy de ne pas trop bousculer la ou les définitions acceptées du mot biais à l'époque. Cette ambigüité sera d'ailleurs, selon nous, levée par David Sackett dans la définition, directement et explicitement emprunté à Murphy et néanmoins « adaptée » (Sackett, 1979, p. 60), qu'il donne du concept dans son article de 1979 :

> « Tout processus à toute étape de l'inférence qui tend à produire des résultats ou des conclusions qui diffèrent systématiquement de la vérité » [49]

Ainsi, Sackett a apporté deux modifications essentielles à la définition de Murphy : il a d'abord ajouté la notion de « conclusions » à celle de résultats, et il a remplacé la notion de « vraies valeurs » par celle de « vérité ». Selon nous, cette définition donnée par Sackett (sur laquelle nous reviendrons par la suite) correspond mieux à la conception du biais selon Murphy que la définition que Murphy donne lui-même. En effet, Murphy, peu après avoir donné sa définition, nous signifie plus clairement le projet du chapitre 15 consacré à la notion de biais : il s'agit « d'explorer systématiquement les multiples voies par lesquelles l'appréhension et l'évaluation de la vérité scientifique pourrait être déformée »[50]. Murphy distingue alors six étapes (« *steps* ») durant lesquelles le « biais pourrait s'introduire dans le processus scientifique » :

1) dans le plan d'expérience (« *design* »),
2) dans l'observation (« *observation* »),
3) dans l'estimation (« *estimation* »),
4) dans la mise à l'épreuve des hypothèses (« *testing hypotheses* »),
5) dans l'interprétation (« *interpretation* »),
6) et enfin dans le compte-rendu (« *reporting* »)[51].

[49] « *Any process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth.* », in Sackett, 1979, p. 60.
[50] « to *explore systematically the several ways in which the apprehension and evaluation of scientific truth may be distorted* », in Murphy, 1976, p. 239.
[51] Murphy, 1976, p. 239.

C'est donc bien une approche à la fois temporelle et logique que Murphy propose dans son ouvrage: pour des raisons de commodité, nous qualifierons cette approche de diachronique, signifiant par là, conformément à l'origine de ce concept qui se situe dans le domaine de la linguistique, qu'il s'agit ici d'une approche qui prend en compte le temps au sens ici où l'inférence décrite par Murphy est un processus certes logique mais qui a aussi des implications matérielles : il s'agit ici pour le médecin-scientifique tout autant de construire une étude scientifique avec un plan précis d'expérience qui permette de tester une hypothèse en estimant un certain nombre de paramètres, que d'être capable de débattre de l'interprétation des résultats ou de faire une lecture critique des articles scientifiques qui relatent ces études, et même d'évaluer non seulement les biais internes à l'étude mais aussi ce qu'on pourrait qualifier de biais externes à l'étude comme ce qu'on appelle aujourd'hui le biais de publication que Murphy décrit dès 1979 (voir Murphy, 1976, p. 259). Il est temps à présent de détailler cette approche originale du concept de biais.

5.1.3 Une approche diachronique du concept de biais

Murphy distingue donc six étapes de l'inférence où un ou plusieurs biais sont susceptibles d'entrer et donc de fausser les résultats. Le premier type de biais que Murphy distingue est le « biais du plan d'étude »[52]. Par « *design* », qu'il emploie en un sens très lâche (« *very loose sense* », in Murphy, 1976, p. 240), Murphy désigne « les moyens que le scientifique met en œuvre pour trouver une réponse à une question » (Murphy, 1976, p. 240). Ce plan se divise en deux sous-étapes : les « méthodes qu'il entend utiliser » et « le matériau auquel il entend les appliquer » (Murphy, 1976, p. 240). Pour résumer la pensée de Murphy, le point central du biais dans le plan d'expérience renvoie au problème de la confusion (« *confounding* », in Murphy, 1976, p. 240) qu'il étudie plus en détail dans le chapitre 16. En effet, Murphy montre, en s'appuyant sur la question de savoir si un médicament particulier influence ou non la pression artérielle dans le cadre d'une expérience où on injecterait ce médicament chez des animaux, qu'il est impossible de déterminer par cette seule expérience si la modification de la pression artérielle est due au médicament, à l'injection ou à l'interaction des deux. De même, selon lui, différentes méthodes peuvent conduire à

[52] « *Bias of design* », in Murphy, 1976, p. 240.

différents résultats : il expose ainsi trois méthodes pour mesurer la durée de vie des globules rouges dans l'organisme, dans l'optique de déterminer combien de temps durent les effets bénéfiques d'une transfusion de sang chez des patients anémiques. Or, chaque méthode donne une durée de vie différente : de trois à quatre semaines selon les études transfusionnelles jusqu'à quatre mois selon la méthode des isotopes radioactifs. Dès lors, cela montre que non seulement les estimateurs sont biaisés dans deux des trois méthodes (la première, celle des transfusions, et la deuxième, la méthode Ashby, qui elle donne une durée de vie entre cinquante et cent jours, soit entre moins de deux mois et un peu plus de trois mois) sont biaisés au sens où elles n'estiment pas exactement (« *accurately* ») la vraie valeur. Mais le biais n'est pas ici simplement la propriété de l'estimateur statistique, mais aussi la propriété de la méthode : en réalité, le « choix de la méthode est aussi une source de biais » (Murphy, 1976, p. 242) car « il n'y a aucun moyen de savoir à l'heure actuelle laquelle de ces estimations est correcte, ni même s'il y en a une » (Murphy, 1976, p. 242). En somme, la manière dont on pose la question détermine à la fois les moyens employés pour y répondre mais aussi la réponse elle-même, et à chacune de ses étapes, des biais peuvent s'introduire.

Murphy passe alors à la deuxième étape de l'inférence, c'est-à-dire au deuxième type de biais qui est le « biais d'observation » (Murphy, 1976, p. 243). Il souligne qu'aujourd'hui, « tout le monde reconnait la réalité selon laquelle les observateurs sont biaisés » (Murphy, 1976, p. 243). Le point sans doute le plus intéressant dans les propos de Murphy est qu'il va établir une sorte de gradation du quantitatif au qualitatif en montrant que les biais qualitatifs sont les plus importants :

> « La taille et l'importance possible du biais peuvent varier du trivial (quand des mesures simples sont faites par des systèmes complètement automatisés) au flagrant [« *gross* »], comme cela se produit en esthétique, dans les jugements de valeur, ou dans l'évaluation psychologique non-quantitative » (Murphy, 1976, p. 244).

Dès lors, dans la mesure où « le biais de l'observateur est d'une importance majeure quand des comparaisons sont effectuées entre les groupes » (Murphy, 1976, p. 244), au sens où « les idées préconçues » de l'observateur ou du patient peuvent créer une « différence systématique sérieuse » (Murphy, 1976, p. 244), il est essentiel de recourir partout où cela est possible à des expériences en double-aveugle, où ni

l'observateur ni le patient ne savent « quel traitement est administré » (Murphy, 1976, p. 245), et où « deux sources importantes de biais sont éliminées » (Murphy, 1976, p. 245).

La troisième étape de l'inférence est celle de l'estimation et elle est étroitement liée selon Murphy à la quatrième étape qui est celle de la « mise à l'épreuve des hypothèses »[53]. En effet, dans les deux premières étapes, il s'agit de collecter les données, et de les collecter de façon correcte ou valide. Une fois les données collectées, « le problème d'effectuer une inférence à partir de ces données se pose » (Murphy, 1976, p. 245), et donc la question de l'inférence statistique. Cette « inférence statistique est de deux grandes sortes : l'estimation et la mise à l'épreuve des hypothèses » (Murphy, 1976, p. 245). Or, ces deux étapes « dépendent de postulats de départ [« *basic assumptions* »], et dans les deux, il peut y avoir un choix de procédures systématiques » (Murphy, 1976, p. 245). En d'autres termes dans les deux cas, il peut être fait usage de « méthodes bayésiennes », c'est-à-dire des méthodes où « des informations préalables ou des convictions » (Murphy, 1976, p. 245) sont incorporées, ce qui peut constituer une source de biais.

Concernant l'estimation tout d'abord, Murphy considère que l'estimateur est en soi une source de biais, même s'il souligne qu'en matière d'estimation, l'absence de biais n'est pas la question « la plus importante » (Murphy, 1976, p. 245) :

> « La source la plus simple de biais à laquelle on a à faire dans l'estimation est celle qui est inhérente à l'*estimateur*, c'est-à-dire à la procédure systématique qui est utilisée pour parvenir à la meilleure *estimation*, ou à la meilleure supposition, de la vraie valeur ». (Murphy, 1976, p. 245).

Ainsi, Murphy montre que de nombreux estimateurs, comme la variance, ne conviennent qu'à des échantillons suffisamment grands, et sont biaisés si l'échantillon est trop petit, ce qui va conduire de façon systématique à une sous-estimation ou à une surestimation de la vraie valeur. Il montre aussi qu'en fonction des « postulats de départ qui sous-tendent l'analyse », des biais peuvent s'introduire, au sens où « des estimations radicalement différentes seront obtenues avec des modèles différents » (Murphy, 1976, p. 247) : ainsi en fonction du postulat de départ dans lequel on se place, par exemple entre la théorie selon laquelle les plaquettes sanguines meurent de vieillesse et celle selon laquelle elles sont détruites de façon indiscriminée en

[53] « *testing hypotheses* » in Murphy, 1976, p. 245.

fonction des besoins de l'organisme, les estimations seront « très différentes » (Murphy, 1976, p. 248), ce qui ne pose pas de problème si l'on utilise le même modèle pour comparer deux ou plusieurs groupes, mais ce qui ne permet pas d'estimer « la vraie valeur de la survie moyenne des plaquettes » (Murphy, 1976, p. 248). Enfin, concernant le problème de l'inférence bayésienne, « l'un des sujets les plus controversés des statistique modernes » (Murphy, 1976, p. 250), Murphy considère que l'incorporation de « convictions préalables » constitue une « source gênante de biais » (Murphy, 1976, p. 250), qui pourrait néanmoins être réglée dans la mesure où il serait possible de présenter les estimations en séparant ce qui relève de « la contribution des données » et ce qui relève de la « contribution des informations préalables » (Murphy, 1976, p. 250). Murphy va d'ailleurs distinguer le biais classique du biais bayésien : en effet, si l'on définit généralement le biais comme une erreur systématique « que de grands échantillons ne corrigent pas nécessairement » (Murphy, 1976, p. 250), le biais dans l'approche bayésienne peut plus ou moins disparaitre si la quantité de données est suffisamment grande.

Concernant à présent le biais dans la mise à l'épreuve des hypothèses, Murphy considère que « les méthodes [pour tester les hypothèses, c'est-à-dire en fait les tests statistiques de signification] doivent être considérées comme des sources importantes de biais » (Murphy, 1976, p. 251). Ainsi, selon lui, « un analyste biaisé pourrait choisir le test qui vient le plus confirmer ses idées préconçues » (Murphy, 1976, p. 252). De la même manière, un « enquêteur biaisé (…), qui échafaude ses hypothèses après avoir établi les faits, pourrait faire son choix parmi les conclusions qui vont se révéler « vraies », en refusant de considérer des tests formels sur des points qui contredisent ses idées préconçues » (Murphy, 1976, p. 254). Dès lors, et cela rejoint la conviction de Murphy que tester des hypothèses est une activité inutile (Murphy, 1976, p. 256), c'est « la totalité du processus de mise à l'épreuve des hypothèses qui est, en un sens, biaisé » (Murphy, 1976, p. 255), puisqu'à la fois l'hypothèse qu'il s'agit de tester et le test qu'on utilise pour mettre à l'épreuve l'hypothèse relèvent finalement non d'une procédure objective mais du choix, nécessairement subjectif, de l'enquêteur.

Le cinquième type de biais, qui correspond à la cinquième étape de l'inférence, est le « biais d'interprétation » (Murphy, 1976, p. 257).

> « En un sens, nous dit Murphy, une fois que l'analyse est complète, la partie scientifique de l'étude est terminée ». (Murphy, 1976, p. 257).

Et pourtant, « rares sont les auteurs qui sont capables de résister à la tentation d'ajouter une discussion » (Murphy, 1976, p. 257). Or cette tentation, souvent légitime, se transforme parfois en un moment où l'auteur sort du strict cadre des résultats de son étude pour se lancer dans des « spéculations » (Murphy, 1976, p. 257). Dès lors, de « nombreuses opportunités nouvelles de biais apparaissent, biais qui sont souvent d'une nature subtile et insaisissable » (Murphy, 1976, p. 257). Pour éviter ce genre de biais, Murphy avance alors un certain nombre de préconisations :

- D'abord, il est important d'épuiser « l'espace des hypothèses » (Murphy, 1976, p. 258), c'est-à-dire de présenter les autres hypothèses possibles qui pourraient expliquer les résultats de l'étude.
- Ensuite, il est important de ne pas « manipuler les cas qui constituent des exceptions aux conclusions générales » (Murphy, 1976, p. 258) afin de renforcer la validité des résultats. Ainsi, « si le traitement est solide, sa valeur sera démontrée par ses résultats, sans qu'il soit nécessaire de plaider pour eux » (Murphy, 1976, p. 258).
- De même, « l'argument *a fortiori* » doit être évité. En effet, dire :« Il est certainement vrai que les membres du groupe traité étaient plus vieux (ou plus gros, ou moins intelligents) que ceux du groupe contrôle, mais cela irait clairement à l'encontre de l'idée que le traitement est efficace, et, si les patients avaient été effectivement comparables, le traitement aurait été encore plus efficace » (Murphy, 1976, p. 258) produit plutôt de la suspicion que de la confiance, car en réalité, étant donné la complexité des phénomènes biologiques, il s'agit là d'une pétition de principe au sens où on ne peut pas savoir ce qu'il se serait passé.
- Enfin, « la dernière et la plus complexe source de biais dans l'interprétation renvoie à l'acceptation des conclusions d'une étude expérimentale ou observationnelle » (Murphy, 1976, p. 259). En effet, nul n'est obligé de se plier aux preuves[54], et les « principes qui devraient gouverner l'attitude du scientifique défient l'analyse ». Néanmoins, il est indubitable que même les scientifiques émettent des « jugements biaisés » (Murphy, 1976, p. 259), donc non objectifs, et que leur « réceptivité » (Murphy, 1976, p. 259) aux preuves et aux arguments pose problème.

[54] « *Acceptance can not be compelled by evidence* » in Murphy, 1976, p. 259.

Enfin, le sixième et dernier type de biais est le « biais dans le compte-rendu » (« *bias of reporting* », in Murphy, 1976, p. 259). En effet, le problème aujourd'hui selon Murphy est que nos « croyances », traditionnellement dérivées de l'expérience, sont « de plus en plus colorées par ce que nous lisons » (Murphy, 1976, p. 259). Or, les éditeurs « sont influencés dans leur choix de la publication de tel ou tel article par le fait que les conclusions de ces articles sont positives ou négatives » (Murphy, 1976, p. 259). Ainsi, les articles où aucune association n'est démontrée, c'est-à-dire où l'hypothèse nulle n'est pas rejetée, ont beaucoup moins de chances d'être publiés. D'ailleurs la majorité des investigateurs qui n'ont trouvé aucune association ne vont même pas prendre la peine de communiquer leurs résultats. Murphy ajoute que « ce biais qui consiste à ne publier que les résultats positifs est plus marqué dans les revues prestigieuses qui circulent le plus » (Murphy, 1976, p. 260). Cela produit même une forme de « cercle vicieux du biais » :

> « Tous ces facteurs conspirent à déformer la croyance générale à propos de ce qui a été démontré dans un cas particulier et, dans la mesure où une telle croyance constitue le fondement de conviction préalables utilisées pour juger des futures publications, un cercle presque clos du biais peut se perpétuer » (Murphy, 1976, p. 260).

Autrement dit, ce à quoi nous invite ici Murphy, et ce à quoi il invite tous les étudiants en médecine et donc les médecins, est une forme de scepticisme radical quant aux connaissances que nous acquérons non seulement à travers des études observationnelles ou expérimentales, mais aussi celles que nous acquérons à travers la lecture des articles publiés dans les grande revues médicales et scientifiques, connaissances qui sont pourtant censées servir de prémisses logiques aux raisonnements des médecins quand ils font un diagnostic, un pronostic ou prescrivent une thérapeutique. La fonction opératoire du biais est ici éminemment critique : il s'agit de mettre en garde les médecins contre toutes les formes d'erreurs, volontaires ou involontaires, mineures ou majeures, qualitatives ou quantitatives, qui sont susceptibles de vicier le raisonnement du médecin et par conséquent de lui faire commettre lui-même des erreurs. C'est pourquoi l'objectif de son ouvrage, réaffirmé dans sa conclusion, consiste à « cultiver un sens commun alerte » (Murphy, 1976, p. 290), c'est-à-dire finalement à cultiver une forme de scepticisme qui doit nous conduire à passer en permanence nos connaissances, anciennes ou nouvelles, à l'examen

(« *skepsis* » en grec). Cela montre que Murphy ne se place pas dans la perspective de l'épidémiologiste, ni non plus dans celle du médecin clinicien qui exerce au lit du patient : il s'agit bien du point de vue du médecin-scientifique qui doit être capable de déceler toutes les failles possibles d'une étude expérimentale ou observationnelle, aussi bien avant la réalisation de l'étude que pendant ou même après l'étude. En d'autres termes, il s'agit pour le médecin d'exercer son esprit critique et une sorte de doute méthodique et permanent, notions auxquelles Murphy ne cesse de faire référence. En somme, le médecin-scientifique, tel qu'il est décrit par Murphy, est passé de l'*examen clinique* (du patient) à l'*examen critique* (d'articles).

5.2 La redéfinition du concept de biais en épidémiologie :

5.2.1 La « Conférence de paix des Bermudes » et la situation épistémologique de l'épidémiologie à la fin des années 1970.

En avril 1978 a lieu aux Bermudes un symposium consacré aux études cas-témoins en épidémiologie, sponsorisé notamment par le *Journal of Chronic Diseases*, qui en fera un numéro spécial (le volume 32), et d'où sera tiré un ouvrage[55], intitulé *L'étude cas-témoins : consensus et controverses*, qui reprend l'essentiel des communications et des discussions qui y ont eu lieu. Cet ouvrage constitue un document historique extrêmement intéressant à plus d'un titre, pour l'histoire de l'épidémiologie en général, et pour notre sujet en particulier, tant le concept de biais est central dans la controverse autour des études cas-témoins.

Intéressant, cet ouvrage l'est d'abord du point de vue de l'historien et du philosophe des sciences, en montrant comment, à force de communications, de commentaires et de discussions, les spécialistes d'une discipline, en l'espèce l'épidémiologie, parviennent par-delà les controverses à parvenir à une forme de consensus, qui porte ici sur le développement de méthodes et de standards propres aux études cas-témoins. Ce document pourrait sans doute constituer une étude de cas à part entière pour voir et montrer la « science en train de se faire », selon l'expression en vogue dans la sociologie des sciences.

[55] Ibrahim, Michel A., *The Case-Control Study: Consensus and Controversy*, New York, Pergamon Press, 1979.

Intéressant, cet ouvrage l'est aussi en ce qu'il oppose deux catégories ou plutôt deux générations d'épidémiologistes dont les parcours, passés et futurs, comme les objectifs poursuivis, diffèrent voire s'opposent : d'un côté, ceux que David Sackett appelle les «épidémiologistes avec un grand E »[56] comme Jerome Cornfield, Abraham Lilienfeld, Philip Sartwell, pour ne citer que ceux dont nous avons étudié les écrits ; de l'autre, la jeune garde de l'épidémiologie, incarnée par Alvan Feinstein, Ralph Horwitz et bien sûr David Sackett qui sont plutôt partisans (et fondateurs) de l'épidémiologie clinique, c'est-à-dire d'une épidémiologie tournée non vers des populations mais vers les patients individuels, d'une épidémiologie qui se préoccupe donc moins de santé publique que de médecine clinique. Plus précisément, la ligne de fracture entre ces deux camps – et le mot de Lilienfeld qui qualifie, dans la discussion finale, ce symposium de « *Bermuda Summit Peace Conference* » (Ibrahim, 1979, p. 137) atteste de la violence de la controverse, qui semble avoir commencé avec la publication d'un article de Feinstein de 1973[57]– porte en réalité sur la validité des études cas-témoins et par là sur la scientificité même de l'épidémiologie. Feinstein répète ainsi sa critique de l'étude cas-témoins dans la conférence, intitulée « Methodologic problems and standards in case-control research »[58], qu'il prononce à l'occasion de ce symposium:

> « Dans des domaines scientifiques tels que la physique, la chimie et la biologie, des standards rigoureux ont été développés pour prouver l'assertion selon laquelle un effet particulier se produit en conséquence d'une cause particulière. (…). Dans le domaine de la médecine clinique, ces standards ont été appliqués à diverses expérimentations conduites pour clarifier une relation de cause à effet (…). En épidémiologie, néanmoins, le genre particulier d'investigation de la relation de cause à effet que l'on appelle « l'étude cas-témoins » n'est pas conduite comme une expérimentation et son efficacité [« *performance* »] aussi bien que son interprétation ne sont pas évaluées selon des standards scientifiques analogues [à la physique, la chimie, la biologie ou la médecine clinique] ». (Feinstein, 1976, p. 35)

[56] Voir l'interview- testament que Sackett a donnée en 2014-2015 (il meurt le 13 mai 2015) : Haynes, R. Brian et Sackett, David L., « An interview with David Sackett, 2014–2015 », 2015, p. 1-104. Pour les passages concernant les « *'big-E' epidemiologists* », voir notamment les pages 19, 20, 26.

[57] Feinstein, Alvan R., « Clinical biostatistics. XX. The epidemiologic trohoc, the ablative risk ratio, and "retrospective" research », *Clinical Pharmacology and Therapeutics*, vol. 14 / 2, avril 1973, p. 291-307.

[58] Feinstein, Alvan R « Methodologic problems and standards in case-control research », *Journal of Chronic Diseases*, vol. 32 / 1-2, 1979, p. 35-41.

En somme, selon Feinstein, l'épidémiologie manque de rigueur scientifique et ne se conforme pas aux critères scientifiques et épistémologiques en vigueur en fondant ses connaissances sur les études cas-témoins, dans la mesure où la méthodologie de ces études, parce qu'elle s'appuierait uniquement sur la « pratique traditionnelle des épidémiologistes » (Feinstein, 1976, p. 35) n'est pas conforme aux standards de la science, comme peut l'être celle de l'expérimentation ou l'essai clinique randomisé. Ainsi, presque vingt ans après la controverse autour de la notion de preuve en épidémiologie ou de la possibilité d'une inférence causale à partir des études observationnelles, controverse qui avait déjà mobilisé un certain nombre de participants à ce symposium comme Lilienfeld ou Sartwell entre la fin des années 1950 et le début des années 1960 dans le contexte du rôle du tabagisme dans le cancer du poumon où la cible était essentiellement les arguments de Joseph Berkson, le statut scientifique de l'épidémiologie et le statut épistémologique des études épidémiologiques apparait toujours menacé. Néanmoins, le contexte a changé : précisément c'est parce que l'épidémiologie, comme les études cas-témoins, est reconnue comme une discipline scientifique à part entière et que ceux qui décident des politiques de santé publique s'appuient sur les résultats qu'elle produit que son statut scientifique pose problème et est remis en cause si jamais ses résultats s'avèrent faux. Walter Spitzer[59], dans la préface de l'ouvrage, souligne ainsi les faits suivants :

> « Les résultats des études cas-témoins ont reçu une attention grandissante durant les deux ou trois dernières décennies non seulement de la part de la communauté scientifique, mais aussi des divers secteurs de la société qui ont

[59] Pour montrer la continuité entre la controverse des années 1950-1960 et celle de la fin des années 1970, il n'est pas anodin de constater que Walter Spitzer fait partie des auteurs, comme Sackett, qui ont démontré empiriquement pour la première fois, en 1978, le fameux biais de Berkson lié à la population hospitalière. Voir Roberts, Robin S., Spitzer, Walter O., Delmore, Terry, Sackett, David L., « An empirical demonstration of Berkson's bias », *Journal of Chronic Diseases*, vol. 31 / 2, 1978, p. 119–128. C'est aussi Spitzer qui écrit la nécrologie d'Alvan Feintein en 2002 dans le *Journal of Epidemiology and Community Health.* Il l'a en effet rencontré en 1968 à la faculté de médecine de Yale et ils ont été longtemps (de 1976 à 1995) co-éditeurs du *Journal of Chronic Diseases*, qu'ils renommèrent en 1988 *Journal of Clinical Epidemiology*. En 2009, c'est Sackett qui écrira la nécrologie de Spitzer dans cette même revue : voir Sackett, David L., « Walter O. Spitzer 1937–2006 », *Journal of Clinical Epidemiology*, vol. 62 / 6, juin 2009, p. 565-566. De même, dans le prologue à son ouvrage de 1985 consacré à l'épidémiologie clinique (Feinstein, Alvan R., *Clinical epidemiology. The architecture of clinical research.*, Philadelphia, PA, WB Saunders, 1985, p. XI), Feinstein remercie non seulement Spitzer et Sackett, ce qui n'est pas surprenant, mais aussi, ce qui l'est plus, Donald Mainland, qui, comme nous l'avons vu n'a cessé de vulgariser le biais de Berkson mais a été aussi celui qui a théorisé le plus précisément le concept de biais dans les années 1950 et qui est à l'origine de la tripartition : cause, hasard, biais.

été affectés par ces résultats de différentes manières. Les stratégies de l'étude cas-témoins ont été entourées de controverses. Ses défenseurs comme ses détracteurs ont exposé de façon ferme leurs opinions divergentes, ce qui a créé un dilemme pour ceux qui font comme pour ceux qui consomment ces recherches. » (Spitzer, Walter O., « Preface », *Journal of Chronic Diseases*, vol. 32 / 1-2, 1979, p. 1)[60].

De même, dans l'article qui conclut l'ouvrage[61], Ibrahim et Spitzer soulignent les « implications politiques profondes » (Ibrahim et Spitzer, 1979, p. 140) qu'ont désormais les études cas-témoins, mais aussi « l'atmosphère intellectuelle de scepticisme » (Ibrahim et Spitzer, 1979, p. 140) qui règne alors au sein de la société en général et de la communauté scientifique en particulier, et qui « empêche d'accepter les résultats des étude sans procéder à l'évaluation critique des méthodes utilisées » (Ibrahim et Spitzer, 1979, p. 140).

Dès lors, cette conférence de paix apparait comme décisive quant à l'avenir non seulement des études cas-témoin, mais aussi de l'épidémiologie : il est urgent pour les épidémiologistes d'élaborer un consensus sur les forces et les faiblesses de ce type d'étude et de se mettre d'accord sur les « priorités méthodologiques pour le futur » (Spitzer, 1979, p. 1)

Intéressant, enfin, cet ouvrage l'est directement pour notre sujet car c'est ici qu'apparait l'article de David Sackett, intitulé « Bias in analytic research », considéré comme séminal sur la question du biais dans la mesure où la définition du biais qui apparait dans les cinq éditions du *Dictionnaire d'épidémiologie* que nous avons pu consulter s'inspire de celle qui est donnée par Sackett dans cet article, à savoir :

[60] Il est probable que Spitzer fasse référence à un certain nombre de controverses qui émaillent l'épidémiologie durant ces années 1970 : il y a d'abord la controverse sur le rôle causal de la réserpine dans le cancer du sein, rapporté par le *Boston Collaborative Drug Surveillance Program*, que M. P. Vessey mentionne d'ailleurs dans le commentaire qu'il fait de l'article de Sackett de 1979 (voir Ibrahim, 1979, p. 64), mais aussi la controverse sur le rôle de l'œstrogène dans le développement du cancer de l'endomètre, où Horwitz et Feinstein jouent un rôle actif à travers la publication d'une série d'articles. Voir par exemple: Horwitz, Ralph I. et Feinstein, Alvan R., « Alternative Analytic Methods for Case-Control Studies of Estrogens and Endometrial Cancer », *New England Journal of Medicine*, vol. 299 / 20, novembre 1978, p. 1089-1094, ou encore: Horwitz, Ralph I. et Feinstein, Alvan R., « Intravaginal Estrogen Creams and Endometrial Cancer: No Causal Association Found », *JAMA*, vol. 241 / 12, mars 1979, p. 1266.

[61] Ibrahim, M. A. et Spitzer, W. O., « The case control study: the problem and the prospect », *Journal of Chronic Diseases*, vol. 32 / 1-2, 1979, p. 139-144.

> « Tout processus à toute étape de l'inférence qui tend à produire des résultats ou des conclusions qui diffèrent systématiquement de la vérité » (Sackett, 1979, p. 60).

Ainsi dans la deuxième édition du *Dictionnaire d'épidémiologie*, qui parait en 1988, le biais est défini de la façon suivante :

> « Déviation des résultats ou des inférences par rapport à la vérité, ou processus menant à une telle déviation »[62]

La même définition sera reprise jusqu'à la quatrième édition en 2001[63]. Un ajout sera néanmoins effectué dans la cinquième édition, publiée en 2008, qui reprend l'élément de systématicité qui apparait dans la définition donnée par Sackett :

> « Déviation systématique des résultats ou des inférences par rapport à la vérité. Processus menant à une telle déviation »[64]

D'un point de vue historique il est surprenant de constater que la définition du biais qui est restée dans l'histoire et qui forme l'intension actuelle du concept de biais ait été donnée non par une des grandes figures de l'histoire de l'épidémiologie, ou comme dirait Sackett par un « Epidémiologiste avec un grand E », mais par Sackett, qui est certes, en plus d'être docteur en médecine, titulaire d'un Master en Santé Publique de la *Harvard School of Public Health*, mais qui dès 1969 (et en fait dès ses études d'épidémiologie à Harvard[65]) a préféré s'orienter vers l'épidémiologie clinique, qu'il définit comme « l'application par un médecin qui prend directement en charge un patient de méthodes épidémiologiques et biométriques dans l'étude des processus diagnostiques et thérapeutiques dans le but de produire une amélioration de la santé »[66]. De plus, comme nous l'avons vu dans la partie précédente, Sackett emprunte directement cette définition, en la modifiant quelque peu, à celui qu'il appelle « le statisticien Tony Murphy » (en réalité Edmond Murphy, que tout le monde appelait Tony, comme d'ailleurs Austin Bradford Hill), qu'il a rencontré vers 1966-1967, et dont la « prose et la présence » l'ont « captivé » (Sackett, 2015, p. 26). Là aussi il est

---

[62] Last, John M., *A dictionary of epidemiology*, 2ème Edition, New York, Oxford University Press, 1988, p. 13

[63] Last, John M. et International Epidemiological Association (eds), *A dictionary of epidemiology*, 4ème édition, New York, Oxford University Press, 2001, p. 14.

[64] Porta, Miquel, et International Epidemiological Association (eds), *A dictionary of epidemiology*, 5ème edition, Oxford ; New York, Oxford University Press, 2008, p. 18.

[65] Voir Last, 1969, p. 126 et Last, 2015, p. 27, où il souligne à quel point l'épidémiologie traditionnelle est non-pertinente (« *irrelevant* ») et vue avec « dédain » par les étudiants en médecine.

[66] Sackett, D. L., « Clinical epidemiology », *American Journal of Epidemiology*, vol. 89 / 2, février 1969, p. 125-128. La définition est à la page 125.

difficile, malgré son doctorat en statistiques, de considérer Murphy comme un « Epidémiologiste avec un grand E ». Il est temps à présent d'étudier plus précisément ce que Sackett a à dire sur le concept de biais.

5.2.2 Le concept de biais dans l'article de David Sackett : de la définition au catalogue.

La grande nouveauté introduite par Sackett réside essentiellement en fait dans l'appendice à son article : c'est en effet dans cet appendice qu'il va proposer le premier « catalogue de biais » en épidémiologie[67], proposition de catalogue qui fait d'ailleurs consensus parmi les participants à cette conférence et qui fait partie des recommandations méthodologiques à suivre dans le futur[68]. Néanmoins, avant de rentrer dans le détail du catalogue, un autre point mérite une attention particulière : en effet, Sackett, entre la définition et le catalogue des biais, détaille les sept « étapes de la recherche » (Sackett, 1979, p. 60) où un biais est susceptible d'intervenir, étapes qui constituent un « aperçu du catalogue » (« *outline of the catalog* », in Sackett, 1979, p. 60). Sackett reprend ainsi l'approche de Murphy, que nous avons qualifiée de diachronique, sauf que le vocabulaire utilisé n'est pas du tout le même que celui de Murphy, et se rapproche plutôt de celui utilisé par Feinstein, notamment la notion de « manœuvre » que Feinstein utilise dans son ouvrage consacré à l'épidémiologie clinique (Feintein, 1985, notamment le chapitre 4). Quelles sont donc ces sept étapes ?

1) La première étape de la recherche où un biais est susceptible de s'introduire est celle où il s'agit de « *s'informer* » ou de « prendre connaissance du domaine »[69], c'est-à-dire de savoir à propos de telle question ce qu'on sait ou ce qu'on ne sait pas. Sackett distingue cinq biais, qui vont des « biais de rhétorique » qui consiste à persuader plutôt qu'à convaincre au « biais du sujet

[67] Sackett précise dans une note que ce catalogue a été initié par une étudiante en épidémiologie clinique nommé JoAnne Chiavetta, qu'il a bénéficié de la contribution de nombreux collègues, notamment John C. Sinclair, mais aussi de la contribution d'autres publications, spécialement l'ouvrage de Murphy, *The Logic of Medicine*, et celui de Feinstein, *Clinicical Judgment*.

[68] Voir Ibrahim, Michel A., et Spitzer, Walter O., « The case-control study: the problem and the prospect », *Journal of Chronic Diseases*, vol. 32 / 1-2, 1979, p. 139-144. La nécessité d'un tel catalogue est affirmée à la page 144.

[69] « *reading-up the field* », in Sackett, 1979, p. 60. C'est Sackett qui souligne.

brûlant » (« *hot stuff bias* ») en passant par le « biais de la littérature du tout-va-bien » (Sackett, 1979, p. 60-61).

2) La deuxième étape renvoie à « spécifier et sélectionner l'échantillon de l'étude » (Sackett, 1979, p. 61). C'est l'étape qui comporte le plus de risques de biais, avec pas moins de vingt-deux biais possibles recensés : il s'agit ici en réalité de ce qu'on appelle le biais de sélection, avec par exemple le « biais du taux d'admission » c'est-à-dire le biais de Berkson.
3) La troisième consiste à « exécuter la *manœuvre* expérimentale (ou l'exposition) » (Sackett, 1979, p. 62[70]). Sackett compte cinq biais, dont celui de « contamination » ou celui « d'observance » (« *compliance* »).
4) Dans la quatrième il s'agit de « *mesurer* les expositions ou les résultats » (Sackett, 1979, p. 62) de la manœuvre. C'est la deuxième source la plus importante de biais avec treize biais énumérés par Sackett qui concernent aussi bien le biais d'information (Sackett liste le « biais d'obséquiosité », le «biais de mémoire », le « biais d'attention », le biais « d'expectation », etc.) que le biais lié à l'instrument. La mesure renvoie clairement au problème de l'estimation (du risque relatif ou de l'odds ratio) et à la juste mesure d'une association.
5) La cinquième est « *l'analyse des* données » (Sackett, 1979, p. 62[71]) : Sackett recense cinq biais, notamment celui qui consiste à sélectionner un test statistique *après* avoir examiné les données.
6) La sixième consiste logiquement à « interpréter l'analyse » (Sackett, 1979, p. 62) des données. Sackett énumère six biais, comme par exemple le « biais de dissonance cognitive » qui consiste paradoxalement à renforcer une croyance en face de preuves contradictoires avec cette croyance, ou encore le « biais » qui consiste à ne pas épuiser l'espace de l'hypothèse (Sackett, 1979, p. 63), Sackett reprenant explicitement les exemples donnés par Murphy.
7) La dernière étape, enfin, porte sur la « publication des résultats » et « ramène à la première étape » (Sackett, 1979, p. 60)., comme l'avait souligné Murphy. Sackett ne donne aucun exemple de biais pour cette étape.

Dans la mesure où cette liste comporte cinquante-six items au total, il serait fastidieux de tous les noter et de les commenter. Le point intéressant est que les deux

[70] C'est Sackett qui souligne.
[71] C'est Sackett qui souligne.

étapes qui comportent le plus de biais, parmi lesquels Sackett sélectionne neuf biais qu'il étudie en détail dans son article, sont celles de la sélection de l'échantillon et celle de la mesure de l'association (qui renvoie essentiellement aux biais d'information), soit les deux problèmes majeurs qui ont occupé les épidémiologistes depuis les années 1940-1950. Sackett précise d'ailleurs que « les biais de mesure sont plus faciles à empêcher et à mesurer que les biais d'échantillonnage » (Sackett, 1979, p. 58), et que des « stratégies efficaces » pour lutter contre ces biais ont été mises en place comme « des interviews aveugles aux diagnostics des sujets » (Sackett, 1979, p. 58), l'établissement de « critères explicites et objectifs pour les expositions et les résultats » (Sackett, 1979, p. 58), et l'obtention « d'informations sur l'exposition à partir de sources indépendantes qui ne sont pas affectées par la mémoire ou par le flux des informations familiales » (Sackett, 1979, p. 58). Pour les biais d'échantillonnage en revanche, le problème est plus difficile, et peut-être insoluble, comme Berkson l'a montré selon Sackett.

La deuxième grande nouveauté introduite par Sackett dans cet article, en plus du « développement continu d'un catalogue annoté de biais » (Sackett, 1979, p. 59), dont il veut bien prendre la responsabilité et qu'il considère comme la « priorité de recherche numéro un » (Sackett, 1979, p. 59), consiste à illustrer « la magnitude et la direction » des différents biais, ainsi que la « description de mesures préventives appropriées » (Sackett, 1979, p. 59). Ainsi, dans le tableau 11 de son article (Sackett, 1979, p. 57), Sackett, liste les neuf biais qu'il étudie plus précisément dans son article (cinq biais d'échantillonnage et quatre biais de mesure) et montre l'effet produit par ses biais sur les « odds relatifs »[72] (Sackett, 1979, p. 57), à la fois dans les études cas-témoins et dans les études de cohorte. Certains biais produisent ainsi systématiquement une surestimation de l'odds ratio, comme le biais de mémoire par exemple dans une étude cas-témoin, tandis que d'autres biais vont conduire soit à une sous-estimation soit à une surestimation du risque relatif. Certains sont propres aux études cas-témoins et ne s'appliquent pas aux études de cohorte, d'autres sont communs aux deux types d'études.

De même, le tableau 12 (Sackett, 1979, p. 58) de son article va reprendre les mêmes biais et préciser à chaque fois s'ils peuvent être ou non empêchés et s'ils peuvent être ou non mesurés : par exemple, le biais de prévalence-incidence ne peut

[72] Nous supposons que Sackett parle ici de risque relatif, mais il pourrait tout aussi bien s'agir d'odds ratio.

pas être prévenu, et ne peut que partiellement être mesuré dans les études cas-témoins, alors qu'il peut être prévenu et mesuré dans les études de cohorte.

Ces deux tableaux sont symptomatiques selon nous du renversement complet de perspective quant à la question du biais en épidémiologie : en effet, là où il s'agissait pour l'épidémiologiste des années 1950-1960 de parvenir à établir une relation de causalité en se fondant sur son jugement qui combinait les données de son enquête, les connaissances biologiques ou cliniques acquises par ailleurs, et divers tests qui permettaient d'établir – et c'était là l'objet essentiel de l'étude et le travail fondamental de l'épidémiologiste – que l'association n'était pas due à un ou plusieurs biais, biais dont la fonction essentielle consistait à soumettre à la critique les résultats de l'enquête ; à l'inverse, pour l'épidémiologiste de la fin des années 1970, le biais devient un objet d'étude à part entière dont il s'agit de s'occuper en priorité, dans la mesure où précisément le biais est susceptible d'intervenir à n'importe quelle étape de la recherche ou de l'inférence, y compris avant et après l'enquête épidémiologique elle-même, ce qui n'aurait eu aucun sens pour les épidémiologistes des années 1950. D'où la nécessité d'un catalogue de biais, qui permet tout autant de les nommer, et donc par là de les faire exister en tant qu'objet d'étude, de les classer en fonction de l'étape de l'enquête où l'épidémiologiste se situe, ce catalogue servant aussi d'outil pour celui-ci, à l'instar de la « *check-list* » des critères de validité développée par Campbell et Stanley dans leur ouvrage de 1967. De même les deux tableaux remplissent une fonction utilitaire ou instrumentale similaire : il ne s'agit plus alors de les nommer et de les catégoriser mais d'en voir ou d'en montrer les effets, d'être capable de mesurer ces effets et aussi éventuellement d'empêcher que ces biais ne se produisent ou du moins de déterminer s'il est possible de les prévenir et de les empêcher. En somme il s'agit d'être capable d'identifier les biais, et pour cela Sackett pose deux conditions dans la discussion qui succède à son article :

- D'abord il faut « identifier les circonstances dans lesquelles il a produit un réel changement dans l'estimation des odds relatifs » (Ibrahim, 1979, p. 67).
- Ensuite, il faut « fournir une explication convaincante, en termes à la fois méthodologiques et biologiques, de la manière dont il opère » (Ibrahim, 1979, p. 67).

Or, cela ne peut être fait selon Sackett que de façon rétrospective, c'est-à-dire après avoir fait l'étude épidémiologique, quand on apprend que la conclusion qu'on

avait tirée de l'étude est en réalité invalide, ce qui peut prendre un certain temps. C'est en ce sens que Sackett enjoint la communauté épidémiologique de procéder à « l'élucidation empirique de la dynamique et des résultats de ces biais » (Sackett, 1979, p. 59) et fait de cette élucidation empirique la deuxième priorité de recherche. Selon lui, en effet :

> « Les méthodologistes ont pendant trop longtemps ignoré la responsabilité qui leur incombe de mesurer l'occurrence et la magnitude des biais, comme le montrent les trente années qui se sont écoulées entre la description du biais du taux d'admission [le biais de Berkson] et sa première démonstration empirique. Nous avons été justement critiqués pour cela, et nous devons nous mettre au travail » (Sackett, 1979, p. 59).

La fonction opératoire du concept de biais en épidémiologie apparait alors de façon on ne peut plus explicite : il s'agit bien de la principale menace à la validité des études épidémiologiques observationnelles, et donc à la scientificité de l'épidémiologie, en tant que cette menace concerne et questionne directement la méthodologie de ces études et par là l'épistémologie même de la discipline. C'est pourquoi la troisième priorité de recherche spécifiée par Sackett concerne « le développement de standards méthodologiques pour les études cas-témoins » (Sackett, 1979, p. 59), standards ou critères qui existent par exemple pour les essais cliniques randomisés. En effet, le risque est que, suite à l'inflation des études cas-témoins qui caractérise l'épidémiologie des années 1960-1970, la présence de nombreuses études défectueuses et mal conçues conduise à « un rejet intégral de cette approche par une communauté scientifique enflammée » (Sackett, 1979, p. 59). De même, et c'est là la quatrième et dernière priorité de recherche selon Sackett, il est décisif de valider « le rôle propre des études cas-témoins dans les prises de décision en matière de clinique et de prise en charge » (Sackett, 1979, p. 59), c'est-à-dire de spécifier précisément le ou les domaine(s) où une étude cas-témoin est utile et où elle ne l'est pas. Sackett cite notamment Sartwell qui considère qu'elles ne sont pas adaptées à l'évaluation des actions préventives ou thérapeutiques vis-à-vis des maladies, ou encore pour étudier les maladies dont l'incidence est forte et la durée brève, et soutient qu'elles « ne devraient jamais être utilisées pour construire de larges politiques cliniques sans preuves additionnelles provenant d'études analytiques de cohorte » (Sackett, 1979, p. 59).

5.2.3 La tripartition du concept de biais : sélection, information et confusion.

Un dernier point, et non des moindres, mérite enfin d'être examiné quant aux conséquences de cette conférence pour l'histoire de l'épidémiologie, qui apparait dans son article conclusif, co-écrit par Michael Ibrahim et Walter Spitzer, et intitulé : « L'étude cas-témoin : problèmes et perspectives »[73]. C'est en effet dans cet article qu'apparait pour la première fois (en tout cas, à notre connaissance) la tripartition du biais en trois classes de biais : biais de sélection, biais d'information et biais de confusion. En effet, si les notions de biais de sélection et de biais d'information sont présentes depuis les années 1950, la notion de confusion (« *confounding* ») a quant à elle été généralement mise à part comme un problème particulier et distinct de celui du biais. Murphy, qui a pourtant sérieusement élargi l'extension du concept de biais, traite ainsi de la question de la confusion dans un chapitre distinct, qui suit celui consacré au biais. Certes, lorsque Mainland fait, dès les années 1950, du concept de biais l'opposé de ceux de hasard et de cause, il considère implicitement que la question de la confusion rentre dans la catégorie de biais, dans la mesure où le biais désigne alors toute explication alternative à celle en termes de causalité ou de hasard qui pourrait expliquer les résultats d'une étude. Pourtant, il s'agit bien là d'un processus d'explicitation, au sens où ce qui était latent devient patent, qui mérite d'être noté au moins d'un point de vue historique et ce d'autant plus que cette catégorisation constitue aujourd'hui un classique des manuels d'épidémiologie et détermine la manière dont les étudiants en épidémiologie et donc les futurs épidémiologistes vont aborder le concept de biais, et le prisme conceptuel au travers duquel ils vont le penser.

Tout d'abord, Ibrahim et Spitzer vont donner une définition du biais, qui n'est pas celle donnée par Sackett ou Murphy, mais la définition de l'édition de 1974 du *Webster's New Collegiate Dictionary*. Un biais y est défini comme :

> « Une erreur systématique introduite dans l'échantillonnage ou dans la mise à l'épreuve (« *testing* ») en sélectionnant ou en encourageant un résultat ou une réponse plutôt que d'autres » (Ibrahim et Spitzer, 1979, p. 141)

[73] Ibrahim, M. A. et Spitzer, W. O., « The case control study: the problem and the prospect », *Journal of Chronic Diseases*, vol. 32 / 1-2, 1979, p. 139-144.

Cette définition, par sa référence à des notions psychologiques comme celle d'encouragement ou celle de réponse nous semble marquée par la psychologie sociale, dont nous avons vu qu'elle avait été la première, dès les années 1950, à faire du concept de biais un concept essentiel en le définissant, en commençant à dresser un catalogue des différents biais, et en en faisant le cœur de la discussion méthodologique, notamment la méthodologie des sondages d'opinion. Ibrahim et Spitzer s'en tiennent alors à une version minimale des sources de biais qui selon eux résident dans la phase du « plan d'expérience et/ou de l'analyse » (Ibrahim et Spitzer, 1979, p. 141) des études cas-témoins. Après l'avoir défini et en avoir donné les sources, Ibrahim et Spitzer vont alors proposer de catégoriser les biais en « trois classes majeures » (Ibrahim et Spitzer, 1979, p. 141), sans pour autant justifier cette catégorisation, ni même prétendre qu'elle fasse consensus parmi les participants à la conférence, comme l'indique une note au bas de la page 139.

La première catégorie est donc celle du biais de sélection, qui est « dû à une sélection différentielle des sujets soit dans le groupe des cas, soit dans le groupe des témoins » et qui produit « une distorsion de la mesure de l'association ou de l'effet » (Ibrahim et Spitzer, 1979, p. 141). Ils donnent alors quatre exemples de biais de sélection :

- Le « biais de *prévalence-incidence*[74] », étudié notamment par Sackett dans son article (Sackett, 1979, p. 51, qu'il appelle aussi le « biais de Neyman ») qui peut être introduit en raison « de facteurs sélectifs de survie parmi les cas prévalents » (Ibrahim et Spitzer, 1979, p. 141).
- Vient ensuite le fameux « biais de *Berkson*[75] », lié à des « facteurs sélectifs dans l'admission à l'hôpital » des patients (Ibrahim et Spitzer, 1979, p. 141).
- Enfin apparait le « biais de *détection* » ou « biais de *diagnostic* »[76], théorisé par Feinstein dans son article de 1974 consacré au rapport entre tabagisme et cancer du poumon[77] et étudié à nouveau lors de sa conférence des Bermudes (Feinstein, 1979, p. 38-39), et qui renvoie à la « détection

[74] Ce sont Ibrahim et Spitzer qui soulignent.
[75] Ce sont Ibrahim et Spitzer qui soulignent.
[76] Ce sont Ibrahim et Spitzer qui soulignent.
[77] Feinstein, Alvan R. et Wells, Carolyn K., « Cigarette smoking and lung cancer: the problems of "detection bias" in epidemiologic rates of disease. », *Transactions of the Association of American Physicians*, vol. 87, 1974, p. 180-185.

sélective des variables d'exposition parmi les cas en raison de la performance des procédures de diagnostic, disproportionnellement haute par rapport à celle des témoins » (Ibrahim et Spitzer, 1979, p. 141).

- Enfin il y a aussi le « biais de *non-réponse* », dû à « des facteurs sélectifs dans le refus de participer à l'étude » (Ibrahim et Spitzer, 1979, p. 141[78]). Les auteurs ajoutent que « toute autre exclusion ou inclusion systématique de sujets dans le groupe de cas ou dans celui des témoins » (Ibrahim et Spitzer, 1979, p. 141) pourrait produire un biais de sélection.

Ces biais sont bien connus des épidémiologistes, tout comme l'importance et la difficulté d'éviter le biais de sélection, ce qui fait dire à Ibrahim et Spitzer que « les biais de sélection constituent les problèmes les plus difficiles à éviter dans les études cas-témoins » (Ibrahim et Spitzer, 1979, p. 141). Si certains peuvent être détectés ou diminués (comme le biais de non-réponse) ce n'est malheureusement pas le cas de tous.

La deuxième catégorie est celle de « biais d'information » (Ibrahim et Spitzer, 1979, p. 142), qu'Ibrahim et Spitzer vont définir comme « une erreur de catégorisation qui peut se produire si l'information qui porte sur la variable d'exposition ou sur l'état pathologique est inconnue ou inexacte [« *inaccurate* »] » (Ibrahim et Spitzer, 1979, p. 142). Ils citent alors comme exemples de biais d'information « la vérification de l'exposition au médicament en se fondant seulement sur l'histoire, la faible remémoration par les témoins des variables d'exposition (et la meilleure remémoration par les cas), et la recherche intensive par les enquêteurs de l'exposition parmi les cas (et beaucoup moins intense parmi les témoins) » (Ibrahim et Spitzer, 1979, p. 142). Tous ces phénomènes peuvent produire des « erreurs sérieuses dans la classification des sujets » (Ibrahim et Spitzer, 1979, p. 142), ce que Mainland, en son temps, avait appelé « *a mislabelling* », c'est-à-dire une « erreur d'étiquetage » (Mainland, 1958a, p. 647).

Enfin, et c'est là le point le plus original de la classification des biais opérée par Ibrahim et Spitzer, les auteurs de l'article vont fusionner la question du biais avec le problème de la confusion, en faisant de la confusion une catégorie à part entière. Le biais de confusion est ainsi défini comme « un *mélange [*« *mixing* »][79] de l'effet de la

[78] Ce sont Ibrahim et Spitzer qui soulignent.
[79] Ce sont Ibrahim et Spitzer qui soulignent.

variable d'exposition avec les effets d'autres variables (confondantes) », mélange qui va « produire une distorsion de la mesure de l'effet ou de l'association » (Ibrahim et Spitzer, 1979, p. 142). Les auteurs citent alors comme facteurs de confusion « l'âge », « le genre, le statut socio-économique, la situation familiale, le dosage et la durée de la prise médicamenteuse, l'observance, le degré de gravité de la maladie, et la comorbidité » (Ibrahim et Spitzer, 1979, p. 142), qui constituent les facteurs de confusion les plus communs, mais dont la liste peut s'avérer très longue. Ibrahim et Spitzer citent alors les outils traditionnels comme l'appariement (« *matching* ») ou plus modernes comme la « stratification » ou « l'analyse multi-variée » pour « isoler l'effet de l'exposition des effets des facteurs confondants » (Ibrahim et Spitzer, 1979, p. 142).

Ainsi, au moment où Murphy et Sackett présentent une taxinomie, que nous avons qualifiée de diachronique, qui renvoie aux différentes étapes de l'inférence, Ibrahim et Spitzer préfèrent quant à eux une taxinomie que nous pourrions, par opposition, qualifier de synchronique. Or, c'est bien cette taxinomie en trois classes qui sera retenue par les épidémiologistes, plutôt que celle proposée par Murphy ou par Sackett : par exemple, Kleinbaum, Kupper, et Morgenstern dans leur ouvrage intitulé *Epidemiologic research. Principles and quantitative methods*[80], publié en 1982, soit quatre ans après la tenue de cette conférence, reprennent la classification en biais de sélection, biais d'information et biais de confusion dans le chapitre 10, consacré à des considérations générales sur la « validité » (Kleinbaum, Kupper et Morgenstern, 1982, p. 183-193). Ils citent d'ailleurs dans la bibliographie de ce chapitre aussi bien l'ouvrage d'Ibrahim consacré à l'étude cas-témoin (Ibrahim, 1979) que l'article de David Sackett (Sackett, 1979) qui y figure, ce qui montre que les auteurs ont bien choisi cette classification en trois catégories, plutôt que l'approche de Sackett. Cette classification se retrouve aussi dans l'ouvrage de Rothman et Greenland, intitulé *Modern Epidemiology*[81], dont la deuxième édition parait en 1998. Dès lors, la question qui se pose est de savoir si le concept épidémiologique de biais (celui d'Ibrahim et Spitzer, de Kleinbaum, Kupper et Morgenstern, ou encore de Rothman et Greenland) est bien le même que le concept médical de biais (celui de Murphy et Sackett) : en

[80] Kleinbaum, David G., Kupper, Lawrence L. et Morgenstern, Hal, *Epidemiologic research. Principles and quantitative methods*, Belmont, CA, Lifetime Learning Publications., 1982.

[81] Rothman, Kenneth J. et Greenland, Sander, *Modern Epidemiology*, 2ème, Philadelphia, PA, Lippincott Williams & Wilkins, 1998. Voir précisément le chapitre 8 (p. 115-134), intitulé "Précision et validité dans les études épidémiologiques"

d'autres termes, si c'est le même mot est-ce bien le même concept ? Plus précisément, la différence de classification traduit-elle une différence de définition du concept de biais, ou bien simplement une différence de perspective ? Autrement dit, la multiplication des définitions du concept de biais et des classifications des différents biais permet-elle de maintenir l'unité du concept de biais ?

5.3 De l'unité du concept de biais.

5.3.1 Une différence de perspective : biais, vérité, validité.

La présentation du texte de Sackett à laquelle nous venons de procéder pourrait laisser penser que la question de la définition du concept de biais est désormais réglée, dans la mesure où ce sera peu ou prou la définition donnée par Sackett qui deviendra la définition standard du biais, que l'on retrouvera, sous des formes quelque peu modifiées, dans tous les dictionnaires d'épidémiologie. Pourtant, et même si cette affirmation est vraie dans une large mesure, le consensus autour de cette définition n'est clairement pas en vigueur au moment où Sackett présente sa définition et son catalogue de biais. Si cela est somme toute logique, les participants ne pouvant pas se mette d'accord sur quelque chose qui n'a pas encore été soumis à la discussion, il n'est sans doute pas inutile pour notre propos de montrer à quel point ce problème du biais, et la question de savoir ce que c'est, comment on le définit, comment on le mesure et comment on le prévient, suscite encore la controverse. De plus il convient de se demander si la définition qui est donnée par Sackett – et la question pourrait tout aussi bien s'appliquer à la définition de Murphy – , en voulant à tout prix intégrer toutes les significations du mot biais qui semblent disponibles à l'époque, et en donnant une liste pléthorique de biais, ne finit pas par dissoudre l'unité ou l'essence même du concept de biais, dont nous avons pu montrer qu'elle s'est construite autour de la notion d'erreur et de la systématicité de l'erreur, au sens où par opposition aux erreurs aléatoires qui finissent par s'annuler, les erreurs systématiques, elles, ne s'annulent pas mais s'accumulent. Surtout, par-delà la définition même du biais, la différence de classification entre par exemple celle de Sackett d'un côté et celle d'Ibrahim et Spitzer de l'autre met à mal l'unité supposée de la définition. Pour le dire autrement, y a-t-il un quelconque rapport entre le « biais de rhétorique » dont parle Sackett dans son article

(Sackett, 1979, p. 60), qui consiste à émouvoir le lecteur de l'article plutôt qu'à le convaincre par des arguments rationnels, et le biais de Berkson, qui renvoie au taux d'admission différentiel des patients à l'hôpital en fonction de leur(s) pathologie(s) ?

Posée ainsi, cette question appelle évidemment une réponse négative et pourtant les choses sont loin d'être aussi simples. En réalité, notre thèse est que la différence d'approche entre par exemple la classification de Sackett et celle de Kleinbaum reflète moins une différence de conception du biais qu'une différence de perspective sur le biais mais aussi sur l'épidémiologie, ce qui justifie la distinction que nous proposons entre un concept médical et un concept épidémiologique de biais. Sackett, comme précédemment Murphy, se place en effet non du point de vue de l'épidémiologiste mais du point de vue du clinicien ou de « l'épidémiologiste clinique » (« *clinical epidemiologist* ») comme il se nomme lui-même (Sackett, 1969, p. 125), et qu'il oppose à « l'épidémiologiste d'enquête » (« *survey epidemiologist* »). Dans cette perspective, l'épidémiologiste clinique, bien que formé aux biostatistiques et à l'épidémiologie « *continue à fournir une prise en charge directe au patient* » (Sackett, 1969, p. 125)[82]. Dès lors, il a besoin pour établir un diagnostic, un pronostic ou décider d'une thérapeutique de pouvoir s'appuyer sur une information ou sur des données solides, c'est-à-dire non seulement valides d'un point de vue formel mais vraies d'un point de vue matériel. Or, cette source de données provient essentiellement des différents articles publiés dans les revues médicales et scientifiques, qui renvoient eux-mêmes à différentes études effectuées par différents enquêteurs selon des méthodologies elles-mêmes différentes et plus ou moins bonnes. Dès lors, il ne s'agit pas pour lui d'effectuer une inférence causale sur le rôle de l'exposition à tel facteur dans le développement de telle maladie, ou sur l'efficacité d'un médicament, mais bien de faire une inférence diagnostique ou thérapeutique à partir de ce qu'il sait ou de ce que les études montrent ou prétendent montrer. C'est pourquoi la classification de Sackett s'intéresse non seulement à l'étude elle-même, par exemple au problème de la sélection de l'échantillon ou de la mesure de l'association, mais aussi à ce qu'il y avant (l'état du sujet) et après l'étude (le problème de la publication), en somme au contexte de l'étude qui détermine d'une certaine manière et la question posée et la réponse apportée. En ce sens, c'est bien à « l'évaluation critique » (« *critical appraisal* ») non seulement de l'étude elle-même, mais aussi de l'article qui en est le

[82] C'est Sackett qui souligne.

résultat, et du contexte dans lequel il est publié, que le clinicien doit procéder, avant et afin de procéder à un diagnostic ou de prendre une décision thérapeutique. En ce sens, le parcours qui mène de l'épidémiologie clinique des années 1960-1970 à l'*Evidence-Based Medicine*, des années 1990, en passant par la médiation de l'évaluation critique (ou de la « lecture critique d'articles », comme l'appelle Murphy) est parfaitement logique : dans l'*Evidence-Based Medicine* en effet, il s'agit de fournir au clinicien les moyens méthodologiques d'évaluer le niveau de preuves fourni par un ou plusieurs articles, comme le montre la définition de l'*Evidence-Based Medicine* :

> « La médecine fondée sur les preuves est l'utilisation consciencieuse, explicite et judicieuse des meilleures preuves disponibles dans la prise de décision en matière de prise en charge de patients individuels. La pratique de la médecine fondée sur les preuves signifie l'intégration de l'expertise clinique individuelle et de la meilleure preuve clinique externe disponible grâce à la recherche systématique »[83].

Ainsi les « preuves » que pourrait apporter les études épidémiologiques sont soumises en dernier ressort à leur utilisation par le clinicien dans un contexte de prise en charge d'un patient. Il est donc essentiel, et même vital, que lesdites preuves soient vraies, c'est-à-dire correspondent à la réalité, par exemple à la réalité de l'efficacité d'un traitement médicamenteux.

A l'inverse, la perspective de « l'épidémiologiste d'enquête », pour reprendre la distinction de Sackett, est tout entière tournée vers l'étude épidémiologique elle-même : il s'agit essentiellement pour lui de s'assurer que son étude est valide, c'est-à-dire qu'elle démontre ce qu'elle prétend démontrer, ce qui ne peut certes jamais être garanti, mais ce que le respect de principes et de critères méthodologiques dans l'enquête, et donc le respect d'une certaine procédure, ont pour fonction précisément d'assurer. En ce sens plus qu'une conception sémantique de la vérité, au sens d'une adéquation entre les énoncés et la réalité, l'épidémiologiste a essentiellement besoin d'une conception syntaxique de la vérité, qui s'appuie sur la cohérence des énoncés entre eux, mais aussi la cohérence de ces énoncés avec ce que l'épidémiologiste sait par ailleurs au plan biologique ou scientifique en général. D'où selon nous,

[83] « *Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research*. » in Sackett, D. L., Rosenberg, W. M., Gray, J. A. *et al.*, « Evidence based medicine: what it is and what it isn't », *British Medical Journal, vol*. 312 / 7023, janvier 1996, p. 71-72.

l'importance de la notion de validité de l'étude, et plus spécifiquement de la notion de validité interne, qui est en épidémiologie l'exacte antithèse de celle de biais.

5.3.2 Le biais et la question de la validité interne

En effet, les discussions sur le concept de biais vont à partir des années 1980, et jusqu'à aujourd'hui, se focaliser sur la question de la validité, les deux concepts – validité et biais – étant considérés comme antithétiques. De fait, dès la conférence des Bermudes de 1978 (et l'on pourrait remonter au moins jusqu'à Berkson, comme le soulignent Kleinbaum Kleinbaum, Kupper et Morgenstern[84]), la question qui est posée à propos des biais en épidémiologie est de savoir si la multitude de biais qui peut potentiellement affecter une étude épidémiologique « constitue une barrière insurmontable à des découvertes valides » (Ibrahim et Spitzer, 1979, p. 144). Les auteurs de l'article concluent en soulignant que cette « barrière est considérable [« *formidable* »] mais pas insurmontable » (Ibrahim et Spitzer, 1979, p. 144), et que le scepticisme à l'égard des études cas-témoins qui a cours à l'époque vient plutôt d'un « usage inapproprié de l'étude cas-témoin », ce qui ne constitue néanmoins pas « une raison suffisante pour abandonner ce type d'enquête » (Ibrahim et Spitzer, 1979, p. 144). Cette référence à la validité des découvertes montre bien qu'en dernier ressort la question du biais en épidémiologie renvoie à la question de la structure logique de l'étude, à sa méthodologie, donc plutôt à ce qu'on pourrait appeler la modalité *de dicto* de la vérité qu'à la modalité *de re* de la vérité, ou, en d'autres termes, à son aspect formel ou syntaxique plutôt qu'à son aspect matériel ou sémantique.

D'ailleurs dans les manuels d'épidémiologie qui fleurissent dans les années 1980 (la première édition du livre de Rothman, *Modern Epidemiology,* parait en 1986), la notion de validité est opposée à celle de précision dans le cadre d'un problème qui porte non pas sur la vérité mais sur la mesure et donc sur l'estimation d'un effet via les estimateurs que sont le risque relatif ou l'odds ratio. Kleinbaum, Kupper et Morgenstern distinguent ainsi deux types d'exactitude (« *accuracy* ») ou plutôt deux sources d'inexactitude. L'erreur aléatoire va alors poser un problème de précision (« *precision* ») alors que l'erreur systématique va poser un problème de validité (Kleinbaum, Kupper et Morgenstern, 1982, p. 185). Ainsi, l'erreur aléatoire « apparait

[84] Kleinbaum, Kupper et Morgenstern, 1982, p. 185.

comme une différence entre l'estimation calculée à partir des données de l'étude et le paramètre qui est en réalité estimé » tandis que l'erreur systématique « se produit quand il y a une différence entre ce que l'estimateur estime en réalité et la vraie mesure de l'effet étudié » (Kleinbaum, Kupper et Morgenstern, 1982, p. 185-186). Pour clarifier cette idée, les auteurs proposent alors une analogie avec une cible :

> « Le problème de la validité consiste à savoir si oui ou non on vise la bonne cible ; celui de la précision concerne la variation entre les différents tirs d'un même tireur, étant donné la cible qui est effectivement visée » (Kleinbaum, Kupper et Morgenstern, 1982, p. 185)

Cette validité se dédouble elle-même en « validité interne » et « validité externe », en fonction de la population sur laquelle porte l'inférence. Kleinbaum, Kupper et Morgenstern distinguent ainsi trois types de population :

- La population étudiée : il s'agit de « l'ensemble des individus à partir desquels les données de l'étude ont été obtenues » (Kleinbaum, Kupper et Morgenstern, 1982, p. 187).
- La population-cible : il s'agit de « l'ensemble des individus qui intéressent spécialement l'enquêteur [« *of restricted interest* »], à sein duquel il va sélectionner son échantillon (…) et à propos duquel il souhaite faire une inférence statistique en rapport avec l'objectif de l'étude » (Kleinbaum, Kupper et Morgenstern, 1982, p. 187).
- La population externe : il s'agit de « l'ensemble des individus que n'a pas étudiés spécialement l'enquêteur [« *to which the study has not been restricted* »] mais à propos desquels il souhaite tout de même généraliser les découvertes de l'étude » (Kleinbaum, Kupper et Morgenstern, 1982, p. 187).

Validité interne et externe se définissent alors de la façon suivante :

> « La *validité interne* concerne la validité des inférences qui portent sur la population-cible, en utilisant des informations tirées de la population étudiée. (…). Le terme *validité externe*, quant à lui, concerne les inférences faites à propos de la population externe, donc au-delà de l'intérêt restreint de l'étude » (Kleinbaum, Kupper et Morgenstern, 1982, p. 187).[85]

[85] Rothman, Greenland et Lash, 1998, reprennent la même distinction.

En d'autres termes, nous retrouvons ici l'opposition théorisée par Campbell et Stanley[86] entre validité interne et validité externe, qui définissaient la validité interne comme « le minimum exigé pour qu'une expérience soit interprétable : les traitements expérimentaux font-ils une différence dans ce cas expérimental spécifique ? » et considéraient que la validité externe posait « la question de la généralisabilité : à quels populations, paramètres, variables dépendantes, mesures, l'effet observé peut-il être généralisé ? »[87].

Pourtant, l'objectif d'une étude épidémiologique n'est évidemment pas tout à fait comparable à celui d'une étude en matière d'éducation car il s'agit ici de mesurer l'effet d'une exposition à un facteur, ou bien celui d'un médicament. Le terme de validité, opposé à celui de précision, dont les négatifs sont les notions d'erreur systématique et d'erreur aléatoire renvoie ainsi, dans un contexte où il s'agit de donner une estimation, directement à la notion non pas d'inférence mais à celle de mesure. C'est pourquoi Rothman, Greenland et Lash définissent le but général de l'étude épidémiologique de la façon suivante :

> « Le but général d'une étude épidémiologique consiste habituellement dans l'exactitude [« *accuracy* »] de l'estimation ».[88]
>
> Plus précisément :
>
> « L'objectif d'une étude épidémiologique est d'obtenir une estimation valide et précise de la fréquence d'une maladie ou de l'effet d'une exposition sur la survenue d'une maladie dans la population-source de l'étude » (Rothman, Greenland et Lash, 2008, p. 128).

Cette question de la mesure est donc centrale dans le projet scientifique de l'épidémiologie et, en ce sens, c'est l'étude épidémiologique qui apparait comme un instrument de mesure, instrument dont la calibration ne peut être assurée précisément que par une attention soutenue à la méthodologie et donc au plan ou à l'architecture même de l'étude : le biais apparait alors comme la principale menace à la validité de la mesure, ce qui nous ramène au problème soulevé par Thomas Bayes dans le cadre de l'astronomie où le mauvais réglage d'un instrument risque de provoquer des erreurs qui vont aller en s'accumulant et en s'aggravant. En somme, l'épidémiologie des

[86] Kleinbaum, Kupper et Morgenstern, font d'ailleurs explicitement référence à l'ouvrage de Campbell et Stanley. Voir Kleinbaum, Kupper et Morgenstern, 1982, p. 44-45.
[87] Campbell et Stanley, 1967, in Leplège, Bizouarn et Coste, 2011, p. 76-77.
[88] Rothman, Kenneth J., Greenland, Sander et Lash, Timothy L., *Modern Epidemiology*, 3ème édition, Lippincott Williams & Wilkins, 2008, p.128

années 1980 opère une sorte de retour aux sources en se focalisant sur la notion de plan d'étude (« *design of experiment* »), phénomène qui se perpétue dans l'épidémiologie actuelle. Aussi Kenneth Rothman peut-il dire en 2007 :

> « Aujourd'hui, nos défis ont l'air plus centrés sur la méthodologie ou la biologie, car nous concentrons toute notre attention à résoudre des biais insolubles ou à expliquer des résultats contradictoires »[89]

Ce soin apporté à la méthodologie des études épidémiologiques et donc à l'épistémologie de l'épidémiologie est selon nous – et c'est la thèse qui est défendue notamment par Mark Parascandola[90] et Olga Amsterdamska[91] – lié au statut épistémologique incertain de l'épidémiologie, dû notamment aux méthodes qu'elle utilise qui sont essentiellement observationnelles et non-expérimentales, et qui serait la source d'une sorte de complexe d'infériorité des épidémiologistes face aux autres scientifiques. La vive controverse du début des années 1960 sur la question de la preuve épidémiologique que nous avons étudiée précédemment[92] et la charge violente de Feinstein en 1979 contre l'absence de standards méthodologiques en matière d'étude cas-témoins attestent de la permanence de ce procès en scientificité (ou en absence de scientificité) que doit subir l'épidémiologie. En retour néanmoins, cette menace permanente de disqualification scientifique a conduit et conduit encore les épidémiologistes à une réflexivité critique extrêmement féconde qui explique pour l'essentiel pourquoi la notion de biais est devenue progressivement un objet d'étude central de l'épidémiologie. C'est pourquoi nous pensons que c'est dans cette discipline que toutes les menaces possibles contre une induction valide ont été le plus développées, par cette attention apportée entre autres au concept de biais, et qu'en ce sens, comme le dit Parascandola, « les chercheurs d'autres disciplines ont deux ou trois choses à apprendre de l'épidémiologie » (Parascandola, 1998, p. 320). En effet, cette question de la validité de la mesure débouche naturellement sur la question de la validité de l'inférence causale, question là aussi éminemment problématique comme nous l'avons montré[93]. Or la leçon que l'on peut tirer de l'épidémiologie et de son

[89] Rothman, K. J., "Commentary: Epidemiology still ascendant", *International Journal of Epidemiology,* 2007; 36, p. 708–10.
[90] Voir par exemple: Parascandola, Mark, « Epidemiology: Second-rate science? », *Public health reports*, vol. 113 / 4, 1998, p. 312-320.
[91] Amsterdamska, Olga, « Demarcating Epidemiology », *Science, Technology, & Human Values*, vol. 30 / 1, janvier 2005, p. 17-51.
[92] Voir la section 4.1.2. du présent travail.
[93] Voir la section 4.3. du présent travail.

histoire est précisément qu'on ne peut jamais être sûr de la validité d'une preuve scientifique ou d'une inférence causale, alors même qu'il est possible d'affirmer la validité d'une étude. Rothman et Greenland disent ainsi, dans un article consacré à la question de la causalité en épidémiologie qui résume assez bien l'attitude scientifique particulière de l'épidémiologiste :

> « De la même manière qu'on ne peut utiliser aucun critère causal pour établir la validité d'une inférence, il n'y aucun critère utilisable pour établir la validité d'une donnée ou d'une preuve. (…) Mais bien qu'il n'y ait aucun critère absolu pour affirmer la validité d'une preuve scientifique, il reste néanmoins possible d'affirmer la validité d'une étude. Ce qui est requis est plus que l'application d'une liste de critères. Au contraire, il faut pratiquer une critique méticuleuse dans le but d'obtenir une évaluation quantifiée de toutes les erreurs qui affectent l'étude »[94]

Ainsi pour Rothman et Greenland, le vrai problème de toute étude scientifique n'est pas de savoir s'il y des erreurs, mais de savoir combien :

> « Le véritable problème est de quantifier l'erreur » (Rothman et Greenland, 2005, p. 150).

Ces deux dernières citations illustrent selon nous le projet central de l'épidémiologie moderne, et son souci constant au cours au cours de son développement : la quantification des biais.

5.3.3 Unité et diversité du concept de biais

Une dernière question reste en suspens : y a-t-il une unité du concept de biais, unité à la fois au sens de continuité de signification au cours de l'histoire, et unité au sens logique au sens où les différents biais et types de biais renverraient tous à une sorte d'essence ou du moins à une communauté de significations ? Autrement dit, comme nous l'avons formulé de façon abrupte, y a-t-il un quelconque rapport entre le « biais de rhétorique » dont parle Sackett dans son article (Sackett, 1979, p. 60), et le biais de Berkson ? En d'autres termes encore, peut-on considérer que ce que nous

[94] Rothman, K. J., et Greenland, S., « Causation and causal inference », *American Journal of Public Health*, 95 (Suppl. 1), 2005, p. 144 - 150. L'extrait se situe à la page 150.

avons appelé le concept épidémiologique et le concept médical de biais sont identiques, partagent des propriétés communes ou bien sont contradictoires ?

A première vue, il semble difficile de concilier les deux approches, l'une consistant à considérer les biais sous la forme de processus discrets qui ne sont pas forcément des erreurs mais tout aussi bien des effets d'un certain phénomène comme par exemple les effets d'un certain langage ou les effets du choix de quelques éditeurs quant à la publication des articles, l'autre se focalisant sur les aspects méthodologiques d'une étude donnée en adoptant une approche par classes (les trois classes de biais que sont la sélection, l'information et la confusion), et insistant plutôt sur la continuité des effets et la perméabilité des classes (un biais de sélection pouvant devenir par exemple un biais de confusion). De même, la définition de ce qu'est un biais varie en fonction de la perspective dans laquelle on se place : le biais chez Sackett ou Murphy affecte directement l'inférence et vient distordre la vérité, tandis que le biais chez Kleinbaum, Kupper et Morgenstern ou bien chez Rothman, Greenland et Lash affecte la validité interne de l'étude. Dans le dernier cas, il s'agit de fournir une estimation de la vraie valeur, et le biais nous en écarte ; dans le premier cas il s'agit de donner des résultats ou des conclusions qui sont vraies en l'absence de biais. Enfin, le biais épidémiologique porte uniquement sur le plan de l'étude, tandis que le biais médical inclut non seulement l'étude mais aussi ce qui vient avant l'étude et ce qui vient après : or, on ne comprend pas bien en quoi par exemple le biais de publication, c'est-à-dire le fait par exemple que les études qui concluent à une association entre un facteur et une maladie aient beaucoup plus de chances d'être publiées que celles qui concluent à l'absence d'association, puisse remettre en cause la validité de l'étude elle-même et donc la vérité de ses conclusions et de ses résultats. En somme, plus les définitions et les catalogues des biais se multiplient, plus il est difficile de concevoir ce qui les réunit sous un même mot.

Ces contradictions disparaissent si l'on prend soin d'examiner dans le détail d'abord l'histoire du concept mais aussi les textes afin de voir ce qui fait l'unité du concept derrière la multiplicité des définitions et des occurrences : à l'origine, en effet, et dans la quasi-totalité des définitions qui jalonnent l'histoire de ce concept, le biais se définit d'abord comme une erreur, erreur dont la propriété essentielle est d'être non pas aléatoire mais systématique. C'est selon nous cette notion d'erreur systématique qui forme le noyau dur du concept, au sens où le concept de biais se désintègre si

l'une de ces deux notions disparait. Plus précisément, c'est la propriété de systématicité qui est intéressante d'un point de vue épistémologique. En effet, si l'on prend par exemple le concept de biais chez Sackett, la notion d'erreur n'apparait pas et est remplacée par celle beaucoup plus vague de « processus » (Sackett, 1979, p. 60), tandis que celle de systématicité est maintenue (« *differ systematically from the truth* »[95]). Mais comment cette systématicité peut-elle être assurée alors même qu'il n'y a pas de rapport entre les différents biais ? Autrement dit, comment un biais de rhétorique, qui ne peut être d'ailleurs considéré comme une erreur, peut-il être ou devenir systématique ? La question n'a tout simplement pas de sens.

La contradiction est surmontée si l'on se place d'un point de vue plus général : le point central en effet dans la présentation que fait Sackett des biais se situe dans les différentes étapes de l'inférence qu'il distingue, chaque étape étant potentiellement une source de biais, et plus précisément dans la relation des étapes entre eux. En effet, à la fin de la septième et dernière étape (le biais de publication) Sackett indique entre crochets : « *and back to (1)* », c'est-à-dire retour à la première étape qui est celle où le chercheur s'informe de l'état du domaine ou du sujet. Il y a donc bien un cercle, et un cercle vicieux de l'erreur, ou comme le dit Murphy « un cercle fermé du biais » (Murphy, 1976, p. 260), le biais de publication influençant les croyances générales des scientifiques sur ce qui a été démontré dans un cas particulier (et qui est peut-être faux), ces croyances formant elles-mêmes « la base des convictions préalables pour juger des futurs travaux publiés » (Murphy, 1976, p. 260), et ainsi de suite. Ainsi c'est la circularité qui produit la systématicité, au sens où comme chez Galton ou Weldon, les erreurs ou les variations (ici, les variations de croyances, croyances potentiellement erronées) vont s'accumuler au fil du temps et sortir des limites au point de créer une nouvelle espèce de croyance, non pas par un processus rationnel mais par un simple processus stochastique assimilable, par analogie, à une sorte de dérive génétique.

C'est pourquoi nous considérons que ce qui fait ultimement l'unité du concept de biais est moins la notion d'erreur que celle de systématicité, à la fois au sens logique et au sens historique, dans la mesure où il permet d'assurer la continuité entre l'idée du biais et le mot, puis entre le mot et le concept de biais, entre ses origines qui se situent dans le calcul des probabilités au XVIIIe siècle, chez Thomas Bayes ou Thomas

[95] Sackett, 1979, p. 60. Nous soulignons.

Simpson, où apparait l'idée; l'application de ce calcul des probabilités à la théorie de l'évolution chez Galton ou Weldon, où apparait le mot; jusqu'à l'application des statistiques à la médecine via l'épidémiologie, mais aussi aux sciences sociales, où est théorisé le concept de biais, conçu ainsi comme une variation ou une erreur systématique.

CONCLUSION

Au terme de ce parcours historique de plus d'un siècle pour comprendre ce que signifie le concept de biais et sa fonction opératoire en épidémiologie, il apparait que cette fonction, si elle a varié en fonction des contextes disciplinaires (théorie de l'évolution, statistiques, épidémiologie), consiste essentiellement en une fonction critique qui porte sur la scientificité même de l'épidémiologie. En effet, ce concept, qui a d'abord un sens psychologique trivial de préjugé, devient progressivement un concept scientifique au sein de l'épidémiologie, au fur et à mesure d'ailleurs que l'épidémiologie devient elle-même une discipline scientifique. Nous pouvons distinguer quatre significations différentes de ce concept, ou plutôt quatre opérations permises par ce concept, qui jalonnent l'histoire de sa formation et de sa déformation qui sont évidemment étroitement imbriquées en pratique mais qu'il nous faut distinguer :

- D'abord, le concept de biais, en son sens psychologique de préjugé, de préférence ou d'idiosyncrasie, apparait comme l'antithèse de la notion d'objectivité. Particulièrement étudiée dans les sciences sociales des années 1950 à travers la question des sondages d'opinion (et ce sont ces sciences sociales qui font du biais un concept scientifique), cette acception du mot prend une importance capitale en épidémiologie dans la mesure où les études épidémiologiques font intervenir dans leur architecture de nombreux acteurs (patients, médecins, intervieweurs, expérimentateur, investigateur) qui sont tous influençables. Ce n'est pas sans rappeler ce que Paolo Vineis appelle dans son article consacré à l'histoire du biais[1], le « préjugé de l'observateur » qui inclut « l'influence de la théorie sur l'observation » (Vineis, 2002, p. 156). En fait dans le plan d'une étude épidémiologique, qui est essentiellement observationnelle, tous les êtres humains intervenant dans l'expérience sont susceptibles de biaiser l'étude, c'est-à-dire de fausser les résultats. Le plan d'expérience a alors essentiellement pour fonction de garantir une certaine objectivité qu'on peut qualifier de mécanique ou de procédurale, et la lutte contre les biais – et l'on reconnait ici de nombreux biais d'information – apparait essentiellement

[1] Vineis, Paolo, « History of bias », *Sozial-und Präventivmedizin*, vol. 47 / 3, 2002, p. 156–161; repris dans Morabia, Alfredo, *A History of Epidemiologic Methods and Concepts*, Boston, MA, Birkhaüser Verlag, 2004, p. 327-336. La citation est à la page 156.

comme une lutte contre la subjectivité. En ce sens, dans cette acception psychologique, la fonction opératoire du biais consiste à débusquer toute trace de subjectivité afin de garantir l'objectivité de l'épidémiologie.

- Parallèlement, le concept de biais prend un sens statistique avec le développement des statistiques mathématiques sous l'égide de Karl Pearson et de Ronald Fisher : il renvoie alors à une erreur systématique de mesure qui n'est pas seulement lié à l'instrument comme le souligne Vineis (Vineis, 2002, p. 156), mais qui est aussi lié à l'observateur, qui est progressivement d'ailleurs, en astronomie comme en médecine, considéré comme un instrument qu'il faut calibrer, comme le dit Feinstein en 1963. Le biais devient alors, dans le cadre de l'invention par Fisher des plans d'expérience, la propriété d'un estimateur, au même titre que la convergence, l'efficacité ou la robustesse. Un estimateur biaisé est ainsi un estimateur qui ne mesure pas correctement la valeur d'une variable donnée au sens où il n'est pas centré par exemple sur son espérance. Ainsi, un plan d'expérience non randomisé mais organisé de façon systématique (et en fait subjective) par l'expérimentateur risque de produire une estimation qui s'écarte systématiquement de la vraie valeur. C'est ce sens que l'on retrouve dans les manuels d'épidémiologie des années 1980 à travers l'opposition entre la notion de biais et celle de validité, plus spécifiquement celle de validité interne d'une étude : une étude biaisée est ainsi une étude qui ne respecte pas les principes méthodologiques au sens où la mesure de l'association entre deux variables qu'elle produit est susceptible d'être éloignée de la vraie valeur de la mesure de l'association, au sens où elle est susceptible de dire qu'il y a une association alors qu'il n'y en a pas, et inversement.
- Ce sens est à mettre directement en rapport avec une autre acception du concept de biais dont l'antithèse est ici la notion de preuve : en effet, si la notion de preuve épidémiologique a été au centre des controverses sur la scientificité de l'épidémiologie dans les années 1960, cela tient au plan de l'étude qui est essentiellement observationnel, plan observationnel auquel a été et continue d'être opposé le plan expérimental, qui serait le seul plan authentiquement scientifique, du fait de l'appel à la randomisation et de la

possibilité pour l'expérimentateur de manipuler les variables à sa guise. Dans le sillage de l'analyse de Campbell et Stanley, qui définissent le biais comme une menace à la validité du plan, les penseurs de l'épidémiologie clinique et de l'*evidence-based medicine* vont établir une hiérarchisation des plans d'expérience en fonction de leur niveau de preuves, c'est-à-dire *in fine* de la quantité et de la variété de biais potentiels présents dans l'étude. Aussi un essai clinique randomisé est-il plus probant qu'une étude cas-témoins précisément dans la mesure où l'essai clinique randomisé est moins susceptible d'être biaisé par le recours à la randomisation qui permet justement d'éviter le biais de sélection, au double-aveugle ou à un groupe placebo, les deux permettant d'éviter le biais d'information. Biais et preuves sont ainsi inversement proportionnels.

- Enfin, la dernière acception renvoie à l'interprétation qu'il est possible de donner aux résultats d'une étude : le facteur d'exposition est-il causal ou non ? Cette question est évidemment étroitement liée à celle de la validité de l'étude c'est-à-dire à l'absence ou à la minimisation des biais. Néanmoins, elle dépasse aussi cette question de la validité dans la mesure où même une étude avec des preuves valides ne suffit en général pas à affirmer le lien de causalité en soi. Il semble donc falloir s'en remettre au jugement et à l'expérience de l'épidémiologiste – ce qui nous ramène au problème de la subjectivité –, épidémiologiste qui peut néanmoins s'appuyer sur un certain nombre de critères tels que les neuf critères définis par A.B. Hill (force de l'association, convergence des résultats des différentes études, spécificité de l'association, relation temporelle, etc.) qui ont d'ailleurs plus une vocation heuristique et pratique que véritablement critique et théorique. L'opération permise par le concept de biais est alors plus générique : le biais désigne toutes les explications alternatives possibles aux résultats de l'étude autre que l'explication en termes de causalité et de hasard, c'est-à-dire au final l'ensemble des biais possibles, y compris le biais de confusion, qui entre assez tardivement comme une catégorie à part entière de biais.

Ainsi, dès lors que le concept de biais s'oppose tout à la fois à l'objectivité, à la validité de la mesure ou de la preuve, et à la causalité, il est logique qu'il ait fini par être opposé par Sackett au concept même de vérité, et constituer une menace à ce

processus de l'établissement de la vérité qu'est l'inférence, dont chaque étape est désormais menacée par d'innombrables biais.

Cette évolution du concept de biais ne peut néanmoins se comprendre que dans le contexte d'une évolution globale des sciences qui a profondément changé notre manière de concevoir le monde, mais aussi notre manière de concevoir les sciences, évolution qui est liée précisément à ce que les épistémologues du groupe de Bielefeld, entre autres, ont appelé la « révolution probabiliste ». Nous sommes ainsi passés d'une vision strictement déterministe des phénomènes, telle qu'elle est défendue par exemple par Claude Bernard au XIXe siècle, à une vision beaucoup plus indéterministe où le hasard ne renvoie plus seulement à notre ignorance des causes mais à la structure même du monde, où le hasard n'est donc plus seulement subjectif mais objectif[2] : c'est l'univers lui-même qui, pour reprendre l'expression de Popper, apparait comme incertain.

En ce sens, en étudiant le concept de biais, nous avons cherché à montrer les conséquences de la révolution probabiliste dans le domaine de la médecine. Selon nous, en effet, la médecine et l'épidémiologie constituent le lieu où la notion d'erreur est sans doute la plus décisive, parce que vitale : une erreur médicale peut en effet entrainer la maladie ou la mort d'un individu, et une erreur systématique celles de plusieurs individus ou dizaine d'individus. En matière de santé publique, dont l'épidémiologie est l'instrument scientifique par excellence, les enjeux d'une erreur systématique se situent quant à eux à l'échelle d'une population. C'est sans doute la raison principale qui explique cet intérêt pour la notion de biais : pour que l'épidémiologie et la médecine – et cette médecine fondée sur l'épidémiologie qu'est l'épidémiologie clinique – puissent devenir scientifiques, une réflexion approfondie sur la notion d'erreur, qu'elle soit aléatoire ou systématique, était nécessaire, et apparait même comme la condition de possibilité de l'émergence de cette médecine scientifique. La valeur cognitive du concept de biais dans ce cadre est donc non pas de prémunir l'épidémiologiste et le médecin contre le risque d'erreur systématique, mais de lui rappeler en permanence que ce risque existe et qu'il n'est pas compensé mais en réalité augmenté par la masse considérable d'informations qu'il a aujourd'hui

[2] L'opposition entre une approche fréquentiste ou objective des probabilités et une approche subjective (autrement dit, bayésienne) apparait d'ailleurs en toile de fond de notre étude sur le biais et n'est sans doute pas étrangère à l'hésitation permanente entre une approche subjective et une approche objective du biais. Mais cela mériterait une étude particulière qui nous aurait mené hors de notre sujet.

à sa disposition. D'où ce rappel à la vertu philosophique du doute méthodique et du scepticisme, c'est-à-dire l'examen critique des informations et des connaissances, scepticisme néanmoins nécessairement limité dans le temps puisque la suspension du jugement est chose impossible pour le médecin comme pour l'épidémiologiste, la décision et l'action étant impératives.

Néanmoins, si ce travail s'inscrit, toutes proportions gardées, dans la continuité des travaux du groupe de Bielefeld (qui portent la période historique qui va de 1800 à 1930) et d'une certaine manière les prolonge en s'intéressant aux effets de la révolution probabiliste sur la médecine et de l'épidémiologie du XXe siècle, les lecteurs les plus sagaces auront sans doute remarqué que s'il y a continuité historique, il n'en est pas de même au plan épistémologique. En effet, nombre de ces auteurs, comme Lorraine Daston ou Theodore Porter, dans le sillage des travaux de Thomas Kuhn (qui signe d'ailleurs l'article inaugural de l'ouvrage consacré à la révolution probabiliste[3]) défendent une approche de l'épistémologie que l'on qualifie en général d'externaliste, au sens où cette épistémologie s'intéresse plus au contexte de l'activité scientifique, c'est-à-dire aux facteurs psychologiques, sociaux, culturels, institutionnels, politiques ou encore économiques de cette activité scientifique. A l'inverse, nous avons adopté pour notre part une approche internaliste de l'épistémologie en faisant le choix d'étudier un concept – celui de biais – au sein d'une discipline – l'épidémiologie –, et si nous avons sacrifié quelquefois à intégrer des éléments contextuels comme des aspects biographiques, ou institutionnels ou encore sociaux au sens large comme dans la controverse sur le tabagisme et le cancer du poumon, il s'agissait pour nous plutôt d'éclairer d'un point de vue historique certaines relations objectives entre des personnes (en l'espèce des épidémiologistes) ou entre des personnes et des institutions (par exemple entre Berkson et le lobby du tabac) que de réduire la position ou les arguments scientifiques de ces individus à ces relations. Pour le dire autrement, comme nous l'avons souligné auparavant[4], en ce qui concerne Berkson et ses liens avec l'industrie du tabac, il nous semble que ce n'est pas parce que Berkson était lié à l'industrie du tabac qu'il niait toute relation causale entre le tabagisme et le cancer du poumon, mais parce qu'il niait toute relation causale entre le tabagisme et le cancer du poumon qu'il a fini par s'allier avec l'industrie du tabac. Plus précisément, ce qu'il

[3] Kuhn, Thomas S., « What are Scientific Revolutions? » in in Krüger, Lorenz, Daston, Lorraine et Heidelberger, Michael (eds), *The Probabilistic Revolution*, Volume 1: Ideas in History, p. 7-22.
[4] Voir la section 4.1.1. du présent travail.

niait en réalité n'était pas l'association causale entre tabagisme et cancer du poumon, mais la capacité qu'avaient les études épidémiologiques à prouver cette relation causale, preuve que selon lui seule une étude expérimentale (ou éventuellement un tribunal d'experts indépendants qu'il appelait de ses vœux, et qui sera finalement incarné dans le Comité consultatif auprès du *Surgeon General* des Etats-Unis) était à même d'apporter. Ce qui nous a semblé important et même décisif, ce n'est donc pas tant les positions institutionnelles de chacun des acteurs de cette histoire de l'épidémiologie moderne que leurs positions proprement philosophiques et épistémologiques ; et c'est pourquoi nous avons souhaité nous placer sur le terrain des concepts et de l'argumentation proprement dite en montrant comment les épidémiologistes agissaient en véritables scientifiques mais aussi en véritables épistémologues en élaborant des hypothèses, en forgeant des théories et des concepts, et en défendant dans des discussions passionnées et passionnantes leurs thèses, sans jamais d'ailleurs céder à des explications de type psychologique et sociologique. Bien évidemment, sans les revues scientifiques, les colloques internationaux ou les conférences aux Bermudes, ces discussions n'auraient pas pu avoir lieu, et il ne s'agit pas ici de disqualifier ou de critiquer les approches externalistes car nous sommes convaincus que les deux approches, externalistes et internalistes, sont complémentaires, dans la mesure où il s'agit précisément de s'inscrire dans la tradition de l'épistémologie historique qui, tout en donnant l*a priorité* grammaticale à l'aspect proprement philosophique, n'en est pas moins de l'histoire.

Le concept de biais se prête d'ailleurs idéalement à ce type d'approches puisqu'il montre comment l'erreur vient se nicher au cœur même de l'activité scientifique, qu'il s'agisse d'une activité de recueil et d'analyse de données, d'observation ou de mesure de phénomènes, ou d'inférence statistique ou causale, et donc indépendamment du contexte, même si l'aspect psychologique est toujours présent en toile de fond. Il renvoie ainsi à la réflexivité des scientifiques, en l'espèce des épidémiologistes, réflexivité dont nous avons montré qu'elle constituait une caractéristique essentielle des épidémiologistes et même de l'épidémiologie elle-même, dont la scientificité a été et continue d'être contestée, justement parce qu'elle est une science d'observation et qu'en tant que tel le processus même d'observation est toujours susceptible d'être biaisé, c'est-à-dire systématiquement erroné. En ce sens, en étudiant le concept de biais, notre objectif consistait à examiner les conditions

de possibilité de la production scientifique, c'est-à-dire ici les conditions de possibilité épistémologiques (et non sociologiques, ou institutionnelles, etc.) de l'avènement de l'épidémiologie comme science à part entière, la condition de possibilité fondamentale de cet avènement consistant précisément non pas à éliminer les biais mais à les identifier comme objet d'étude principal, à travers un processus de nomination, de définition, de catégorisation, de classification et finalement de quantification de ces biais. En d'autres termes, c'est en se constituant comme « science du biais » que l'épidémiologie a pu, au prix d'une bataille épistémologique toujours renouvelée, devenir une science à part entière et participer par là à l'émergence d'une médecine elle aussi scientifique.

Enfin, un des objectifs de ce travail, bien qu'implicite, consistait à éclairer non seulement l'épistémologie de l'épidémiologie moderne, mais aussi son histoire. En effet, comme le relèvent Joël Coste et Alain Leplège dans un éditorial[5] de la *Revue d'Epidémiologie et de Santé Publique,* intitulé : « Pour l'épistémologie et l'histoire de l'épidémiologie. », « l'épidémiologie ne dispose pas aujourd'hui de manuel d'épistémologie ni même d'ouvrage d'histoire de référence » en raison du « caractère récent de son évolution scientifique » (Coste et Leplège, 2009, p. 317). Si nous ne prétendons pas avoir produit ici ni un manuel d'épistémologie ni un ouvrage d'histoire de référence, nous espérons néanmoins avoir apporté notre modeste contribution à l'histoire et à l'épistémologie de l'épidémiologie.

[5] Coste, Joël et Leplège, Alain « Editorial. Pour l'épistémologie et l'histoire de l'épidémiologie. », *Revue d'Epidémiologie et de Santé Publique*, vol. 57 / 5, octobre 2009, p. 317-318.

BIBLIOGRAPHIE

SOURCES PRIMAIRES

ARTICLES

Aldrich, John, « Correlations genuine and spurious in Pearson and Yule », *Statistical science*, 1995, p. 364–376.

Aoyama, Hirojiro, « On the interviewing bias », *Annals of the Institute of Statistical Mathematics*, vol. 5 / 1, 1953, p. 73–76.

Berkson, Joseph, « Tests of significance considered as evidence », *The Journal of the American Statistical Association*, vol. 37, 1942, p. 325-335.

Berkson, Joseph, « Limitations of the Application of Fourfold Table Analysis to Hospital Data. », *Biometrics Bulletin*, vol. 2, 1946, p. 47-53.

Berkson, Joseph, « The Statistical Study of Association between Smoking and Lung Cancer », *Proceedings of the Staff Meetings. Mayo Clinic*, vol. 30, 1955, p. 319-348.

Berkson, Joseph, « Smoking and Lung Cancer: Some Observations on Two Recent Reports », *Journal of the American Statistical Association*, vol. 53 / 281, mars 1958, p. 28-38.

Berkson, Joseph, « The statistical investigation of smoking and cancer of the lung. », *Proceedings of the Staff Meetings. Mayo Clinic*, vol. 34 / 8, avril 1959, p. 206-224.

Berkson, Joseph, « Mortality and marital status. Reflections on the derivation of etiology from statistics », *American Journal of Public Health and the Nations Health*, vol. 52 / 8, 1962, p. 1318–1329.

Berkson, Joseph, « Smoking and Lung Cancer », *The American Statistician*, vol. 17 / 4, octobre 1963, p. 15-22.

Burch, Paul R.J., « Smoking and Lung Cancer: The Problem of Inferring Cause », *Journal of the Royal Statistical Society. Series A (General)*, vol. 141 / 4, 1977, p. 437-477.

Box, G. E. P., « Statistical design in the study of analytical methods », *Analyst*, vol. 77 / 921, 1952, p. 879–891.

Bross, Irwin D. J., « Statistical criticism », *Cancer*, vol. 13 / 2, mars 1960, p. 394-400.

Campbell, D. T., « Factors relevant to the validity of experiments in social settings », *Psychological Bulletin*, vol. 54 / 4, juillet 1957, p. 297-312.

Campbell, Donald T., « Systematic error on the part of human links in communication systems », *Information and Control*, vol. 1 / 4, 1958, p. 334–369.

Campbell, Donald T. et Ferber, Robert, « Bias in Mail Surveys », *Public Opinion Quarterly*, vol. 13 / 3, 1949, p. 562.

Canguilhem, Georges, « Le concept et la vie », *Revue Philosophique de Louvain*, vol. 64 / 82, 1966, p. 193-223.

Cantril, Hadley, « How Accurate Were the Polls? », *The Public Opinion Quaterly*, vol. 1 / 1, janvier 1937, p. 97-109.

Chevry, Gabriel, « L'Institut international de statistique (I.I.S.) », *Economie et statistique*, vol. 13 / 1, 1970, p. 63-65.

Clausen, John A. et Ford, Robert N., « Controlling Bias in Mail Questionnaires », *Journal of the American Statistical Association*, vol. 42 / 240, décembre 1947, p. 497.

Cochran, W. G., « The Planning of Observational Studies of Human Populations », *Journal of the Royal Statistical Society. Series A (General)*, vol. 128 / 2, 1965, p. 234.

Cornfield, Jerome, « A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix », *Journal of the National Cancer Institute*, vol. 11 / 6, juin 1951, p. 1269-1275.

Cornfield, Jerome, « Principles of Research », *Statistics in Medicine*, vol. 31 / 24, octobre 2012, p. 2760-2768.

Cornfield, Jerome, Haenszel, William, Hammond, Ernst C., Lilienfeld, Abraham, Shimkin, Michale B., et Wynder, Ernst L., « Smoking and lung cancer: recent evidence and a discussion of some questions », *Journal of the National Cancer Institute*, vol. 22 / 1, janvier 1959, p. 173-203.

Cornfield, Jerome et Haenszel, William, « Some Aspects of Retrospective Studies », *Journal of Chronic Diseases*, vol. 11 / 5, mai 1960, p. 523–534.

Crossley, Archibald M., « Straw polls in 1936 », *Public Opinion Quarterly*, vol. 1 / 1, 1937, p. 24–35.

Delgado-Rodriguez, M, « Bias », *Journal of Epidemiology & Community Health*, vol. 58 / 8, août 2004, p. 635-641.

Denoix, P. F., Schwartz, D. et Anguera, G., « L'enquête française sur l'étiologie du cancer broncho-pulmonaire. Analyse détaillée. », *Bulletin De l'Association Francaise Pour l'Etude Du Cancer*, vol. 45 / 1, mars 1958, p. 1-37.

Doll, R., « Mortality in relation to smoking: 50 years' observations on male British

doctors », *British Medical Journal*, vol. 328 / 7455, juin 2004, p. 1519-0.

Doll, Richard, « Etiology of Lung Cancer », *Advances in Cancer Research*, vol. 3, janvier 1955, p. 1-50.

Doll, Richard et Hill, A. Bradford, « Smoking and carcinoma of the lung », *British Medical Journal*, vol. 2 / 4682, 1950, p. 739-748.

Doll, Richard et Hill, A. Bradford, « Tabagisme et cancer du poumon: rapport préliminaire », *Bulletin de l'Organisation mondiale de la Sante*, 1999, p. 185-197.

Doll, Richard et Hill, A. Bradford, « Study of the Aetiology of Carcinoma of the Lung », *British Medical Journal*, vol. 2 / 4797, 1952, p. 1271-1286.

Doll, Richard et Hill, A. Bradford, « The mortality of doctors in relation to their smoking habits », *British Medical Journal*, vol. 1 / 4877, 1954, p. 1451–1455.

Doll, Richard et Hill, A. Bradford, « Lung cancer and other causes of death in relation to smoking », *British Medical Journal*, vol. 2 / 5001, 1956, p. 1071-1081.

Doll, Richard et Hill, A. Bradford, « Mortality in relation to smoking: ten years' observations of British doctors », *British Medical Journal*, vol. 1 / 5395, 1964, p. 1399–1410.

Doll, Richard, Peto, Richard, Wheatley, Keith [*et al.*], « Mortality in relation to smoking: 40 years' observations on male British doctors », *British Medical Journal*, vol. 309 / 6959, 1994, p. 901–911.

Dorn, Harold F., « Some Applications of Biometry in the Collection and Evaluation of Medical Data », *Journal of Chronic Diseases*, vol. 1 / 6, juin 1955, p. 638-664.

Dorn, Harold F., « The mortality of smokers and nonsmokers », *Proc Soc Stat Sect American Statistical Association*, 1958, p. 34–71.

Dorn, Harold F., « Tobacco consumption and mortality from cancer and other diseases », *Public health reports*, vol. 74 / 7, 1959, p. 581-593.

Dubost, J. et Durandin, G., « 1° Opinion publique et attitudes », *L'année psychologique*, vol. 52 / 2, 1952, p. 594–613.

Feinstein, Alvan R., « The basic elements of clinical science », *Journal of Chronic Diseases*, vol. 16 / 11, 1963a, p. 1125–1133.

Feinstein, Alvan R., « Boolean Algebra and Clinical Taxonomy: Analytic Synthesis of the General Spectrum of a Human Disease », *New England Journal of Medicine*, vol. 269 / 18, 1963b, p. 929-938.

Feinstein, Alvan R., « Taxonomy and logic in clinical data », *Annals of the New York*

Academy of Sciences, vol. 161 / 1, 1969, p. 450–459.

Feinstein, Alvan R., « Quality of data in the medical record », *Computers and Biomedical Research*, vol. 3 / 5, 1970, p. 426–435.

Feinstein, Alvan R., « Clinical biostatistics. X. Sources of “transition bias” in cohort statistics. », *Clinical Pharmacology and Therapeutics*, vol. 12 / 4, août 1971, p. 704-721.

Feinstein, Alvan R., « Clinical biostatistics. XX. The epidemiologic trohoc, the ablative risk ratio, and “retrospective” research », *Clinical Pharmacology and Therapeutics*, vol. 14 / 2, avril 1973, p. 291-307.

Feinstein, Alvan R., « Clinical biostatistics. XXX. Biostatistical problems in “compliance bias” », *Clinical Pharmacology and Therapeutics*, vol. 16 / 5, novembre 1974, p. 846-857.

Feinstein, Alvan R « Methodologic problems and standards in case-control research », *Journal of Chronic Diseases*, vol. 32 / 1-2, 1979, p. 35-41.

Feinstein, Alvan R., « An Additional Basic Science for Clinical Medicine: IV. The Development of Clinimetrics », *Annals of Internal Medicine*, vol. 99 / 6, décembre 1983, p. 843-848.

Feinstein, Alvan R. et Horwitz, Ralph I., « A critique of the statistical evidence associating estrogens with endometrial cancer », *Cancer research*, vol. 38 / 11 Part 2, 1978, p. 4001–4005.

Feinstein, Alvan R. et Horwitz, Ralph, I., « Methodologic Standards and Contradictory Results in Case-Control Research », *The American Journal of Medicine*, vol. 66 / 4, avril 1979, p. 556-564.

Feinstein, Alvan R. et Landis, J. Richard, « The role of prognostic stratification in preventing the bias permitted by random allocation of treatment », *Journal of Chronic Diseases*, vol. 29 / 4, 1976, p. 277–284.

Feinstein, Alvan R. et Wells, Carolyn K., « Cigarette smoking and lung cancer: the problems of “detection bias” in epidemiologic rates of disease. », *Transactions of the Association of American Physicians*, vol. 87, 1974, p. 180-185.

Fisher, Ronald A., « Statistical Tests of Agreement between Observation and Hypothesis », *Economica*, juin 1923, p. 139.

Fisher, Ronald A., « Theory of Statistical Estimation », *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 22 / 05, juillet 1925, p. 700.

Fisher, Ronald A., « The Arrangement of Field Experiments », *Journal of the Ministry of*

Agriculture of Great Britain, vol. 33, 1926, p. 503-513.

Fisher, Ronald A., « The Logic of Inductive Inference », *Journal of the Royal Statistical Society*, vol. 98 / 1, 1935, p. 39-82.

Fisher, Ronald A., « Alleged dangers of cigarette-smoking. Letter to the Editor. », *British Medical Journal*, vol. 2, juillet 1957, p. 43.

Fisher, Ronald A., « Dangers of cigarette-smoking », *British Medical Journal*, vol. 2 / 5039, août 1957, p. 297-298.

Fisher, Ronald A., « Lung cancer and cigarettes. Letter to the Editor », *Nature*, vol. 182, juillet 1958, p. 108.

Fisher, Ronald A., « Cancer and Smoking. Letter to the Editor. », *Nature*, vol. 182, août 1958, p. 108.

Fisher, Ronald A., « Cigarettes, Cancer and Statistics », *The Centennial Review*, vol. 2 / 2, 1958, p. 151-166.

Fisher, Ronald A., « Inhaling », in *Smoking: the cancer controversy: some attempts to assess the evidence*, Edinburgh, Oliver and Boyd, 1959, p. 45-47, [En ligne : http://www.med.mcgill.ca/epidemiology/hanley/c609/material/FisherOnSmokingAndCancer.pdf].

Fisher, Ronald A., « Statistical Methods and Scientific Induction », *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 17 / 1, 1955, p. 69-78.

Fisher, Ronald A., « The Nature of Probability », *The Centennial Review*, vol. 2 / 3, Et 1958, p. 261-274.

Fisher, Ronald A., « The place of the design of experiments in the logic of scientific inference », *Sankhyā: The Indian Journal of Statistics, Series A*, 1965, p. 33–38.

Fisher, Ronald A., « Note on Dr. Berkson's criticism of tests of significance », *International Journal of Epidemiology*, vol. 32 / 5, octobre 2003, p. 692-692.

Froggatt, P. et Nevin, N. C., « The "law of ancestral heredity" and the Mendelian-ancestrian controversy in England, 1889-1906 », *Journal of Medical Genetics*, vol. 8 / 1, mars 1971, p. 1-36.

Galton, Francis, « On blood-relationship », *Proceedings of the Royal Society of London*, vol. 20 / 130-138, 1871, p. 394–402.

Galton, Francis, « Typical Laws of Heredity », *Nature*, vol. 15, 1877, p. 492-495, 521-514, 532-533.

Goldsmith, Maurice, « The Government Social Survey », *Nature*, vol. 161 / 4093, avril

1948, p. 573-574.

Greenberg, B. G., « Why Randomize? », *Biometrics*, vol. 7 / 4, décembre 1951, p. 309.

Guyatt, Gordon, Cairns, John, Churchill, David [*et al.*], « Evidence-based medicine: a new approach to teaching the practice of medicine », *Journal of the American Medical Association*, vol. 268 / 17, 1992, p. 2420–2425.

Hammond, E. Cuyler, « The association between smoking habits and death rates », *American Journal of Public Health and the Nations Health*, vol. 48 / 11, 1958, p. 1460–1468.

Hammond, E. Cuyler et Horn, Daniel, « The relationship between human smoking habits and death rates: a follow-up study of 187,766 men. », *Journal of the American Medical Association*, vol. 155 / 15, août 1954, p. 1316-1328.

Hammond, E. Cuyler et Horn, Daniel, « Smoking and death rates-report on forty-four months of follow-up of 187,783 men I. Total mortality », *Journal of the American Medical Association*, vol. 166 / 10, mars 1958, p. 1159-1172.

Herrera, L., « Bias in the allocation of treatments by random numbers », *Science*, vol. 122 / 3174, 1955, p. 828–829.

Hill, A. Bradford, « I.—The aim of the statistical method », *The Lancet*, vol. 229 / 5914, 1937, p. 41–43.

Hill, A. Bradford, « The clinical trial », *British Medical Bulletin*, vol. 7 / 4, 1951, p. 278–282.

Hill, A. Bradford, « Observation and Experiment », *New England Journal of Medicine*, vol. 248 / 24, juin 1953, p. 995-1001.

Hill, A. Bradford (1953b), « Assessment of therapeutic trials », *Transactions of the Medical Society of London*, 68, 132, 1953.

Hill, A. Bradford et Doll, Richard, « Lung Cancer and Tobacco. The BMJ's Questions », *British Medical Journal*, vol. 1 / 4976, mai 1956, p. 1160-1163.

Hill, A. Bradford, « The environment and disease: association or causation? », *Proceedings of the Royal Society of Medicine*, vol. 58 / 5, 1965, p. 295-300.

Hill, A. Bradford et Doll, Richard, « Mortality of British doctors in relation to smoking: Observations on coronary thrombosis », *National Cancer Institute monograph*, janvier 1966, p. 205-268.

Hill, A. Bradford, « Memories of the British streptomycin trial in tuberculosis: the first randomized clinical trial », *Controlled clinical trials*, vol. 11 / 2, 1990, p. 77–79.

Horwitz, Ralph I. et Feinstein, Alvan R., « Alternative Analytic Methods for Case-Control Studies of Estrogens and Endometrial Cancer », *New England Journal of Medicine*, vol. 299 / 20, novembre 1978, p. 1089-1094.

Horwitz, Ralph I. et Feinstein, Alvan R., « Intravaginal Estrogen Creams and Endometrial Cancer: No Causal Association Found », *JAMA*, vol. 241 / 12, mars 1979, p. 1266.

Hyman, Herbert, « Isolation, Measurement, and Control of Interview Effect », *National Opinion Research Center,* University of Chicago, 1953.

Ibrahim, M. A. et Spitzer, W. O., « The case control study: the problem and the prospect », *Journal of Chronic Diseases*, vol. 32 / 1-2, 1979, p. 139-144.

Kish, Leslie, « Some Statistical Problems in Research Design », *American Sociological Review*, vol. 24 / 3, juin 1959, p. 328-338.

Korteweg, R., « The significance of selection in prospective investigations into an association between smoking and lung cancer », *British Journal of Cancer*, vol. 10 / 2, 1956, p. 282.

Kraus, Arthur S., « The Use of Hospital Data In Studying the Association Between a Characteristic And a Disease », *Public Health Reports*, vol. 69 / 12, décembre 1954, p. 1211-1214.

Lasagna, Louis, « The controlled clinical trial: Theory and practice », *Journal of Chronic Diseases*, vol. 1 / 4, 1955, p. 353–367.

Levin, Morton L., « The occurrence of lung cancer in man », *Acta - Unio Internationalis Contra Cancrum*, vol. 9 / 3, 1953, p. 531-541.

Levin, Morton L., « Smoking and cancer: retrospective studies and epidemiological evaluation », *Journal of Chronic Diseases*, vol. 16, 1963, p. 375-381.

Lilienfeld, Abraham M., « Epidemiological methods and inferences in studies of noninfectious diseases: PHR review », *Public health reports*, vol. 72 / 1, 1957, p. 51-60.

Lilienfeld, Abraham M., « The Epidemiologic Method in Cancer Research », *Journal of Chronic Diseases*, vol. 8 / 5, 1958, p. 649–654.

Lilienfeld, Abraham M., « “On the methodology of investigations of etiologic factors in chronic diseases”—some comments », *Journal of Chronic Diseases*, vol. 10 / 1, 1959, p. 41–46.

MacMahon, Brian, « Epidemiologic methods in cancer research. », *The Yale journal of biology and medicine*, vol. 37 / 6, 1965, p. 508-520.

MacMahon, Brian, « Strengths and limitations of epidemiology », in National Research Council (U.S). *The National Research Council in 1979: current issues and studies*, Washington, D.C., National Academy of Sciences, 1979, p. 91-104.

Mainland, Donald, « Problems of Chance in Clinical Work », *British Medical Journal*, vol. 2 / 3943, 1936, p. 221-224.

Mainland, Donald, « Statistics in clinical research: some general principles », *Annals of the New York Academy of Sciences*, vol. 52, 1950, p. 922-930.

Mainland, Donald, « The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease », *American Heart Journal*, vol. 45 / 5, Mai 1953, p. 644-654.

Mainland, Donald, « The rise of experimental statistics and the problems of a medical statistician », *The Yale Journal of Biology and Medicine*, vol. 27 / 1, 1954, p. 1-10.

Mainland, Donald, « Use of Case Records in the Study of Therapy and Other Features in Chronic Disease I. Planning the Survey », *Annals of the Rheumatic Diseases*, vol. 14 / 4, décembre 1955, p. 337-352.

Mainland, Donald, « Safety in numbers », *Circulation*, vol. 16 / 5, 1957, p. 784–790.

Mainland, Donald, « Notes on the planning and evaluation of research, with examples from cardiovascular investigations. Part I », *American Heart Journal*, vol. 55 / 5, 1958, p. 644–655.

Mainland, Donald, « Notes on the planning and evaluation of research, with examples from cardiovascular investigations. Part II », *American Heart Journal*, vol. 55 / 6, 1958, p. 824–837.

Mainland, Donald, « Notes on the planning and evaluation of research, with examples from cardiovascular investigations. Part III », *American Heart Journal*, vol. 55 / 6, 1958, p. 838-850.

Mainland, Donald, « The clinical trial—some difficulties and suggestions », *Journal of Chronic Diseases*, vol. 11 / 5, 1960, p. 484–496.

Mainland, Donald, « The use and misuse of statistics in medical publications », *Clinical Pharmacology and Therapeutics*, vol. 1, août 1960, p. 411-422.

Mainland, Donald et Herrera, Lee, « The risk of biased selection in forward-going surveys with nonprofessional interviewers », *Journal of Chronic Diseases*, vol. 4 / 3, 1956, p. 240–244.

Mantel, Nathan et Haenszel, William, « Statistical aspects of the analysis of data from

retrospective studies of disease », *Journal of the National Cancer Institute*, vol. 22 / 4, 1959, p. 719–748.

Morris, J. N., « Uses of epidemiology », *British Medical Journal*, vol. 2 / 4936, 1955, p. 395-401.

Moser, C. A., « Interview Bias », *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, vol. 19 / 1, 1951, p. 28-40.

Moss, Louis, « The War-time Social Survey », *Public Administration*, vol. 21 / 3, 1943, p. 119-125.

Murphy, Edmond A., « Medical Data and Applied Ethics: Part I », *The Linacre Quarterly*, vol. 36 / 3, 1969, p. 158-164.

Murphy, Edmond A., « Medical Data and Applied Ethics: Part II: The Sources of Data », *The Linacre Quarterly*, vol. 36 / 4, 1969, p. 229-235.

Murphy, Edmond A., « Medical Data and Applied Ethics: Part III: The Interpretation of Evidence », *The Linacre Quarterly*, vol. 36 / 4, 1969, p. 236-241.

Neyman, Jerzy, « On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection », *Journal of the Royal Statistical Society*, vol. 97 / 4, 1934, p. 558-625.

Neyman, Jerzy, « Statistics - Servant of All Sciences », *Science*, vol. 122 / 3166, septembre 1955, p. 401-406.

Neyman, J., Iwaszkiewicz, K. et Kolodziejczyk, St., « Statistical Problems in Agricultural Experimentation », *Supplement to the Journal of the Royal Statistical Society*, vol. 2 / 2, 1935, p. 107.

Pearson, Egon S., « Statistical concepts in the relation to reality », *Journal of the Royal Statistical Society. Series B (Methodological)*, 1955, p. 204–207.

Pearson, Egon S., « Studies in the History of Probability and Statistics. XIV Some Incidents in the Early History of Biometry and Statistics, 1890-94 », *Biometrika*, vol. 52 / 1/2, juin 1965, p. 3-18.

Pearson, K., « Contributions to the Mathematical Theory of Evolution », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 185 / 0, janvier 1894, p. 71-110.

Pearson, Karl, « Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 186, 1895, p. 343-414.

Pearson, Karl, « Mathematical Contributions to the Theory of Evolution. – On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs », *Proceedings of the Royal Society of London (1854-1905)*, vol. 60 / 1, janvier 1896, p. 489-498.

Pearson, Karl, « On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling », in Kotz, Samuel et Johnson, Norman L. (éds.). *Breakthroughs in Statistics*, New York, NY, Springer New York, 1992, p. 11-28.

Pearson, Karl, "Walter Frank Raphael Weldon, 1860-1906", *Biometrika*, 5, p. 1-52, 1906.

Pearson, Karl, « Mathematical contributions to the theory of evolution. XIX. Second supplement to a memoir on skew variation », *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 216, 1916, p. 429–457.

Reuchlin, Maurice, « Utilisation en psychologie de certains plans d'expérience », *L'année psychologique*, vol. 53 / 1, 1953, p. 59-81.

Rice, Stuart A., « Contagious Bias in the Interview: A Methodological Note », *American Journal of Sociology*, vol. 35 / 3, 1929, p. 420–423.

Roberts, Robin S., Spitzer, Walter O., Delmore, Terry, et Sackett, David L., « An empirical demonstration of Berkson's bias », *Journal of Chronic Diseases*, vol. 31 / 2, 1978, p. 119–128.

Rosenthal, Robert, « On the social psychology of the psychological experiment: the experimenter's hypothesis as unintended determinant of experimental results », *American Scientist*, vol. 51 / 2, 1963, p. 268–283.

Rothman, K. J., « Commentary: Epidemiology still ascendant », *Int J Epidemiol* 2007;36, p. 708–10

Rothman, Kenneth J. et Greenland, Sander, « Causation and causal inference in epidemiology », *American Journal of Public Health*, vol. 95 / Supplément 1, 2005, p. S144–S150.

Rubin, Ernest, « Statistical Relationships and Proof in Medicine », *The American Statistician*, vol. 8 / 5, décembre 1954, p. 19-21.

Sackett, David L., « Clinical epidemiology », *American Journal of Epidemiology*, vol. 89 / 2, février 1969, p. 125-128.

Sackett, David L., « Bias in Analytic Research », *Journal of Chronic Diseases*, vol. 32 / 1-2, février 1979, p. 51-63.

Sackett, David L., « Clinical epidemiology: what, who, and whither », *Journal of Clinical Epidemiology*, vol. 55 / 12, 2002, p. 1161–1166.

Sackett, David L., « Walter O. Spitzer 1937–2006 », *Journal of Clinical Epidemiology*, vol. 62 / 6, juin 2009, p. 565-566.

Sackett, David L., Rosenberg, W. M., Gray, J. A. [*et al.*], « Evidence based medicine: what it is and what it isn't », *British Medical Journal, v*ol. 312 / 7023, janvier 1996, p. 71-72.

Sartwell, Philip E., « "On the methodology of investigations of etiologic factors in chronic diseases"—Further comments », *Journal of Chronic Diseases*, vol. 11 / 1, 1960, p. 61–63.

Schwartz, Daniel, « La méthode statistique en médecine: les enquêtes éthiologiques », *Revue de Statistique Appliquée*, vol. 8 / 3, 1960, p. 5–27.

Smith, Harry L. et Hyman, Herbert, « The Biasing Effect of Interviewer Expectations on Survey Result », *Public Opinion Quarterly*, vol. 14 / 3, 1950, p. 491.

Simpson, Thomas, "A letter to the Right Honorable George Earl of Macclesfield, President of the Royal Societyn on the advantage of taking the mean of a number of observations, in practical astronomy.", *Philosophical Transactions of the Royal Society of London,* 49, p. 82-93.

Spitzer, W. O., « The teacher's teacher: a personal tribute to Alvan R Feinstein », *Journal of Epidemiology and Community Health*, vol. 56 / 5, mai 2002, p. 328-329.

Stallones, Reuel A, « The association between tobacco smoking and coronary heart disease », *International Journal of Epidemiology*, vol. 44 / 3, juin 2015, p. 735-743.

Stoetzel, Jean, « Une enquête sur l'opinion publique française », *Revue d'Histoire des Sciences Humaines*, vol. 6 / 1, 2002, p. 155.

Weldon, W. F. R., « The Variations Occurring in Certain Decapod Crustacea.-- I. *Crangon vulgaris* », *Proceedings of the Royal Society of London*, vol. 47 / 286-291, janvier 1889, p. 445-453.

Weldon, W. F. R., « Certain Correlated Variations in *Crangon vulgaris* », *Proceedings of the Royal Society of London*, vol. 51 / 308-314, janvier 1892, p. 1-21.

Weldon, Walter Frank Raphael, « On certain correlated variations in *Carcinus maenas* », *Proceedings of the Royal Society of London*, vol. 54 / 326-330, 1893, p. 318–

329.

White, Colin, « Sampling in medical research », *British Medical Journal*, vol. 2 / 4849, 1953, p. 1284-1288.

White, Colin et Bailar III, J. C., « Retrospective and prospective methods of studying association in medicine », *American Journal of Public Health and the Nations Health*, vol. 46 / 1, 1956, p. 35–44.

Wynder, Ernst L., « Laboratory contributions to the tobacco-cancer problem », *British Medical Journal*, vol. 1 / 5118, 1959, p. 317-352.

Wynder, Ernst L., « An appraisal of the smoking-lung-cancer issue », *New England Journal of Medicine*, vol. 264 / 24, 1961, p. 1235–1240.

Wynder, Ernst L., « Tobacco and health: a review of the history and suggestions for public health policy. », *Public Health Reports*, vol. 103 / 1, 1988, p. 8-18.

Yates, F., « Complex Experiments », *Supplement to the Journal of the Royal Statistical Society*, vol. 2 / 2, 1935, p. 181-247.

Yates, F., « Sir Ronald Fisher and the Design of Experiments », *Biometrics*, vol. 20 / 2, juin 1964, p. 307-321.

Yates, F., « Some examples of biased sampling », *Annals of Eugenics*, vol. 6 / 2, 1935, p. 202–213.

Yerushalmy, Jacob, « Self-selection—a major problem in observational studies », *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, 1972, p. 329–342.

Yerushalmy, Jacob et Palmer, Carroll E., « On the methodology of investigations of etiologic factors in chronic diseases », *Journal of Chronic Diseases*, vol. 10 / 1, 1959, p. 27–40.

Yule, G. Udny, « On the Influence of Bias and of Personal Equation in Statistics of Ill-Defined Qualities », *The Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 36, juillet 1906, p. 325-381.

INTERVIEWS

Haynes, R. Brian et Sackett, David L., « An interview with David Sackett, 2014–2015 », 2015, p. 0-104.

OUVRAGES

Bernard, Claude, *Introduction à l'étude de la médecine expérimentale*, 1ère edition 1865, rééd. Paris, Garnier-Flammarion, 1966

Broadbent, Alex, *Philosophy of epidemiology*, Basingstoke, Hampshire ; New York, Palgrave Macmillan, 2013.

Campbell, Donald Thomas et Stanley, Julian Cecil, *Experimental and quasi-experimental designs for research*, 2. print, Boston, Houghton Mifflin Comp, 1967, 84 p.

Canguilhem, Georges, *Études d'histoire et de philosophie des sciences*, 7. éd. augm., Réimpr, Paris, Vrin, 2002, 430 p., (« Problèmes et controverses »).

Cochran, William Gemmell et Cox, Gertrude M., *Experimental designs*, New York, Wiley, 1950.

Collins, H. M., *Changing order: replication and induction in scientific practice*, London ; Beverly Hills, Sage Publications, 1985, 187 p.

Feinstein, Alvan R., *Clinical epidemiology. The architecture of clinical research.*, Philadelphia, PA, WB Saunders, 1985.

Fisher, Ronald A., *The Design of Experiments*, Edinburgh, Oliver and Boyd, 1935.

Fisher, Ronald A., *Statistical Methods For Research Workers*, 11ème, Edinburgh, Oliver and Boyd, 1950.

Fisher, Ronald A., *Smoking: the cancer controversy: some attempts to assess the evidence*, Oliver and Boyd Edinburgh, 1959.

Gage, N.L (éd.), *Handbook of Research on Teaching*, Chicago, Rand McNally & Company, 1963.

Galton, Francis, *Memories of my Life*, Londres, Methuen and Co, 1908.

Galton, Francis, *Natural Inheritance*, 1ère, Londres, MacMillan, 1889.

General, Surgeon, « Report of the Advisory Committee to the Surgeon General of the Public Health Service », *US Department of Health, Education and Welfare, Public Health Service Publication*, 1964,

[En ligne : https://profiles.nlm.nih.gov/NN/B/B/M/Q/– /nnbbmq.ocr].

Gilovich, Thomas, Griffin, Dale W., Kahneman, Daniel (éds.), *Heuristics and biases: the psychology of intuitive judgment*, Cambridge, U.K. ; New York, Cambridge University Press, 2002.

Greenland, Sander, *Evolution of Epidemiologic Ideas: Annotated Readings and Concepts*, Epidemiology Resources, 1987.

Hill, A. Bradford, *Principles of Medical Statistics*, 5ème, Londres, The Lancet, 1950,

297 p.

Ibrahim, Michel A., *The Case-Control Study: Consensus and Controversy*, New York, Pergamon Press, 1979.

Kleinbaum, David G., Kupper, Lawrence L. et Morgenstern, Hal, *Epidemiologic research. Principles and quantitative methods*, Belmont, CA, Lifetime Learning Publications., 1982.

Kleinbaum, David G., Sullivan, Kevin M. et Barker, Nancy D., *A pocket guide to epidemiology*, New York, Springer, 2007, 281 p.

Kotz, Samuel et Johnson, Norman Lloyd (éds.), *Breakthroughs in statistics*, New York, Springer-Verlag, 1992.

Last, John M., *A dictionary of epidemiology*, 2ème Edition, New York, Oxford University Press, 1988.

Last, John M. et International Epidemiological Association (eds), *A dictionary of epidemiology*, 4ème édition, et, New York, Oxford University Press, 2001.

Lilienfeld, Abraham M. et Stolley, Paul D., *Foundations of epidemiology*, 3ème édition, New York, Oxford University Press, 1994.

MacMahon, Brian, Pugh, Thomas F. et Ipsen, Johannes, *Epidemiologic Methods*, Boston, MA, Little, Brown & Co, 1960.

Mausner, Judith S. et Bahn, Anita K., *Epidemiology: an Introductory text*, Philadelphia, PA: Saunders, 1974, 377 p.

McCall, William, A., *How to experiment in education*, New York, MacMillan, 1923, [En ligne : https://babel.hathitrust.org/cgi/pt?id=mdp.39015062754992;view=1up;seq=1].

Miettinen, Olli S., *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*, New York, John Wiley & Sons, 1985.

Miettinen, Olli S., *Epidemiological Research: Terms and Concepts*, Dordrecht, Springer Netherlands, 2011, [En ligne : http://link.springer.com/10.1007/978-94-007-1171-6].

Murphy, Edmond M., *The Logic of Medicine*, Baltimore, MD, Johns Hopkins University Press, 1976.

Porta, Miquel, et International Epidemiological Association (eds), *A dictionary of epidemiology*, 5ème edition, Oxford ; New York, Oxford University Press, 2008.

Rosenthal, Robert et Rosnow, Ralph L., *Artifacts in Behavioral Research*, New York, Oxford University Press, 2009.

Rothman, Kenneth J. et Greenland, Sander, *Modern Epidemiology*, 2ème, Philadelphia,

PA, Lippincott Williams & Wilkins, 1998.

Rothman, Kenneth J., Greenland, Sander et LASH, Timothy L., *Modern Epidemiology*, 3ème édition, Lippincott Williams & Wilkins, 2008.

Simpson, Thomas, *Miscellaneous Tracts on some Curious, and Very Interesting Subjects in Mechanics, Physical-Astronomy and Speculative Mathematics*, Londres: J. Nourse, 1757.

Weisberg, Herbert I., *Bias and causation: models and judgment for valid comparisons*, Hoboken, N.J, Wiley, 2010, 348 p., (« Wiley series in probability and statistics »).

Yule, George Udny, *An Introduction to the Theory of Statistics*, Londres, C. Griffin and Co., 1911.

SOURCES SECONDAIRES

ARTICLES

Amsterdamska, Olga, « Demarcating Epidemiology », *Science, Technology, & Human Values*, vol. 30 / 1, janvier 2005, p. 17-51.

Armatte, Michel, « La construction des notions d'estimation et de vraisemblance chez Ronald A. Fisher », *Journal de la société française de statistique*, vol. 129 / 1-2, 1988, p. 68-95.

Armitage, P., « Fisher, Bradford Hill, and randomization », *International Journal of Epidemiology*, vol. 32 / 6, décembre 2003, p. 925-928.

Armitage, Peter, « The role of randomization in clinical trials », *Statistics in medicine*, vol. 1 / 4, 1982, p. 345–352.

Armitage, Peter, Doll, Richard, Bodmer, Walter [*et al.*], « Fisher and Bradford Hill: a discussion », *International Journal of Epidemiology*, vol. 32 / 6, décembre 2003, p. 945-948.

Berlivet, Luc, «'Association or Causation?' The Debate on the Scientific Status of Risk Factor Epidemiology, 1947–c. 1965 », *Clio Medica/The Wellcome Series in the History of Medicine*, vol. 75 / 1, 2005, p. 39–74.

Bhopal, Raj, « Paradigms in epidemiology textbooks: in the footsteps of Thomas Kuhn. », *American Journal of Public Health*, vol. 89 / 8, septembre 1999, p. 1162-1165.

Bird, Alexander, « The epistemological function of Hill's criteria », *Preventive Medicine*, vol. 53 / 4-5, octobre 2011, p. 242-245.

Bodmer, W., « RA Fisher, statistician and geneticist extraordinary: a personal view », *International Journal of Epidemiology*, vol. 32 / 6, décembre 2003, p. 938-942.

Bourdieu, « Les doxosophes. », *Minuit*, novembre 1972, p. 26-45.

Bourdieu, Pierre, « L'opinion publique n'existe pas. », *Les temps modernes*, janvier 1973, p. 1292-1309.

Bourdieu, Pierre, « Remarques à propos de la valeur scientifique et des effets politiques des enquêtes d'opinion. », *Pouvoirs, revue française d'études constitutionnelles et politiques*, avril 1985, p. 131-139.

Chalmers, I., « Commentary: The 1944 patulin trial: the first properly controlled multicentre trial conducted under the aegis of the British Medical Research Council », *International Journal of Epidemiology*, vol. 33 / 2, avril 2004, p. 253-260.

Chalmers, I., « Fisher and Bradford Hill: theory and pragmatism? », *International Journal of Epidemiology*, vol. 32 / 6, décembre 2003, p. 922-924.

Chalmers, Iain, « Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments », *International Journal of Epidemiology*, vol. 30, 2001, p. 1156-1164.

Clarke, Mike, « History of evidence synthesis to assess treatment effects: personal reflections on something that is very much alive. », *James Lind Library Bulletin: Commentaries on the history of treatment evaluation*, 2015, [En ligne : http://www.jameslindlibrary.org/articles/history-of-evidence-synthesis-to-assess-treatment-effects-personal-reflections-on-something-that-is-very-much-alive/].

Connelly, James, Hill, Gerry et Millar, Wayne, « " The Great Debate": Smoking, Lung Cancer, and Cancer Epidemiology. », *Canadian Bulletin of Medical History/Bulletin canadien d'histoire de la médecine*, vol. 20 / 1, 2003, p. 367–386.

Coste, Joël et Leplège, Alain « Editorial. Pour l'épistémologie et l'histoire de l'épidémiologie. », *Revue d'Epidémiologie et de Santé Publique*, vol. 57 / 5, octobre 2009, p. 317-318.

Cox, D. R, « Commentary: Smoking and lung cancer: reflections on a pioneering paper », *International Journal of Epidemiology*, vol. 38 / 5, octobre 2009, p. 1192-1193.

Desrosières, Alain, « L'histoire de la statistique comme genre : style d'écriture et usages sociaux », *Genèses*, vol. 39 / 1, 2000, p. 121-137.

Doll, Richard, « Sir Austin Bradford Hill: A personal view of his contribution to epidemiology », *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 1995, p. 155–163.

Doll, Richard, « Fisher and Bradford Hill: their personal impact », *International Journal of Epidemiology*, vol. 32 / 6, décembre 2003, p. 929-931.

Dupâquier, Michel, « William Farr, démographe », *Population (French Edition)*, vol. 39 / 2, mars 1984, p. 339-355.

Fagot-Largeault, Anne, « Les origines de la notion d'essai contrôlé randomisé en médecine », *Les procédures de preuve sous le regard de l'historien des sciences et des techniques (Cahiers d'histoire et de philosophie des sciences)*, 1991, p. 281–300.

Farewell, Vern et Johnson, Anthony, « The first British textbook of medical statistics », 2010, En ligne :

http://www.jameslindlibrary.org/illustrating/articles/farewell-v-johnson-a-2010-the-first-

british-textbook-of-medi.pdf].

Farewell, Vern et Johnson, Tony, « Woods and Russell, Hill, and the emergence of medical statistics », *Statistics in Medicine*, 2010, p. 1459-1476.

Farewell, Vern et Johnson, Anthony, « The origins of Austin Bradford Hill's classic textbook of medical statistics », *Journal of the Royal Society of Medicine*, vol. 105 / 11, 2012, p. 483–489.

Foucault, Michel, « Crise de la médecine ou crise de l'anti-médecine », in Foucault, Michel, *Dits et écrits,* Paris, Gallimard, coll. Quarto, 1994, Tome III, n° 170, p. 40-58.

Gayon, Jean, « L'eugénisme, hier et aujourd'hui », in *Médecine/Sciences*, 15, n°6-7 (juin-juillet 1999), I-VI.

Hacking, Ian, « Equipossibility theories of probability », *The British Journal for the Philosophy of Science*, vol. 22 / 4, 1971, p. 339–355.

Hardy, Anne et Magnello, M. Eileen, « Statistical methods in epidemiology: Karl Pearson, Ronald Ross, Major Greenwood and Austin Bradford Hill, 1900–1945 », in *A*

Harman, Gilbert H., « The Inference to the Best Explanation », *The Philosophical Review*, vol. 74 / 1, janvier 1965, p. 88.

Hart, P. D'Arcy, « A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s », *British Medical Journal*, vol. 319 / 7209, 1999, p. 572-573.

Kaptchuk, Ted J., « Intentional ignorance: a history of blind assessment and placebo controls in medicine », *Bulletin of the History of Medicine*, vol. 72 / 3, 1998, p. 389-433.

Kaptchuk, Ted J., « The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? », *Journal of clinical epidemiology*, vol. 54 / 6, 2001, p. 541–549.

Kaptchuk, Ted J., « Effect of interpretive bias on research evidence », *British Medical Journal*, vol. 326 / 7404, 2003, p. 1453-1455.

Kaptchuk, Ted J. et Kerr, Catherine, E., « Commentary: Unbiased divination, unbiased evidence, and the patulin clinical trial », *International Journal of Epidemiology*, 2004, p. 247-251.

Krieger, Nancy, « Epidemiology and the web of causation: has anyone seen the spider? », *Social Science and Medicine*, vol. 39 / 7, octobre 1994, p. 887-903.

Lilienfeld, David E., « Harold Fred Dorn and the First National Cancer Survey (1937-1939): the founding of modern cancer epidemiology », *American Journal of Public*

Health, vol. 98 / 12, 2008, p. 2150-2158.

Lilienfeld, David E., « Abe and Yak: The Interactions of Abraham M. Lilienfeld and Jacob Yerushalmy in the Development of Modern Epidemiology (1945-1973) », *Epidemiology*, vol. 18 / 4, juillet 2007, p. 507-514.

Magnello, M. Eileen, « Karl Pearson and the Establishment of Mathematical Statistics », *International Statistical Review*, vol. 77 / 1, avril 2009, p. 3-29.

Marcel, Jean-Christophe, « Le premier sondage d'opinion », *Revue d'Histoire des Sciences Humaines*, vol. 6 / 1, 2002, p. 145.

Marks, Harry M., « Confiance et méfiance dans le marché : les statistiques et la recherche clinique (1945-1960) », *Sciences sociales et santé*, vol. 18 / 4, 2000, p. 9-27.

Marks, Harry M., « Rigorous uncertainty: why RA Fisher is important », *International Journal of Epidemiology*, vol. 32 / 6, décembre 2003, p. 932-937.

Morabia, A. et Szklo, M., « Re: Smoking and lung cancer: recent evidence and a discussion of some questions », *International Journal of Epidemiology*, vol. 39 / 6, décembre 2010, p. 1676-1676.

Morabia, Alfredo, « History of the modern epidemiological concept of confounding », *Journal of Epidemiology & Community Health*, vol. 65 / 4, avril 2011, p. 297-300.

Ogien, Albert, « La volonté de quantifier. Conceptions de la mesure de l'activité médicale », *Annales. Histoire, Sciences Sociales*, vol. 55 / 2, 2000, p. 283-312.

Paneth, Nigel, Susser, Ezra et Susser, Mervyn, « Origins and early development of the case-control study: Part 1, Early evolution », *Sozial-und Präventivmedizin*, vol. 47 / 5, 2002, p. 282–288.

Paneth, Nigl, Susser, Ezra et Susser, Mervyn, « Origins and early development of the case-control study: Part 2, The case-control study from Lane-Claypon to 1950 », *Sozial-und Präventivmedizin*, vol. 47 / 6, 2002, p. 359–365.

Parascandola, M., « Commentary: Smoking, birthweight and mortality: Jacob Yerushalmy on self-selection and the pitfalls of causal inference », *International Journal of Epidemiology*, vol. 43 / 5, octobre 2014, p. 1373-1377.

Parascandola, Mark, « Causes, risks, and probabilities: Probabilistic concepts of causation in chronic disease epidemiology », *Preventive Medicine*, vol. 53 / 4-5, octobre 2011, p. 232-234.

Parascandola, Mark, « Epidemiology: Second-rate science? », *Public health reports*, vol. 113 / 4, 1998, p. 312-320.

Parascandola, Mark, « Objectivity and the neutral expert », *Journal of epidemiology and community health*, vol. 57 / 1, 2003, p. 3–4.

Parascandola, Mark, « Skepticism, Statistical Methods, and the Cigarette: A Historical Analysis of a Methodological Debate », *Perspectives in Biology and Medicine*, vol. 47 / 2, 2004, p. 244-260.

Parascandola, Mark, « Two approaches to etiology: the debate over smoking and lung cancer in the 1950s », *Endeavour*, vol. 28 / 2, juin 2004, p. 81-86.

Parascandola, Mark et Weed, D. L., « Causation in epidemiology », *Journal of Epidemiology and Community Health*, vol. 55 / 12, 2001, p. 905–912.

Pearce, Neil, « Commentary: The rise and rise of corporate epidemiology and the narrowing of epidemiology's vision », *International Journal of Epidemiology*, vol. 36 / 4, août 2007, p. 713-717.

Pearce, Neil, « Traditional epidemiology, modern epidemiology, and public health. », *American Journal of Public Health*, vol. 86 / 5, 1996, p. 678–683.

Poole, Charles, « On the Origin of Risk Relativism », *Epidemiology*, vol. 21 / 1, janvier 2010, p. 3-9.

Porta, Miquel, Fernandez, Esteve et Puigdomènech, Elisa, « Book citations: influence of epidemiologic thought in the academic community », *Revista de saúde pública*, vol. 40 / SPE., 2006, p. 50–56.

Porta, Miquel, Vandenbroucke, Jan P., Ioannidis, John P. A. [*et al.*], « Trends in Citations to Books on Epidemiological and Statistical Methods in the Biomedical Literature », *PLoS ONE*, vol. 8 / 5, éd. Alan Hubbard, mai 2013, p. e61837.

Pritchard, Chris, « Inheriting Galton's Statistics: George Darwin, Edgeworth and Weldon », Royal Statistical Society, 2007.

[En ligne : http://www.galton.org/Pritchard/Inheriting– Galtons– Statistics.pdf].

Snoep, J. D., Morabia, Alfredo, Hernandez-Diaz, S. [*et al.*], « Commentary: A structural approach to Berkson's fallacy and a guide to a history of opinions about it », *International Journal of Epidemiology*, vol. 43 / 2, avril 2014, p. 515-521.

Squire, Peverill, « Why the 1936 Literary Digest poll failed », *Public Opinion Quarterly*, vol. 52 / 1, 1988, p. 125–133.

Stigler, Stephen, « How Ronald Fisher became a mathematical statistician », *Mathématiques et sciences humaines. Mathematics and social sciences*, 2006, p. 23–30.

Stigler, Stephen M., « Karl Pearson's Theoretical Errors and the Advances They Inspired », *Statistical Science*, vol. 23 / 2, mai 2008, p. 261-271.

Stolley, Paul D., « When genius errs: RA Fisher and the lung cancer controversy », *American Journal of Epidemiology*, vol. 133 / 5, 1991, p. 416–425.

Susser, Mervyn, « Epidemiology in the United States after World War II: the evolution of technique », *Epidemiologic reviews*, vol. 7 / 1, 1985, p. 147–177.

Susser, Mervyn, « What is a cause and how do we know one? A grammar for pragmatic epidemiology. », *American Journal of Epidemiology*, vol. 133 / 7, avril 1991, p. 635-648.

Susser, Mervyn et Susser, Ezra, « Choosing a future for epidemiology: I. Eras and paradigms. », *American Journal of Public Health*, vol. 86 / 5, 1996, p. 668–673.

Susser, Mervyn et Susser, Ezra, « Choosing a Future for Epidemiology: II. From Black Box to Chinese Boxes and Eco-Epidemiology », *American Journal of Public Health*, vol. 86 / 5, Mai 1996, p. 674-677.

Vandenbroucke, Jan P., « On the rediscovery of a distinction », *American Journal of Epidemiology*, vol. 121 / 5, 1985, p. 627–628.

Vandenbroucke, Jan P., « The history of confounding », in Alfredo Morabia. *A History of Epidemiologic Methods and Concepts*, 1ère, Boston, MA, Birkhaüser Verlag, 2004, p. 313-326.

Vandenbroucke, Jan P., « Commentary: "Smoking and lung cancer"–the embryogenesis of modern epidemiology », *International Journal of Epidemiology*, vol. 38 / 5, octobre 2009, p. 1193-1196.

Vandenbroucke, Jan P. et Pearce, Neil, « 'Causal inference' is necessary but insufficient for causal inference », *The 20th IEA World Congress of Epidemiology (17-21 August 2014, Anchorage, AK)*, WCE, 2014.

Vineis, Paolo, « Causality in epidemiology », in Morabia, Alfredo. *A History of Epidemiologic Methods and Concepts*, 1ère, Boston, MA, Birkhaüser Verlag, 2004, p. 337-349.

Vineis, Paolo, « History of bias », *Sozial-und Präventivmedizin*, vol. 47 / 3, 2002, p. 156–161; repris dans Morabia, Alfredo, *A History of Epidemiologic Methods and Concepts*, Boston, MA, Birkhaüser Verlag, 2004, p. 327-336.

Vineis, Paolo et Michael, Anthony J., « Bias and confounding in molecular epidemiological studies: special considerations », *Carcinogenesis*, vol. 19 / 12, 1998, p. 2063-2067.

White, Colin, « Research on smoking and lung cancer: a landmark in the history of chronic disease epidemiology. », *The Yale Journal of Biology and Medicine*, vol. 63 / 1, 1990, p. 29-46.

Zhang, Fang F., Michaels, Desireé C., Mathema, Barun [*et al.*], « Evolution of epidemiologic methods and concepts in selected textbooks of the 20 th century », *Sozial- und Präventivmedizin/Social and Preventive Medicine*, vol. 49 / 2, avril 2004, p. 97-104.

Zwahlen, M., « Commentary: Cornfield on cigarette smoking and lung cancer and how to assess causality », *International Journal of Epidemiology*, vol. 38 / 5, octobre 2009, p. 1197-1198

OUVRAGES ET CHAPITRES D'OUVRAGES

Armitage, Peter, et Colton, Theodore (éds.), *Encyclopedia of Biostatistics*, Wiley, 2005.

Bowler, Peter J., *Darwin deleted: imagining a world without Darwin*, Chicago; London, The University of Chicago Press, 2013.

Blondiaux, Loïc, *La fabrique de l'opinion : une histoire sociale des sondages*, Paris, Seuil, 1998, (« Science politique »).

Canguilhem, Georges. "Le concept et la vie." Dans : *Études d'histoire et de philosophie des sciences*. Problèmes & Controverses. 1è edition, 1968, 3ème édition Paris : Vrin

Daly, Jeanne, « A Short History of Evidence-Based Medicine », in Victor M. Montori, (éd.). *Evidence-based endocrinology*, éd. Victor M. Montori, Humana Press, Totowa, N.J, 2006, (« Contemporary Endocrinology »), p. 11–24, [En ligne : http://link.springer.com/content/pdf/10.1007/978-1-59745-008-9.pdf#page=21].

Daston, Lorraine, et Galison, Peter, *Objectivity*, New York, Zone Books, 2010; trad. fr. Hélène Quiniou et Sophie Renaut, *Objectivité*, Dijon, Les Presses du réel, 2012.

Gayon, Jean, *Darwin et l'après Darwin : une histoire de l'hypothèse de sélection naturelle*, Paris, Editions Kimé, 1992, 453 p., (« Histoire des idées, théorie politique et recherches en sciences sociales »).

Hippocrate, *Airs, eaux, lieux*, trad. fr. Jacques Jouanna, Paris, Les Belles Lettres, 2003.

Jorland, Gérard, Opinel, Annick et Weisz, George (éds.) *Body counts: medical quantification in historical and sociological perspective / La quantification médicale, perspectives historiques et sociologiques*, Montréal ; Ithaca, McGill-Queen's University

Press, 2005.

Keel, Othmar, *La médecine des preuves : une histoire de l'expérimentation thérapeutique par essais cliniques contrôlés*, Montréal, Québec, Presses de l'Université de Montréal, 2014.

Krüger, Lorenz, Daston, Lorraine et Heidelberger, Michael (eds), *The Probabilistic Revolution*, 1: Ideas in History, Cambridge, Mass., MIT Press, 1987.

Krüger, Lorenz, Daston, Lorraine et Heidelberger, Michael (eds), *The Probabilistic Revolution*, 2: Ideas in the Sciences, Cambridge, Mass., MIT Press, 1987.

Leplège, Alain, Bizouarn, Philippe et Coste, Joël, *De Galton à Rothman: les grands textes de l'épidémiologie au XXe siècle*, Paris, Hermann, 2011.

Marks, Harry, *La médecine des preuves: histoire et anthropologie des essais cliniques : 1900-1990*, Le Plessis-Robinson, Institut Synthélabo pour le progrès de la connaissance, 1999.

Matthews, J. Rosser, *Quantification and the quest for medical certainty*, Princeton, N.J, Princeton University Press, 1995, 195 p.

Morabia, Alfredo, (éd.) *A History of Epidemiologic Methods and Concepts*, Basel, Birkhäuser, 2004

Morange, Michel, *Histoire de la biologie moléculaire*, Paris, Découverte/Poche, 2003.

Moss, Louis, *The Government Social Survey: a History*, London, HMSO, 1991.

Nixon, James, W., *A History of the International Statistical Institute 1885-1960*, The Hague, International Statistical Institute, 1960.

Platt, Jennifer, *The British Sociological Association: a sociological history*, Durham, Sociology press, 2003, 214 p.

Porter, Theodore M., *The rise of statistical thinking, 1820-1900*, Princeton University Press, 1986.

Porter, Theodore M., *Trust in numbers: the pursuit of objectivity in science and public life*, Princeton, N.J, Princeton University Press, 1995.

Stigler, Stephen M., *The History of Statistics: the Measurement of Uncertainty before 1900*, Cambridge, Mass, Belknap Press of Harvard University Press, 1986.

Susser, Mervyn et STEIN, Zena, *Eras in epidemiology: the evolution of ideas*, Oxford ; New York, Oxford University Press, 2009, 352 p.

Valleron, A.-J. (Éd.) *L'Épidémiologie humaine: conditions de son développement en France, et rôle des mathématiques*, Les Ulis : Paris, EDP Science ; Académie des

sciences, 2006, 1 p., (« Rapport sur la science et la technologie », no. 23)
Weisz, George, « From clinical counting to evidence-based medicine », *Body counts: Medical quantification in historical and sociological perspectives*, 2005, p. 377–393.
Whitehead, Frank, « The Government Social Survey », in Bulmer, Martin (éd.). *Essays on the History of British Sociological Research*, Cambridge, Cambridge University Press, 1985, p. 83-100.

THESES

Berlivet, Luc, *Une santé à risques. L'action publique de lutte contre l'alcoolisme et le tabagisme en France (1954-1999)*, Thèse de science politique, Rennes I, 2000.

OUVRAGES ET ARTICLES EN PHILOSOPHIE ET PHILOPHIE DES SCIENCES

Adorno, Theodor W. et Horkheimer, Max, *Dialektik der Aufklärung: philosophische Fragmente*, Amsterdam, 1947. Trad. Fr. Éliane Kaufholz-Messmer, *La dialectique de la raison: fragments philosophiques*, Paris, Gallimard, 2013
Aristote, *Métaphysique*, Tome 1, trad. fr. Jules Tricot, Paris, Librairie J. Vrin, 2000.
Aristote, *Métaphysique.*, Tome 2, trad. fr. Jules Tricot, Paris, Librairie J. Vrin, 1991.
Bloor, David, *Knowledge and social imagery*, 2nd ed, Chicago, University of Chicago Press, 1991; trad. fr. Dominique Ebnöther, *Socio-logie de la loqique ou les limites de l'epistémologie*, Paris, Pandore, 1983
Bouveresse, Jacques, « Tyrannie de la science ou liberté par la science ? », [En ligne : http://www.opuscules.fr/tyrannie-de-la-science-ou-liberte-par-la-science/].
Canguilhem, Georges, *Le normal et le pathologique*, 7. éd, Paris, Quadrige/PUF, 1998, 224 p., (« Quadrige », 65).
Durkheim, Emile, *Les Règles de la méthode sociologique*, Paris, Alcan, 1901.
Fossier, Arnaud et Gardella, Édouard, « Entretien avec Bruno Latour », *Tracés*, février
Feyerabend, Paul, Against *Method*, Londres, Verso, 1975, Trad. Fr. Baudouin Jurdant et Agnès Schlumberger, *Contre la méthode*, Paris, Seuil, 1979.
Foucault, Michel, *Leçons sur la volonté de savoir*, suivi de *Le Savoir d'Œdipe*, Cours au Collège de France (1970-1971), édition établie sous la direction de François Ewald et Alessandro Fontana par Daniel Defert, Paris, Gallimard/Seuil, 2011,

Kuhn, Thomas S, *The Structure of Scientific Revolutions*, 2nd éd., Enlarged, University of Chicago Press, 1970; trad. fr. Laure Meyer, *La structure des révolutions scientifiques*, Paris, Flammarion, 1983.

Nietzsche, Friedrich, *Le Livre du philosophe*, trad. fr. A. K. Marietti, Paris, Aubier-Flammarion, 1969 [1873],

Popper, Karl R., *A world of propensities*, Bristol, Thoemmes, 1990; trad. fr. Alain Boyer, *Un univers de propensions: Deux études sur la causalité et l'évolution*, L'éclat, Paris, 1992.

Quine, W. V., « On the Reasons for Indeterminacy of Translation », *The Journal of Philosophy*, vol. 67 / 6, mars 1970, p. 178.

Rivenc, François (éd.). *Logique et fondements des mathématiques : anthologie (1850 - 1914)*, Paris, Payot, 1992.

Rousseau, Dominique et Morvan, Michel, *L'erreur.*, Paris, Odile Jacob, 2000.

Tarski, Alfred, « The Semantic Conception of Truth and the Foundations of Semantics », *Philosophy and Phenomenological Research*, vol. 4 / 3, mars 1944, p. 341.

TABLE DES FIGURES

INDEX NOMINUM

INDEX RERUM

Titre : **Le concept de biais en épidémiologie.**

Résumé : **Cette thèse, qui s'inscrit dans la tradition méthodologique de l'épistémologie historique, porte sur l'histoire et la formation du concept de biais dans l'épidémiologie moderne. Elle montre que la fonction opératoire du concept de biais est essentiellement critique, au sens où ce concept, que les épidémiologistes opposent au cours de l'histoire aux concepts d'objectivité, de preuve et de causalité, joue un rôle décisif dans la constitution de l'épidémiologie comme science, mais aussi dans l'avènement d'une médecine scientifique. Un éclairage historique et critique est apporté à la définition actuelle du biais, conçu comme une erreur ou un écart systématique par rapport à la vérité, ainsi qu'aux différentes taxinomies des biais qui jalonnent l'histoire de ce concept, dont l'origine se situe chez les fondateurs de la statistique mathématique. Le biais apparait ainsi comme une menace aussi bien à la validité du plan d'expérience d'une étude épidémiologique, et donc à la validité de l'inférence statistique, qu'à la vérité des connaissances médicales acquises à travers les études observationnelles ou expérimentales. En d'autres termes, ce sont les conséquences que la révolution probabiliste a eues sur l'épidémiologie et sur la médecine qui sont ici étudiées, et qui ont conduit les épidémiologistes et les médecins à une forme de scepticisme et même de criticisme envers leurs propres inférences, ce qui donnera naissance au mouvement de la médecine fondée sur des preuves.**

Mots clefs : **Histoire et épistémologie de la médecine, Histoire et épistémologie de l'épidémiologie, Biais, Erreur aléatoire et erreur systématique, Vérité, Validité, Niveaux de preuve, Causalité, Plan d'expérience, Etude cas-témoins, Etude de cohorte.**

Title : **The concept of bias in epidemiology**

Abstract : **This PhD thesis, belonging to the tradition of historical epistemology, deals with the history and the formation of the concept of bias in epidemiology. It shows that the operational function of the concept of bias is essentially critical, in the sense that this concept, used by epidemiologists throughout history as an antonym to both objectivity, causality and evidence, is central to both the construction of epidemiology as a scientific discipline and the advent of scientific medicine. An historical and critical account is given of the actual definition of bias, conceived as a systematic error or deviation from the truth, and to the various taxonomies of bias which marked the history of this concept, whose origin goes back to the founders of mathematical statistics. Bias thus appears as a threat to the validity of the design of an epidemiological study, and to the validity of statistical inference; but also as a threat to the truth of medical knowledge provided by observational and experimental studies. In other words, what is studied here is the consequences of the probabilistic revolution on both epidemiology and medicine, which led epidemiologists and physicians to a kind of scepticism or even criticism about their own inferences, which would ultimately give birth to evidence-based medicine's movement.**

Keywords : **History and Philosophy of Medicine, History and Philosophy of Epidemiology, Bias, Random Error and Systematic Error, Truth, Validity, Hierarchy of Evidence, Causation, Design of Experiment, Case-control Study, Cohort Study.**