Contributions à la génétique et l'épidémiologie des maladies complexes pour une médecine personnalisée

Félix Balazard

Sciences mathématiques de Paris Centre

Thèse de doctorat

Discipline : Mathématiques

Spécialité : Statistique

présentée par

Félix Balazard

Contributions à la génétique et l'épidémiologie des maladies complexes pour une médecine personnalisée

---

dirigée par Gérard Biau et par Pierre Bougnères

Au vu des rapports établis par
MM. Christophe Ambroise et Vivian Viallon

Soutenue le 17 décembre 2018 devant le jury composé de :

M. Christophe Ambroise	Université d'Évry	Rapporteur
M. Gérard Biau	Sorbonne Université	Directeur de thèse
M. Pierre Bougnères	INSERM	Directeur de thèse
M. Raphaël Porcher	Université Paris Descartes	Examinateur
M. Alain-Jacques Valleron	Sorbonne Université	Président du jury
M. Vivian Viallon	Université Claude Bernard	Rapporteur

Laboratoire de Probabilité, Statistique et Modélisation (LPSM)
Université Pierre et Marie Curie
Case courrier 188
4 place Jussieu
75252 Paris Cedex 05

INSERM U1169
Hôpital Bicêtre
78 Rue du Général Leclerc
94270 Kremlin-Bicêtre

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse Gérard Biau et Pierre Bougnères qui m'ont soutenu et fait confiance tout au long de cette thèse. Gérard était toujours réactif et disponible et m'a appris à rédiger rigoureusement. Pierre, en plus de m'accueillir au sein de l'étude Isis-Diab, m'a fait profiter de sa grande culture scientifique et de son énergie. J'adresse les mêmes remerciements à Alain-Jacques Valleron qui a été comme un troisième directeur pour moi.

Je remercie également toute l'équipe d'Isis-Diab qui a accompli un travail énorme de collecte des données. En particulier, Sophie le Fur sut me guider dans les recoins de la base de données. Sophie Valtat fut également d'une grande aide.

J'ai eu la chance de pouvoir collaborer très efficacement avec Raphaël Porcher. Des échanges ponctuelles avec Mark Lathrop, Pierre-Yves Boëlle, Etienne Roquain et Ismaël Castillo m'ont également permis d'orienter mes recherches. Je les en remercie.

Je remercie Vivian Viallon et Christophe Ambroise d'avoir consacré du temps aux rapports de cette thèse.

Je tiens aussi à remercier pour leur gentillesse et leur efficacité les secrétaires des deux laboratoires, Muriel Delacroix puis Léa Etcheverry à l'INSERM et Corinne Van Vlierberghe et Louise Lamart au LPSM.

Mon doctorat n'aurait pas été aussi agréable si je n'avais pas pu partager mes idées avec mes pairs, doctorants ou stagiaires. À l'INSERM, j'ai eu la chance de discuter longuement avec Nicolas, Luc et Adrien. Du côté du LPSM, je remercie Lucie et Yohann d'avoir organisé le séminaire avec moi, Dimby avec qui j'ai partagé mon bureau, ainsi que tous ceux avec qui j'ai eu le plaisir de partager un déjeuner, un café, un verre, un trajet en moto, un logement pour une conférence voire même une partie de laser game. Je tiens aussi à adresser toute ma gratitude aux doctorants de l'association Doc'Up avec qui j'ai pu organiser des actions dont l'horizon temporel était inférieur à trois ans.

Je remercie mes amis qui m'ont permis de m'évader de temps en temps : Alexandre (x2), Romain(x2), Solène, Sylvain, Loïk, Joseph, Stefan, Paul, Benoît, Milan, Titouan, Reda. Mon quotidien pendant ces trois ans a été confortable grâce au cocon de franche camaraderie offert par mes colocs Maxence, Nikolas et Laurent.

Je remercie toute ma famille y compris les derniers arrivés. En particulier, je souhaite remercier mon père pour la relecture de cette thèse. Je remercie aussi ma mère pour son soutien indéfectible.

Finalement, merci Wen de partager ma vie. Le temps passé avec toi est doux et léger.

Contents

Chapter 1

Causes génétiques et environnementales des maladies complexes

Cette thèse a été effectuée sous la supervision de Gérard Biau, professeur au Laboratoire de Probabilité, Statistique et Modélisation (LPSM) à Sorbonne Université et de Pierre Bougnères, pédiatre endocrinologue et professeur de médecine à l'unité 1169 de l'INSERM au Kremlin-Bicêtre.

À l'image de sa supervision, cette thèse est marquée du sceau de l'interdisciplinarité, entre génétique, épidémiologie et statistique. Ce chapitre introductif expose les notions de ces disciplines qui sont utiles à la compréhension de nos travaux. Nous commençons en introduisant notre objet d'étude : les maladies complexes. Nous proposons ensuite un aperçu des concepts de la génétique quantitative. Nous présentons également les principaux modèles d'études d'épidémiologie ainsi que la méthodologie statistique permettant de les analyser. Ceci est illustré par l'étude Isis-Diab qui est l'application centrale de cette thèse. Pour cette étude ambitieuse, l'équipe de Pierre Bougnères à l'INSERM a collecté des données sur des milliers de patients atteints du diabète de type 1, afin de déterminer notamment les causes génétiques et environnementales de cette maladie. Après avoir ainsi décrit le contexte scientifique et institutionnel de ce travail, nous annoncerons les contributions apportées dans cette thèse.

Contents

1.1 Les maladies complexes

Les maladies complexes sont causées à la fois par des facteurs environnementaux et génétiques [Craig et al., 2008]. Cette catégorie recouvre les maladies à l'origine de l'essentiel des coûts de santé des pays développés : maladies cardio-vasculaires, diabète de type 1 et 2, maladie d'Alzheimer, maladie de Parkinson, cancers, maladie de Crohn, colite ulcérative. On les qualifie parfois de maladies multifactorielles. Elles sont définies par opposition aux maladies mendéliennes dont les causes sont simples.

Les maladies mendéliennes sont provoquées par une mutation récessive. Ce qui veut dire que si une personne a une seule copie de l'allèle mutée, elle n'est pas malade mais si elle possède deux copies mutées, elle développe la maladie. Par exemple, la maladie de Tay-Sachs fréquente chez les juifs ashkénazes, les Québécois et les Cadiens, est due à une mutation dans le gène HEXA et entraîne la mort du nourrisson dans les quatre ans après la naissance [Fernandes F. and Shapiro, 2004]. On les appelle mendéliennes en référence aux travaux de Gregor Mendel, un moine autrichien, qui établit par l'étude de la transmission des caractères chez les plantes les lois de l'hérédité en 1866. Étant donné leur gravité et donc la pression de sélection négative qui en résulte, ces maladies sont rares dans la population.

Les maladies complexes au lieu d'être provoquées par un seul gène sont rendues plus probable par la combinaison de milliers de gènes de prédisposition et par des facteurs environnementaux. L'identification de leurs causes peut permettre de prévenir ces maladies : si on peut prédire le risque génétique de développer une maladie, on peut identifier une population à risque et si on comprend les causes environnementales de la maladie, on pourra proposer une intervention à cette population à risque afin d'éviter que la maladie ne se déclenche. Cet objectif est la ligne de mire de cette thèse.

La phénylcétonurie est l'exemple le plus frappant de ce que peut accomplir cette stratégie de prévention. C'est une maladie mendélienne et non pas complexe. Chez les patients, une mutation d'un gène empêche de métaboliser la phénylalanine, un acide aminé. Celle-ci se trouve dans l'alimentation, va s'accumuler dans l'organisme du patient et provoquer un retard mental. En France, la phénylcétonurie est dépistée à la naissance. Afin de limiter les effets de la maladie, on prescrit alors un régime alimentaire pauvre en phénylalanine aux enfants porteurs de la mutation. Ces derniers ont par conséquent un développement cérébral normal.

1.1.1 Le diabète de type 1

L'exemple central de maladie complexe de cette thèse est le diabète de type 1 auquel est consacré l'étude Isis-Diab. Le diabète de type 1 est une maladie auto-immune [Atkinson et al., 2014]

dans laquelle le système immunitaire détruit les cellules β situées dans les îlots de Langherans du pancréas. Ces cellules produisent l'insuline, l'hormone qui permet de faire baisser la glycémie dans le sang en provoquant le stockage du glucose dans le foie, les tissus adipeux et les muscles. En l'absence d'insuline, la glycémie des patients diabétiques n'est pas contrôlée et atteint des valeurs très élevées ce qui provoquent une soif extrême (polydipsie), une sécrétion d'urine fréquentes (polyurie) et une faim intense (polyphagie). La réaction auto-immune commence longtemps avant les premiers symptômes [Achenbach et al., 2005]. On peut suivre l'apparition chez les enfants à risques de divers anticorps anti-îlots qui sont un marqueur de la réaction auto-immune. Cette période asymptomatique peut durer des mois ou des années. Les symptômes apparaissent quand les cellules β ont été détruites en majorité et que celles qui restent ne sont plus en mesure de produire assez d'insuline pour contrôler la glycémie. Ces symptômes conduisent le patient à consulter un médecin qui pourra alors diagnostiquer la maladie.

Heureusement, il est possible de compenser l'absence de production d'insuline par des injections régulières d'insuline. La première injection d'insuline à un patient diabétique a eu lieu en 1922 à Toronto. À l'origine, l'insuline était récupérée dans des pancréas d'animaux d'élevages et c'était donc la version animale de cette hormone qui était utilisée. Depuis les années 1980, la version humaine de l'insuline, produite par des levures modifiées génétiquement, est utilisée [Tattersall, 2009]. L'insulinothérapie nécessite des contrôles fréquents de la glycémie afin d'ajuster la quantité à injecter, une trop grosse dose d'insuline menant à l'hypoglycémie. Ce traitement a permis de donner aux patients diabétiques une espérance de vie comparable au reste de la population. Cependant leur qualité de vie est réduite par l'attention constante qu'ils doivent porter à leur traitement et les complications que toute inattention provoque.

Le diabète de type 1 se déclenche le plus souvent chez l'enfant mais peut également se déclencher chez l'adulte. L'âge médian au diagnostic est de 8 ans et un mois dans l'étude Isis-Diab. L'incidence de la maladie est très variable entre pays, le maximum étant atteint en Finlande avec 60 nouveaux cas pour 100000 enfants de moins de 15 ans [Harjutsalo et al., 2013] alors que la maladie est quasiment absente en Asie. L'incidence mondiale a augmenté de 3% par an pendant la décennie 1990-1999 [DIAMOND Project Group, 2006] mais semble avoir atteint un plateau en Finlande depuis 2005 [Harjutsalo et al., 2013]. En France, l'incidence a été de 18 pour 100000 enfants de moins de 15 ans entre 2013 et 2015 [Piffaretti et al., 2017]. C'est le double de l'incidence mesurée sur la décennie 1990-1999 (9 pour 100000) ce qui est cohérent avec une croissance annuelle comprise entre 3% et 4%. Si l'incidence reste constante au niveau de 2013-2015, la prévalence de la maladie sera à terme de 0.25%.

1.2 La génétique des maladies complexes

1.2.1 L'ADN, support moléculaire de l'hérédité

L'information génétique d'un organisme est contenue dans le noyau de chaque cellule de cet organisme sous la forme de molécules d'ADN (Acide DésoxyriboNucléique). C'est une molécule double-brin qui est constituée par une succession de bases choisies parmi les 4 possibles : A (adénine), T (thymine), C (cytosine) et G (guanine). Les deux brins sont complémentaires : face à chaque A se trouve un T et face à chaque C se trouve un G. On considérera toujours le même brin qui est fixé par convention. L'information génétique se résume donc à une suite de

bases parmi les 4 possibles. On parle de paire de bases pour se référer au caractère double-brin de l'ADN. Chez l'homme, il y a un total de 3 milliards de paires de bases divisé en 23 paires de chromosomes, de très longues molécules d'ADN. Parmi ces 23 paires de chromosomes, il y a 22 chromosomes autosomaux numérotés de 1 à 22 par ordre décroissant de taille et une paire de chromosomes sexuels.

Dans chaque paire de chromosome autosomal, il y a un chromosome qui vient de la mère et un du père de l'enfant. Ces chromosomes sont eux-mêmes une mosaïque des chromosomes des grands-parents. Alors que les chromosomes sont reproduits à l'identique lors de la mitose (la division cellulaire), ce n'est pas le cas lors de la méiose (la production des gamètes : spermatozoïdes et ovules). Au cours de cette dernière, des entrecroisements se produisent entre les chromosomes homologues (qui appartiennent à la même paire) et des bouts de chromosome sont échangés. On appelle ce phénomène la recombinaison. Elle induit une plus grande diversité génétique et influe sur les corrélations entre variants génétiques. Les variants qui sont à une faible distance en paire de base ont peu de chances d'être séparés par la recombinaison. Par conséquent, ces variants vont être très corrélés dans la population. On appelle cette corrélation liée à la distance chromosomique le déséquilibre de liaison [Slatkin, 2008].

Parmi les trois milliards de paire de bases du génome humain, l'immense majorité est identique chez tous les êtres humains. On appelle SNP (prononcé snip) pour Single Nucleotide Polymorphism les variants relativement communs dans la population, par exemple si l'allèle mineur est présent chez au moins 1% de la population. Un SNP correspond à une localisation dans le génome où il y a deux génotypes possibles disons A et C. Comme il y a deux chromosomes homologues avec chacun deux possibilités, un SNP peut prendre trois valeurs : A/A, A/C ou C/C qu'on peut coder 0, 1 ou 2 en comptant le nombre de copies de l'allèle mineur (la moins fréquente) présent dans la paire de chromosomes.

Le principe d'Hardy-Weinberg [Hartl and Clark, 1997, p. 48] nous dit que si la reproduction se fait de manière aléatoire et que la mutation est neutre au niveau adaptatif, la fréquence des trois possibilités sera p^2, $2p(1-p)$ et $(1-p)^2$ où p est la proportion de l'allèle majeur dans la population. On parle d'équilibre d'Hardy-Weinberg quand un SNP respecte cette distribution. Le respect de cet équilibre est un moyen de contrôler la qualité des données génétiques.

Si on considère plusieurs variants, le codage en 0, 1 ou 2 ne permet pas de distinguer les deux situations présentées dans la figure 1.1. Deux SNPs hétérozygotes se situent sur le même chromosome par exemple le chromosome 3. Le chromosome 3 est en fait une paire de chromosomes homologues et les deux possibilités sont que les allèles mutants sont sur un seul chromosome ou sur les deux chromosomes distincts de la paire de chromosomes. Les séquences sur les deux chromosomes homologues sont appelées haplotypes et l'information de phase est la connaissance des haplotypes au lieu du génotype. C'est une information biologique importante [Tewhey et al., 2011] que nous utiliserons dans le chapitre 2. Par exemple, si les deux allèles d'un gène ont une mutation non-sens distincte, une condition appelée hétérozygotie composée, il n'y aura pas d'expression du gène. En revanche, si les deux mutations étaient sur le même chromosome, il y aurait toujours un allèle fonctionnel. C'est un exemple simple où l'information de phase est cruciale. L'information de phase est liée à la présence de deux chromosomes dans chaque paire et non à la complémentarité des deux brins d'ADN.

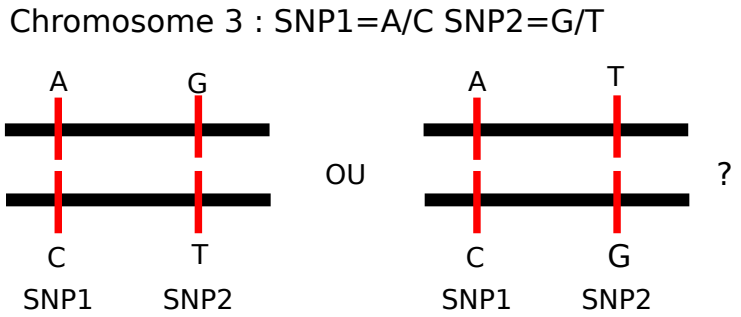

Figure 1.1: Les valeurs des deux SNPs ne permettent pas de distinguer entre les deux paires d'haplotypes possibles.

1.2.2 La génétique par les phénomènes

L'étude de la génétique fut amorcée pendant le dix-neuvième siècle bien longtemps avant qu'on ait une idée des bases moléculaires de l'hérédité. Cette étude se basa donc nécessairement sur les conséquences de la génétique, c'est-à-dire l'observation des caractères héréditaires et de leur transmission.

Si l'approche de Mendel permit de comprendre les lois de l'hérédité pour des traits simples comme la couleur des yeux, une autre approche de l'hérédité fut développée par Francis Galton. Touche-à-tout brillant, Galton s'intéressa à l'hérédité à cause de ses convictions eugénistes [Gillham, 2001]. Au lieu de s'intéresser à des traits catégoriels, il s'intéressa à des traits continus comme la taille. Lors de l'exposition internationale de la santé qui eut lieu à Londres en 1884, il ouvrit un laboratoire anthropométrique où les familles en goguette se faisaient mesurer sous toutes les coutures. Il analysa ensuite la corrélation entre la taille des parents et celle des enfants dans son article "Regression towards mediocrity in hereditary stature" [Galton, 1886]. Il constata que si les tailles des enfants étaient corrélées avec celle de leurs parents, les enfants étaient plus proches de la moyenne que leurs parents d'où le titre : régression vers la médiocrité, c'est-à-dire la moyenne. Cet article a également eu une grande postérité dans le domaine de la statistique puisque le mot régression, comme régression linéaire, vient de là.

On peut comprendre ce phénomène de la régression vers la moyenne grâce au modèle suivant: supposons que le phénotype P de moyenne nulle est la somme d'une contribution génétique G et d'un contribution environnementale E. On suppose également que G et E sont indépendants dans la population et qu'ils suivent des lois normales centrées et de variance σ_G^2 et σ_E^2. On a alors

$$P = G + E \sim \mathcal{N}(0, \sigma_G^2 + \sigma_E^2).$$

Si on suppose pour simplifier qu'il n'y a qu'un parent par enfant i.e. que l'enfant est un clone du parent mais avec un environnement E' indépendant, son phénotype sera $P' = G + E'$ et on définit l'héritabilité [Visscher et al., 2008] du phénotype par

$$H^2 = \mathrm{Corr}(P', P) = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}.$$

Comme (P', P) est un vecteur gaussien, $\mathbb{E}[P'|P]$ est la projection orthogonale de P' sur l'espace engendré par le vecteur 1 et P puis $\mathbb{E}[P'|P] = H^2 P$. Ainsi, le coefficient de régression de P' sur P est l'héritabilité. L'héritabilité mesure la proportion dans une population de la variance d'un trait dû à la génétique. Si elle est proche de 0, le trait ne dépend que peu de la génétique et si elle est proche de 1, toute la variation dans la population s'explique par des différences génétiques. C'est donc une grandeur qui dépend de la population et de l'autre source de variation, l'environnement.

Les approches de Galton et de Mendel semblaient proposer deux mécanismes distincts de l'hérédité : le premier étant un mélange continu entre les caractères des deux parents et le deuxième une combinaison de particules d'hérédité. En 1918 alors qu'il débutait sa carrière, Ronald Fisher réconcilia les deux approches en proposant que les traits continus soient le résultat de nombreux variants obéissant aux lois mendéliennes [Visscher and Walsh, 2017]. C'est également dans cet article que le mot de variance fut introduit ainsi que l'analyse de la variance.

Il prit également en compte le fait que les enfants ne sont pas des clones d'un parent unique mais sont le produit de la méiose puis de la fécondation qui mélangent les variants génétiques et créent des combinaisons originales de ces variants. Cela veut dire que les interactions au niveau d'un même variant, ce qu'on appelle la dominance, ou les interactions entre différents variants, appelées épistasie, ne sont pas transmises aux enfants. Ainsi si on estime la variance expliquée par la génétique grâce à la ressemblance entre parents et enfants, on ne peut estimer que la variance due aux contributions additives de chaque variant. On appelle cette quantité h^2 l'héritabilité au sens étroit par opposition à H^2 l'héritabilité au sens large.

La notion d'héritabilité fut développée par des eugénistes (Galton, Fisher) qui voulaient améliorer la qualité des gènes de la population humaine. Cela conduisit à un mouvement international pour l'eugénisme qui provoqua notamment divers programmes de stérilisations forcées de personnes jugées nuisibles dans les pays nordiques [Broberg and Roll-Hansen, 2005] et aux États-Unis [Lombardo, 2011] qui continuèrent jusque dans les années 70. Cela inspira également de nombreux aspects des atrocités nazies.

Cette même notion fut appliquée, avec plus de rigueur et moins de conséquences néfastes, en agriculture pour guider la sélection végétale et animale. L'héritabilité d'un trait exprime sa sensibilité à la sélection : si seuls les animaux ayant un phénotype $P = p$ se reproduisent, la moyenne des phénotypes de leurs enfants sera $\mathbb{E}[P'|P = p] = h^2 p$. Connaître l'héritabilité d'un trait permet donc de prévoir la réponse à la sélection : si on sélectionne les animaux qui vont se reproduire sur P, va-t-on ou non améliorer le trait P de la génération suivante ? Le succès de cette notion en agriculture explique que l'extension de la notion d'héritabilité aux maladies humaines complexes fut faite par un généticien spécialisée en élevage [Falconer, 1965]. Une maladie étant un trait discret et non un trait continu, cette extension propose de supposer l'existence d'un trait continu sous-jacent responsable de la maladie (liability of disease) qui s'il dépasse un seuil provoque la maladie. Cela correspond au modèle probit et son utilisation dans ce cadre, plutôt que le modèle logistique, vient de ce lien disciplinaire fort avec la génétique animale.

Une objection sur le rôle de la génétique sur la ressemblance entre parents mérite ici d'être mentionnée. Comment séparer dans les ressemblances entre frères et sœurs les contributions de l'environnement familial partagé de celles de la génétique ? Cela est particulièrement pertinent pour les traits comportementaux pour lesquels l'éducation joue a priori un rôle important. Les

études de jumeaux apportent une réponse convaincante à cette remarque. Les vrais jumeaux (monozygotes) et les faux jumeaux (dizygotes) partagent leur environnement familial de la même manière mais les vrais jumeaux partagent tout leur génome alors que les faux jumeaux en partagent seulement la moitié. Ainsi pour le diabète de type 1, les études de jumeaux en Finlande montrent que parmi les paires de jumeaux monozygotes avec au moins un malade, 27% est concordante pour la maladie (i.e. les deux sont malades) alors que cette même proportion n'est que de 4% pour les jumeaux dizygotes [Hyttinen et al., 2003]. La même étude conclut que l'héritabilité du diabète de type 1 est de 88% en Finlande. Cette concordance de 27% est peut-être même sous-estimée à cause de la limite de la période de suivi. En suivant plus longtemps des jumeaux monozygotes, une concordance de 65% a été observée [Redondo et al., 2008].

1.2.3 La génétique révélée

Au cours de la seconde moitié du vingtième siècle, des progrès considérables en biologie moléculaire permirent d'approcher progressivement de la lecture directe de l'ADN. Ces progrès ne permirent de faire avancer la connaissance des causes génétiques des maladies complexes qu'à partir des années 2000. Le diabète de type 1 est une exception. Ainsi dans les années 70, l'association entre la région du chromosome 6 codant pour HLA (Human Leukocyte Antigen) et le diabète de type 1 fut découverte [Singal and Blajchman, 1973]. HLA, la version humaine du complexe majeur d'histocompatibilité, venait d'être découverte pour son rôle dans le rejet des greffes et plus généralement dans l'immunité [Terasaki, 2007]. Si l'association entre le diabète de type 1 et HLA a pu être détectée si tôt, c'est parce qu'HLA a une grande influence sur le diabète de type 1. Comme les généticiens le découvriront plus tard, la plupart des associations entre maladies et variants génétiques sont de très petite tailles avec des rapports de côtes (odd-ratio) proches de 1.

La technologie qui permit d'importants progrès dans l'identification des variants génétiques associés aux maladies complexes sont les puces ADN (DNA microarray) qui atteignirent un niveau de performance et de coût raisonnables dans les années 2000 [Bumgarner, 2013]. Les puces ADN permettent d'avoir un premier aperçu de la variation génétique humaine en lisant le génotype d'un individu aux niveaux des SNPs les plus communs. Le nombre de SNPs dans une puce varie entre quelques centaines de milliers et plusieurs millions. Un individu est dit génotypé si une puce ADN a été utilisée pour connaître ses SNPs.

Il ne faut pas confondre génotypage et séquençage. Le séquençage consiste à lire l'ensemble des paires de bases d'un génome humain. La différence est qu'en plus de connaître les SNPs, ces variants assez communs dans la population, le séquençage donne accès aux variant rares, des variants qui sont présents uniquement chez quelques individus. Le séquençage a permis de grandes avancées dans le diagnostic et parfois le traitement des maladies mendéliennes [Jamuar and Tan, 2015].

La technologie des puces ADN fut utilisée dans les études d'association à l'échelle du génome (GWAS, Genome-Wide Association studies). Elles consistent à génotyper des milliers de patients et des milliers de témoins sains et de tester pour chaque SNP sur la puce si il est associé avec la maladie. Un seuil de significativité très bas est choisi afin de corriger pour les tests multiples. La première grande étude GWAS est le Wellcome Trust Case Control Consortium 1 (WTCCC1) [Burton et al., 2007]. Pour cette étude, 2000 patients pour chacune de 7 maladies complexes

et 3000 témoins partagés furent génotypés. L'étude identifia 24 associations significatives au total. De plus, les données de cette étude sont accessibles aux chercheurs en faisant la demande et elles ont servies au travail que je présente dans cette thèse dans le chapitre 2. Depuis, les études GWAS ont permis d'identifier plus de vingt mille associations entre des SNPs et des traits [Welter et al., 2014].

Néanmoins, ces beaux succès étaient assombris par un paradoxe. Maintenant que l'on pouvait voir les variations génétiques dans la population, on s'attendait à ce qu'elles expliquent la variance due à la génétique, c'est-à-dire l'héritabilité. Or en combinant les variants significatifs pour les études GWAS, on n'expliquait qu'une faible partie de l'héritabilité. On appela ce problème l'héritabilité manquante [Manolio et al., 2009]. Cette héritabilité manquante peut être due en partie à des variations génétiques qui ne sont pas couvertes par les puces ADN comme les variants rares. Une explication d'ordre plus méthodologique est de dire que la correction pour les tests multiples utilisée sert à contrôler les faux positifs et non à expliquer un maximum de variance. Une réponse à ceci fut de développer des méthodes pour estimer la variance d'un trait expliquée par une puce ADN en entier et non pas uniquement les variants associés significativement [Yang et al., 2011, Speed et al., 2017]. Ces méthodes ont permis de montrer qu'une grande partie de la variance des traits complexes pouvait être expliquée par les puces ADN.

Cependant, ces approches ne permettent pas de faire de prédiction du risque génétique. Si on veut pouvoir prévenir les maladies complexes, il est important que l'on puisse identifier les personnes à risque de développer la maladie. Ce sujet est traité dans le chapitre 2. En ce qui concerne le diabète de type 1, le typage HLA a permis d'identifier des personnes à risque et d'essayer de prévenir le diabète chez ces personnes avant même que les puces ADN soient disponibles.

1.3 Causalité en épidémiologie

1.3.1 Essai randomisé contrôlé et prévention du diabète de type 1

Ronald Fisher mit en exergue la notion de randomisation dans le cadre de ses travaux en agronomie [Hall, 2007]. Les agronomes cherchent à savoir quelle est la variété de blé qui donne le meilleur rendement. Malheureusement, si on plante une variété dans un champ et une autre variété dans un champ différent, la différence qu'on observe entre le rendement des variétés peut être due aux champs aussi bien qu'à la variété. Ainsi, le premier champ peut avoir un sol plus riche et une meilleure exposition au soleil qui fait que le blé pousse mieux. La variété de blé est *confondue* avec le champ et on ne peut donc pas conclure sur l'influence de la variété. On dit que le champ est une variable de confusion.

Afin de pouvoir connaître le rendement de la variété, il faut faire en sorte que les deux variétés de blé soient comparés sur des champs de même qualité. Il n'est cependant pas évident de quantifier la qualité d'un champ. Comment faire alors pour que la comparaison entre les deux variétés soit informative ? La randomisation résout ce problème en utilisant le hasard. On divise les champs en de nombreuses parcelles et on attribue au hasard une variété à chaque parcelle. Si les parcelles sont assez nombreuses, les conditions seront identiques pour les deux variétés et on pourra donc comparer leur rendement.

La même idée s'applique aux essais cliniques. Une entreprise pharmaceutique qui veut prouver que son nouveau médicament est plus efficace que l'ancien traitement doit faire un essai thérapeutique randomisé. Cela consiste à attribuer au hasard le traitement que recevra un patient quand il est inclus dans l'essai clinique. Aux variétés de blé de l'exemple précédent correspondent les deux traitements considérés et aux champs correspondent les patients. Les patients sont tous différents et certains auraient guéri de toute façon alors que d'autres avaient peu d'espoir d'en réchapper. En allouant le traitement au hasard, on s'assure que les deux traitements sont testés sur des populations identiques. Ainsi, le résultat de l'essai permettra de dire si le nouveau traitement est préférable à l'ancien traitement. Le résultat d'un essai randomisé est considéré comme le plus haut niveau de preuve dans la recherche biomédicale.

Les essais thérapeutiques sont utilisés afin de déterminer le traitement qui sera le meilleur en moyenne. L'intérêt porté à la médecine personnalisée conduit à utiliser les mêmes données afin de trouver le traitement le plus adapté à chaque patient. Cette idée est développée dans le chapitre 6.

Pour ce qui concerne la prévention du diabète de type 1, plusieurs essais cliniques de prévention ont eu lieu. Il faut distinguer la prévention primaire, qui vise à empêcher l'apparition de l'autoimmunité, de la prévention secondaire qui cherche à empêcher la progression vers le diabète des personnes ayant des anticorps anti-îlots et de la prévention tertiaire qui cible les patients récemment diagnostiqués [Skyler, 2013]. Le diabète de type 1 étant une maladie à la prévalence faible, un essai thérapeutique de prévention dans la population générale aurait besoin d'une taille d'échantillon gigantesque afin d'avoir une puissance statistique suffisante. Afin de se ramener à des tailles d'échantillon plus raisonnable, il faut sélectionner une population à risque élevée. C'est une motivation importante pour développer des prédictions du risque génétique comme nous le ferons dans le chapitre 2. Ainsi, les essais de prévention primaire ont été effectués dans des populations à risques définis par leur type HLA, par la présence d'un parent atteint par la maladie ou les deux.

Le plus ambitieux des essais de prévention primaire est TRIGR (Trial to Reduce IDDM in the Genetically at Risk–IDDM voulant dire insulin-dependent diabetes mellitus c'est-à-dire le diabète de type 1). Cet essai a testé si la substitution du lait de vache par un lait dont la caséine a été hydrolysée affecte la survenue d'auto-immunité contre les îlots ou la progression vers le diabète de type 1. Pour cela, 5000 nouveau-nés dont un parent au premier degré était atteint de la maladie ont été identifiés. Le typage HLA a ensuite permis de sélectionner 2000 participants qui possédaient l'un des 5 types HLA à haut risque [Åkerblom et al., 2011]. Ces participants ont été randomisés et ont reçu une des deux versions de lait. Aucune différence significative n'a été observée entre les deux groupes, ni pour l'apparition de l'auto-immunité [Knip et al., 2014] ni pour la progression vers le diabète de type 1 [Knip et al., 2018]. Dans les deux bras de l'essai, 8% des participants ont développé le diabète après un temps médian de suivi de 11 ans et demi.

Un tel essai thérapeutique par sa durée et son échelle représente un coût économique très important. Il est donc crucial d'avoir un traitement candidat aussi convaincant que possible avant de débuter un essai thérapeutique. Un traitement peut être prometteur si des expériences animales ou cellulaires permettent de comprendre la biologie qui sous-tendrait un effet. Une autre source de preuve est l'épidémiologie grâce à des études observationnelles. Nous nous concentrerons sur ces dernières dans le cadre de ce chapitre.

En simplifiant un peu, les études observationnelles peuvent suivre deux modèles : l'étude

prospective de cohorte où l'on va mesurer les expositions de participants puis observer s'ils développent par la suite la maladie et des études rétrospectives cas-témoin où l'on recrute des patients déjà malades dont on compare les expositions passés à un groupe témoin choisi de manière adéquate. Ces études ne sont pas randomisées et sont donc vulnérables au problème de la confusion. Ce problème a motivé de nombreux développements méthodologiques sur lesquels nous nous pencherons par la suite.

1.3.2 Études prospectives de cohorte

Une étude prospective de cohorte consiste donc à suivre des participants et à voir qui va développer la maladie. Le caractère prospectif de ces études se traduit dans le fait que les expositions sont mesurées en même temps qu'elles ont lieu. On va donc de l'exposition vers la maladie. Ceci permet d'éviter le biais de rappel qui peut affecter les études rétrospectives : le fait d'être atteint par une maladie peut influencer la manière dont on se remémore et dont on interprète les expositions passées [Coughlin, 1990]. Comme les expositions ne sont pas randomisés, le coût de mesurer une exposition de plus est très faible. De telles études permettent donc d'étudier de nombreuses expositions en même temps.

Comme les participants ne sont pas malades au début de l'étude, la même remarque que pour les essais randomisés est valable quand la maladie est peu fréquente : pour qu'il y ait assez de malades dans l'échantillon à la fin de l'étude, il faudrait une énorme taille d'échantillon. La solution est la même : il faut sélectionner des participants à haut risque de développer la maladie.

Pour le diabète de type 1, la plus ambitieuse étude du genre est TEDDY (The Environmental Determinants of Diabetes in the Young) qui n'est pas achevée pour le moment [TEDDY Study Group, 2007]. Le TEDDY study group rassemble 6 centres cliniques dont trois aux États-Unis, un en Finlande, un en Suède et un en Allemagne. Les plans initiaux de l'étude prévoyaient de dépister 360000 nouveaux-nés pour détecter des types HLA à haut-risque. Parmi ces nouveaux-nés, 18000 devraient être à risque et 7800 devraient consentir à participer à l'étude. Un suivi jusqu'à l'âge de 15 ans est prévu. Les expositions mesurées couvrent tout ce qui est potentiellement lié à la maladie : de l'alimentation aux maladies infectieuses.

1.3.3 Études rétrospectives cas-témoin : exemple de l'étude Isis-Diab

Nous arrivons maintenant aux études rétrospectives cas-témoin. Dans une telle étude, les patients sont recrutés alors qu'ils sont déjà malades et un groupe de participants non malades, les témoins, est également recruté. On demande alors à tous les participants de répondre à des questions sur leurs expositions. Ces questions portent sur la période précédant la maladie pour les patients et sur une période comparable pour les témoins. C'est le modèle d'étude qui est le plus simple à réaliser notamment quand la maladie est peu fréquente. Mais c'est également le modèle le plus vulnérable à différents biais.

C'est le plus simple à réaliser car contrairement aux études précédentes, les patients sont recrutés alors qu'ils sont déjà malades. Cela permet d'avoir une grande proportion de malades dans l'échantillon et donc d'atteindre une même puissance statistique avec moins de participants. C'est également plus simple car il n'y a pas de suivi à assurer. Tous les événements importants

se sont déjà produits, que ce soit la maladie ou les expositions, et il y a juste à récolter les informations.

Ce modèle est plus vulnérable à différents biais. On a évoqué précédemment le biais de rappel lié au caractère rétrospectif. Un autre problème de taille est la définition du groupe témoin. En effet, si le groupe témoin est choisi dans une population différente de celle dont viennent les patients, les différences entre patients et témoins risquent d'être dus principalement à cela et non aux causes de la maladie.

L'étude Isis-Diab qui est l'application centrale de cette thèse est une étude cas-témoin sur le diabète de type 1. Nous la décrivons plus bas. La partie environnementale de cette étude a consisté à envoyer un long questionnaire environnemental aux patients sur leur environnement avant le diagnostic du diabète. Le groupe témoin a été conçu afin de maximiser les ressemblances entre patients et témoins: il a été demandé aux familles des patients qui ont accepté de remplir le questionnaire environnemental de recruter parmi les amis d'enfance du patient deux témoins. Ainsi, on peut espérer que les témoins partagent certaines caractéristiques (e.g. région d'origine, classe sociale) du patient qui les a recrutés. Le lien entre un patient et les témoins qu'il a recrutés constitue un appariement naturel entre les patients et les témoins de cette étude.

Nous nous tournons maintenant vers les méthodes statistiques qui permettent de contrôler le problème de la confusion que ce soit en tirant parti de cet appariement ou en ajustant sur les variables de confusion mesurées.

1.3.4 Appariement et ajustement sur variables de confusion

L'appariement est le regroupement des observations dans des strates. Ces strates sont censées correspondre à de mêmes valeurs des variables de confusion. Ainsi, en basant les comparaisons uniquement sur les différences internes aux strates, les variables de confusion ne peuvent pas influencer la conclusion de notre étude. Dans l'exemple du questionnaire environnemental de l'étude Isis-Diab, les strates sont les groupes constitués par chaque patient ainsi que le ou les témoins qu'il a recrutés. Ainsi chaque strate peut avoir deux ou trois observations et il faut donc des méthodes qui puissent prendre en compte une taille variable de strate. Nous verrons dans le chapitre 3 un passage en revue de ces méthodes ainsi qu'un résultat d'équivalence asymptotique entre deux de ces méthodes : le test de Hotelling apparié et la régression logistique conditionnelle. Dans le chapitre 4, nous utiliserons ces méthodes pour analyser les données de l'étude Isis-Diab en prenant en compte l'appariement.

L'intérêt de l'appariement, quand il traduit une structure des données, est qu'il permet de contrôler pour des variables de confusion sans avoir à mesurer ces variables ni à les expliciter. Quand il n'y a pas de telle structure, il faut mesurer les variables que l'on soupçonne être des variables de confusion. Il faut ensuite se servir de ces mesures afin de compenser leur effet. La méthode la plus simple pour faire cela est d'ajouter les variables de confusion en tant que variables explicatives dans une régression linéaire. Si on cherche à savoir si un traitement ou une exposition T influe sur une variable de sortie Y mais que l'on pense que X est une variable de confusion qui influe à la fois sur T et sur Y, on effectuera la régression linéaire de Y sur T et X. Si l'on trouve que le coefficient de T est significativement non nul, on dira que Y est associé à T en contrôlant pour X. La régression linéaire consiste à projeter le vecteur $Y \in \mathbb{R}^n$ sur l'espace engendré par le vecteur constant égal à 1, T et X. Cela prend donc en compte

les corrélations entre T et X. Cette méthode permet également de contrôler pour plusieurs variables de confusion à la fois.

Une autre approche, plus riche conceptuellement, a été développée par Rosenbaum and Rubin [1983b]. Ces auteurs se placent dans le cadre des issues potentielles. Supposons que T notre traitement est binaire ($T = 0$ ou 1). Les issues potentielles (Y^1, Y^0) sont les valeurs de l'issue, i.e., de la variable de sortie Y, si le traitement T valait 0 ou 1. Bien sûr, seule une de ces deux valeurs est effectivement observée. Cependant cette formulation nous permet de commencer à réfléchir à des questions contre-factuelles : "Si ce patient avait reçu l'autre traitement, aurait-il eu une meilleure issue ?" et non plus uniquement à des questions factuelles : "Les patients qui ont reçu le traitement $T = 1$ ont-ils eu de meilleures issues ?".

L'outil qui va nous permettre de répondre à ces questions contre-factuelles est le score de propension. Le score de propension est la probabilité de recevoir le traitement si l'on connaît les variables de confusion $X = (X_1, \cdots, X_d)$:

$$e(X) = \mathbb{P}(T = 1|X).$$

Le score de propension capture toute l'information que X contient sur T. Par conséquent, conditionnellement à $e(X)$, X et T sont indépendants : $X \perp\!\!\!\perp T|e(X)$.

Afin de pouvoir répondre à des questions contre-factuelles, il faut supposer que toutes les variables de confusion font partie de X. Cela se traduit par l'indépendance entre les issues potentielles et l'allocation du traitement sachant X: $(Y^1, Y^0) \perp\!\!\!\perp T|X$. Cette hypothèse est l'hypothèse d'ignorabilité de l'allocation du traitement. Si on ajoute à cette hypothèse qu'aucun participant n'est sûr de recevoir le traitement ou de ne pas le recevoir, i.e., $0 < e(X) < 1$, on parle d'ignorabilité forte.

Sous l'hypothèse d'ignorabilité forte de l'allocation du traitement et si on se place à $e(X)$ fixé, les différences d'issues observées sont des estimations non-biaisées des différences d'issues potentielles pour cette valeur d'$e(X)$:

$$\mathbb{E}[Y^1|T = 1, e(X)] - \mathbb{E}[Y^0|T = 0, e(X)] = \mathbb{E}[Y^1 - Y^0|e(X)].$$

Par conséquent, afin de savoir si le traitement a un effet, il suffit de comparer uniquement des participants avec des scores de propension proches. Pour faire cela, il faut tout d'abord estimer ce score de propension. Une fois celui-ci estimé, il y a plusieurs manières de procéder [Austin, 2011]. On peut apparier des participants ayant reçu des traitements différents sur leur score de propension pour ensuite utiliser les méthodes adaptées aux données appariées. On peut estimer l'effet du traitement moyen grâce à la pondération par l'inverse du score de propension. On peut également stratifier tout l'échantillon sur le score de propension. Dans le chapitre 4, nous ferons cela dans une seconde analyse de l'étude environnementale Isis-Diab afin de pouvoir inclure les nombreux patients qui n'ont pas recruté de témoins.

Dans le cas d'un essai randomisé, le score de propension $e(X) = \mathbb{P}(T = 1|X)$ est une constante, souvent 1/2. Ainsi quand on compare la moyenne du groupe traité et du groupe témoin, on estime bien une différence d'issues potentielles. L'appariement ou la stratification sur le score de propension reviennent donc à dire qu'on a fait un certain nombre de petits essais cliniques avec une proportion variable de participants traités et un recrutement non aléatoire dans chaque strate. En utilisant des méthodes d'analyse qui ne prennent en compte les différences

qu'à l'intérieur de chaque strate, on aura effectivement évacué la possibilité d'une confusion de nos résultats.

1.3.5 Causalité et modèles graphiques

La causalité est un concept fondamental dans notre compréhension du monde. Cependant, la statistique n'a pas de définition formelle de celle-ci. Dans un cours typique de statistique, la causalité est évoquée pour dire ce qu'elle n'est pas : corrélation n'est pas causalité. Afin d'argumenter pour le caractère causal d'une association, le praticien utilise des arguments qualitatifs. Le principal argument est l'origine des données : quand il s'agit d'un essai randomisé, on a le droit de conclure à des liens causaux. Cette prééminence de la randomisation est la raison qui fait que nous cherchons à l'émuler avec les scores de propension quand nous avons des données observationnelles.

Dans son livre Causalité [Pearl, 2009], Pearl propose une définition mathématique de la causalité qui donne un rôle central à l'intervention. Il définit une intervention comme étant l'action d'un opérateur extérieur au système étudié et qui fixe la valeur d'une variable de ce système. La force de l'essai randomisé est que les traitements sont des interventions ; la randomisation sert seulement à assurer que l'allocation de cette intervention n'est influencée par rien. Quand nous évoquions précédemment les questions contre-factuelles, il s'agissait déjà de comprendre l'effet d'une intervention.

Pearl défend l'idée que les liens causaux sont des hypothèses qui ne sont pas purement statistiques : les liens causaux ne portent pas uniquement sur la distribution jointe des variables mais également sur leur mécanisme de génération. Il propose d'encoder les liens causaux avec des graphes dirigés acycliques et des équations structurelles. Considérons le graphe de la figure 1.2. Chaque nœud du graphe correspond à une variable aléatoire et les arêtes orientées sont les liens causaux qui relient ces variables. L'absence d'arête entre deux sommets revient à dire que les variables correspondantes sont indépendantes. La loi jointe des variables A, B, C et D peut alors se factoriser sous la forme suivante :

$$\mathbb{P}(A, B, C, D) = \mathbb{P}(D|B, C)\mathbb{P}(C|A, B)\mathbb{P}(B)\mathbb{P}(A).$$

De manière générale, on considère des variables $X_1, \cdots, X_d$. On définit les parents de X_i comme les sommets qui ont des arêtes dirigés vers le sommet X_i. On note pa_i l'ensemble des parents de X_i. L'absence de cycle dirigé permet de factoriser la loi jointe des X_i :

$$\mathbb{P}(X_1, \cdots, X_d) = \prod_{i=1}^{d} \mathbb{P}(X_i|\mathrm{pa}_i).$$

Il reste ensuite à définir les $\mathbb{P}(X_i|\mathrm{pa}_i)$ et c'est là que les équations structurelles vont servir. Ce sont des équations qui décrivent la loi de X_i en fonction des valeurs de ses parents.

D'autres factorisations de la distribution jointe sont possibles pour un même graphe si on ne prend pas en compte la direction des arêtes. Cependant la factorisation ci-dessus traduit la structure causale et nous permet de définir ce qu'est une intervention et son effet. Pour dire qu'une intervention fixe la valeur de X_j à x_j, on note $\mathrm{do}(X_j = x_j)$. Une intervention casse les liens causaux qui déterminaient précédemment X_j et un opérateur extérieur fixe la valeur de

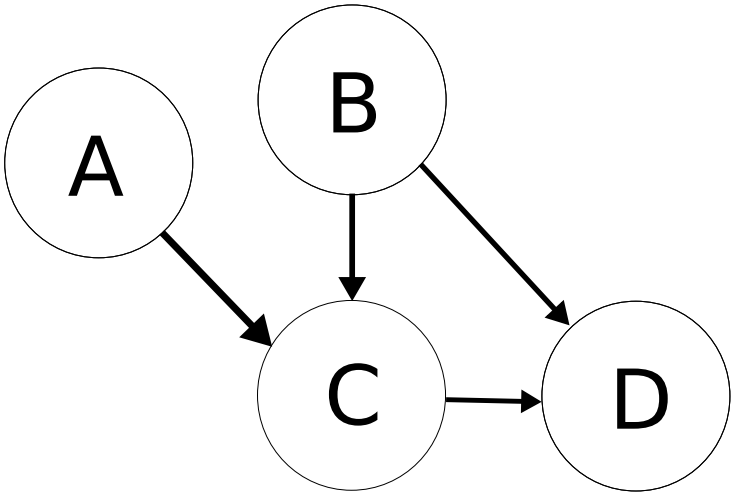

Figure 1.2: Exemple de graphe acyclique dirigé.

cette variable. La loi des X_i sous l'intervention $\text{do}(X_j = x_j)$ prend la forme d'une factorisation tronquée :

$$\begin{aligned}\mathbb{P}(X_1, \cdots, X_m|\text{do}(X_j = x_j)) &= \prod_{i \neq j} \mathbb{P}(X_i|\text{pa}_i) \text{ si } X_j = x_j \\ &= 0 \text{ sinon.}\end{aligned}$$

Par rapport à l'expression précédente, on a enlevé le facteur $\mathbb{P}(X_j|\text{pa}_j)$. Cela traduit bien que l'intervention soustrait X_j à l'influence de ses parents.

Cette définition permet de calculer les effets d'une intervention quand on connaît la structure causale. Dans la pratique, une question importante est de déterminer cette structure à partir de données [Maathuis et al., 2010], c'est ce qu'on appelle le problème inverse. Les modèles graphiques et leur utilisation causale forment un domaine de recherche actif à la fois d'un point de vue théorique et d'un point de vue appliqué.

Ces aspects graphiques de la causalité sont présents dans le chapitre 5 où nous parlerons d'étude gène-environnement portant uniquement sur les cas. Ce chapitre fera intervenir la notion de collisionneur. Un collisionneur est un sommet vers lequel plusieurs arêtes pointent. Par exemple, les nœuds A et B pointent vers le sommet C de la figure 1.2 et C est donc un collisionneur. Quand on conditionne sur un collisionneur, on crée une corrélation entre les parents du collisionneur, c'est ce qu'on appelle le biais de collision. Les gènes et l'environnement provoquent la maladie qui est donc un collisionneur. En ne considérant que des cas, on conditionne sur le collisionneur et on crée donc une corrélation entre la génétique et l'environnement.

1.3.6 Randomisation mendélienne

L'un des développements récents les plus importants de l'épidémiologie se trouvent dans un autre modèle d'étude qui croise la génétique et l'environnement: la randomisation mendélienne. Elle

permet d'étudier grâce à la génétique les liens entre phénotypes intermédiaires et phénotypes d'intérêt afin de déterminer si ces liens sont causaux ou de simples corrélations [Davey Smith and Hemani, 2014].

Illustrons ceci avec un exemple. Après avoir observé que le bon cholestérol (phénotype intermédiaire) était associé à un risque plus faible d'avoir des maladies cardiaques (phénotype d'intérêt) [Miller et al., 1977], l'industrie pharmaceutique a développé des médicaments pour augmenter le bon cholestérol. Malheureusement, aucun essai thérapeutique ne démontra de bénéfice à augmenter le bon cholestérol [Barter et al., 2007, Schwartz et al., 2012].

L'idée de la randomisation mendélienne est de concevoir les gènes qui ont été associés à la concentration de bon cholestérol comme des médicaments qui sont attribués au hasard à la naissance. Ainsi, le groupe qui a un gène qui augmente le bon cholestérol est le groupe "traité" et le reste de la population étudiée est le groupe "témoin". En comparant le taux de maladies cardiovasculaires chez les personnes "traités" et les autres, on peut déduire si l'augmentation du bon cholestérol est effectivement protecteur. En 2012, une étude de randomisation mendélienne sur ce sujet a montré qu'augmenter le bon cholestérol ne protège pas des maladies cardiovasculaires [Voight et al., 2012]. L'association observée entre le bon cholestérol et le risque diminué de maladies cardiovasculaires n'est donc pas causal.

La randomisation mendélienne imite ainsi un essai clinique afin de décortiquer les liens causaux entre des phénotypes intermédiaires comme le cholestérol et des phénotypes d'intérêt, souvent des maladies. La seule limite est donc de trouver un gène qui provoque ce dont on veut étudier les conséquences.

Ce modèle d'étude est identique à la méthode des variables instrumentales très utilisée en économétrie. Revenons à la figure 1.2. Admettons qu'on cherche à déterminer si le lien de C vers D est causal. Une variable instrumentale est une variable qui est un parent de C, qui n'influe sur D qu'à travers son influence sur C et qui est indépendante de toute autre source de confusion. Dans notre graphe, A est le seul instrument valide. Pour reprendre notre exemple, l'instrument A est le gène, C est le bon cholestérol et D est la maladie. La génétique fournit de bons instruments car un gène sera en général indépendant des autres variables de confusion. La deuxième hypothèse peut être contrôlée si on a une compréhension précise de la fonction du gène choisi comme instrument.

1.4 Étude Isis-Diab

L'étude Isis-Diab a été initiée par Pierre Bougnères et Alain-Jacques Valleron et pilotée par Sophie Le Fur et Sophie Valtat. L'ambition initiale était d'étudier le diabète de type 1 sur de multiples fronts : à la fois les causes de la maladie, qu'elles soient génétiques, environnementales ou épigénétiques, et également les complications qu'entraîne la maladie. Dans cette thèse, il ne sera question que de la partie de l'étude concernant l'élucidation des causes génétique et environnementale du diabète de type 1.

Le recrutement des patients, qui a commencé en 2007, s'est fait par la constitution d'un réseau de centres d'endocrinologie couvrant une grande partie du territoire français : le groupe collaboratif Isis-Diab. Chaque centre participant, après le consentement du patient, l'inclut dans la base de données et des échantillons sanguins sont prélevés et conservés dans une biobanque.

L'étude environnementale a consisté à envoyer à partir de 2010 trois longs questionnaires avec plus de 850 questions à 6618 patients. En plus de remplir leur exemplaire du questionnaire, il leur a été demandé de faire remplir deux questionnaires par des témoins, des amis d'enfance du patient. 1769 patients et 1085 témoins ont renvoyé le questionnaire complété. Parmi ces réponses, il y a 251 trios complets (un patient et deux témoins appariés), 451 duos (un patient et un témoin apparié), 1077 patients seuls et 152 témoins seuls.

Étant donné le faible taux de réponse, un questionnaire plus court fut préparé afin d'obtenir quand même des informations des patients qui n'avaient pas répondu au grand questionnaire. Ainsi, quand les patients ne renvoyaient pas le grand questionnaire, un petit questionnaire d'une cinquantaine de questions leur était envoyé. 1321 patients ont renvoyé ce petit questionnaire mais seulement 187 témoins y répondirent. Parmi ces réponses, on comptent seulement 36 trios complets et 60 duos.

La partie génétique de l'étude consista à génotyper des patients. Il était initialement prévu de génotyper également les témoins de l'étude environnementale mais très peu acceptèrent de participer à la partie génétique de l'étude. Le génotypage se fit au Centre National de Génotypage alors dirigé par Mark Lathrop. Le recrutement de l'étude eut lieu alors que la technologie des puces ADN s'améliorait très vite. Par conséquent, le génotypage se fit d'abord sur une puce de 370 000 SNPS, puis 610 000 SNPs et enfin 4,5 millions de SNPs. Ainsi, il y a 806 patients génotypés sur la puce Illumina HumanCNV370-Duo, 876 génotypés sur la puce Illumina Human610-Quad et 934 génotypés sur la puce Illumina Human Omni 5 Exome.

Pour mettre à profit ces données hétérogènes, il est heureusement possible de deviner les valeurs manquantes de puces ADN *in silico* grâce à Impute 2, logiciel qui tire parti du déséquilibre de liaison [Howie et al., 2009]. Ceci fut effectué par Yoichiro Kamatani sous la supervision de Mark Lathrop à l'université McGill à Montréal. Malheureusement, les données de la plus petite puce après imputation avaient de nombreux problèmes de qualité de données et nous avons dû les exclure.

Dans le chapitre 5, nous nous intéresserons aux données croisées gène-environnement. Il y a au total 831 patients pour lesquels on a des données génétiques et environnementales (grand ou petit questionnaire) et qui sont de plus d'origine européenne. La restriction sur l'origine sert à s'assurer de la validité du score de risque génétique.

1.5 Contributions

Après ce tour d'horizon de la génétique, de l'épidémiologie et des méthodes statistiques pour les données observationnelles ainsi qu'un aperçu de l'étude Isis-Diab, nous pouvons nous tourner vers les apports de cette thèse. Cette dernière est structurée en 5 chapitres. Le chapitre 2 porte sur la prédiction du risque génétique, notamment en utilisant les haplotypes et est disponible sur *PeerJ preprints*. Le chapitre 3 concerne un résultat d'équivalence asymptotique entre deux méthodes d'analyse de données appariés et est disponible sur *arXiv*. Le chapitre 4, co-écrit avec Sophie Le Fur, Sophie Valtat, Alain-Jacques Valleron, Pierre Bougnères et le groupe collaboratif Isis-Diab, présente l'analyse du questionnaire environnemental d'Isis-Diab et a été publié dans *BMC Public Health*. Le chapitre 5, co-écrit avec Sophie Le Fur, Pierre Bougnères et Alain-Jacques Valleron, traite des interactions et du biais de collision dans les données gènes-environnement des études

patients seuls et a été soumis (disponible sur *BioRxiv*). Le chapitre 6, co-écrit avec Gérard Biau, Philippe Ravaud et Raphaël Porcher, parle d'évaluation de la personnalisation du traitement en prenant en compte la politique de personnalisation et a été soumis (disponible sur *arXiv*). Les chapitres peuvent être lus indépendamment sauf le chapitre 5 pour lequel il est préférable d'avoir lu le chapitre 2 et 4. Les forêts aléatoires, un algorithme d'apprentissage statistique décrit au début du chapitre 2, sont également utilisées dans les chapitre 4 et 6.

Chapitre 2

Le chapitre 2 commence par une présentation de deux algorithmes d'apprentissage statistique: la régression lasso et les forêts aléatoires. Nous proposons ensuite une méthode qui prend en compte l'information de phase dans la prédiction de risque génétique : PH (Prediction with Haplotypes). Cette méthode d'apprentissage en plusieurs étapes combine la régression lasso et les forêts aléatoires. Elle capture les interactions locales dans des haplotypes courts et combine les résultats linéairement. Nous la comparons avec des variantes de notre méthode sur les données du WTCCC1. Nos résultats indiquent qu'une légère amélioration peut être obtenue en prenant en compte la structure métrique de l'ADN mais qu'il n'est probablement pas nécessaire d'utiliser l'information de phase pour cela.

Dans ce chapitre, nous reproduisons également le score de prédiction génétique pour le diabète de type 1 proposé dans Wei et al. [2009] et entraîné sur les données du WTCCC et nous l'appliquons aux patients d'Isis-Diab.

Chapitre 3

Le chapitre 3 commence par un aperçu des différentes méthodes permettant d'analyser les données appariées ainsi que leur condition d'application. Nous montrons ensuite un résultat d'équivalence asymptotique entre deux de ces méthodes : le test du score de la régression logistique conditionnelle et le test de Hotelling apparié.

L'appariement des patients et des témoins est pris en compte dans l'analyse des données des questionnaires environnementaux de l'étude Isis-Diab dans le chapitre suivant.

Chapitre 4

Dans le chapitre 4, nous exposons l'analyse des questionnaires environnementaux de l'étude Isis-Diab. Nous présentons deux analyses complémentaires sur le grand questionnaire afin de tester l'influence de l'environnement. La première analyse est basée sur l'appariement naturel qui découle de l'organisation de l'étude. Cela exclut de nombreux patients sans témoins appariés et également des témoins sans patients. La deuxième analyse permet de les prendre en compte. Dans cette deuxième analyse, un score de propension est estimé et sert à stratifier l'échantillon. Nous portons une grande attention aux biais potentiels qui pourrait affecter l'étude. Nous étudions en particulier des problèmes liés à l'âge au diagnostic des patients et à l'âge de référence des témoins correspondants.

En plus de cette analyse déjà publiée sur le grand questionnaire, ce chapitre inclut une réplication de l'analyse appariée sur les données du petit questionnaire.

Chapitre 5

Le chapitre 5 commence par la présentation du modèle d'étude patients-seuls pour les interactions gène-environnement. Nous présentons un cadre de simulation afin de déterminer l'importance du biais de collision dans les données patients-seuls gène-environnement. Cela nous permet de montrer que les conclusions d'une étude patients-seuls sur les interactions entre les gènes BRCA et l'environnement dans le cancer du sein sont affectés par le biais de collision [Moorman et al., 2010]. Nous proposons ensuite un nouveau modèle d'étude sur le même genre de données qui s'appuie sur la prédiction de risque génétique et le biais de collision afin de confirmer des associations environnementales et que nous nommons DAC (Disease As Collider). Nous illustrons DAC sur Isis-Diab en analysant les corrélations entre risque génétique développé au chapitre 2 et questionnaire environnemental présenté au chapitre 4. En modifiant notre cadre de simulation, nous estimons la puissance statistique de DAC. Ces simulations montrent que DAC a très peu de puissance dans le cadre d'Isis-Diab et montrent dans quelle circonstance cette méthodologie pourrait être utile.

Chapitre 6

Le chapitre 6 porte sur autre approche de la médecine personnalisée : la personnalisation des traitements. Un essai thérapeutique permet de comparer l'effet moyen de deux traitements $T = 1$ et $T = 0$ sur une variable d'intérêt Y. Si un des deux traitements, disons $T = 1$, est significativement meilleur que l'autre, alors il sera donné à tous les patients. Cependant, il est possible qu'il y ait une hétérogénéité dans la réponse au traitement et qu'une grande proportion des patients répondent mieux au traitement 0. Afin de profiter de cette hétérogénéité, il faut donc prédire à partir de covariables X quels vont être les patients qui vont profiter du traitement alternatif. On cherche donc à déterminer une politique de traitement, c'est-à-dire décider à qui on va attribuer le traitement 1 et le traitement 0. Une politique est une fonction $\texttt{pol} : x \mapsto \texttt{pol}(x) \in \{0, 1\}$. On note $\Delta(X) = \mathbb{E}(Y^1|X) - \mathbb{E}(Y^0|X)$ la différence d'effet des traitements pour un patient avec les covariables X. On s'intéresse au bénéfice de la politique $\texttt{pol}$ par rapport au statu quo :

$$\Theta(\texttt{pol}) = \mathbb{E}_X[-\Delta(X)\mathbb{1}_{\texttt{pol}(X)=0}].$$

Cette quantité mesure le bénéfice global qui découle de la personnalisation. Bien sûr la politique va également être estimée et on aura donc un $\hat{\texttt{pol}}$. En utilisant des arguments Bayésiens, nous proposons une estimation de $\Theta(\hat{\texttt{pol}})$ ainsi que des bornes de crédibilité. Ceci nous amène à définir une nouvelle politique: la Borne inférieure max (max lower bound $\hat{\texttt{mlb}}$). Cette politique permet de détecter un bénéfice significatif de la personnalisation aussi souvent que possible. Les propriétés de cette politique sont établis par des résultats théoriques ainsi que des simulations. Nous appliquons notre méthode à un essai thérapeutique de l'aspirine complémentaire dans le traitement des crises cardiaques et nous identifions un bénéfice à personnaliser, c'est-à-dire à ne pas donner d'aspirine à certains patients.

Chapter 2

Genetic risk prediction for complex diseases

Abstract Genome-wide association studies (GWAS) have uncovered thousands of associations between genetic variants and diseases. Using the same datasets, prediction of disease risk can be attempted. Phase information is an important biological structure that has seldom been used in that setting. We propose here a multi-step machine learning method that aims at using this information. Our method captures local interactions in short haplotypes and combines the results linearly. We show that it outperforms standard linear models on some GWAS datasets. However, a variation of our method that does not use phase information obtains similar performance. Source code is available on github at https://github.com/FelBalazard/Prediction-with-Haplotypes. We replicate the genetic risk estimation of Wei et al. [2009] on the Isis-Diab patients.

Contents

2.1 Introduction

Genome-wide association studies (GWAS) have used micro-array technology to genotype hundreds of thousands of single nucleotide polymorphisms (SNPs) in thousands of patients and controls. The main goal of those studies has been to identify associations between SNPs and diseases that could help understand the genetic component of the disease. The methodology used for this purpose relies on univariate hypothesis tests with corrections for multiple testing. GWAS have unraveled over one thousand new SNP disease associations [Welter et al., 2014].

A potential clinical utility of GWAS is implementation of personalized medicine. For example, genetic risk prediction could be useful for prevention of complex diseases. Unfortunately, the SNPs found significant in GWAS are not sufficient in aggregate to be used for prediction of disease status [Manolio et al., 2009]. However, it is not necessary that each variable passes a stringent p-value threshold to be useful in a multivariate setting. In order to increase predictive power, one can use a lenient significance threshold to preselect SNPs before combining them.

One way of combining SNPs has been to add the univariate effect sizes of each pre-selected variant to form a polygenic risk score. As those variants are not independent, corrections have to be made, most notably to take into account linkage disequilibrium [Chatterjee et al., 2016]. On a first impression, this approach seems to ignore that machine learning is a field of research dedicated in part to the problem of predicting from data. However, there are valid reasons for the popularity of this approach. Besides their simplicity, polygenic risk scores can be computed using only summary GWAS data that are easily available and do not require individual level data. They can therefore leverage the information available in the largest meta-analyses for which access to individual level data is limited.

This lack of access for researchers to the largest GWAS datasets is due to concern over privacy and medical confidentiality and is often enshrined in the consent form that participants signed before joining a study. International efforts to enhance the sharing of data are underway [Global Alliance for Genomics and Health, 2016] and recent high-profile initiatives, such as the UK Biobank, have made data sharing central in their endeavor [Allen et al., 2014]. As the main contribution of this chapter was written (preprint available in May 2016) before the UK Biobank made available its genetic dataset of 500000 participants (summer 2017), we will comment on the subsequent evolution of the field at the end of the discussion of this chapter.

In this chapter, we will focus on applying machine learning to predict genetic risk of disease. This implies access to individual level data. In previous work, algorithms such as support vector machine or lasso regression have been used to predict type 1 diabetes [Wei et al., 2009] and Crohn's disease [Wei et al., 2013]. There is often still a preselection step to allow for manageable computation time. Optimization efforts were made to allow L^1-penalized linear regression with square-hinge loss to be run over the whole dataset [Abraham et al., 2012]. This was applied to celiac disease [Abraham et al., 2014].

The methodology used in those articles is to apply general purpose machine learning algorithm to GWAS datasets. The biological structure of genetic data is therefore not taken into account. A first example of such a structure is distance inside chromosomes measured in base pairs. In [Botta et al., 2014], the T-trees method was introduced to capture interactions inside small blocks of nearby SNPs as well as between blocks. The rationale is that SNPs that are next to each other are more likely to impact the same function and therefore to interact (in

the statistical sense) together. The T-Trees method is a variation on the random forests (RF) [Breiman, 2001b] algorithm tailored to focus on local interaction between SNPs. It can also assess the importance of individual SNPs as well as the importance of blocks of SNPs. It is very successful in increasing predictive performance compared to RF or linear methods. It also identifies new associations between loci and disease.

Phase information (cf. figure 1.1 and corresponding text) is the second structure that we will use in our design of a machine learning algorithm tailored to GWAS data. It complements chromosomal distance. It is reasonable to expect that two SNPs that are physically on the same chromosome and not too distant are more likely to interact than if they are not. Interaction, here and throughout this chapter, is understood in the statistical sense as a departure from linear effects. Previous work has used haplotypes of two contiguous SNPs with a methodology similar to polygenic risk scores for prediction of Crohn's disease [Kang et al., 2011]. Their results are suggestive of the interest of haplotypes in a predictive context. Phase information is not available using micro-array technology but computational methods have been developed to allow phase imputation [Delaneau et al., 2013]. They have limited accuracy which means that only short haplotypes should be used.

Up to this point, we only discussed the potential interest of haplotypes regarding prediction accuracy but haplotypes are also interesting for heritability. Heritability quantifies the proportion of phenotypic variance explained by genetic factors. It is estimated through family studies. It can give upper bounds for prediction accuracy [Wray et al., 2010]. A distinction exists between broad-sense heritability and narrow-sense heritability [Visscher et al., 2008]. The latter only includes additive effects while the former also includes interaction terms such as dominance and epistasis. The rationale behind this distinction is that when estimating heritability with pedigrees, one only estimates narrow-sense heritability. Dominance and epistasis are lost due to genetic mixing. However, interactions in haplotypes are actually part of narrow-sense heritability as they are shared among all members of the same family. Haplotypes are the support of heredity and not single SNPs. Narrow-sense heritability includes additive effects of *haplotypes*. Of course, long haplotypes are broken by recombination but short haplotypes are seldom concerned. This may sound counter-intuitive but the interactions inside haplotypes are part of additive heritability.

Considering interaction in haplotypes is more general than the idea that for each association signal at a locus there is a causal variant responsible for it. If there is a causal variant that is not part of the typed SNPs but that is associated with a particular haplotype, capturing interaction in haplotype should recover this variant's effect better than relying on unphased data. If the variant is only in a subset of the haplotype, the effect will be diluted but will still be captured more precisely. Moreover, it is possible that there exists haplotypic effects not linked to a single variant.

We begin this chapter by introducing two popular machine learning algorithms: logistic regression with lasso penalty as well as random forests that we will use throughout this dissertation. The main contribution of this chapter is to introduce a multi-step machine-learning method –noted PH for Prediction with Haplotypes– that captures interactions in short haplotypes centered around association signal, then combines the results using Lasso regression. This can be seen as logistic regression by blocks. In order to know what phase information adds to the analysis, we also applied a similar method on genotypes and not haplotypes. We

also adapt our method to capture dominance effect between the two haplotypes at a same loci. We compare our method and its two variations to lasso regression with preselection on GWAS datasets made available by the Wellcome Trust Case Control Consortium (WTCCC) [Burton et al., 2007]. Finally, we replicate the externally-validated genetic risk score designed by Wei et al. [2009] on the Isis-Diab genetic data.

2.2 Machine learning

In this section, we briefly describe the machine learning methods used in PH: lasso logistic regression and random forests. They are part of supervised learning, the subset of machine learning devoted to prediction. A more detailed presentation of those techniques is available in the monograph [Friedman et al., 2001].

We introduce a few notations: we have n observations (in our case participants) of d variables (for example, SNPs) that we can summarize in an n by d matrix $X = (x_{ij})$. The value of variable X_j for observation i is x_{ij}. We also have a binary response variable $Y = (y_i)$ that we want to predict using X. In our case, it is disease status and $y_i = 1$ if individual i is diseased and 0 otherwise.

2.2.1 Lasso logistic regression

In logistic regression, the posterior probability of being a case or a control is modeled by a linear combination of the variables :

$$\log \frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d.$$

The vector $\beta = (\beta_0, \beta_1, ..., \beta_d)$ of weights is chosen to maximize the likelihood of the training data. When the dimension d becomes large compared to n, the maximum likelihood estimate will closely fit to the training data but have no predictive power on test data. This problem is called overfitting. To address this issue, several penalization procedures have been proposed. They force the model to be simpler and keep predictive power [Friedman et al., 2001, p. 61-73]. We will use L^1 penalization also known as Lasso [Tibshirani, 1996a] as it has the nice additional property of sparsity: some variable's coefficients will be assigned to 0 which makes the model more interpretable.

The function $x \in (0, 1) \mapsto \log(\frac{x}{1-x})$ is called the logit function. It is the link function in logistic regression. It maps $(0, 1)$ onto $(-\infty, \infty)$.

2.2.2 Decision trees and random forests

Decision trees are non-linear machine learning algorithms. They can capture interactions between variables and are easily interpretable. They are represented by a binary tree with a binary test of the form $X_j > c$ on each node [Friedman et al., 2001, p. 305-316].

However, single decision trees are often poor classifiers. They also are very unstable as all splits are conditioned by the first split. To take advantage of this instability, the random forests (RF) algorithm was designed [Breiman, 2001b]. It consists in randomizing the growth process

of the tree, growing many independent such random trees and aggregating the results of all the trees [Friedman et al., 2001, p. 587-604]. It is very effective in increasing predictive accuracy. While the algorithm is less interpretable than logistic regression, an importance score can be attributed to variables to rank their relevance to prediction accuracy. RF's use in computational biology and its challenges are reviewed in [Boulesteix et al., 2012].

One of the ways that the trees are randomized is that they use bootstrap versions (subsets obtained by sampling with replacement) of the training data to train different trees. This means that a specific observation will not be used to train all the trees. For the trees that did not use the observation, we say the observation is out-of-bag. For each observation, we can look at all the trees for which it is out-of-bag and aggregate the predictions of those trees for the observation. This allows to have predictions on the training set that should behave similarly to predictions on the test set, i.e., without overfitting. This will be critical in our setting.

2.3 Prediction with Haplotypes

The diploid nature of the genome is an important structure left mostly unused in earlier attempts at genetic risk estimation. It is challenging to use this information from a machine learning point of view. Indeed, once phasing is performed, we have the same set of variables twice but with different values and a metric structure. Interactions inside short haplotypes are what we aim to exploit thanks to phase information.

2.3.1 Algorithm

A preliminary step is to use Shapeit 2 [Delaneau et al., 2013] to obtain estimates of haplotypes.

The first step of the algorithm is performing an univariate test of association of SNPs to disease on the unphased training set. This is done using PLINK [Purcell et al., 2007]. This is to work with a computationally manageable number of variables. We define blocks around the most associated SNPs. Those blocks consist of all the SNPs (not only highly associated ones) under a fixed distance in kb (thousand of base pairs) from the associated (or central) SNP as shown in Fig.2.1. The window size L_w is an important hyper-parameter with biological significance. The order of magnitude we used for L_w is 10 kb. Blocks are allowed to overlap but the central SNP of a block must be outside of the other blocks. Therefore, a highly associated SNP will not be used to define a block if its distance with a more highly associated SNP is smaller than the half window size, i.e., the SNP is already included in a block. Besides reduced computation, the motivation of centering the blocks on associated SNPs compared to using a fixed grid as in [Botta et al., 2014] is to be able to capture the important interactions. If two interacting nearby SNPs fall by misfortune on both sides of the border between two blocks, their interaction will not be captured. It seems reasonable to assume that the locally highest associated SNP will be part of the local interaction if there is one. The number of blocks N_b is another hyper-parameter.

Inside a block, we want to capture interactions inside haplotypes. For each observation, there are two haplotypes and therefore, we have two times the same set of variables with different values. We treat each haplotype as a distinct observation and attribute to it the response variable of the individual it belongs to. We train random forests on the haplotypes of the training set and this gives us an estimated probability that the haplotype belongs to a diseased person. This

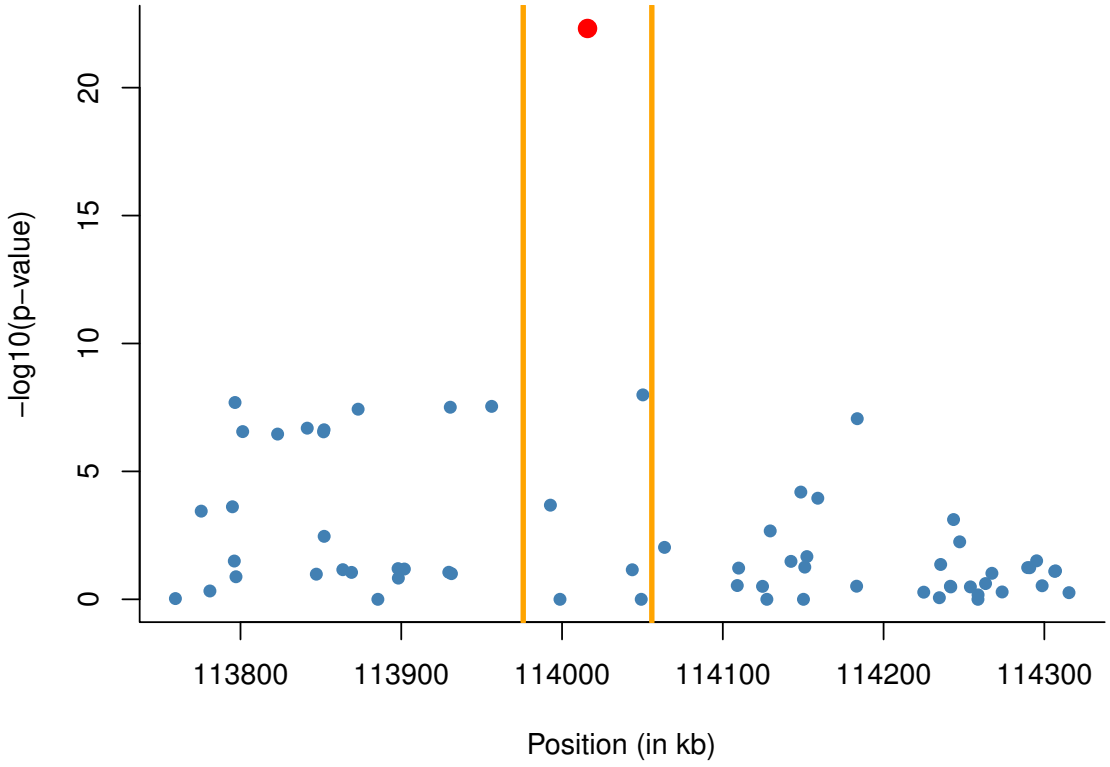

Figure 2.1: **Block definition around associated SNP** Blocks include all SNPs at distance smaller than L_w of the central SNP.

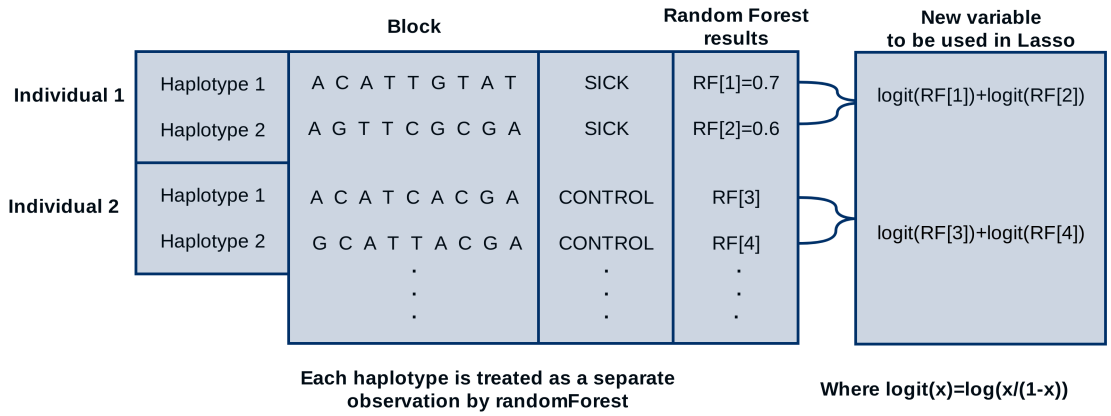

Figure 2.2: **Interactions inside haplotypes captured with Random Forests.** Inside a block, each haplotype is treated as a distinct observation. The SNPs in the block are the input variables used to train a RF predictor. The results are then combined into a variable that summarizes the information contained in the block. The abbreviation kb stands for kilobases, i.e., thousands of base pairs.

estimated probability is the out-of-bag estimate for haplotypes belonging to the training set and the prediction using the full forest for the test set. The Gini impurity was used as node-splitting criterion. The default value for classification was used for the m_{try} parameter. We tried different values for N_{leaf} the minimum number of data points in a leaf. At the level of the haplotype, we are interested in estimating probabilities and not in classification, therefore the mean (over the forest) predicted probability was used as the method of aggregation of results. As the computation for one block are independent from the computation in the other blocks, computation was parallelized.

Every individual has two estimated probabilities of being sick that come from the two independent haplotypes. We combine those by adding the logit of probability of being sick for the two haplotypes. This gives us a new variable that is the evidence of being sick given the haplotypes in the block. There is one such variable for each block. This step combines the results for the two haplotypes in a principled way and it builds a variable that is homogeneous to logistic regression. These two steps are illustrated in Fig. 2.2.

For each block, we obtain one variable that summarizes the information we obtained from it. We use those N_b variables as predictors of disease using Lasso regression. We train the Lasso regression on the block variables obtained for the training set. Using the trained regression model, we predict on the block variables obtained for the validation set. The full procedure with emphasis on training set and test set separation is summarized in Fig. 2.3.

Variation of the method The two variations of the method we considered differ from PH only in the computation inside blocks illustrated in Fig 2.2.

The first variation of PH is designed to look at whether phase information increases predictive

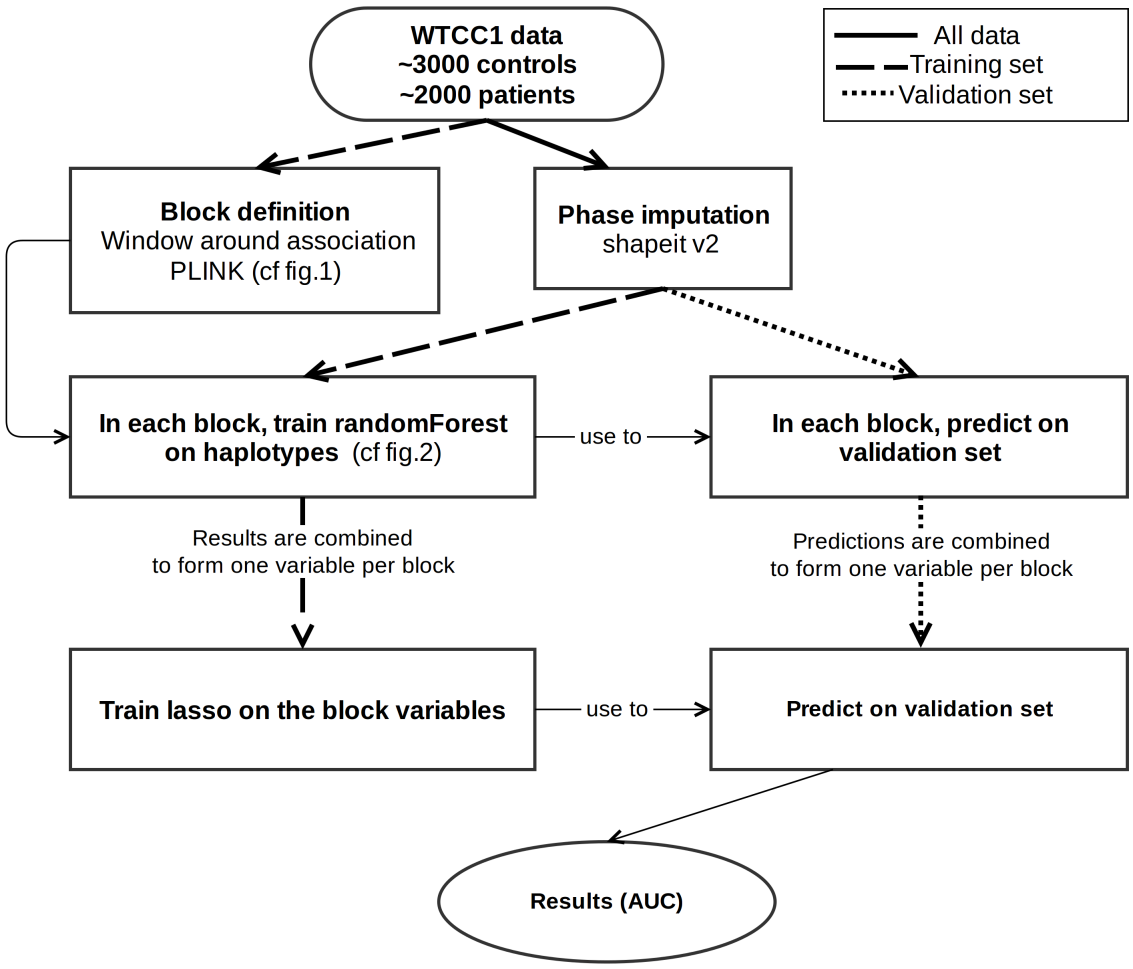

Figure 2.3: **Pipeline of the method** The different kinds of line indicate the separation between training set and validation set.

accuracy or if the same information can be captured using SNPs. It is the closest variation of PH not using haplotypes. Block definition stays the same but inside the block, we train random forests on SNPs instead of using haplotypes. We only have one result and we compute its logit to create a new variable in the same way as before. This variation is no longer capturing only additive heritability as it can potentially capture dominance effect. We call it PwoH for Prediction without Haplotypes.

The second variation we consider aims at capturing dominance effect. Dominance is understood in a broad sense as interaction between the two haplotypes of the same loci. Inside blocks, we train random forests on pairs of haplotypes instead of single haplotypes by concatenating the haplotype of the homologous chromosome. Each individual is thus still represented by two observations varying only by the order in which the two haplotypes appear. There are therefore twice the number of variables inside each block compared to PH. We call it PHd for Prediction with Haplotypes and dominance.

Comparison point We compared the three variations of the method to lasso regression with preselection. First, the N most associated SNPs in the training set were selected. Lasso regression is then fitted to the training set. The penalization parameter is selected through cross-validation in the training set. The resulting regression model is then used to predict on the validation set. The number N of preselected variables is a hyper-parameter.

Implementation details The source code is a mix of bash, R and python scripts, uses Plink [Purcell et al., 2007] and Shapeit 2 [Delaneau et al., 2013] and is available on github at https://github.com/FelBalazard/Prediction-with-Haplotypes. The glmnet R package was used for lasso regression [Friedman et al., 2010]. The python machine learning package scikit-learn was used for random forests [Pedregosa et al., 2011].

2.3.2 Datasets and protocol

We tested our method on the GWAS datasets made available by the WTCCC and first described in [Burton et al., 2007]. The WTCCC data collection contains 17000 genotypes, composed of 3000 shared controls and 2000 cases for each of 7 complex diseases: bipolar disorder (BD), Crohn's disease (CD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). Individuals were genotyped with the Affymetrix GeneChip 500K Mapping Array Set and are described by about 500,000 SNPs (before the application of quality control filters).

Quality control (QC) is important for GWAS datasets. Corrupt variables can allow for almost perfect discrimination while not respecting Hardy-Weinberg Equilibrium (HWE) [Botta et al., 2014]. We first excluded the exclusion lists for individuals and SNPs used in [Burton et al., 2007] and provided with the data. Then, for each disease, an exclusion list was defined for SNPs that were missing in more than 5% of the individuals (patients and controls), that had a minor allele frequency smaller than 0.1% or that had a p-value for HWE smaller than 10^{-6} for controls or smaller than 10^{-10} for patients.

With Shapeit 2, phasing accuracy increases with sample size [Delaneau et al., 2013]. To achieve maximum accuracy, we phased all the 17000 patients and controls together excluding

only the intersection of all disease specific exclusion lists for SNPs. We then used the disease specific exclusion list to obtain each phased disease dataset with proper exclusions.

The predictive performance of all methods were assessed by the area under the ROC curve (AUC). We performed 10-fold cross-validation and averaged the AUCs over the 10 folds. The same 10 folds are used for the different methods to limit variability.

2.3.3 Results

In this section, we present our results on the seven WTCCC datasets. We first investigate the importance of two hyperparameters on the CD dataset. We then use parameters that obtained good performance on the CD dataset to evaluate predictive performance and influence of window size on the 7 datasets.

2.3.4 Influence of the hyperparameters

Concerning lasso regression with preselection, we had one hyperparameter to select: the number N of pre-selected SNPs. We tried the values 500, 1000 and 1500 on all diseases. The best result for all diseases except BD was obtained for $N = 500$. We use the values obtained for $N = 500$ in the following. The results are available in the online supplementary material of the preprint [Balazard, 2016].

On the CD dataset, we studied the influence of two hyperparameters of PH and its variants: the minimum number of data points in a leaf N_{leaf} and the number of blocks N_b.

For N_{leaf}, the values 1, 5, 10, 15, 25, 50, 100 were assessed. The results (available as online supplementary material [Balazard, 2016]) imply that the choice of $N_{leaf} = 5$, the standard value for regression, could not be improved upon notably by another choice of value for this parameter. We chose $N_{leaf} = 5$ for subsequent analysis.

For N_b, we tried the values 300, 500, 700. The results (available as online supplementary material [Balazard, 2016]) were similar for all three values. We chose $N_b = 500$ for subsequent analysis.

2.3.5 Predictive performance and influence of the window size

Given the biological significance of window size L_w, we studied its influence on all diseases. The values 10kb, 20kb, 30kb, 40kb, 60kb, 80kb, 100kb and 150 kb were tried. Results are shown together with the result for preselection and lasso in Fig.2.4 and are also available in the online supplementary material. When the window is too large, prediction accuracy is impaired. This is true except for T1D for which performance seems stable.

For CAD, T2D and to a lesser extent HT and RA, the best performance is obtained for intermediate values of the window size. The optimal value is 60kb for CAD, 40kb for T2D and 20kb for HT and RA.

Lasso slightly outperforms our methods on three out of seven datasets: BD, CAD and HT. This shows that our methods do not always recover all the information contained in the central SNPs. For RA and T1D, performances are very similar for all methods. However, for CD and T2D, our methods outperform lasso for most values of window size considered.

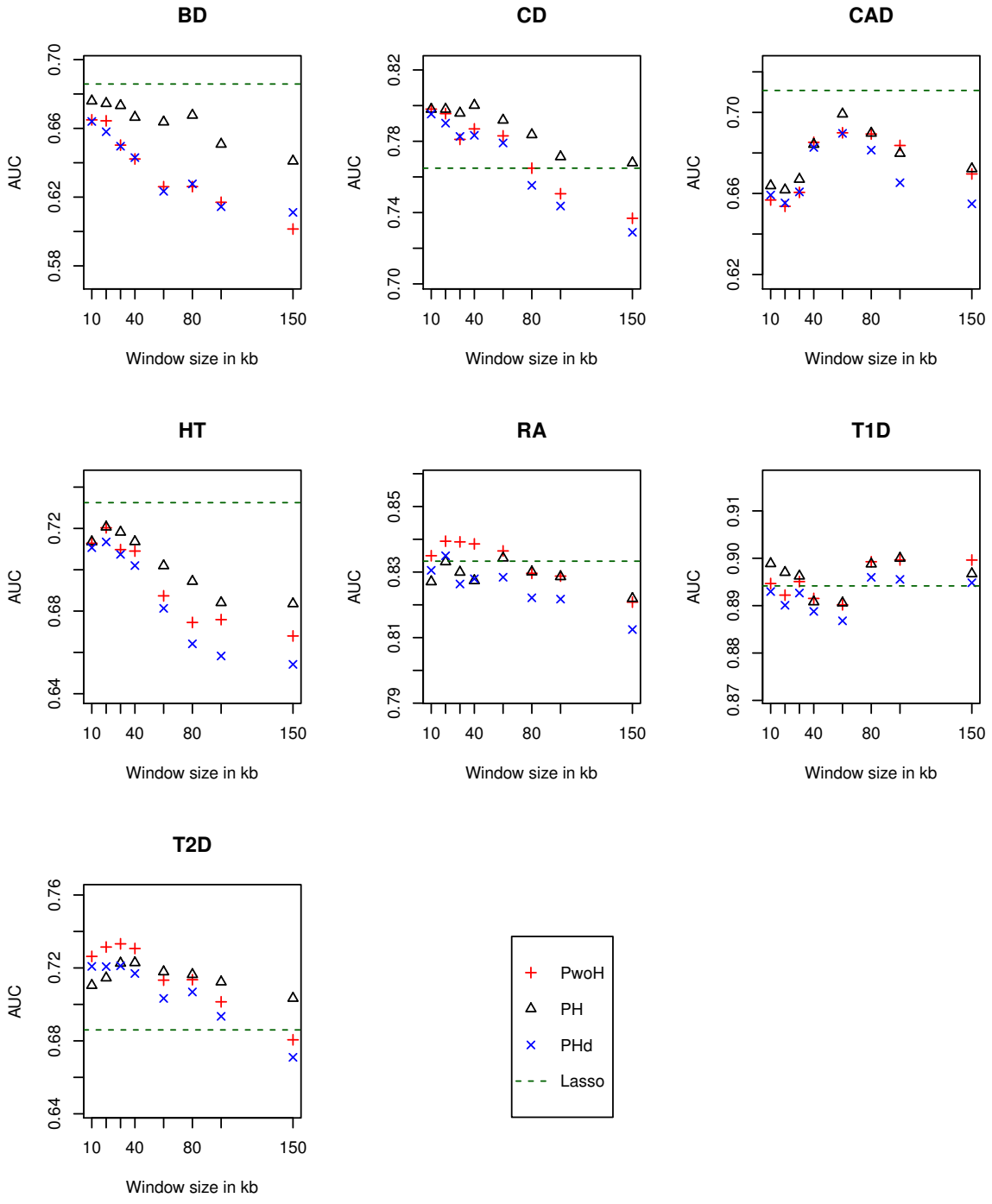

Figure 2.4: **Predictive performance and influence of the window size** Results for PwoH, PH and PHd on all diseases are displayed depending on the window size. The green line shows the performance obtained by preselection with lasso.

The three variants obtained similar performances. PHd does not outperform PH, it fails to capture any dominance effect. PH does not consistently outperforms PwoH. The decrease in performance with increasing window size is true for all three variants. However, for large window sizes, PH outperforms or equals the other variants.

2.4 Replication of genetic risk prediction of type 1 diabetes on cases of the Isis-Diab study

To quantify the genetic risk of T1D in patients of the Isis-Diab cohort, we used a genetic risk estimator as close as possible from the one in Wei et al. [2009]. In this article, the authors defined a genetic risk estimator that obtained an AUC of 0.84 on a Canadian and an American dataset. Our estimator was trained on the same data: the Wellcome Trust Case Control Consortium 1 (WTCCC1) T1D study (1963 cases and 2938 controls) [Burton et al., 2007]. It was used to predict on 1491 cases from Isis-Diab. Cases from the WTCCC1 studies on the non-autoimmune diseases type 2 diabetes, hypertension, coronary artery disease and bipolar disorder totaling 7670 individuals were used as validation controls. We refer to this group as cases of non-autoimmune diseases (CNAD).

Genome wide genotyping and imputation of the Isis-Diab data. Among the 1491 Isis-Diab patients for whom genotype data were available, 817 patients were genotyped with Illumina Human610-Quadv1B (610 000 SNPs) microarrays, and 673 patients with Illumina Human Omni 5 Exome microarrays (4 500 000 SNPs). Genome-wide genotyping was performed on bar-coded LIMS (Laboratory Information Management System) tracked samples using two different Illumina microarrays (Human610-Quadv1B and HumanOMNI5-4v1B). BeadChips were processed within an automated BeadLab at the Centre National de Génotypage as per the manufacturer's instructions. Samples were subject to strict quality control criteria including assessment of concentration, fragmentation and response to PCR. A total of 20 μl of DNA aliquoted to a concentration of 50 ng/μl was used for each array. In the discovery phase, genome-wide genotypes were used for controlling the quality of the samples. First, individuals with call rates <95% or duplicates and individuals who were possibly non-European were removed. By using this filtered sample set, we calculated quality control statistics, and SNPs with call rates <98% or SNPs with a Hardy-Weinberg equilibrium test p-value $< 10^{-6}$ or SNPs with a minor allele frequency <1% were excluded.

Finally, 517 864 SNPs (Human610-Quadv1B) and 3 309 261 SNPs (HumanOMNI5-4v1B) were used for imputation analysis. Imputation was done using IMPUTE v2, following the instructions provided by its authors [Howie et al., 2009]. Full sequence data from the phase I 1000 Genomes Project was used [1000 Genomes Consortium et al., 2012]. Only SNPs with an info metric over 0.8 for both chips were kept for analysis.

Quality control and missing data. The same quality control was used for SNPs as in Wei et al. [2009] to filter SNPs: missing rate<5%, Hardy-Weinberg Equilibrium p-value $> 10^{-3}$ and minor allele frequency>5% in the training set. Additionally, SNPs had to have missing rate<5%

in the Isis-Diab study to be included. The remaining missing data in the training set, the CNAD controls and the Isis-Diab patients were imputed by sampling randomly the training set.

Preselection. SNPs who passed quality control were selected if their training set association p-value was under 10^{-5} which was the tied best-performing threshold in the original paper. This resulted in a set of 505 SNPs.

Algorithm. The machine learning method achieving the best performance in the original paper is Support Vector Machine (SVM). It belongs to a family of methods called kernel methods [Shawe-Taylor and Cristianini, 2004]. SVM with the default radial kernel was trained on the training set. The estimator was then used to predict on the validation set: Isis-Diab patients and the CNAD. AUC on the validation set was computed. The e1071 package implementation of SVM was used.

Calibration. AUC measures the separation between the two classes obtained by a machine learning algorithm. However, AUC depends only on the ranking of the predictions and not on its numerical value. For our purpose, in particular for the estimation of power, we need the genetic risk to encode probabilities. Therefore, we perform calibration of our risk estimate on the validation set: the SVM estimated probabilities are replaced by probabilities estimated by a logistic regression of disease outcome on the logit of the SVM output.

Genetic risk. The genetic risk estimation trained on the WTCCC1 yielded an AUC of 0.86 when evaluated on Isis-Diab patients and CNAD controls. This value is intermediate between the AUC of 0.89 obtained in cross-validation on the WTCCC1 data and the AUC of 0.84 obtained on North-American cohorts. This may be due to the use of controls from the same study as the training set. The ROC curve, calibration plot and density plot of the genetic risk are presented in figure 2.5. The estimator is well calibrated except for the highest intervals. Given the larger proportion of controls compared with cases, those intervals also contain the least observations.

2.5 Discussion

In this chapter, we have focused on the design of genetic risk prediction. Our main contribution is the design of a method to try and capture interactions inside haplotypes. This implies a different setting than is customary in machine learning. Variables have two values for each observation and there is a metric structure to be taken into account. PH is designed to take advantage of those structures.

PH outperformed standard lasso regression on two datasets but not on all of them. This is suggestive of haplotypic effects in Crohn's disease and type 2 diabetes. The result for CD is reminiscent of the results of [Kang et al., 2011] as well as the two new loci discovered in the CD dataset by [Botta et al., 2014]. On the other hand, in three datasets, PH was slightly less accurate than lasso regression. This might be due to the non-standard multi-step design that results in some loss of information.

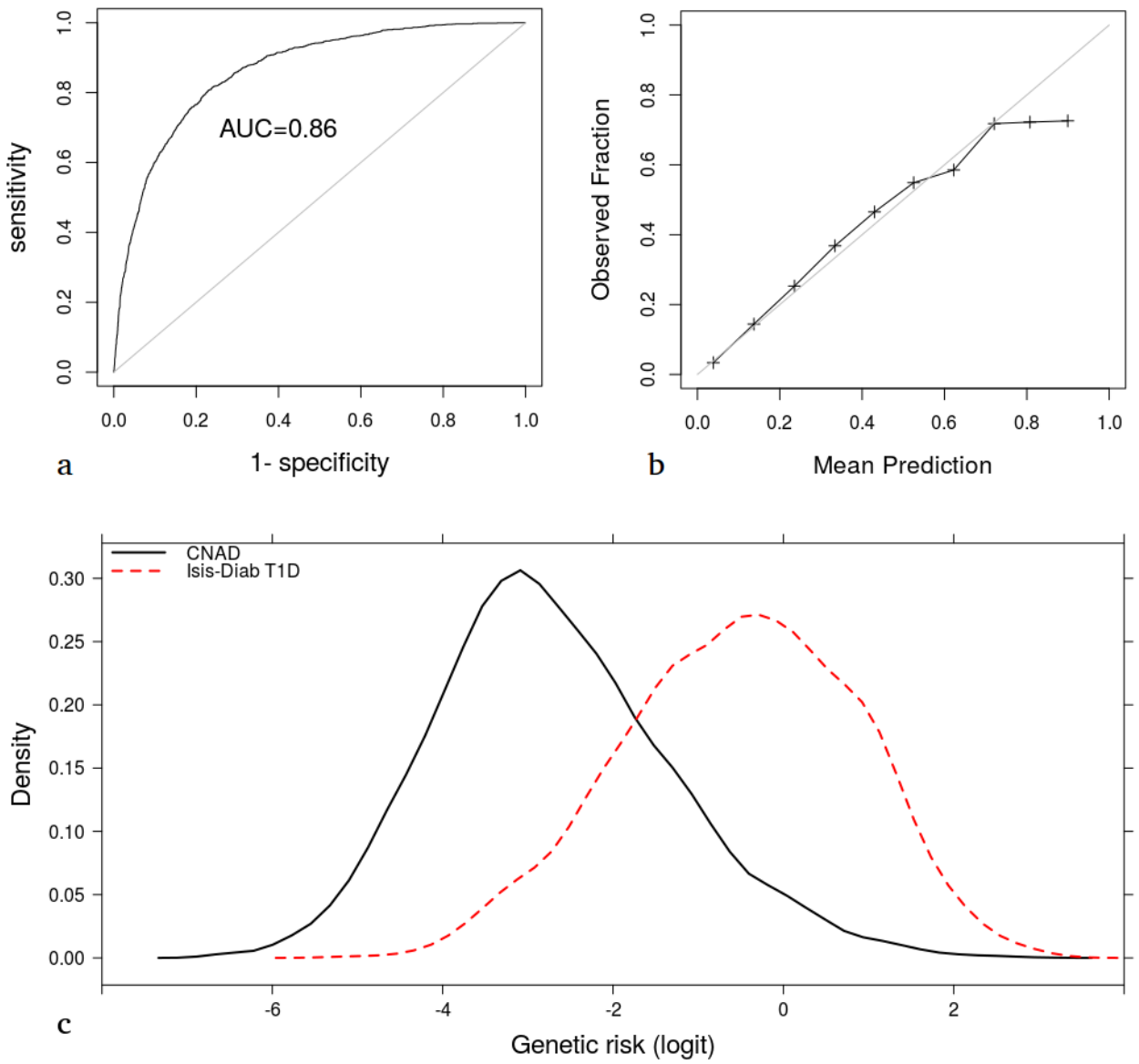

Figure 2.5: **Genetic risk estimation on the CNAD and Isis-Diab patients.** a: Receiver operating curve (ROC) of the estimator. The AUC is given. b: Calibration plot after calibration of the risk estimator. The range of values taken by the estimator is divided in 10 bins of equal length. The average prediction is plotted against the actual proportion of cases in each bin. c: Density plot on the logit scale of the risk estimate of the CNAD and Isis-Diab patients.

We noted that interactions inside haplotypes are a part of narrow-sense heritability. Our results therefore show that some of the missing heritability can be explained by the lack of consideration for interaction inside haplotypes. For type 2 diabetes, the proportion of genetic variance explained went from 40% to 66% (for prevalence $K = 20\%$ and heritability $h^2 = 0.30$) using the online calculator accompanying Wray et al. [2010]. For Crohn's disease, this proportion went from 16% to 22% (for $K = 1\%$ and $h^2 = 0.8$).

These estimates of explained heritability and all of the above AUCs are optimistic due to various study specific quality problems that result in overestimation of predictive performance as shown in the drop in out-of-study performance in Wei et al. [2009] and our replication. The limited availability of comparable datasets is therefore a hindrance to progress in this area of research.

PwoH obtains similar performance than PH on all datasets. This means that even if there are haplotypic effects, it is not necessary to perform phase imputation to capture them. This supports the idea that it is sufficient to capture local interactions using genotype to recover haplotypic effects. PHd did not outperform PH. This suggests that dominance effects are not an important part of genetic risk for complex diseases.

Small window sizes obtained the best performances while larger window sizes led to decreased performance. This is consistent with the results in [Botta et al., 2014] and supports the idea that local interactions are important in the genetic architecture of disease.

We also replicated an existing estimation of genetic risk for type 1 diabetes in the Isis-Diab patients. Besides offering further validation of the method, this will be useful in chapter 5 where we will cross genetic risk and environmental factors.

Since the UK Biobank has made available a dataset of 500000 genotyped participants, there has been a flurry of work on genetic prediction. A lasso regression to predict height trained on this dataset has achieved high predictive accuracy [Lello et al., 2017]. However, polygenic risk scores are still popular as they can leverage all pre-existing studies and are only validated on UK Biobank [Inouye et al., 2018].

Genetic risk estimation is expected to progressively find more uses in research as well as in the clinic [Torkamani et al., 2018]. Concerning type 1 diabetes, improved risk stratification for prevention trials and prospective studies using polygenic score has been recently validated in the TEDDY study [Bonifacio et al., 2018]. Other uses of polygenic scores for type 1 diabetes are the discrimination between monogenic and type 1 diabetes [Patel et al., 2016] or between type 1 and type 2 diabetes in young adults [Oram et al., 2016]. The risk estimations in these articles have similar performance to that of Wei et al. [2009].

Acknowledgement. This chapter makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under awards 076113 and 085475.

Chapter 3

Asymptotic equivalence of paired Hotelling test and conditional logistic regression

Abstract Matching, the stratification of observations, is of primary importance for the analysis of observational studies. We show that the score test of conditional logistic regression and the paired Student/Hotelling test, two tests for paired data, are asymptotically equivalent.

Contents

3.1 Introduction

Observational studies are the main source of information in many areas of science such as epidemiology or the social and political sciences [Rosenbaum, 2002]. Such studies aim at estimating the effect of a treatment or predictor T on an outcome Y and they differ from randomized experiments by the absence of random allocation of treatment. This lack of randomization is due to ethical or economical considerations and as a consequence, observational studies are subject to confounding.

Matching is the grouping of observations in strata. It is an important method to control for bias in observational studies: in presence of confounding, matching on the values of the confounder removes bias [Rubin, 1973] or when there are several confounders, other techniques based on matching can be used [Rosenbaum and Rubin, 1983a]. Matching can also guide the data collection of an observational study. This has been the case for the Isis-Diab study as we will see in the next chapter. In that study, disease cases were recruited at participating centers. The recruitment of controls was not as straightforward as the recruitment of cases and

had to aim and limit confounding. Consequently, for each case, controls were chosen among friends or neighbors whose –potentially unobserved– confounders are similar to that of the case. As the case and his controls have been chosen for their similarity, they cannot be considered independent and must be grouped in a strata.

Once a matching is obtained, adapted statistical procedures are needed. A good reference on statistical methods for matched data is the monograph by Breslow and Day [1981]. For strata limited to pairs of one case and one control (or before and after treatment) and when the predictor T is a binary variable, the so-called McNemar test can be used [McNemar, 1947]. In this simple case, there are only 4 possibilities for each pair. The data can be summarized in the following two-by-two contingency table.

Table 3.1: **McNemar test**

		$Y=1$	
		$T=0$	$T=1$
$Y=0$	$T=0$	a	b
	$T=1$	c	d

In the table above, $a+d$ is the number of pairs of a case and a control concordant for the predictor T and $b+c$ is the number of discordant pairs. The statistic ξ_{mc} of the McNemar test depends only on the numbers of discordant pairs b and c and has the following form:

$$\xi_{mc} = \frac{(b-c)^2}{b+c}.$$

Under the null hypothesis of independence between Y and T, ξ_{mc} follows the χ^2 distribution asymptotically [McNemar, 1947]. This test statistic has the advantage of being computationally simple.

However, in many settings, we have access to more than one control for each case. This is the case in the Isis-Diab study where each patient was asked to recruit two controls. A test for strata of arbitrary size and a binary predictor T was proposed in Mantel and Haenszel [1959] and is now referred to as the Cochran-Mantel-Haenszel (CMH) test. It also provides an estimate of the odd-ratio associated with the predictor.

If the predictor T is not binary but continuous, appropriate tests are available if the matching is a pairing i.e., each case has exactly one control. Let us denote the vector of differences between cases and controls $Z_i = T_{i1} - T_{i2}$ where $T_{i\ell}$ is the predictor of the ℓ-th observation in strata i. If Z can be assumed to be normally distributed, a Student test can be applied on Z to decide if its mean is zero. The resulting test is called the paired Student test [McDonald, 2009, p. 180-185]. The paired Hotelling test is the generalization of the paired Student test when there are several predictors, i.e, $Z_i \in \mathbb{R}^p$. When the normality assumption is not respected, a non-parametric alternative is available via the Wilcoxon signed rank test.

None of the above procedures allow for analysis of continuous variables with arbitrary strata size. This was made possible by conditional logistic regression (CLR) first introduced in Breslow et al. [1978]. CLR is however not limited to continuous variables and can be applied in all settings of the previous tests. As CLR is based on logistic regression, it is more flexible than the previous tests and allows for multivariate analyses.

When different statistical tests are available for the same data, it is desirable to prove that their results are similar. For procedures based on maximum likelihood such as CLR, the three tests of significance for predictors (the Wald test, the likelihood ratio test and the score test) have been shown to be asymptotically equivalent [Engle, 1984]. This result of asymptotic equivalence is with respect to the null hypothesis as well as to a sequence of local alternatives.

For the analysis of matched data with a binary predictor, the available tests have been shown to be identical. When the strata are pairs, both the CMH test and the McNemar test can be applied and their statistic is identical [Agresti and Kateri, 2011, p. 413]. When the strata size is arbitrary, CMH and CLR can be applied and the CMH statistic is identical to the score test statistic of CLR [Day and Byar, 1979] and therefore if the strata are pairs, the McNemar statistic will be identical to the score test statistic of CLR.

When the strata are pairs and the predictor (resp. predictors) is continuous, the paired Student test (resp. paired Hotelling test) and CLR can be applied. However, to the best of our knowledge, there are no results comparing the two procedures in that case. The objective of this chapter is to prove that the paired Hotelling test is asymptotically equivalent to the score test of CLR and therefore also to the Wald and likelihood ratio tests.

We start by deriving the exact form of the two test statistics in section 2. Finally, we present our result and prove it in section 3.

3.2 Test statistics

Let $Y_{i\ell}$ be the label of the ℓ-th individual of the i-th stratum and $T_{i\ell} \in \mathbb{R}^p$ its continuous predictor value. The paired Hotelling test can be applied when stratas are pairs of discordant observations and therefore $\ell \in \{1, 2\}$. Without loss of generality, we can assume that $Y_{i1} = 1$ and $Y_{i2} = 0$.

We are interested in testing the null hypothesis:

$$H_0 : \text{There is no association between } Y \text{ and } T \tag{3.1}$$

against the alternative hypothesis:

$$H_1 : \text{There is an association between } Y \text{ and } T.$$

To test this null hypothesis in this setting, we have two options: the paired Hotelling test whose test statistic we denote ξ_{hot} or CLR whose score test statistic we denote ξ_{sc}. We now derive the expressions of those two test statistics. As we will show below, the paired Hotelling test and CLR depend only on the vector of differences between pairs. We denote as before $Z_i = T_{i1} - T_{i2}$ the difference between pairs.

Paired Hotelling test. The hypotheses tested by the paired Hotelling test are expressed by the parameter $\boldsymbol{\mu} = (\mathbb{E}[Z_j])_{j\in\{1,\dots,p\}}$, the mean of the difference between pairs. Using this parameter, the hypothesis (6.4) becomes $H_0 : \boldsymbol{\mu} = 0$.

Let the unbiased sample covariance be defined as:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})(Z_i - \bar{Z})^\top,$$

where $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ and $*^\top$ is the transposition operator. The statistic of the test is:

$$\xi_{\text{hot}} = n\bar{Z}^\top C^{-1}\bar{Z}. \tag{3.2}$$

When Z follows a centered Gaussian distribution, the distribution of ξ_{hot} is called the Hotelling distribution.

Conditional Logistic Regression score test. In logistic regression, the likelihood of an observation, given the parameters $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, equals:

$$\mathbb{P}(Y = y|T) = \frac{e^{y(\alpha+\boldsymbol{\beta}^\top T)}}{1 + e^{\alpha+\boldsymbol{\beta}^\top T}},$$

with $y \in \{0, 1\}$.

Logistic regression can take into account stratification by having a different constant term α_i for each stratum. However, when the strata are small e.g., pairs, this leads to a large number of parameters and biases parameter estimation [Breslow and Day, 1981, p. 249-251]. Conditional logistic regression addresses the problem by conditioning the likelihood on the number of cases in each stratum. This eliminates the need to estimate the α_i. In the case of pairings, when the first observation is a case and the second is a control, the conditional likelihood of the i-th strata is:

$$\begin{aligned}
&\mathbb{P}(Y_{i1} = 1, Y_{i2} = 0|T_{i1}, T_{i2}, Y_{i1} + Y_{i2} = 1) \\
&= \frac{\mathbb{P}(Y_{i1} = 1|T_{i1})\mathbb{P}(Y_{i2} = 0|T_{i2})}{\mathbb{P}(Y_{i1} = 1|T_{i1})\mathbb{P}(Y_{i2} = 0|T_{i2}) + \mathbb{P}(Y_{i1} = 0|T_{i1})\mathbb{P}(Y_{i2} = 1|T_{i2})} \\
&= \frac{\frac{\exp(\alpha_i+\boldsymbol{\beta}^\top T_{i1})}{1+\exp(\alpha_i+\boldsymbol{\beta}^\top T_{i1})} \times \frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}^\top T_{i2})}}{\frac{\exp(\alpha_i+\boldsymbol{\beta}^\top T_{i1})}{1+\exp(\alpha_i+\boldsymbol{\beta}^\top T_{i1})} \times \frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}^\top T_{i2})} + \frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}^\top T_{i1})} \times \frac{\exp(\alpha_i+\boldsymbol{\beta}^\top T_{i2})}{1+\exp(\alpha_i+\boldsymbol{\beta}^\top T_{i2})}} \\
&= \frac{\exp(\boldsymbol{\beta}^\top T_{i1})}{\exp(\boldsymbol{\beta}^\top T_{i1}) + \exp(\boldsymbol{\beta}^\top T_{i2})} \qquad \text{(The } \alpha_i \text{ have been eliminated.)} \\
&= \frac{1}{1 + \exp(\boldsymbol{\beta}^\top (T_{i2} - T_{i1}))}.
\end{aligned}$$

In the general case, when there is k cases in a strata of size m, with the cases being the first k observations, the conditional likelihood of a strata can be written:

$$\mathbb{P}(Y_{ij} = 1 \text{ for } j \leq k, Y_{ij} = 0 \text{ for } j > k|T_{i1}, \ldots, T_{im}, \sum_{j=1}^{m} Y_{ij} = k) = \frac{\exp(\sum_{j=1}^{k} \boldsymbol{\beta}^\top T_{ij})}{\sum_{J\in\mathcal{C}_k^m} \exp(\sum_{j\in J} \boldsymbol{\beta}^\top T_{ij})},$$

where $\mathcal{C}_k^m$ is the set of all subsets of size k of the set $\{1, \ldots, m\}$.

To obtain the full conditional likelihood, we multiply over each stratum and take the log to obtain the log-likelihood $L(\boldsymbol{\beta}, Z)$:

$$L(\boldsymbol{\beta}, Z) = -\sum_{i=1}^{n} \log(1 + e^{-\boldsymbol{\beta}^\top Z_i}).$$

The score $s(\boldsymbol{\beta}, Z)$ is then defined by:

$$s(\boldsymbol{\beta}, Z) = \frac{\partial L}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}, Z) = \sum_{i=1}^{n} \frac{Z_i}{\exp(\boldsymbol{\beta}^\top Z_i) + 1}.$$

If $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, the covariance matrix of $s(\boldsymbol{\beta}_0, Z)$ is Fisher's information matrix:

$$\mathcal{I}(\boldsymbol{\beta}_0) = \left(\mathbb{E}[-\frac{\partial^2 L}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k}(\boldsymbol{\beta}_0, Z) | \boldsymbol{\beta}_0] \right)_{j,k \in \{1,\dots,p\}} = \left(\sum_{i=1}^{n} \frac{Z_{ij} Z_{ik} \exp(\boldsymbol{\beta}_0^\top Z_i)}{(\exp(\boldsymbol{\beta}_0^\top Z_i) + 1)^2} \right)_{j,k \in \{1,\dots,p\}}.$$

To test the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, the score test statistic is:

$$\xi_{\text{sc}} = s(\boldsymbol{\beta}_0, Z)^\top \mathcal{I}(\boldsymbol{\beta}_0, Z)^{-1} s(\boldsymbol{\beta}_0, Z)$$

Expressed using $\boldsymbol{\beta}$, hypothesis (6.4) becomes $H_0 : \boldsymbol{\beta} = 0$. As shown above, the null hypothesis of the paired Hotelling test is expressed using $\boldsymbol{\mu}$, the mean of the difference between pairs, rather than $\boldsymbol{\beta}$, the regression coefficient. There is no general correspondence between the two parameters. However, the two null hypotheses that are of interest in practice are the same: $\boldsymbol{\mu} = 0 \iff \boldsymbol{\beta} = 0$.

For $\boldsymbol{\beta} = 0$, the score, Fisher's information matrix and the test statistic become:

$$s(0, Z) = \frac{n}{2} \bar{Z}, \qquad \mathcal{I}(0, Z) = \frac{1}{4} \sum_{i=1}^{n} Z_i Z_i^\top$$

and

$$\xi_{\text{sc}} = n \bar{Z}^\top \left(\frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top \right)^{-1} \bar{Z}. \tag{3.3}$$

Now that we have expressed analytically ξ_{hot} and ξ_{sc}, we state our result in the following section.

3.3 Asymptotic equivalence

In this section, we introduce the sequence of local alternatives and state the result of asymptotic equivalence between the two tests.

We adopt the framework of Engle [1984] for asymptotic equivalence. The motivation for considering a sequence of local alternatives is that any reasonable test will have the right size and will reject a fixed alternative when the number of observations becomes large. The sequence of local alternatives considers deviations from the null that approach the null as sample size increases. This allows to compare tests more precisely.

We model the sequence of local alternatives by a triangular array of observations $(Z_i^{(n)})_{n \in \mathbb{N}, i \in \{1,\dots,n\}}$ that respects the following assumption. Let $\boldsymbol{\delta} \in \mathbb{R}^p$ and Σ be a positive definite matrix of dimension p.

(H3.1) *For all $n \in \mathbb{N}$, $i \in \{1, \dots, n\}$, we have $\mathbb{E}[Z_i^{(n)}] = \frac{\boldsymbol{\delta}}{\sqrt{n}}$ and $\text{Cov}[Z_i^{(n)}] = \Sigma$. In addition, $W_i^{(n)} = Z_i^{(n)} - \frac{\boldsymbol{\delta}}{\sqrt{n}}$ are independent and identically distributed.*

The triangular array of **Assumption 3.1** is the sequence of local alternatives. The parameter $\boldsymbol{\delta}$ controls the deviation from the null. The null hypothesis is the special case of the sequence of local alternatives when $\boldsymbol{\delta} = 0$. The $\sqrt{n}$ in the denominator of the deviation shrinks the deviation towards the null as the sample size increases. This particular power of n ensures that the test statistics are well-behaved as they converge in distribution.

We denote $\xi_{\text{sc},n}$ (resp. $\xi_{\text{hot},n}$) the statistic of the score test (resp. of the paired Hotelling test) associated with $Z^{(n)}$, the n-th line of the triangular array of **Assumption 3.1**. In the same fashion as for the Student distribution and the χ^2 distribution, the Hotelling distribution with degrees of freedom p and $n-1$ and the χ^2_p distribution are different for finite samples but they are the same asymptotically. Therefore, $\xi_{\text{sc},n} - \xi_{\text{hot},n} \xrightarrow{P} 0$ is enough to guarantee asymptotic equivalence. Our theorem is more precise as it includes the convergence rate and limiting distribution of this difference. We note the convergence in distribution d.

Theorem 3.1. *Under the null or a sequence of local alternatives, the paired Hotelling test and the score test of CLR are asymptotically equivalent. More precisely, under* ***Assumption 3.1****, we have:*

$$n(\xi_{\text{sc},n} - \xi_{\text{hot},n}) \xrightarrow{d} K$$

where $K = (\boldsymbol{\delta} + V)^\top \Sigma^{-1} \left(\Sigma - (\boldsymbol{\delta} + V)(\boldsymbol{\delta} + V)^\top\right) \Sigma^{-1}(\boldsymbol{\delta} + V)$ *with* $V \sim \mathcal{N}(0, \Sigma)$ *and* $\mathcal{N}$ *refers to the normal distribution.*

Examples of distribution of K for $p = 1$, $\Sigma = \sigma^2 = 1$ and different values of $\boldsymbol{\delta}$ are shown in figure 3.1. We see that for small values of $\boldsymbol{\delta}$, K is concentrated between 0 and 1/4. As $\boldsymbol{\delta}$ grows, the distribution shifts to large, negative values.

For the practitioner, the result of **Theorem 3.1** will not affect the choice of test but ensures that similar conclusions are reached regardless of that choice. The result also implies that CLR can be considered a generalization of the paired Hotelling or Student test when the conditions for the latter do not apply.

Proof of Theorem 3.1 To facilitate the notation, we omit the dependence in n in most quantities, writing Z instead of $Z^{(n)}$. The quantity we want to estimate is:

$$n(\xi_{\text{sc},n} - \xi_{\text{hot},n}) = n^2 \bar{Z}^\top (\tilde{\mathcal{I}}^{-1} - C^{-1}) \bar{Z} = n^2 \bar{Z}^\top \tilde{\mathcal{I}}^{-1} (C - \tilde{\mathcal{I}}) C^{-1} \bar{Z}, \tag{3.4}$$

where $\tilde{\mathcal{I}} = \frac{1}{n} \sum_{i=1}^n Z_i Z_i^\top$.

The Lindeberg-Feller Central Limit Theorem for triangular arrays [Hall and Heyde, 1980, Theorem 3.2 p. 58] applied on the $W_i^{(n)}$ gives:

$$\sqrt{n}\bar{Z} \xrightarrow{d} (\boldsymbol{\delta} + V), \tag{3.5}$$

with $V \sim \mathcal{N}(0, \Sigma)$. The weak law of large numbers for triangular arrays implies:

$$C \xrightarrow{P} \Sigma. \tag{3.6}$$

Combining (3.5) and (3.6), we see:

$$\tilde{\mathcal{I}} = \frac{1}{n} \sum_{i=1}^n Z_i Z_i^\top = \frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(Z_i - \bar{Z})^\top + \bar{Z}\bar{Z}^\top \xrightarrow{P} \Sigma \tag{3.7}$$

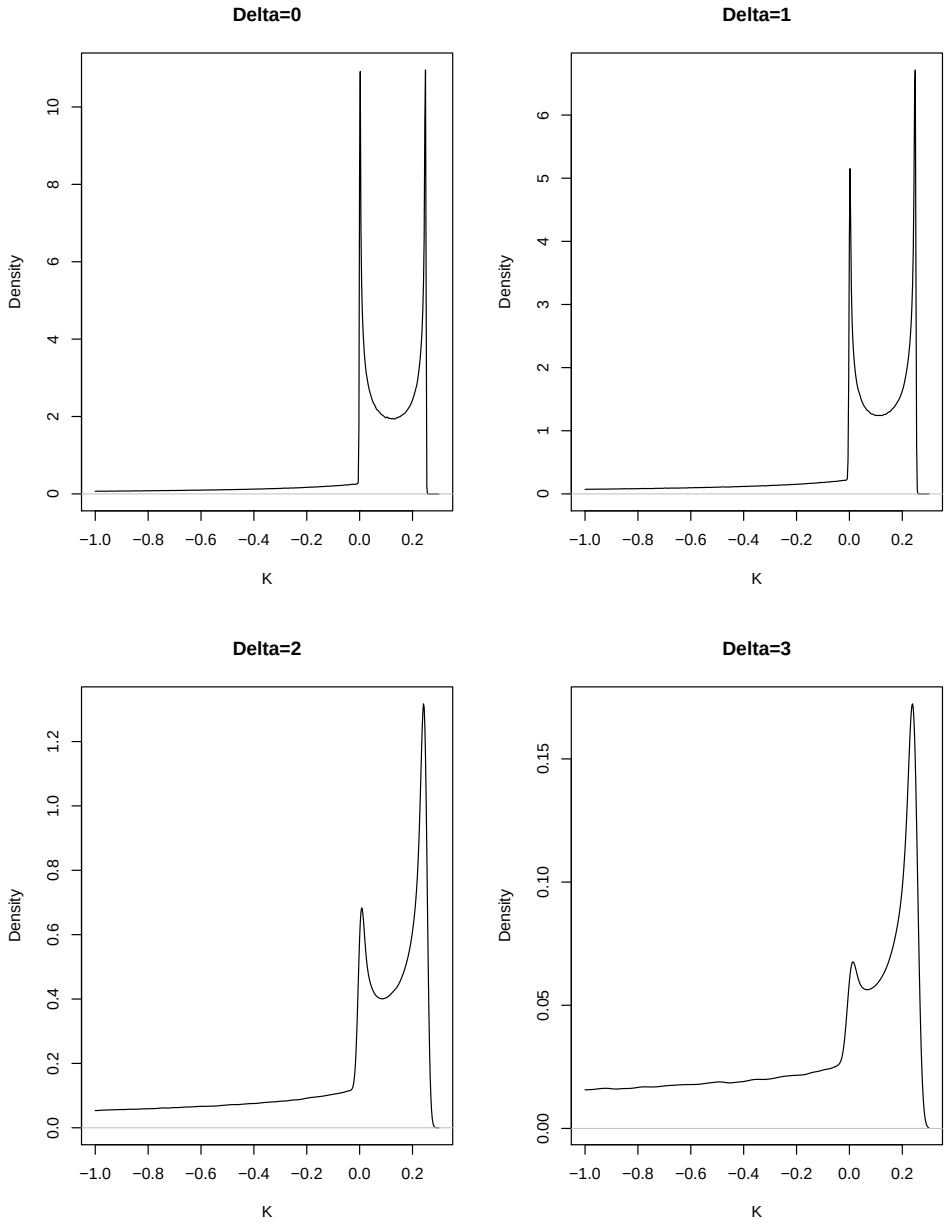

Figure 3.1: **Distribution of K.** Distribution of the limit variable for $p = 1$, $\Sigma = \sigma^2 = 1$ and $\boldsymbol{\delta} \in \{0, 1, 2, 3\}$.

and, using the second identity in (3.7), we obtain:

$$n(C - \tilde{\mathcal{I}}) = C - n\bar{Z}\bar{Z}^\top \xrightarrow{d} \Sigma - (\boldsymbol{\delta} + V)(\boldsymbol{\delta} + V)^\top.$$

Applying Slutsky's lemma to all of the above, we conclude:

$$n(\xi_{\mathrm{sc},n} - \xi_{\mathrm{hot},n}) \xrightarrow{d} K$$

where $K = (\boldsymbol{\delta}+V)^\top \Sigma^{-1} \left(\Sigma - (\boldsymbol{\delta} + V)(\boldsymbol{\delta} + V)^\top\right) \Sigma^{-1}(\boldsymbol{\delta}+V)^\top$ with $V \sim \mathcal{N}(0, \Sigma)$. This concludes the proof of **Theorem 3.1**.

□

Chapter 4

Association of environmental markers with childhood type 1 diabetes mellitus revealed by questionnaires on early life exposures and lifestyle in a case-control study

Abstract The incidence of childhood type 1 diabetes (T1D) incidence is rising in many countries, supposedly because of changing environmental factors, which are yet largely unknown. To unravel environmental markers associated with T1D, we conducted a case-control study. Cases were children with T1D from the French Isis-Diab cohort. Controls were schoolmates or friends of the patients. Parents were asked to fill a 845-item questionnaire investigating the child's environment before diagnosis. The analysis took into account the matching between cases and controls. A second analysis used propensity score methods. We found a negative association of several lifestyle variables, gastroenteritis episodes, dental hygiene, hazelnut cocoa spread consumption, wasp and bee stings with T1D, consumption of vegetables from a farm and death of a pet by old age. We attempted to replicate some of those results using a shorter questionnaire. The found statistical association of new environmental markers with T1D calls for investigation of new environmental areas.

Trial registration: Clinical-Trial.gov NCT02212522. Registered August 6, 2014.

Contents

4.1 Introduction

The current rise in T1D incidence [DIAMOND Project Group, 2006] is attributed to environmental causes to which genetically predisposed children are increasingly exposed, but epidemiology has delivered more questions than robust answers. Dissecting the environment is a daunting task, with paramount difficulties for extracting relevant information from multiple known and unknown exposures occurring during childhood. The fact that childhood T1D occurs early in life allows restraining the environmental analysis to the few years encompassing intrauterine life, infancy and childhood. A classical way of doing this is using retrospective questionnaires, but the questions are necessarily limited to selected areas of child life and answers may be biased by parental recall. Environmental comparison between cases and controls can also be prospective. To achieve this given the low prevalence of T1D, it is necessary to study a genetically at risk population, for example positivity for HLA screening in the TEDDY study [TEDDY Study Group, 2007]. Another way of avoiding recall-related problems is to use registries [D'angeli et al., 2010]. However, registries are more limited in their scope than a questionnaire. Another difficulty inherent to any environmental approach is that participants are not aware of many exposures. Collecting biological samples to characterize the "exposome" [Rappaport et al., 2014] of T1D children also has several drawbacks, since blood parameters may be modified as a consequence of T1D not as a causal component, and are confined to the only environmental parameters that leave a long living trace in patient's blood, i.e. a minority of exposures.

Over the recent years, suspicion has almost exclusively focused on infectious agents and nutrition in the early years of life [Egro, 2013, Knip and Simell, 2012, Forlenza and Rewers, 2011]. Enteroviruses have been the subject of numerous studies and have remained the most often suspected environmental contributors to T1D [Green et al., 2004, Yeung et al., 2011]. In contrast, infections have been considered as protective from T1D according to the hygiene hypothesis, which postulates that the increase in autoimmune T1D could be due to the decrease of early infections [Chapman et al., 2012, Bach, 2002] or lack of parasites [Gale, 2002]. This has been shown in the isogenic NOD mice model [Bach, 2002, Like et al., 1991], but epidemiological evidence in humans, who are exposed to different infectious agents and have a wide genetic variation, is still pending. Studies attempting to relate infectious episodes with T1D have yielded contrasted results [Schneider and von Herrath, 2014]. Respiratory infections in the first year of life have been shown to increase the risk of seroconversion to islet autoimmunity (IA) in the BABYDIET cohort and in the MIDIA study [Beyerlein et al., 2013, Rasmussen et al., 2011]. On the other hand, they were not associated with T1D in the DAISY cohort [Snell-Bergeon

et al., 2012]. Gastrointestinal illnesses at precise periods were associated with higher risk of IA in the same study. More recently, the gut microbiome has been investigated in search of a bacterial composition that could be associated with T1D [Vatanen et al., 2016].

Nutrition has been the other focus of environmental research for T1D. Overfeeding and the ensuing increase of beta cell functional activity for producing more insulin has been suspected to favor autoimmunity towards the beta cells (the overload hypothesis) [Dahlquist, 2006]. Meta-analyses have found that early weight gain [Harder et al., 2009] or obesity [Verbeeten et al., 2011] showed a modest association with T1D. Vitamin D supplementation studied through questionnaires has been suggested to protect from T1D [Dong et al., 2013], but this has not been confirmed when 25-hydroxyvitamin D levels in plasma were studied [Simpson et al., 2011]. Since vitamin D supplementation of infants is generalized in French infants since the 70s, vitamin D deficiency is not likely to be a driver of increasing T1D incidence.

Several dietary interventions have attempted to prevent T1D. The TRIGR trial tested whether substitution of cow's milk by casein hydrolysate formula affects the occurrence of IA or progression to T1D. No significant difference has been observed between the two groups for the appearance of IA [Knip et al., 2014] or the progression to T1D [Knip et al., 2018]. The possibility that exclusive breast-feeding or late introduction to cow's milk is associated with a modest protection is supported by a meta-analysis of observational studies [Cardwell et al., 2012]. A few other nutrients have been studied. An older age at first introduction to gluten showed no protective effect in the BABYDIET study [Hummel et al., 2011]. Omega-3 fatty acids seemed to be associated with a slightly reduced risk of islet autoimmunity in the DAISY cohort [Norris et al., 2007] but the pilot study that was then performed did not show significant protection [Chase et al., 2015]. Nicotinamide did not modify the progression to T1D in children with IA in the ENDIT trial [Gale et al., 2004]. Other prevention trials are underway [Skyler, 2013]. Early nutrition is a favorable field of investigation through randomized trials since a vast number of factors can be manipulated experimentally.

The BABYDIAB and the DAISY cohort have found that IA often appears in the first years of life preceding clinical diagnosis T1D by several months or years [Barker et al., 2004, Ziegler et al., 1999], which stress a potential predisposing role for early environmental exposures. This has inspired our approach for screening early life events that could be associated with environmental differences between cases and controls, including a number of infectious and nutritional exposures that can be reliably recalled by parents.

Our study is a tentative and still limited step for moving environmental research from hypothesis-driven to more data-driven approaches. A comparable move has occurred in the 90s when genetic research has switched its candidate gene approach of complex diseases, notably T1D, to interrogate the complete genome variation blindly with genome wide association studies (GWAS) with the aim of unraveling disease markers [Welter et al., 2014] that could secondarily lead to true genetic causation [Farh et al., 2015, Stamatoyannopoulos, 2016]. Environment wide association studies (EWAS) [Patel et al., 2010] or exposome association studies [Rappaport et al., 2014] will likely allow researchers to investigate children environment on a vast scale without making a priori hypotheses. Such approaches will remain limited because a myriad of environmental markers will escape investigation, while genomic variation is finite. In this respect, our current 845-item questionnaire can only be viewed as a preliminary proof-of-concept approach for scanning the environment of a child. It is indeed limited by the number of

questions that have been selected to describe this environment, by the recall errors that could be made by the parents of the cases and controls, and by the number of participants who agree to spend two hours filling a complex questionnaire. False positivity is an expected weakness of this approach, but careful statistical analysis can provide a list of environmental markers for which false discoveries are controlled.

A shorter questionnaire was designed to try and increase the response rate of the study. We tried to replicate the results obtained on the long questionnaire on the shorter questionnaire.

4.2 Main study on the long questionnaire

4.2.1 Questionnaire

The questionnaire was built by a group of academics composed of obstetricians and pediatricians specialized in pediatric infectious diseases, nutrition, and lifestyle. Their task was to define the environment of pregnant women, neonates, infants and young children, by enumerating all aspects that they thought a mother will likely be able to recall years later. A group of mothers of young children (living in urban or rural environment) were also asked to participate. A first questionnaire of nearly 1,000 questions was built and tested across 100 young mothers. Only questions that could be answered rapidly were kept, because we considered that the speed of the answers would favor spontaneity and minimize recall errors and bias. The questionnaire was also tested in 30 mothers of young children with recently diagnosed T1D and 30 mothers of children who had declared T1D five to ten years before. Only questions that had a comparable recall score in the two groups of mothers were kept in an effort to eliminate questions that could not be easily recalled. The final questionnaire contained 576 main questions and 845 items when counting sub-questions about the environment including 90 questions about pregnancy, 25 concerning the delivery and early post-natal life, 20 about early childhood, 75 on the subject's medical life, 60 on nutrition, 40 on housing, 30 on daycare, 30 on leisure and trips, 80 on contact with animals and 60 on family members' environment. Depending on mothers, the time to fill the questionnaire ranged form 90 to 120 minutes. A PDF version of it in French is available as online additional file 1 of the published version of this chapter [Balazard et al., 2016].

4.2.2 Data collection

The Isis-Diab cohort is a large multi-centric cohort of T1D patients in France which recruitment started in 2007. Starting in March 2010, three copies of the questionnaire were sent to the parents of 6618 T1D patients enrolled in the cohort during the month following their inclusion in the study. Parents were asked to fill the questionnaire regarding the exposures and events having taken place in their child's life before the clinical onset of T1D. They were also asked to enroll as controls two of their friends having an unaffected child of the same age. The 6144 parents having provided a phone number were contacted once during the week following the questionnaire sending. If the questionnaire was not returned within 3 months, parents received a reminder by mail. 1769 cases (i.e. 27% of the patients to which a questionnaire was sent) and 1085 controls returned the questionnaire. 241 cases provided two controls, 451 cases provided one control, and 1077 cases provided no control. 152 controls were not associated to a case

that returned his questionnaire. All the questionnaires completed by patients and controls were seized by a private provider (numerical input for all the « checkbox » responses, and dual manual entry for handwritten responses). Patients living in areas with higher economic deprivation were less likely to respond [Le Fur et al., 2014]. The questionnaire investigated the period preceding diagnosis of the disease. Matched controls were asked to fill the questionnaire with respect to the age at which the patient had been diagnosed. We will refer to this age as the reference age.

4.2.3 Pre-analysis treatment

A computerized treatment was designed to code categorical questions into binary variables and to allow analysis of sub-questions. After the pre-analysis treatment, 845 variables were available for analysis. In order for effect sizes to be on a similar scale even though we have binary questions and ordinal questions with up to 5 different levels. For example, consumption of cola drinks frequency was quantified on five levels from never to several times a day. All variables were scaled to be between 0 and 1. In this way, the effect size for ordinal variables corresponds intuitively to the odds ratio between the two extreme responses. The encoding of the variables were modified so that the directionality of the effects be intuitive. A description of the 845 variables is available in the online additional file 2 [Balazard et al., 2016].

4.2.4 Exclusions

We excluded from the analysis the questionnaires where more than 50% of the questions were left unanswered. As our questionnaire was designed to quantify a child's environment, we included only participants whose reference age was between 0.5 years and 15.5 years. To minimize recall errors, we excluded participants for whom the delay between diagnosis and questionnaire reception was greater than 10 years. We used primary school attendance as another marker of the quality of recall: we excluded participants who reported that their child attended primary school before 5.5 years. In the next section, we use a questionnaire-based prediction model for age to justify this exclusion. Using the same prediction model, we consider a second exclusion of outliers for predicted age. We report which results are significantly affected by this further exclusion. For the first analysis, we excluded participants without matched counterparts, i.e. patients with no matched control or controls with no matched patient. The matched analysis then compared 469 patients with 624 matched controls. We also performed a propensity analysis without using the matching. We only excluded participants with no available postal code or parents' profession as these variables were used to control for bias. This resulted in a sample of 1151 patients and 689 controls. The processes of exclusion and sample definition are summarized in figure 4.1.

4.2.5 Analytical procedures

Matched analysis: We used methods that take matching into account and allow for variable size of the matched strata: either one patient and one control or one patient and two controls. For questions with binary responses, we performed Cochran-Mantel-Haenszel tests and for ordinal responses, we performed conditional logistic regression [Breslow and Day, 1981]. In both

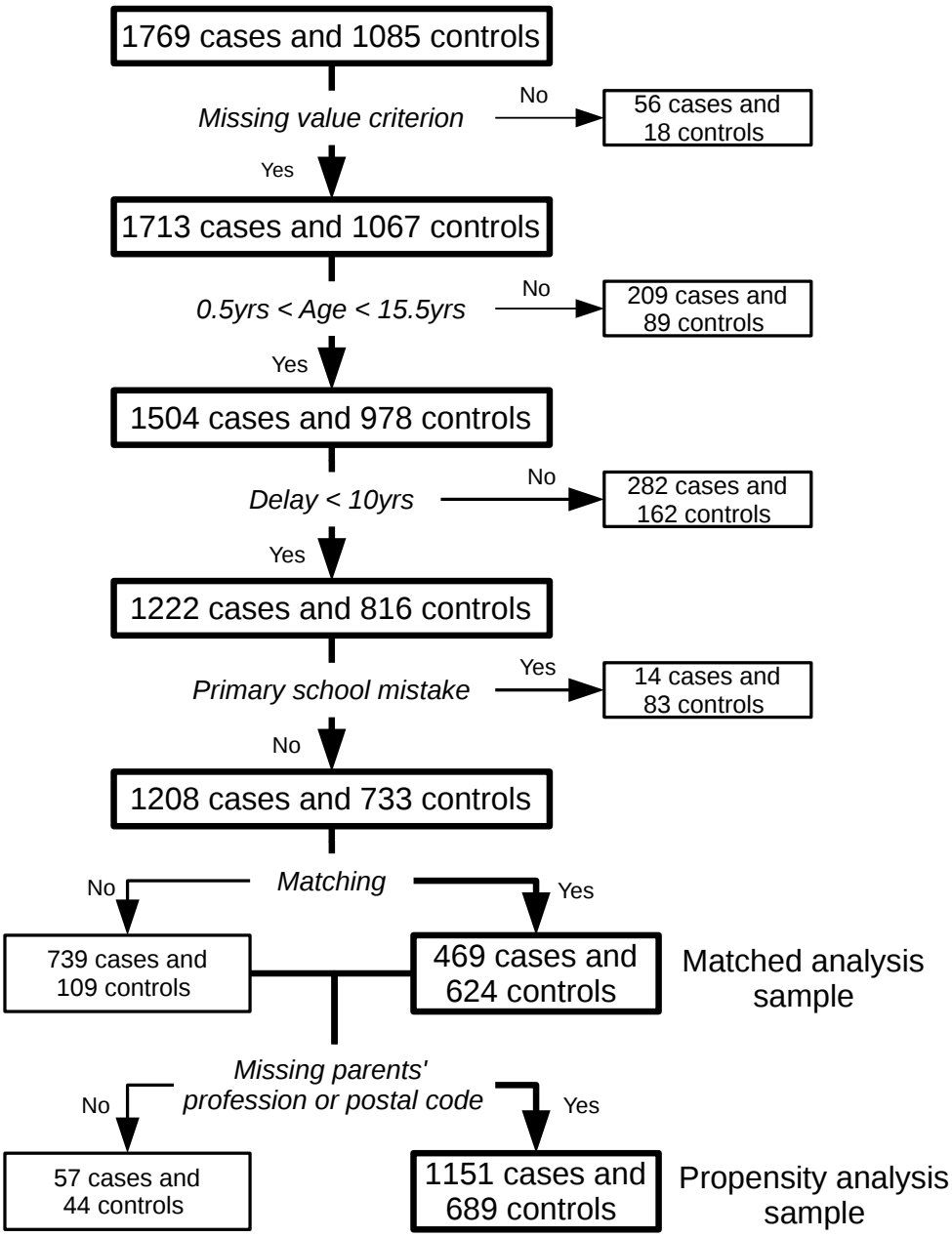

Figure 4.1: **Flowchart of the samples definition.** Missing value criterion is verified if at least half the questionnaire was filled. Delay refers to the time between diagnosis and questionnaire reception. Participants have made the primary school mistake if they answered that they went to primary school even though their reference age is smaller than 5.5 years. The two samples on which analyses were performed are in the bottom right corner.

cases, we used the strata defined by the matching and the disease status as outcome. To avoid convergence problems, we excluded variables with a standard deviation smaller than 0.1.

Propensity analysis: In this second analysis, we used stratification on the propensity score [Austin, 2011] to control for bias. Propensity score methods allow to control for bias by comparing participants with a similar probability of treatment (here the response to a question) given the covariates defined below. Random forests [Friedman et al., 2001, p. 587-604] is a popular machine learning algorithm praised for its state-of-the-art predictive performance. Furthermore, it provides a reliable prediction on the training set called out-of-bag estimate which is not prone to overfitting. We used the randomForest package in R [Liaw and Wiener, 2002]. We trained a random forests regression to predict the treatment status using as predictors reference age, socio-economic status, urban/rural environment and study center. We then defined the propensity score as the out-of-bag estimate of the random forests. We then stratified our sample in 10 strata according to deciles of the propensity score and performed a Cochran-Mantel-Haenszel test (respectively a conditional logistic regression) between the question of interest if it was binary (respectively if it was ordinal) and disease status. We again excluded variables whose standard deviation was smaller than 0.1.

Covariate description. The following covariates were used to define the propensity score for the propensity analysis:

- **Age.** The reference age was written on the first page of the questionnaire as an integer number of years that corresponds to a rounding of the patient's age at diagnostic. In both analyses, we used non-rounded patient's age at diagnostic for both the patient and his matched controls.

- **Socio-economic status.** Socio-economic status was assessed using the hand-written professions of parents. It was encoded as an ordinal variable taking value 0, 1 or 2 where 0 corresponds to blue-collar workers, 1 to intermediate professions and 2 to upper class. Among the 1840 participants of the propensity analysis, 837 were classified as 0, 725 as 1, and 278 as 2.

- **Urban/farm environment.** Using the postal code of the participants obtained through the questionnaire, two variables defined at the level of the patient's "commune" (town) of residence were used to quantify whether the participants lived in an urban or rural area. Those variables are the urban units index (as a code reflecting the size of the commune's urban area) and the percentage of farmers in the active population. Those two variables came from anonymous public databases (French Quetelet Network (http://www.reseau-quetelet.cnrs.fr), via the Centre Maurice Halbwachs –Archives de Données Issues de la Statistique Publique (http://www.cmh.greco.ens.fr/adisp.php)) and were dated in 2007 (census closest to the date that patients started to receive the environmental questionnaire). Environment was also controlled by the recruitment center e.g. the hospital or pediatric endocrinology practice that recruited the patient: each center with more than 30 participants was coded as a distinct binary variable.

	Matched analysis sample		Propensity analysis sample	
	Cases	Controls	Cases	Controls
Number	469	624	1151	689
Age (years)	7.5 (4.2;10.5)	7.7(4.6;10.5)	7.6(4.2;10.6)	7.8(4.6;10.5)
Delay (years)	3.0(1.0;5.2)	2.9(1.1;5.2)	2.9(0.9;5.6)	3.0(1.1;5.4)
Missing data (%)	4.4(2.7;6.8)	3.9(2.5;6.0)	4.9(3.1;7.5)	3.8(2.4;5.9)

Table 4.1: **Characteristics of cases and controls in the two samples.** Age is the reference age and delay is the time between the diagnosis date and the questionnaire reception. The values displayed are the median value and the first and third quartile between parentheses.

Correction for multiple tests. To control for multiple testing, we used the Bonferroni correction which allows to control the family-wise error rate at 5%. For the matched analysis, as we consider that it is of better quality than the propensity analysis, we also considered the more lenient false discovery rate [Benjamini and Hochberg, 1995] for a level of 5%. We report the list of variables that passes both the FDR threshold for the matched analysis and the Bonferroni threshold for the propensity analysis. This provides better control over false positives than considering only one of the two thresholds. We also report results for variables associated with T1D in the literature.

4.2.6 Results

We give in table 4.1 the characteristics of the two samples on which we perform analysis.

Matched analysis. For convenience, the variables have been labeled in the figures. Correspondence between labels and precise description of variables are available in the online additional file 2 [Balazard et al., 2016]. Figure 4.2 presents a volcano plot where both the effect size and the significance of answers to each question are displayed. We also display in blue the Bonferroni-Holm threshold for multiple testing, this means that we control the family-wise error rate at 5% for the list of variable over the blue line. The more lenient threshold for a false discovery rate of 5% is displayed in red. Questions that pass this threshold are labeled in the plot. Exact sample size, p-value, estimate and confidence interval for each variable are available as online additional file 2 [Balazard et al., 2016]. Three questions showed that cases more often had a relative with T1D and are excluded from the plots and discussion.

Propensity analysis. Results are also available in the online additional file 2 [Balazard et al., 2016]. They are shown in figure 4.3.

Comparison The result of the two analyses are summarized in figure 4.4 and in table 4.2.

Social variables and markers of outdoor life are negatively associated with T1D: club attendance, playing with friends during the week-end, going to the pool at a friend's house, winter sports and going often to the beach. Going often to the beach was sensitive to the age-related exclusion considered in the next section. Club attendance was also partially affected. Patients had less gastroenteritis before T1D diagnosis.

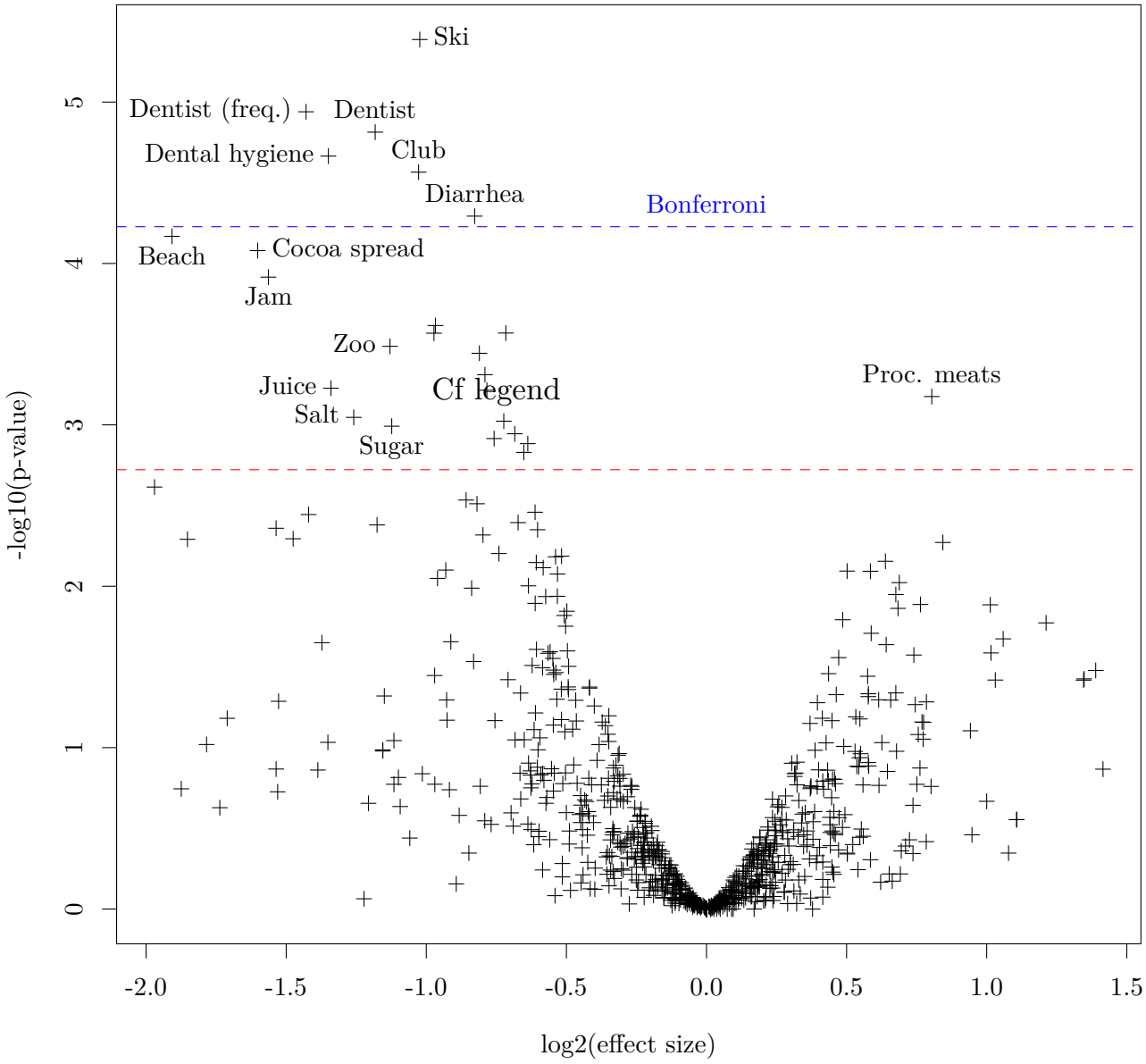

Figure 4.2: **Volcano plot for the matched analysis.** The x-axis shows the effect size with protective factors on the left and risk factors on the right. The y-axis indicates the significance. The higher line indicates the Bonferroni threshold while the lower line shows the more lenient threshold for 5% of false discovery rate. The unlabeled variables above the FDR threshold are from most significant to least: week-ends with other children, taste for sugar as a baby, death of a pet from old age, vegetables from farm, home-made delicatessen, stings (mainly wasps and bees), siblings before birth, friend's pool, plane, fresh exotic fruits, vegetables from a rural market during pregnancy.

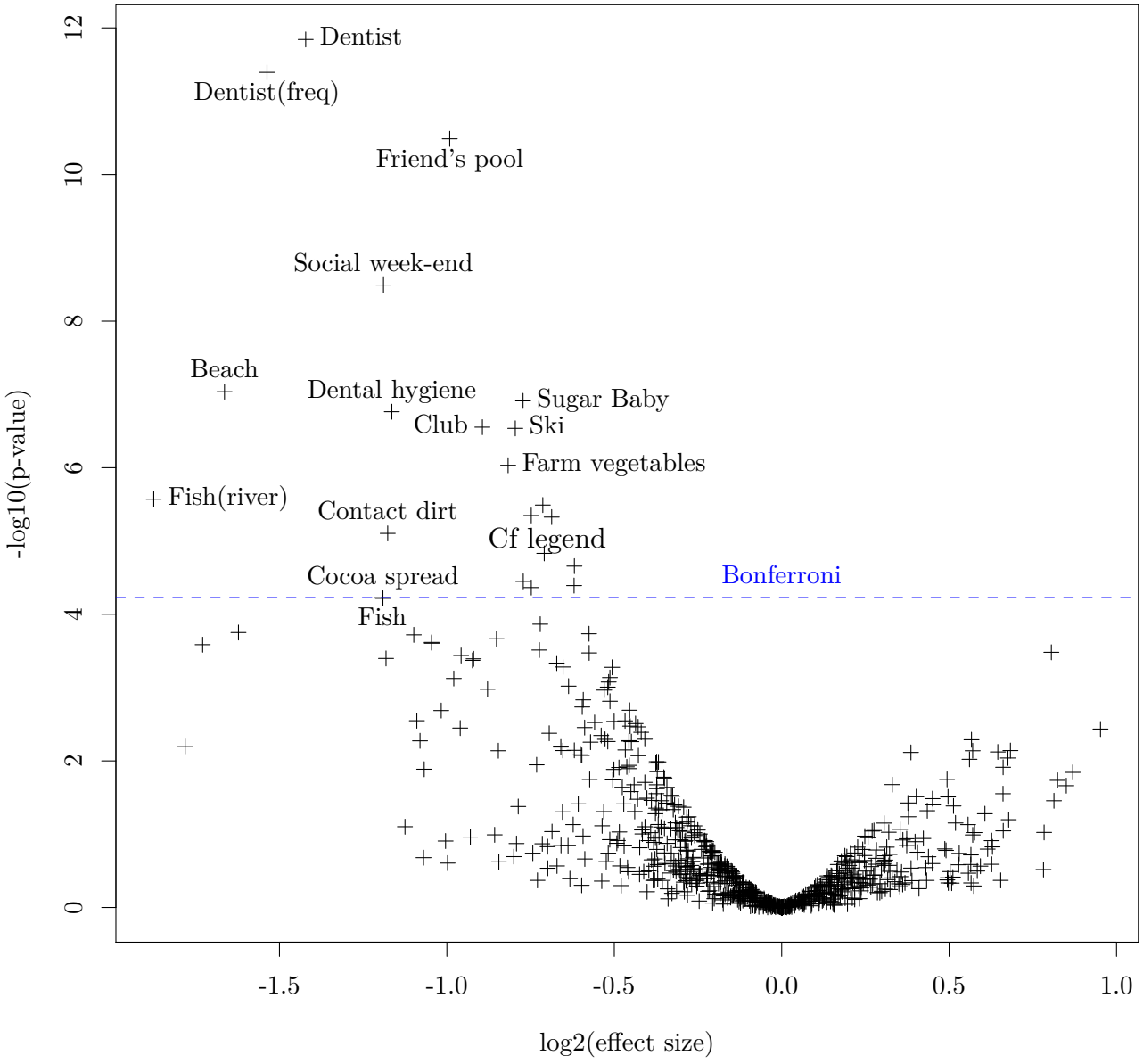

Figure 4.3: **Volcano plot for the propensity analysis.** The x-axis shows the effect size with protective factors on the left and risk factors on the right. The y-axis indicates the significance. The horizontal line indicates the Bonferroni threshold. The unlabeled variables above the threshold are from most significant to least: fruits from a farm or a family garden during childhood, stings, diarrhea, diarrhea during winter, contact with cats in the neighborhood, pet shop, swimming pool during pregnancy and death of a pet of old age.

		Matched analysis			Propensity analysis		
Label	Levels	Missing	Size	(CI)	Missing	Size	(CI)
Club†	2	1%/ 2%	0.49	(0.35;0.68)	2%/2%	0.54	(0.42;0.68)
Social week-end	2	1%/ 1%	0.51	(0.36;0.73)	1%/1%	0.44	(0.33;0.58)
Friend's pool	2	1%/0%	0.62	(0.47;0.82)	1%/0%	0.5	(0.41;0.62)
Ski	2	1%/2%	0.49	(0.36;0.67)	2%/2%	0.58	(0.47;0.71)
Beach*	4	3%/2%	0.27	(0.14;0.51)	3%/2%	0.32	(0.20;0.49)
Diarrhea	2	5%/5%	0.56	(0.43;0.74)	7%/5%	0.62	(0.51;0.76)
Cocoa spread	5	1%/1%	0.33	(0.19;0.57)	0%/1%	0.44	(0.29;0.66)
Sugar baby	2	2%/3%	0.61	(0.47;0.79)	3%/2%	0.59	(0.48;0.71)
Dental hygiene†	3	0%/0%	0.39	(0.25;0.6)	1%/0%	0.45	(0.33;0.61)
Dentist*	2	3%/1%	0.44	(0.3;0.64)	3%/3%	0.37	(0.28;0.49)
Dentist (freq.)*	4	2%/3%	0.37	(0.24;0.58)	2%/1%	0.34	(0.25;0.47)
Stings	2	3%/3%	0.58	(0.43;0.79)	3%/3%	0.6	(0.48;0.74)
Pet's death	2	14%/12%	0.51	(0.35;0.73)	13%/11%	0.6	(0.47;0.76)
Farm vegetables	2	1%/1%	0.57	(0.42;0.77)	1%/1%	0.57	(0.45;0.71)
Exclusive breastfeeding	2	2%/2%	0.88	(0.68;1.15)	2%/2%	0.77	(0.63;0.94)
Respiratory infections	2	5%/4%	0.87	(0.68;1.12)	6%/4%	0.89	(0.73;1.1)

Table 4.2: **Effect sizes for significant variables and pending risk factors.** Effect sizes are odd ratios for binary variables and correspond to odd ratio between extreme responses for ordinal variables. Percentage of missing data are split between patients and controls. Factors from the literature are at the end of the table. *: variables affected by further age-related exclusion. †: variables affected by the further exclusion for the propensity analysis only.

Hazelnut cocoa spread consumption and sweet eating as a baby were both negatively associated with T1D.

Three variables were closely connected to dental hygiene. The variable "dental hygiene" is an ordinal variable quantifying the frequency at which the participants brush their teeth. The two variables "dentist" and "dentist (freq.)" are a binary and an ordinal variable quantifying the number of dentist visit attended by the participant. Future T1D patients attended the dentist less and brushed their teeth less as well. The association for dentist attendance was very sensitive to the further exclusion considered in the next section. Dental hygiene was also partially affected.

The patients reported having been stung less than controls. "Stings" refers to the question: Was the subject stung by an animal who left a clear spot (red spot, painful or not)? with four propositions for the responsible animal: a wasp, a bee, another insect or a fish. Mosquitoes, spiders and ticks were the subjects of separate questions. Wasp and bee stings were the most common stings. Patients less often had the experience of having a pet die of old age. Patients ate less vegetables coming from a farm or a family garden.

Factors studied in the literature. We compared the results of our study with the few risk factors that have been suspected to be associated with T1D in the studies cited in the introduction. Breastfeeding was investigated by two questions in the questionnaire: whether the

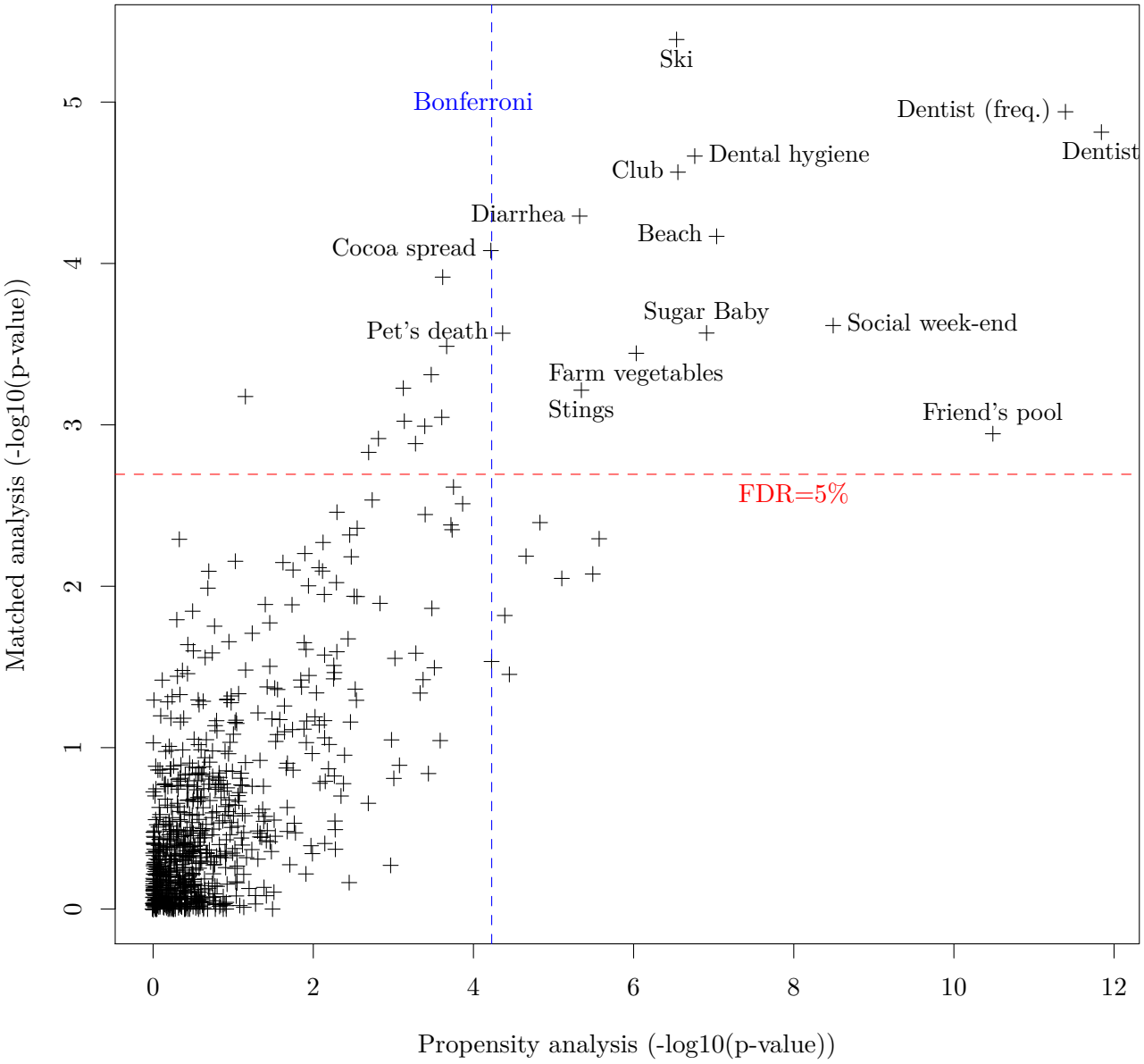

Figure 4.4: **Comparison of the results of the two analysis.** -log10(p-value) of the two analysis plotted against each other. The most associated variables in both analysis are in the top right corner. The Bonferroni threshold for the propensity analysis is the vertical line. The false discovery rate threshold for the matched analysis is the horizontal line. A more lenient statistical control is used for the matched analysis as it is less prone to bias. All variables passing both thresholds are labeled.

subject had been breastfed at all and the duration of exclusive breastfeeding. In the matched analysis, neither questions were significant at the nominal level but in the propensity analysis, the duration of exclusive breast-feeding was found to be highly protective. Any breastfeeding was also protective with nominal significance. Lower respiratory infections were not associated with risk of T1D in our analyses. Vitamin D supplementation for the mother after birth was not associated with T1D in either analysis.

4.3 Further inquiry on age-related bias

A first analysis of the data showed a significant association between primary school attendance and disease. Patients and controls were expected to be matched on age. A closer look at the data showed that a fraction of participants, especially controls, reported attendance to primary school before 6 years old. Admission in primary school is rarely allowed before 6 years old in France so it seemed like mistakes. We interpreted this mistake to be a marker that the whole questionnaire had been filled without respect to the reference age. In the previous section, we excluded those participants where the mistake occurred. To assess if the primary school mistake was a one time error or if it impacted the rest of the questionnaire, we trained a random forests on the questionnaire to predict reference age. Prediction results show that the exclusion was justified. We then consider an additional exclusion based on the prediction of age.

4.3.1 Prediction model for age

We trained random forests to predict age using the questionnaire. The rationale is that as a participant advances in age, he experiences more diverse environments. This allows to try and predict age using the answers to the questionnaire. While this prediction is by nature approximate, large differences between reference age and predicted age are suspected to reveal inadequate filling of the questionnaire.

We trained our model on the dataset obtained after the primary school mistake. We excluded the question regarding primary school from our model as we wanted to know if the rest of the questionnaire was influenced by a mistake on that question. We also excluded the 295 variables with more than 5% of missing data. After the exclusions, the overall missing rate was 2%. We performed a simple imputation of the remaining missing data using the na.roughfix function of the randomForest package [Liaw and Wiener, 2002]. We then trained a random forests regression on the completed dataset. The default parameters for regression were used.

The trained random forests out-of bag estimate accounted for 63% of the reference age variance in the training set. It did not predict ages above 12. This is understandable since the questionnaire is centered on early childhood and therefore no questions allow to distinguish a 15 years old from a 12 years old. We then used the model to predict the age of those who had wrongly answered the primary school question. Figure 4.5 justifies the exclusion of the 14 cases (1%) and 83 controls (10%) who made the primary school mistake as their predicted age is in general much larger than their reference age.

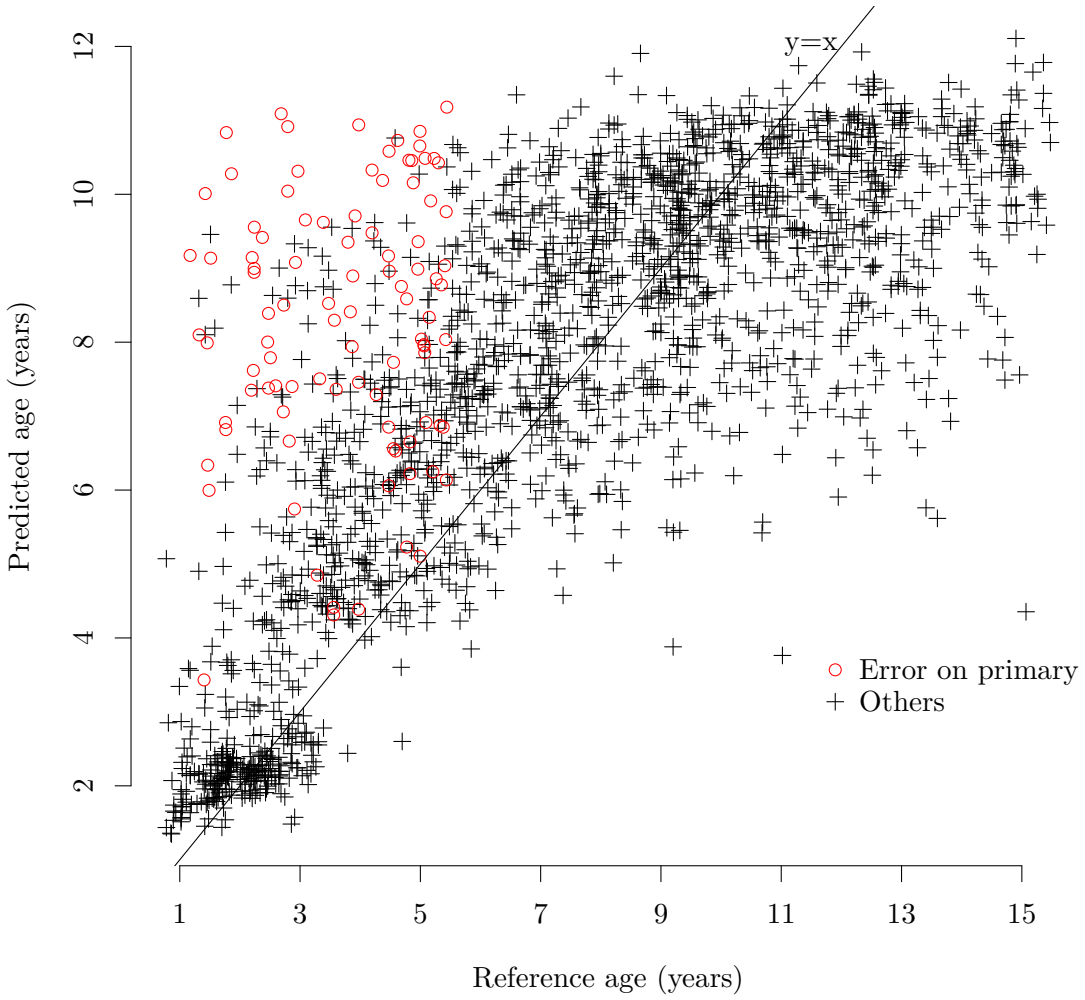

Figure 4.5: **Predicted age for primary school mistake participants.** The predicted age is plotted against the reference age. The predicted age is the out-of-bag estimate for the participants who did not make the primary school mistake (+). The predicted age is the prediction of the entire forest for the ones who made the primary school mistake (circles). The line y=x corresponds to a perfect prediction.

4.3.2 Additional age-related exclusion

In this section, using the same prediction model for age as above, we consider a further exclusion of participants whose predicted age is much larger than their reference age but did not make the primary school mistake. The excluded participants are again disproportionately controls. This exclusion is not as justified as there is no way to know if a mistake has indeed happened or if the participant simply had many experiences early in life. We study how the results are impacted by this further exclusion.

In the remaining participants, 16 patients (1%) and 39 controls (5%) had a difference between predicted age and reference age larger than 4 years. While those participants did not make the primary school mistake, we considered the possibility that they nevertheless filled the questionnaire not paying much attention to the reference age. This is supported by the disproportion between cases and controls in this set. However, there is no definite way of knowing if there has been a mistake or not in this case and this is why we considered presented the results without this exclusion in the previous section. Figure 4.6 shows the set of excluded participants. We then defined once again two datasets in the same way as in the main text. The modified exclusion process is summarized in the flowchart of figure 4.7. We refer to the resulting datasets as modified. The datasets defined in the main text are referred to as original.

On the list of the original significant results, we performed the same analysis as described in the main text in the modified datasets to evaluate the impact of the exclusion. In order to determine if the drop in significance was simply due to smaller sample size, we defined an empirical distribution of p-values under random exclusion. We randomly select a subsample of the corresponding original dataset (matched or propensity) with the same sample size as the modified dataset. As missing data does not influence the results of the tests, for each variable, we exclude the missing data before determining the sample size. For the propensity analysis, having the same sample size means having the same number of patients and the same number of controls. For the matched analysis, having the same sample size means having the same number of strata of the same type: one patient and one control or one patient and two controls. In this random subsample, we then perform the same analysis as in the main text. This gives a p-value. We then repeat this process 10000 times to obtain an empirical distribution of p-value under random subsampling. This allows us to test if the new p-value obtained for the modified dataset can be attributed to smaller sample size or not. The p-value of that test is then the proportion of the distribution that has a larger p-value than the modified p-value.

4.3.3 Results

The significance of the drop compared to random exclusion for both p-values is shown in table 4.3. The exclusion affected significantly the two dentist variables and the beach variable for both analysis. The exclusion affected significantly for the propensity analysis the dental hygiene variable and the club variable. The drop in p-value is shown in figure 4.8. The full results are available in online additional file 4 [Balazard et al., 2016].

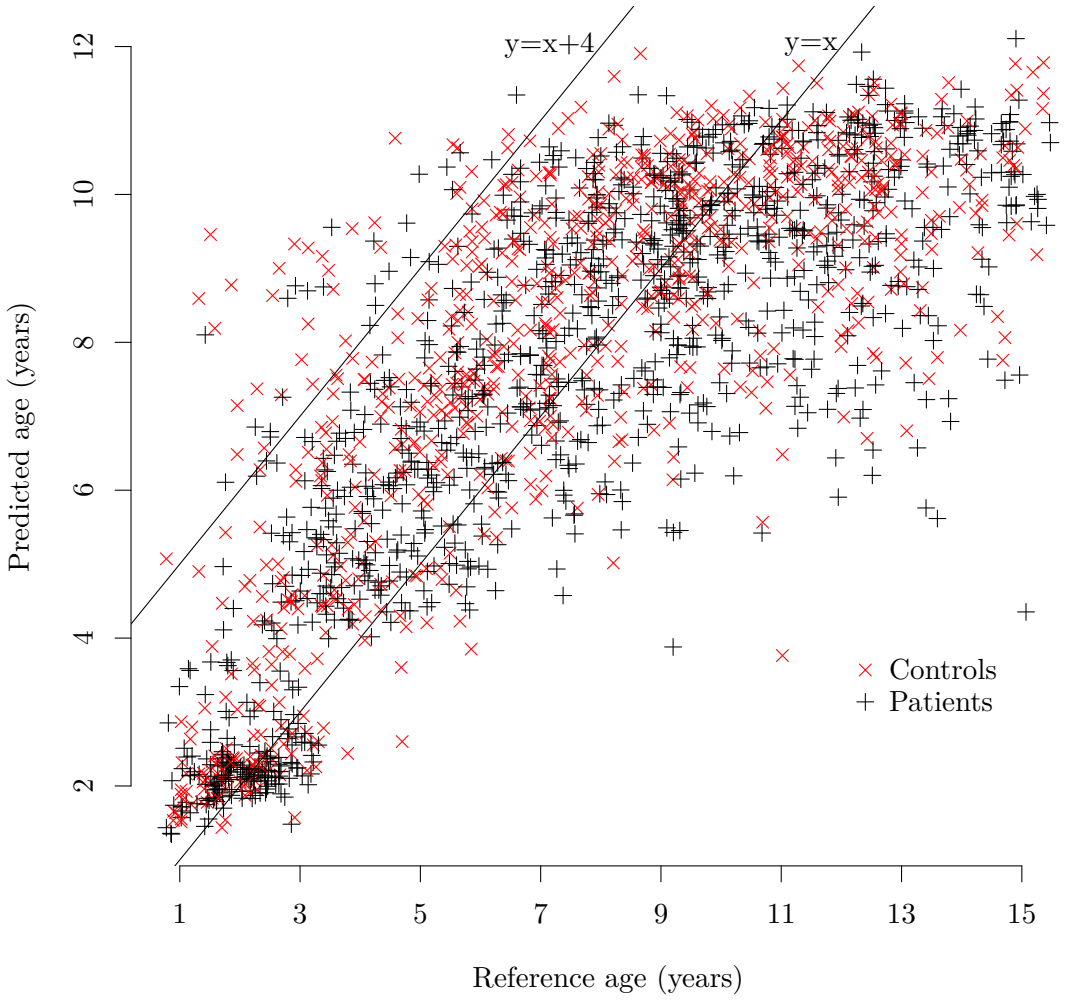

Figure 4.6: **Predicted age much larger than reference in a subset.** The out-of-bag estimate of age is plotted against the reference age. The line y=x corresponds to a perfect prediction. Over the line y=x+4, the participants are excluded.

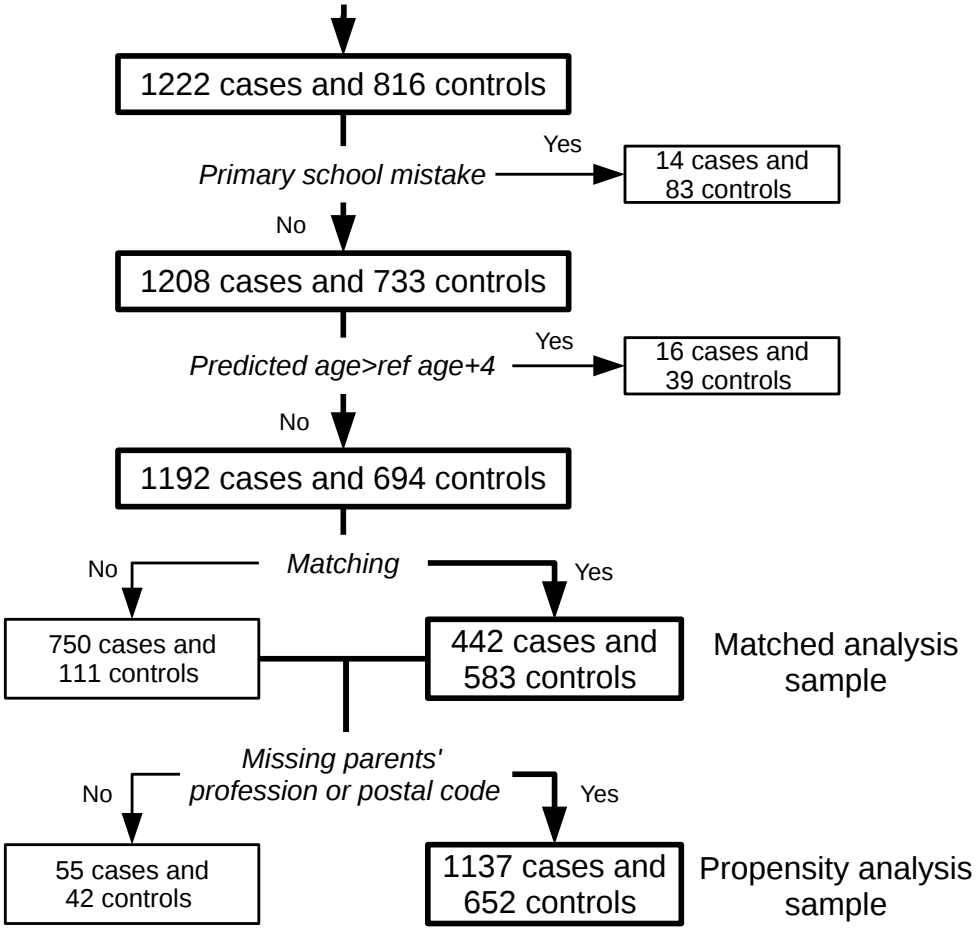

Figure 4.7: **Flowchart of modified exclusions and sample definition.**

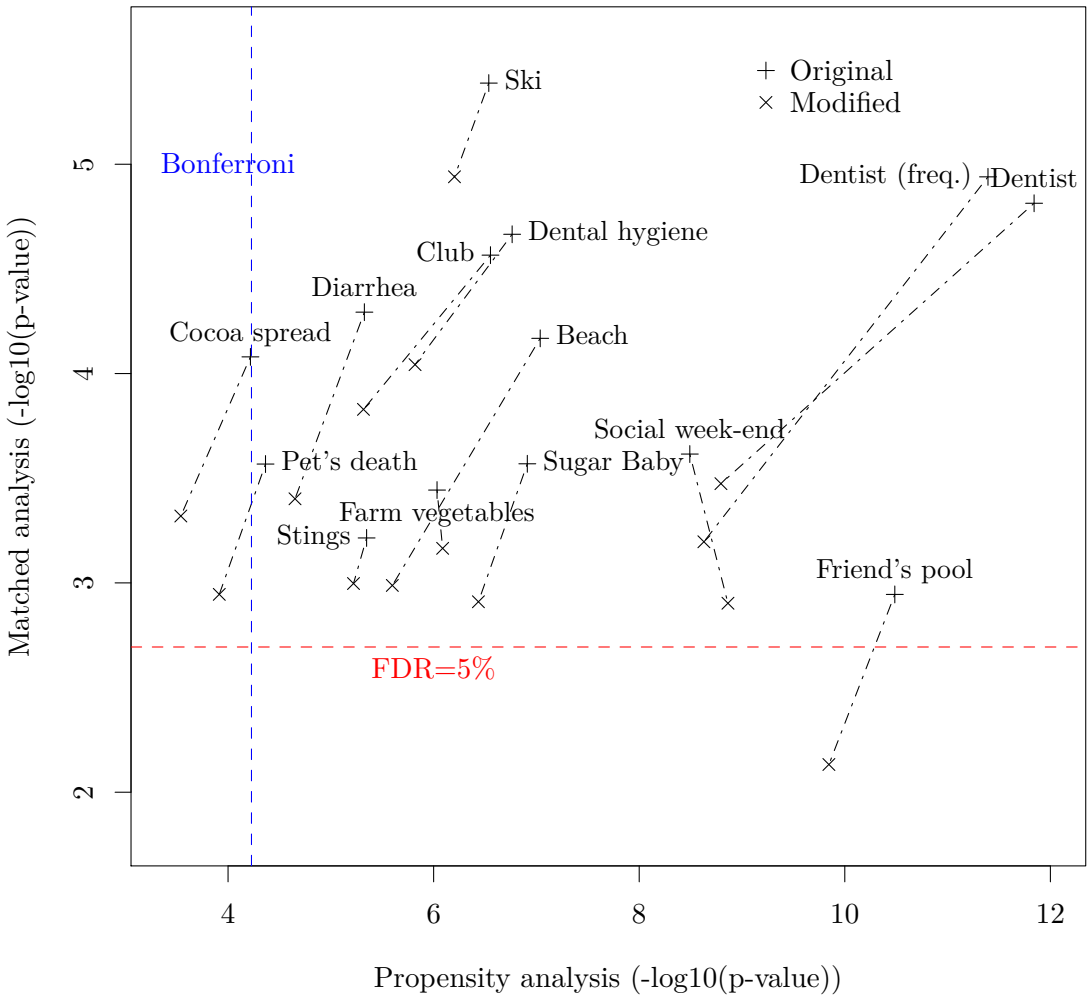

Figure 4.8: **Drop of p-values after the exclusion.** The red and blue line are the same threshold as in figure 4.4. The original and modified p-value of the same variable are connected by a dotted line.

Label	p-value (Matched)	p-value (Propensity)
Sugar baby	0.2175	0.1801
Dentist	0.0013	0
Dentist (freq.)	0	0
Dental hygiene	0.1209	0.0028
Diarrhea	0.1599	0.0867
Stings	0.2477	0.3417
Farm vegetables	0.4555	0.7016
Cocoa spread	0.1026	0.2002
Ski	0.4433	0.2292
Beach	0.0002	0.0042
Friend's	0.0112	0.0611
Club	0.0664	0.0003
Social week-end	0.0876	0.729
Pet's death	0.1316	0.124

Table 4.3: **Significance of the drop in p-value compared to a random exclusion.** These p-values are obtained by subsampling 10000 times the original datasets. A p-value of 0 means that no occurrence of the subsampling resulted in a p-value as large as that obtained by the exclusion.

4.4 Replication on a short questionnaire

When patients did not respond to the long questionnaire, a shorter questionnaire of 49 questions was sent for them to fill. This questionnaire was designed while the study was underway and partial results were used to choose the included questions. As a result, 7 of the 22 variables deemed significant in the analysis presented in section 4.2 are among the 49 questions of the shorter questionnaire. The 7 variables are the answers to the following questions concerning the period before diagnosis:

- Sugar baby: "As a baby, did the patient like baby food jars containing sweet foods more than the ones without sweet foods?",
- Dentist: "Had the patient gone to the dentist?",
- Dental hygiene: "How many times a day did the patient brush his teeth?",
- Diarrhea: "Did the patient experience severe diarrhea accompanied by vomiting?",
- Cocoa spread: "Did the patient eat hazelnut cocoa spread?",
- Ski: "Had the patient been to winter sports?" and
- Club: "Did the patient attend a club with other children (sports, music,...)?".

We replicated the matched analysis on those 7 questions. After excluding patients and controls with no matched counterpart, there were 96 patients and 132 controls. We performed

Variable	p_{null}	p_{alter}
Sugar Baby	0.017	0.70
Dentist	0.44	0.020
Dental hygiene	0.19	0.10
Diarrhea	0.60	0.0077
Cocoa spread	0.28	0.0755
Ski	0.009	0.618
Club	0.0004	0.978

Table 4.4: Results of the replication.

conditional logistic regression on this sample. As a direction of association had been found in the main analysis, we performed unilateral tests for significance. The unilateral p-values of the tests are in the column labeled p_{null} of table 4.4. A small value means that there is an association in the same direction as for the long questionnaire.

In addition to this, we evaluated whether the new data was compatible with the data we observed in the main analysis. This was done by sub-sampling the matched analysis dataset of the main analysis to obtain as many duos and trios as in the new dataset (after excluding missing data). We then performed on this sample the same analysis as in the previous paragraph which gives us one p-value. This procedure was repeated 10000 times which gives us a distribution of p-values under the assumption that the new data comes from the distribution of the old dataset. We then consider the proportion of these p-values that are less significant, i.e., larger, than the observed value. This proportion is a p-value to test the null hypothesis that the new data comes from the same distribution as the old data against the alternative hypothesis of independence between the variable and the disease. These p-values are in the column labeled p_{alter} of table 4.4

These results are interpreted as follows. Variables that have p_{null} small and p_{alter} large are confirmed by the new data: the observed data are far from the null hypothesis and are in the middle of the distribution obtained by sub-sampling the old data. Therefore the negative association of taste for sugar as a baby and experience of winter sports are replicated. This is also true for club attendance and in fact the association observed is more important than expected.

Variables that have p_{null} large and p_{alter} small are negated by the new data. This is the case of diarrhea and dentist attendance.

Variables that have p_{null} and p_{alter} of the same order of magnitude are neither confirmed nor negated by the new data. This is the case for consumption of cocoa spread and dental hygiene. Both distributions (null and alternative) are close and the new data falls between the two. The effect size of these variables is maybe overvalued.

4.5 Discussion

While our statistical analysis indicates that playing with friends during week-ends or going to the pool at a friend's house, experience of winter sports and club attendance were all negatively

associated to childhood T1D (and replicated in the short questionnaire for winter sports and club attendance), we have not attempted to interpret these protective associations.

We also found a negative association of gastroenteritis and T1D that was however negated in the short questionnaire. Gut microbiology is an area highlighted by this association. The results of the DAISY study suggests a complex relationship with gastroenteritis [Snell-Bergeon et al., 2012].

As sugar consumption is strongly present as a nutritional caveat in the minds of parents having a child with T1D, we suspected that the replicated negative association between "appetite for sugar as a baby" and T1D could be due to recall bias. However, with respect to a possible recall bias, sugary products such as cola drinks or chocolate show no association with T1D. This gives credibility to the found negative association for hazelnut cocoa spread. Furthermore, hazelnut cocoa spread remains significant after adjustment for appetite for sugar as a baby: in the matched sample, fitting a conditional logistic regression to both variables gives an estimate for cocoa spread of 0.36 $(0.20, 0.64)$ instead of 0.33 (0.19,0.57), suggesting that the result for cocoa spread is not affected by recall bias. Hazelnut cocoa spread contains a large proportion of palm oil thus a high content of tocotrienol. In murine models, tocotrienol was shown to affect NLPR3 and NF-$\kappa\beta$ [Kim et al., 2016, Kuhad et al., 2009], which may play a role in T1D pathogenesis [Hu et al., 2015, Evans et al., 2003]. We found that items related with dental hygiene, such as frequency of teeth brushing and dentist attendance, were negatively associated with T1D although they were sensitive to a further exclusion and dentist attendance was negated in the replication. Again, we have not attempted to interpret this protective association in our current state of knowledge.

Wasp and bee stings also showed a significant association with T1D, but the meaning of this observation remains to be found.

Death of pet by old age was negatively associated with T1D. This was a subquestion of death of a pet which was nominally significant in both analysis. Another subquestion, death of a cat, was also associated. We offer no interpretation.

Eating vegetables from a farm or a family garden was negatively associated with T1D. The analogue question for fruits passed the Bonferroni threshold in the propensity analysis and was also nominally significant in the matched analysis. These associations might be connected to contact with dirt which was also significant in the propensity analysis and nominally significant in the matched analysis. Again, we offer no interpretation.

While many exposures and events have remained out of reach of our questionnaire because they were not detectable or escaped parental memory, the novel protective associations that were found cannot be entirely false positive findings. They may open new areas of investigation for T1D environmental research and should not be dismissed more than yet biologically inexplicable SNP associations generated by GWAS. However they will only be of interest if they can be confirmed in other childhood T1D cohorts.

Acknowledgement. We thank Alain Fourreau, Adeline Guégan, Gaël Leprun and Valérie Jauffret for mailing the questionnaires and entering the responses. We thank the participants and their parents for their time. We acknowledge the collective effort of the Isis-Diab collaborative group which is composed of the following physicians: : Dr Dominique Thevenieau, Dr Corinne Fourmy Chatel, Dr Rachel Desailloud, Dr Hélène Bony-Trifunovic, Dr Pierre-Henri

Ducluzeau, Prof Régis Coutant, Dr Sophie Caudrelier, Dr Armelle Pambou, Dr Emmanuelle Dubosclard, Dr Florence Joubert, Dr Philippe Jan, Dr Estelle Marcoux, Dr Anne-Marie Bertrand, Dr Brigitte Mignot, Prof Alfred Penformis, Dr Chantal Stuckens, Dr Régis Piquemal, Prof Pascal Barat, Prof Vincent Rigalleau, Dr Chantal Stheneur, Dr Sylviane Fournier, Prof Véronique Kerlan, Dr Chantal Metz, Dr Anne Fargeot-Espaliat, Prof Yves Reznic, Dr Frédérique Olivier, Dr Iva Gueorguieva, Dr Arnaud Monier, Dr Catherine Radet, Dr Vincent Gajdos, Dr Daniel Terral, Dr Christine Vervel, Dr Djamel Bendifallah, Dr Candace Ben Signor, Dr Daniel Dervaux, Dr Abdelkader Benmahammed, Dr Guy-André Loeuille, Dr Françoise Popelard, Dr Agnès Guillou, Prof Pierre-Yves Benhamou, Dr Jamil Khoury, Dr Jean-Pierre Brossier, Dr Joachim Bassil, Dr Sylvaine Clavel, Dr Bernard Le Luyer, Prof Pierre Bougnères, Dr Françoise Labay, Dr Isabelle Guemas, Prof Jacques Weill, Dr Jean-Pierre Cappoen, Dr Sylvie Nadalon, Dr Anne Lienhardt-Roussie, Dr Anne Paoli, Dr Claudie Kerouedan, Dr Edwige Yollin, Prof Marc Nicolino, Prof Gilbert Simonin, Dr Jacques Cohen, Dr Catherine Atlan, Dr Agnès Tamboura, Dr Hervé Dubourg, Dr Marie-Laure Pignol, Dr Philippe Talon, Dr Stéphanie Jellimann, Dr Lucy Chaillous, Dr Sabine Baron, Dr Marie-Noëlle Bortoluzzi, Dr Elisabeth Baechler, Dr Randa Salet, Dr Ariane Zelinsky-Gurung, Dr Fabienne Dallavale, Dr Etienne Larger, Dr Marie Laloi-Michelin, Dr Jean-François Gautier, Dr Bénédicte Guérin, Dr Laure Oilleau, Dr Laetitia Pantalone, Dr Céline Lukas, Dr Isabelle Guilhem, Dr Marc De Kerdanet, Dr Marie-Claire Wielickzo, Dr Mélanie Priou-Guesdon, Dr Odile Richard, Dr François Kurtz, Dr Norbert Laisney, Dr Déborah Ancelle, Dr Guilhem Parlier, Dr Catherine Boniface, Dr Dominique Paris Bockel, Dr Denis Dufillot, Dr Berthe Razafimahefa, Dr Pierre Gourdy, Prof Pierre Lecomte, Dr Myriam Pepin-Donat, Dr Marie-Emmanuelle Combes-Moukhovsky, Dr Brigitte Zymmermann, Dr Marina Raoulx, Dr Anne Gourdin, Dr Catherine Dumont and Dr Michèle Chambon.

Chapter 5

Interactions and collider bias in case-only gene-environment studies

Abstract Case-only design for gene-environment interaction (CODGEI)'s interpretation relies on the rare disease assumption. When this assumption is not respected, a negative association due to collider bias appears between gene and environment. We propose a framework for simulation of disease occurrence in a source population that allows to estimate the influence of collider bias in CODGEI. Collider bias offers an alternative interpretation to the results of CODGEI in a published study on breast cancer. In a more speculative part of this chapter, we introduce Disease As Collider (DAC), a new case-only methodology that leverages collider bias to discover environmental factors using genetic risk estimation: a negative correlation between genetic risk and environment among cases provides a signature of a genuine environmental risk marker. We illustrate DAC in 831 type 1 diabetes cases. Our simulation framework allows to estimate the power of DAC. In the current context, DAC is underpowered.

Contents

5.1 Introduction

Case-only gene-environment studies are attractive since data are often easily available in cases. It also means that the selection of controls, a sensitive process, can be avoided. Case-only design for gene-environment interaction (CODGEI) allows to study gene-environment interactions in

	G=0	G=1
E=0	a	b
E=1	c	d

Table 5.1: **Gene-environment data in a case-only setting.** The letters stand for the number of cases in the corresponding category.

this setting [Piegorsch et al., 1994, Khoury and Flanders, 1996]. An assumption on the joint distribution of G and E in the general population is needed to compensate for the loss of information induced by observing only cases. The natural assumption is independence between G and E in the general population [Albert et al., 2001, Gatto et al., 2004]. Under this assumption, CODGEI uses case-only data to identify gene-environment interaction. Specifically, if both G and E are binary traits as shown in table 5.1, the cross-product ratio (CPR) ad/bc computed from the case-only data is an estimator of the interaction risk ratio

$$RR_I = \frac{\mathbb{P}(D=1|G=1,E=1)}{\mathbb{P}(D=1|G=0,E=1)} \Big/ \frac{\mathbb{P}(D=1|G=1,E=0)}{\mathbb{P}(D=1|G=0,E=0)}.$$

Indeed, following Schmidt and Schaid [1999], we denote $p_{ij} = \mathbb{P}(G=i,E=j|D=1)$ and we have :

$$\begin{aligned}
\mathbb{E}[\text{CPR}] = \mathbb{E}[\frac{ad}{bc}] &\approx \frac{p_{00}p_{11}}{p_{10}p_{01}} \\
&= \frac{\mathbb{P}(D=1|G=1,E=1)}{\mathbb{P}(D=1|G=0,E=1)} \Big/ \frac{\mathbb{P}(D=1|G=1,E=0)}{\mathbb{P}(D=1|G=0,E=0)} \\
&\times \frac{\mathbb{P}(G=1,E=1)}{\mathbb{P}(G=0,E=1)} \Big/ \frac{\mathbb{P}(G=1,E=0)}{\mathbb{P}(G=0,E=0)} \\
&= RR_I \times OR_{GE}
\end{aligned}$$

where OR_{GE} measures the association of G and E in the general population and is therefore 1 when $G \perp\!\!\!\perp E$. The RR_I measures the departure from multiplicative risk ratios. However, there is another measure of interaction that we will focus on in this chapter: the interaction odd-ratio OR_I. It measures the departure from multiplicative odd-ratios, i.e., an interaction in the logistic model, and equals :

$$OR_I = RR_I \times \frac{\mathbb{P}(D=0|G=0,E=1)}{\mathbb{P}(D=0|G=1,E=1)} \Big/ \frac{\mathbb{P}(D=0|G=0,E=0)}{\mathbb{P}(D=0|G=1,E=0)} \tag{5.1}$$

This is the interaction term that is estimated by a case-control study.

In its initial formulation by Piegorsch et al. [1994], CODGEI was based on an additional assumption (reformulated by Schmidt and Schaid [1999]) that the disease is rare at all levels of gene and environment: for all g and e,

$$\mathbb{P}(D=1|G=g,E=e) \ll 1.$$

When the rare disease assumption is verified, the second factor in equation (5.1) is 1 and therefore $OR_I = RR_I$ and the CPR estimates the interaction odd-ratio OR_I.

Schmidt and Schaid [1999] then go on to evaluate the influence of deviations from the rare disease assumption on the mismatch between OR_I and RR_I. Their conclusion is that RR_I can be substantially smaller than OR_I under large deviations from the assumption. This conclusion can be misinterpreted: since ratios are on a multiplicative scale, an underestimation would mean that RR_I is closer to 1 compared to OR_I. In the Figure 3 of their article, you can see that when $OR_I = 1$, we have $RR_I < 1$: when there is no interaction in the logistic model, an inverse interaction will be detected by CODGEI, i.e., the null hypotheses of no interaction represent different situations for OR_I and RR_I. As we have CPR $= RR_I$, the inverse interaction in this situation corresponds to a negative association between G and E among cases.

As was already noted in Cole et al. [2009], this negative association and the corresponding mismatch between OR_I and RR_I is due to collider bias. Collider bias (or collider-stratification bias) is the negative correlation that appears between two causes (G and E) when conditioning on their shared consequence (the collider, in our case D) [Cole et al., 2009]. It can mislead epidemiological investigation [Gage et al., 2016, Greenland, 2003]. A classic example is Berkson's bias in which two diseases are negatively associated in a hospitalized population even though they are independent in the general population [Berkson, 1946, Snoep et al., 2014]. In this example, the collider is hospitalization, the shared consequence of both diseases. By looking only at cases in the hospital, i.e., by conditioning on hospitalization, a negative correlation appears between the two diseases. This principle is illustrated in a and b of figure 5.1. However, it is not necessary for the environmental factor to be a cause for collider bias to appear. If the environmental factor of interest is simply correlated with a causal factor for the disease, collider bias will appear as shown in c and d of figure 5.1.

Our main contribution in this chapter is to introduce a simulation framework of disease occurrence when there is no interaction on the odd-ratio scale. We can then obtain a distribution of case-only datasets that are subject to collider bias. This allows to test against the null hypothesis of no interaction on the odd-ratio when the rare disease assumption is not respected.

To document an example where different conclusions are reached when the null hypothesis is multiplicative odd-ratios instead of multiplicative risk ratios, we searched the literature for a study that applied CODGEI in a situation where the rare disease assumption is not respected. The study that best suited our criterion is the Genetic and Environmental Modifiers of BRCA1/BRCA2 Study (GEMS) [Moorman et al., 2010]. It deviates strongly from the rare disease assumption as it considers interactions between the highly penetrant BRCA1/2 and environment in breast cancer, the most common cancer in women. Using our simulation framework, we show that the conclusions of the study are changed depending on the null hypothesis considered.

In the example of CODGEI and also more generally in epidemiology, collider bias is seen as a nuisance that hinders understanding. In section 5.3 we propose a change of view-point in order to harness collider bias in service of epidemiology. To do this, we need to maximize collider bias and therefore deviate as much as possible from the rare disease assumption. This can be achieved if, instead of considering one genetic variant at a time, we consider genetic risk predictions that use many variants to estimate as accurately as possible the genetic risk. Indeed the individuals with the worst combination of variants will have a non-negligible risk of disease. As we have seen in chapter 2, Genome-Wide Association studies (GWAS) datasets have been used to estimate genetic risk using statistical learning.

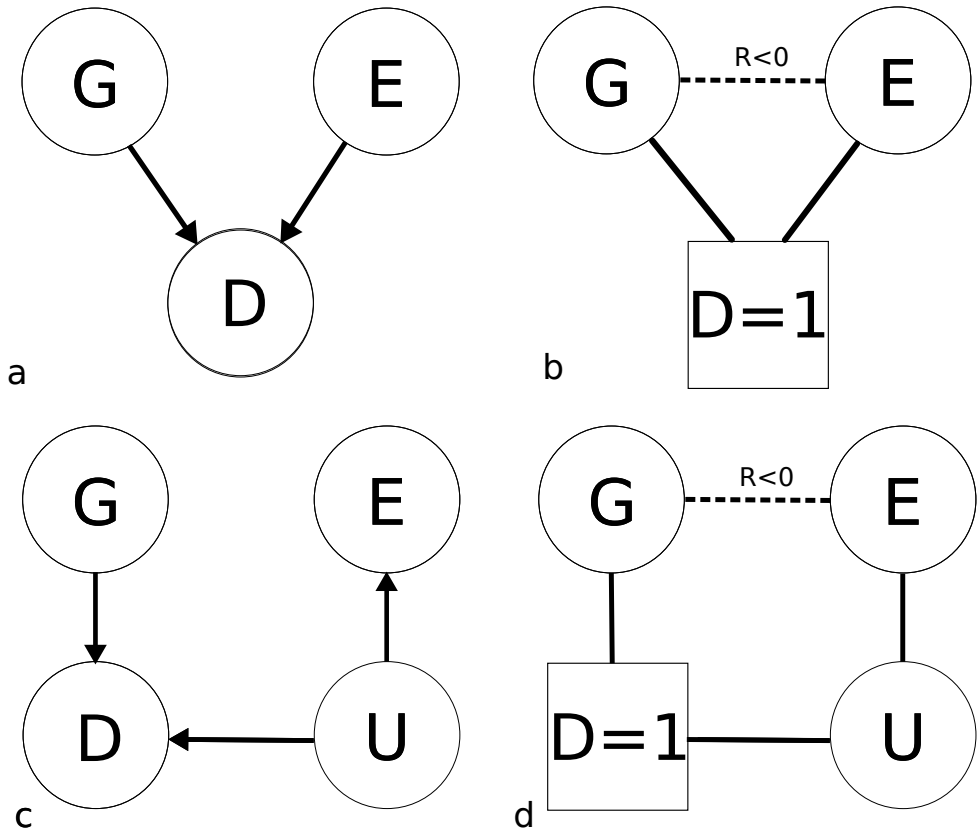

Figure 5.1: **Collider bias in case-only gene-environment data.** a: In the general population, disease is a consequence of both genetic and environmental causes. Depending on the environmental factor considered, we can assume independence between gene and environment. b: When considering only cases, i.e., conditioning on the disease, a negative correlation R appears between genetics and the environmental factor. c and d: If the environmental factor E is only a marker for the unobserved U that is the true environmental cause, collider bias appears nevertheless between G and E.

Under certain assumptions, after conditioning on disease, i.e., by considering only cases, a negative association found between the genetic risk and an environmental candidate will signal a true association between the environmental marker and the disease. We refer to this methodology as Disease As Collider (DAC). To sum up, DAC allows to detect or confirm a putative environmental marker by looking for an association between this marker and genetic risk in case-only data. We illustrate DAC on a subset of genotyped cases from the Isis-Diab case-control study of T1D that we described in chapter 4 and we adapt our simulation framework to evaluate the power of DAC.

5.2 Influence of collider bias in a case-only study

5.2.1 Simulation framework

In this section, we are interested in devising a test for case-only data against the null hypothesis of no interaction on the odd-ratio scale. To do this, we want to simulate disease occurrence in a source population under two assumptions: independence between G and E and no interaction on the odd-ratio scale.

We first recall a few definitions. The logistic model transfers probabilities in $[0,1]$ to log odd-ratios in $\mathbb{R}$ thanks to the logit function $\text{logit}(x) = \log(x/(1-x))$. We refer to the target set of the logit function as the logit scale. As most of our calculations are made on the logit scale, we will write risk instead of risk on the logit scale throughout.

For an individual with genome G and environment E that are both binary and in the absence of an interaction on the logit scale, the total risk $R(G,E) = \text{logit}(P(D=1|G,E))$ is

$$R(G,E) = \beta_0 + \beta_1 G + \beta_2 E.$$

One would expect β_1 (resp. β_2) to be equal to $\text{logit}(P(D=1|G=1))$ (resp. $\text{logit}(P(D=1|E=1))$). However, because of the non-collapsibility of odd-ratios, this is not always the case [Guo and Geng, 1995, Greenland et al., 1999]. Fortunately, we will see that this does not appear to be an issue in our illustration and that therefore this is a reasonable choice for β_1 and β_2. We will later circumvent the need to choose β_0.

For a sample size of N cases, the source population consists of N/K individuals, K being the prevalence of the disease. To allocate disease status in this synthetic population, we need to define a distribution for G and for E. We describe the choice of distributions used for our illustration below. Once both distributions are defined, we then attribute to each individual in the population its genotype G and its environment E by drawing independently from those distributions. This uses the assumption of independence of genetic risk and environmental factor in the population. We then compute the total risk R as $R = \beta_0 + \beta_1 G + \beta_2 E$. To decide whether an individual with gene G and environment E has the disease, we draw a uniform variable U on $[0,1]$ and we could then define the disease variable D:

$$D = 1 \text{ if } U \leq \mathbb{P}(D=1|G,E)$$
$$D = 0 \text{ if } U > \mathbb{P}(D=1|G,E).$$

This approach would yield a different number of cases in each simulation. To always have N cases, we compute $R - \text{logit}(U)$ and define the top N individuals for that sum as the cases ($D = 1$). This also allows not to choose β_0 and we therefore compute $R = \beta_1 G + \beta_2 E$.

Finally, when the simulated sample has been defined, we compute from it the CPR and store it. We then repeat the procedure the desired number of times and obtain a distribution of the CPR. With the resulting empirical distribution of CPR under the null hypothesis of no interaction on the logit scale, we can define a rejection region as the complementary of the 95% confidence interval of the CPR.

5.2.2 Influence of collider bias in GEMS

We investigate if the four significant associations reported in the GEM study [Moorman et al., 2010] could be explained by collider bias. The four associations are BRCA1 and alcohol use (yes vs no), BRCA1 and parity (nulliparous vs 3 children or more), BRCA2 and parity (nulliparous vs 2 children) and BRCA2 and age at menarche (before 11 vs after 14). Given the direction of the main effects in the literature for the three risk factors, the RR_I are in the direction expected under absence of interaction on the logit scale.

We therefore implement our simulation framework in the precise setting of the study to find out if the results could be attributed to collider bias, i.e., if changing the null hypothesis would change the results of the tests. We use the sample size and the number of carriers relevant for the 4 comparisons. We retrieve relative risks from the literature and then compute the corresponding OR. We choose a RR of 1.32 for alcohol use vs no alcohol [Collaborative group on Hormonal Factors in Breast Cancer et al., 2002], a RR of 1.29 for nulliparous vs 2 children, a RR of 1.54(=1.29/0.84) for nulliparous vs more than 3 children [Ewertz et al., 1990] and a RR of 1.05^{-4} for menarche after 14 years old vs menarche before 11 [Collaborative group on Hormonal Factors in Breast Cancer et al., 2012]. We adjust the distribution of the risk factor in the source population in order to obtain in simulated cases the observed distribution in cases.

On the genetic side, there is a single variant: either BRCA1 or BRCA2. The distribution in the general population of these variants is simply the prevalence of the mutations. The risk on the logit scale can be obtained from the prevalence of breast cancer in the general population and in carriers of the mutation. We choose a prevalence of breast cancer of 12%, a prevalence of breast cancer among carriers of BRCA1/2 of 60% [Chen and Parmigiani, 2007] and therefore the OR for BRCA was 11. We choose a prevalence of 0.1% for BRCA1 and 0.2% for BRCA2 [Malone et al., 2006]. The OR for each effect, genetic and environmental are presented in table 5.3.

The procedure presented above has to be adapted to the precise setting of the GEM study. Indeed, all cases were not included in the original study: the authors included all cases carrying BRCA1/2 and a number of non-carrier cases as a comparison group. In consequence, the prevalence of BRCA1/2 is much higher in the study than in the population of cases. To take this into account, we add an additional step to the simulation after the allocation of disease status: we define a variable for inclusion in the study. All carrier cases are included in the study and the rest of the sample size is filled at random from non-carrier cases. We then use only the cases who are included in the study. This means that the source population is larger than N/K. We adjust the size of the source population to obtain in average the observed fraction of carriers

Interaction	Median CPR (CI) under H0	Reported CPR
BRCA1-alcohol	0.86 (0.63;1.20)	0.65
BRCA1-parity	1.26 (0.86;1.85)	1.54
BRCA2-parity	1.15 (0.78;1.71)	1.54
BRCA2-menarche	1.11 (0.71;1.73)	1.65

Table 5.2: **Results of 10000 simulations under the null hypothesis of no interaction on the logit scale for the 4 significant associations in the GEM Study.** The reported CPR is reproduced from the original paper.

Interaction	BRCA OR	Median	Environmental OR	Median
BRCA1-alcohol	11	10.9	1.38	1.38
BRCA1-parity	11	10.7	0.61	0.61
BRCA2-parity	11	11	0.74	0.74
BRCA2-menarche	11	11.1	0.80	0.80

Table 5.3: **Collapsibility of genetic and environmental effects.** We compare the parameter used in the multivariate model with the median of the distribution of the univariate estimate for the genetic and environmental effects in 10000 simulations.

in the simulated samples.

The results are presented in Table 5.2. The median and a 95% CI for the CPR under the null hypothesis of multiplicativity on the odd-ratio scale is presented alongside the CPR adjusted for age and center from the original article. In the 4 cases, there is a shift away from 1 of the median CPR and the reported CPR falls in the 95% CI. For all four variables, the null hypothesis $H_0 : OR_I = 1$ is not rejected.

To evaluate if non-collapsibility is an issue in our setting, we performed in each simulation univariate logistic regressions for E and for G in a case-control sample drawn from our source population. The case-control sample included all cases and as many controls. This then gives a distribution of estimated odd-ratios over all simulations. As shown in table 5.3, the median of that distribution was very close to the value used in the multivariate definition of risk. Non-collapsibility therefore does not appear to be an issue here.

5.3 Disease as collider

In this section, we consider whether collider bias could be harnessed in service of epidemiology. We propose a new methodology : Disease as collider (DAC) that aims to confirm a putative environmental risk marker by looking for an association with genetic risk in case-only data. In order for collider bias to be the only phenomenon present in case-only data, we need to assume that G is independent from E and that there is no interaction on the odd-ratio scale. In other words, we place ourselves under the same assumptions that were made in the previous section. These assumptions are now our model assumptions and not a null hypothesis.

Compared with the previous section, we are not considering a single genetic variant but rather a genetic risk estimation $R_g(G)$ such as we saw in chapter 2. This will allow us to be as

far as possible from the rare disease assumption and therefore maximize collider bias and the statistical power of DAC.

Under our model assumptions, if E is a genuine environmental marker, there is an association between E and the genetic risk R_g in cases due to collider bias. Our method consists simply in estimating R_g in cases and then on testing for association between R_g and E using standard tests (such as a linear regression t-test) while controlling for potential confounders. When a significant association is found, association of E with the disease D is supported by DAC. The association that appears because of collider bias is a negative association. Therefore, DAC predicts that the cases the most at risk genetically are the least at risk because of environment. When a putative direction of association has been established, one can perform one-sided tests. This is the case when DAC is applied to confirm findings from a case-control association.

5.3.1 Illustration of DAC on the Isis-Diab study

We applied DAC to the cases of the Isis-Diab study. To quantify the genetic risk of T1D, we used the genetic risk estimator presented in subsection 2.4 that was designed to be as close as possible from the one in Wei et al. [2009]. We applied our method on the 7 environmental variables presented in subsection 4.4 that were significant in the long questionnaire of the Isis-Diab study and also present in the sort questionnaire. A total of 2959 cases filled a questionnaire: 1713 cases filled the long questionnaire and 1246 the short questionnaire. Finally, 831 cases of European descent had both genetic data and environmental data from a questionnaire. This subset constitutes the dataset on which we apply DAC.

Genetic risk and age at diagnosis. A potential source of dependence between genetic risk and environmental factors is age at diagnosis. For T1D, the MHC region, the region that affects genetic risk the most, is also associated with age at diagnosis [Howson et al., 2012, Caillat-Zucman et al., 1992]. Of course, age at diagnosis has a strong impact on the experiences that a child has had before diagnosis and therefore the environment of cases as measured by a questionnaire. Consequently, we assessed association between genetic risk and age at diagnosis using linear regression on the 1491 Isis-Diab cases of European descent for whom genetic risk was available.

Regression of age on genetic risk yielded a negative association. In average, cases with a genetic risk increased by one standard deviation were younger at diagnosis by 3.5 months (CI= $[-5.7, -1.3]$, $p = 2 \times 10^{-3}$). This motivates the control for age at diagnosis in the main analysis.

Main analysis. Our main analysis is testing for association between environmental factors and genetic risk. This association was assessed while controlling for age at diagnosis. We used a generalized additive model (GAM) in order for the dependence between environmental factor and age to be captured by a smooth function. The environmental factor was regressed on the genetic risk and a smooth function of age. Association with genetic risk was tested using the standard Student test provided by the fitted GAM. The tests were one sided as explained above. In our case, the associations with disease are negative and therefore the association between genetic risk and environmental factor is expected to be positive: the most at risk genetically

Variable	Missing data	DAC p-value	Effect size in simulation	Power
Sugar baby	4%	0.09	0.59	7%
Dentist	4%	0.80	0.37	10%
Dental hygiene	2%	0.032	0.39	8%
Diarrhea	6%	0.66	0.56	7%
Cocoa spread	0.4%	0.77	0.33	7%
Ski	1%	0.22	0.49	8%
Club	1%	0.60	0.49	8%

Table 5.4: **Results of the DAC method for confirmation of 7 variables from the Isis-Diab case-control study.** The effect size used in simulations is the farthest from 1 in the two analysis (matched and propensity) on the long questionnaire presented in chapter 4.

should be more exposed to protective factors. The mgcv package was used for GAM [Wood, 2012].

Results are summarized in table 5.4. Dental hygiene is nominally associated with genetic risk in the expected direction in Isis-Diab cases. However, this does not take into account correction for multiple testing. Other variables do not show association with genetic risk.

In the next subsection we adapt the simulation framework of the previous section to estimate the power of DAC. The estimation of power for DAC in the Isis-Diab data showed that DAC has power under 10% for every variable. Given the low power of DAC herein, the nominally significant result for dental hygiene is almost as unlikely under the alternative than under the null. This low power estimate shows that DAC is not informative in the Isis-Diab data.

5.3.2 Power estimation for DAC

As we are under the same assumptions as in section 5.2, simulations to estimate the power of DAC will follow the same framework. However, instead of computing the CPR for each simulation, we perform a regression of the environmental factor on the genetic risk in the cases and obtain our quantity of interest: a one-sided p-value. Our estimator of power is then the proportion of p-values under the threshold 0.05.

A second difference is that we have replaced the binary G by a genetic risk estimation. The distribution of genetic risk in the general population is a mixture of the distribution of genetic risk in the controls and in the cases. If we denote $\mathcal{L}(X)$ the distribution of X, we have that:

$$\mathcal{L}(R_g) = (1-K)\mathcal{L}(R_g^{\text{controls}}) + K\mathcal{L}(R_g^{\text{cases}}).$$

In practice, we use the distributions of genetic risk obtained in section 2.4. This genetic risk estimation has an AUC of 0.86 and has been calibrated in order to represent probabilities. We sample N genetic risks from the genetic risks of Isis-Diab cases and we sample the rest from the genetic risks of the controls (cases of non-auto immune diseases in the WTCCC1 data).

Power estimation for Isis-Diab. We apply this simulation framework to estimate power of DAC on the precise setting of the Isis-Diab data. The prevalence K is set to 0.2%, a reasonable

estimate of the prevalence of T1D in France. For each variable, the sample size is set to the sample size available after excluding missing data.

To define the environmental factor's distribution for the Isis-Diab data, we need to choose an effect size for each environmental factor. In order to do this, we take into account the results of the case-control study presented in chapter 4. Two analyses were performed for the long questionnaire on the case-control data. Between the two resulting effect size estimate for each variable, we chose the most extreme, i.e., the most favorable scenario for power. To have a power estimate as precise as possible, we need to obtain in the simulated case population the observed distribution of the environmental factor in cases. As before, to achieve this, we adjust the distribution in the source population to obatin in simulated cases the observed distribution in cases. The results are shown in table 5.4 alongside the results of the analysis.

Power estimation in generic scenarii. We also consider more general scenarii, to evaluate the potential of DAC in other situations. We evaluate the influence of prevalence and also of prediction accuracy of genetic risk. To do this for prevalence, genetic risk is left untouched and we set prevalence to 0.2%, 0.6% or 1% and sample size to 500, 1500, 3000 or 5000. The three prevalences correspond to the prevalence of T1D in France for the lowest, T1D in Finland for the intermediate value and high estimation of prevalence of celiac disease for the highest [Gujral et al., 2012].

Concerning the influence of prediction accuracy of the genetic risk, we set prevalence at 0.2%, we modified the genetic risk estimate to have an AUC of 0.88, 0.90 or 0.92 and we set the sample size to 500, 1500, 3000 or 5000. The genetic risk distribution with modified AUC is obtained by adding to the risk of cases a constant chosen to obtain the desired AUC. The estimate of risk in cases and controls is then calibrated again to correspond to probabilities. Concerning the definition of the environmental factor, we choose an odd-ratio of 3 which is a large but plausible effect size for epidemiology and we choose the most favorable distribution of the environmental factor in the cases, i.e., the one with the most variance: an evenly split binary variable. The distribution in the source population is the desired one in cases weighted by the inverse of the relative risk.

The results of the power estimation in generic scenarii are presented in figure 5.2. Power increases with sample size, prediction accuracy of the genetic risk and prevalence of the disease. With a prevalence of 0.2% and an AUC of 0.86, power is very limited. Even if our sample size had been 5000 cases and despite the favorable assumptions made on the effect size and the distribution of the environmental factor in cases, power would be only 26%. Given this low power, the results on the Isis-Diab data are not informative. Our estimation show that power depends strongly on prevalence of the disease. For a disease with prevalence of 1%, 80% power is attained for a sample size under 3000.

5.4 Discussion

CODGEI has been proposed in 1994 to uncover gene-environment interaction using case-only data [Piegorsch et al., 1994]. Here, we have proposed a simulation framework to quantify the sensitivity of this design to the rare disease assumption. We also propose DAC as a new method-

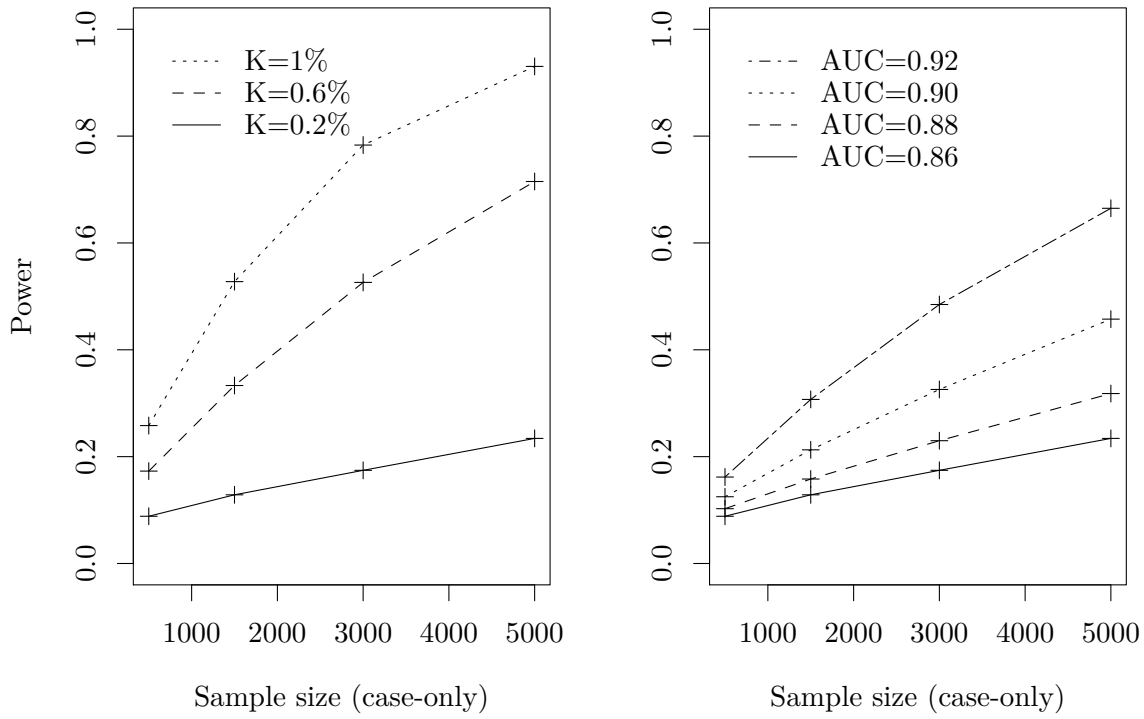

Figure 5.2: **Power of the DAC methodology in different settings.** The environmental factor's odd-ratio is set at 3 and the environmental factor in cases is evenly split. Left panel: Influence of the prevalence of the disease on power. The AUC of the genetic risk estimator remains at 0.86. Right panel: Influence of the genetic risk accuracy (AUC) on power. The prevalence of the disease remains at 0.2%

ology for analysis of case-only data; it allows to discover or confirm associations of environmental factors with disease using genetic risk estimation. Ideally, DAC should be used after a standard environmental case-control study to confirm findings. Indeed, it has modest power but it brings additional evidence to an environmental association with disease that is not liable to the same biases as the case-control study, e.g, the choice of controls.

As important information is missing in the case-only setting, assumptions need to be made to be able to draw conclusions from case-only data. Both DAC and CODGEI rely on an assumption of independence between gene and environment. This is reasonable but deviation from independence should be kept in mind as an alternative explanation for a positive result. While certain genes affect certain exposures such as alcohol consumption [Adkins et al., 2015], coffee consumption [Cornelis et al., 2015] or smoking [Furberg et al., 2010], there is a priori for independence between most genes and most environmental factors. We stress that the only independence needed for DAC is between the aggregated genetic risk score and the environmental factor: DAC does not require independence between each SNP and the environmental factor. When the environmental factor has genetic determinants and case-control data is available, Mendelian randomization [Davey Smith and Hemani, 2014] will be more informative as it allows to substantiate causal claims.

Under the assumption of independence, two phenomenons are present in the case-only gene-environment data: interactions and collider bias. The presence of collider bias depends on the distance to the rare disease assumption. Collider bias can be seen as the gap between two measures of interaction: OR_I and RR_I. This gap is modest in most situations: for example, interaction between BRCA1/2 and oral contraceptives in ovarian cancer [Modan et al., 2001] are only marginally affected by collider bias despite the high penetrance of BRCA1/2 (results not shown). However, associations between risk factors and BRCA1/2 in breast cancer in the GEM Study [Moorman et al., 2010] are not significant under the null hypothesis $H_0 : OR_I = 1$. When the disease is common and there is highly penetrant variants, CODGEI should be applied with caution and collider bias should be considered as an alternative explanation for a significant negative association between gene and environment among cases. If the prevalence of the disease and the main effects of both genetic variant and environmental factor are known, the simulation framework that has been described here allows to test the presence of interaction on the logit scale.

If there is no interaction between genetic risk and environment, only collider bias is left and DAC can be applied. This means that DAC is dependent on an assumption of absence of interaction. Indeed, interactions between genetic risk and environmental factor are problematic for DAC. A negative interaction strengthens the negative association that DAC tries to uncover but makes the findings less actionable as the people at highest genetic risk would respond less to intervention on the environmental factor (if the factor is a cause and not a mere marker). A positive interaction cancels the negative association that DAC tries to uncover despite increasing the prevention potential of the factor. This is a notable caveat to DAC as interactions between an aggregated genetic risk and environmental factors have been detected in relation to obesity [Tyrrell et al., 2017] and must be present in other settings as well.

As noted by Schmidt and Schaid [1999], under the rare disease assumption, collider bias is negligible. Their theoretical argument for absence of collider bias at low prevalences is in accordance with the results of power estimation. These power estimations show that DAC

can be successful in higher prevalence situations, with large sample sizes and better genetic risk estimation. However, in more common diseases, genetic risk estimation typically obtains sensibly weaker results and the prospective cohort design is more feasible. Nevertheless, DAC needs stronger prevalences of the disease to achieve reasonable power. This could be obtained in countries where T1D has a high prevalence such as Finland or on more frequent diseases such as celiac disease. DAC also underscores the importance for epidemiology of having a genetic risk estimation as predictive as possible.

Given the prevalence of T1D in France, DAC is underpowered in the setting of the Isis-Diab study. Nevertheless, the application of our method to these data illustrates the practical considerations that go into applying DAC such as the problem of confounding by age at diagnosis. Furthermore, it allowed to base our power estimations on an actual predicted genetic risk distribution.

Chapter 6

Personalized Treatment Strategies

Abstract Personalizing treatment according to patient's characteristics is at the core of stratified or precision medicine. There has been a recent surge of statistical methods aiming at identifying so-called optimal treatment strategies, i.e., strategies that assign a given treatment to a patient according to his/her characteristics. However, when data from a randomized controlled trial are used to estimate the optimal treatment strategy, it is not straightforward to estimate and test the benefit of the estimated strategy as compared to not personalizing treatment. In this context, we propose a principled approach for the estimation of the benefit of an estimated treatment strategy, accounting for its uncertainty. This leads to formalizing a strategy that we term the max lower bound strategy. Numerical simulations are used to show it allows proper type I error rate control and coverage probabilities. The approach is extended to multiple covariates using machine learning techniques. It is then applied to the data of a large randomized trial in acute ischemic stroke. We show that, although aspirin is beneficial on average, there is a small proportion of patients (6.8%) for which it could be detrimental.

Contents

6.1 Introduction

Personalized—or stratified, or precision—medicine consists in differentially treating patients based on their individual characteristics. If adapting the treatment to the patient is not a new idea in itself, it has attracted wide-ranging attention since the 2010s, in particular thanks to progresses in understanding of genetics and molecular biology. Personalized medicine has been considered to have the potential to radically advance patient care by improving prevention and treatment efficacy while avoiding side effects [Hamburg and Collins, 2010, US Food and Drug Administration, 2013]. If this concept has successfully transformed the treatment of certain diseases, such as chronic myeloid leukemia, where imatinib has drastically changed the outcome of patients with BCR-ABL mutation [Capdeville et al., 2002], or metastatic melanoma, with vemurafenib for patients harboring a BRAF V600E mutation [Chapman et al., 2011], it has also been suggested that the hopes of personalized medicine were not matched by evidence [Khoury, 2010, Wang et al., 2014, Vivot et al., 2015]. Indeed, limitations have been noted in how personalized medicine is developed. Of particular concerns are the quasi-exclusive reliance on genetic alterations to define biomarkers ignoring other potentially important characteristics [Ziegelstein, 2015], the emphasis on prognostic rather than predictive markers [Mandrekar and Sargent, 2009, Simon et al., 2009], and the widespread use of trials with an enrichment design [Freidlin and Korn, 2014, Vivot et al., 2016]. In fact, enrichment trials do not allow to assess the medical yield of a biomarker-based strategy as compared to a traditional treatment strategy, and make it virtually impossible to identify combinations of biomarkers to efficiently guide treatment selection.

To improve the efficacy of personalized medicine and circumvent aforementioned limitations, reliable approaches to classify patients who respond to a given treatment better than to another one are therefore needed. Several statistical methods have been developed to derive combinations of markers predicting improved response to treatment using data from randomized clinical trials (RCT, Cai et al., 2011, Foster et al., 2011, Lipkovich et al., 2011, Zhao et al., 2012, 2013, Kang et al., 2014, Zhou et al., 2017), as well as observational studies [Qian and Murphy, 2011, Zhang et al., 2012, Zhao et al., 2015, Shen and Cai, 2016, Shen et al., 2017, Künzel et al., 2017].

Let us consider a potential outcome framework, where it is assumed that each patient is associated with a vector (Y^0, Y^1) representing the outcome that would be observed under each treatment option so that the outcome is $Y = Y^0\mathbb{1}_{T=0} + Y^1\mathbb{1}_{T=1}$. Assuming that higher values of Y are beneficial, it would be natural to give treatment 1 to patients with $Y^1 \geq Y^0$ and treatment 0 to those with $Y^1 < Y^0$. Since both are never observed together, and cannot be known before administering the treatment, the approaches cited above mostly attempt to relate Y^0 and Y^1 to a set of covariates X representing the patient's characteristics. If we let $\Delta(X) = \mathbb{E}(Y^1|X) - \mathbb{E}(Y^0|X)$ and if this quantity is known, then giving treatment 1 to

individuals with $\Delta(X) \geq 0$ and treatment 0 to individuals with $\Delta(X) < 0$ yields an optimal treatment strategy in that it maximizes the expectation of the outcome over the population [Zhang et al., 2012].

A treatment strategy—also termed treatment regime [e.g., Zhang et al., 2012], individualized treatment rule [e.g., Shen and Cai, 2016], or policy [e.g., Kang et al., 2014]—consists in formalizing a rule determining which treatment a patient should receive according to his/her covariates. To define the benefit of using a treatment strategy, i.e., of personalizing treatment, one needs to compare the expectation of the outcome in the population under this treatment strategy to what would be obtained under the usual or reference treatment strategy [Janes et al., 2014]. What should be the reference strategy is a complex issue, but we will simply consider here that there exists a treatment that is at some point viewed as the best treatment option for a given disease. Let us assume that one RCT compares a new treatment to this "old" one. If the new treatment is significantly superior to the old one in terms of average outcome, then the reference treatment strategy would be to now recommend treating all patients with the new treatment. In contrast, if the new treatment is not significantly better, then the reference treatment strategy is to treat all patients with the old treatment. In order to deal with both situations, we will simply refer throughout the manuscript to the reference treatment ($T = 1$), and the other treatment will be called the alternative treatment ($T = 0$). Later on, we will call the subset of patients for whom a treatment strategy recommends the alternative rather than the reference treatment the "personalized set".

In their comprehensive work on how to evaluate the performance of personalized treatment strategies, Janes et al. [2014] have proposed a plug-in estimator of the average gain under the optimal treatment strategy, and used bootstrap to obtain the corresponding percentile confidence intervals. They show that their estimators have good properties when an improvement is present (that is, there are patients who have better outcome under the alternative treatment than under the reference one) but warn the reader not to use their estimators to test for the presence of an improvement. While testing for presence of an improvement is not the focus of their work, they suggest to use a composite test on the linear regression coefficients. In a previous work, Shuster and van Eys [1983] proposed to divide the range of X in regions of superiority of one treatment over the other, and a region of uncertainty where there is no significant difference between the treatment effects. This naturally allows to test for the benefit of personalizing treatment based on the covariates X.

The starting point of this chapter is to note that, in practice, the optimal treatment strategy is not known, and therefore any actual personalization will be dependent on an estimated treatment strategy. Thus, our main objective is to propose a principled way to estimate the benefit of personalization of an estimated strategy. This allows us to study the choice of strategy and, accordingly, to design a new strategy with maximal guarantee on its gain. We also provide a statistical test for presence of improvement under the estimated policy and show its good properties.

In Section 6.2, we adopt a Bayesian framework to estimate the quantities of interest and their credible quantiles. We start Section 6.3 by illustrating the problems posed by the naïve plug-in estimator of the optimal treatment strategy. The study of the links between statistical guarantees at the individual and the aggregated levels allows us to consider more general forms of policies. We then propose a new treatment strategy, which we call the max lower bound

strategy, dealing with the highlighted problems. We study in Section 6.4 the test naturally associated with our strategy, and show that it defines a personalized set whenever presence of improvement under the optimal treatment strategy is detectable. We verify in Section 6.5 that estimation under the max lower bound strategy yields nominal coverage probability. Our approach is extended in Section 6.6 to other types of outcomes (binary and censored) and to the multivariate case, using machine learning techniques. Finally, we illustrate the procedure in Section 6.7 by showing that aspirin is detrimental to some patients after ischemic stroke, using data from a large RCT.

6.2 Strategy-aware estimation

It is assumed throughout that each triplet patient/response/treatment is modeled by a random vector (X, Y, T), where X is a vector of covariates taking values in $\mathcal{X} \subset \mathbb{R}^d$ (the patient's characteristics), T is the treatment (alternative = 0, reference = 1), and $Y = Y^0\mathbb{1}_{T=0} + Y^1\mathbb{1}_{T=1}$ is the patient's outcome, depending on the treatment. Of course, the variables Y^0 and Y^1 are never observed simultaneously—for a given value $t \in \{0, 1\}$ of T, we only have access to Y^t.

In essence, a treatment strategy is the choice of the treatment (either reference or alternative) that a particular patient will receive, given his/her personal characteristics. A treatment strategy can therefore be represented by a function $\texttt{pol} : \mathcal{X} \to \{0, 1\}$, which assigns a possible treatment 0 or 1 to each possible realization x of X. For instance, the strategy consisting in giving the reference treatment to everyone could be termed $\texttt{ref} : x \mapsto 1$. In the context of personalized medicine, we are looking for a more complex strategy, which takes into account each patient's characteristics. In particular, the optimal treatment strategy introduced earlier can therefore be written $\texttt{opt}(x) = \mathbb{1}_{\Delta(x)\geq 0}$, where $\Delta(x) = \mathbb{E}(Y^1|X = x) - \mathbb{E}(Y^0|X = x)$. A general strategy (or policy), say $\texttt{pol}$, can be characterized by its average outcome, $\mathbb{E}Y^{\texttt{pol}}$, as well as the gain in average outcome as compared to using the reference strategy $\texttt{ref}$, that is:

$$\Theta(\texttt{pol}) \stackrel{\text{def}}{=} \mathbb{E}Y^{\texttt{pol}} - \mathbb{E}Y^1 = \mathbb{E}_X[-\Delta(X)\mathbb{1}_{\texttt{pol}(X)=0}],$$

where $\mathbb{E}_X$ is the expectation under $\mathbb{P}_X$, the probability distribution of X.

Taking the optimal strategy $\texttt{opt}$ gives

$$\Theta(\texttt{opt}) = \mathbb{E}_X[-\Delta(X)\mathbb{1}_{\Delta(X)<0}], \tag{6.1}$$

the quantity used in Janes et al. [2014], who have further proposed additional measures for the benefit of personalization, such as the proportion of marker-negative (or positive) patients:

$$P_{\text{neg}} = \mathbb{P}(\Delta(X) < 0),$$

and the average benefits of no treatment among marker-negatives:

$$B_{\text{neg}} = \mathbb{E}[-\Delta(X)|\Delta(X) < 0],$$

so that $\Theta(\texttt{opt}) = P_{\text{neg}}B_{\text{neg}}$. We note in passing that $\Theta(\texttt{opt}) \geq 0$, and that $P_{\text{neg}} > 0$ is a necessary condition to have $\Theta(\texttt{opt}) > 0$. Of course, for a general strategy $\texttt{pol}$, $\Theta(\texttt{pol})$ can have an arbitrary sign.

In practice however, the distribution of (X, Y, T) is unknown, and so is the optimal strategy `opt`. Fortunately, we have access to an i.i.d. sample corresponding to a RCT $\mathcal{D}_n = (X_i, Y_i, T_i)$, $1 \leq i \leq n$, where each triplet (X_i, Y_i, T_i) is distributed as the generic (X, Y, T). So, for each i, X_i represents the patient's characteristics that we want to use to personalize treatment, Y_i is the observed continuous outcome, and $T_i \in \{0, 1\}$ is the treatment allocated in the trial. We consider the potential outcome framework introduced earlier, i.e., $Y_i = Y_i^0$ if $T_i = 0$ and $Y_i = Y_i^1$ if $T_i = 1$, and assume throughout that X_i is independent of T_i (randomized trial).

In this section, and until Subsection 6.6.2, we make the assumption that X is a bounded scalar (that is, $d = 1$, with $\mathcal{X} = \text{support}(\mathbb{P}_X) = [x_0, x_1]$, $-\infty < x_0 < x_1 < \infty$), and that data arise from a linear model. The linear model assumes that Y^t, $t = 0, 1$, can be expressed as a sum of an intercept β_0, a prognostic term with coefficient β_1, an average treatment effect term with coefficient β_2, an interaction term between treatment and the covariate X with coefficient β_3, and some independent Gaussian noise ε with variance $\sigma^2 > 0$:

$$Y^t = \beta_0 + \beta_1 X + \beta_2 t + \beta_3 X t + \varepsilon, \quad t = 0, 1. \tag{6.2}$$

We will suppose, without loss of generality, that X is centered ($\mathbb{E}_X X = 0$), so that β_2 represents the average treatment effect, $\mathbb{E}Y^1 - \mathbb{E}Y^0$. Note that this implies that $0 \in [x_0, x_1]$. It is also assumed that the variance σ^2 is known for simplicity. If we do not make this assumption, we have to use the t-distribution instead of the Gaussian distribution, which adds complication for no additional insights.

For each patient i with covariate X_i, the difference in outcome between the reference treatment and the alternative treatment is simply

$$\Delta(X_i) = \mathbb{E}[Y_i^1 | X_i] - \mathbb{E}[Y_i^0 | X_i] = \beta_2 + \beta_3 X_i.$$

If the coefficients were known exactly, then we would choose to give the reference treatment 1 to all patients with $\Delta(X_i) \geq 0$ (favoring the treatment with the best outcome on average when $\Delta(X_i) = 0$ but the other choice would be equivalent) and give the alternative treatment 0 to patients with $\Delta(X_i) < 0$. The average gain between this perfect optimal treatment strategy and the reference strategy is then the one given in equation (6.1). We note that if $\beta_3 = 0$, i.e., if there is no interaction, then $\Delta(X_i)$ is constant, equal to the average treatment effect β_2. We also see that if $\beta_2 < 0$, then there are always some patients who would benefit from personalization as $\Delta(0) = \beta_2 < 0$ and $0 \in [x_0, x_1]$. On the other hand, when $\beta_2 \geq 0$, the presence of an interaction, i.e., $\beta_3 \neq 0$, is not sufficient for the existence of patients benefiting the alternative treatment. Indeed, if $\beta_2 \geq 0$, the existence of $x \in [x_0, x_1]$ such that $\Delta(x) < 0$ is equivalent to $-\beta_2/\beta_3 > x_0$ when $\beta_3 > 0$, and to $-\beta_2/\beta_3 < x_1$ when $\beta_3 < 0$.

In practice, we do not have access to the true values of β_2 and β_3, and have to rely on some estimation procedure. This implies that there will be uncertainty in our estimation, and therefore uncertainty in the associated strategy. An important contribution of this chapter is to estimate the average gain under the estimated strategy and <u>not</u> the perfect, but unknown, strategy `opt`.

Let us denote by $\hat{\beta} \stackrel{\text{def}}{=} (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ the standard least square estimators of $\beta \stackrel{\text{def}}{=} (\beta_1, \beta_2, \beta_3, \beta_4)$. We let $\mathbb{P}_\beta$ be the sampling probability of ϵ or, in other words, the sampling probability of Y

given X and T. Under $\mathbb{P}_\beta$, we have $\hat{\beta} \sim \mathcal{N}(\beta, \Sigma)$, where $\Sigma = \sigma^2(Z^\top Z)^{-1}$ and Z is the design matrix

$$Z = \begin{bmatrix} 1 & X_1 & T_1 & X_1T_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n & T_n & X_nT_n \end{bmatrix}.$$

While $\hat{\beta}_2$ is the estimated effect of the reference treatment versus the alternative treatment, this does not warrant that $\hat{\beta}_2 \geq 0$. As we discussed in the introduction, we could have $\hat{\beta}_2 \leq 0$ if the new treatment improved outcomes compared to the old treatment but not significantly. The reference treatment is then set to the old treatment, and the new treatment is the alternative one.

The presence of an interaction is a necessary condition for improvement due to personalization, and the associated test will be ubiquitous throughout. In order to make the discussion more transparent, we formally define this test below. Notation $\Sigma_{i,j}$ means the element in the i, j position of the covariance matrix Σ, and q_α is the α-quantile of the standard normal distribution.

Definition 1 (Interaction test)**.** *To test against the null hypothesis $H_0^{\text{interact}} : \beta_3 = 0$, the test statistic is $\hat{\beta}_3/\sqrt{\Sigma_{3,3}}$. Its distribution is $\mathcal{N}(0,1)$ under $\mathbb{P}_\beta$. When the alternative hypothesis is $H_1^{\text{interact}} : \beta_3 \neq 0$, the rejection region is $\{|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}} > q_{1-\alpha/2}\}$. If the alternative hypothesis is $H_1^{\text{interact}} : \beta_3 > 0$ (respectively, $H_1^{\text{interact}} : \beta_3 < 0$), the rejection region is $\{\hat{\beta}_3/\sqrt{\Sigma_{3,3}} > q_{1-\alpha}\}$ (respectively, $\{\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha\}$).*

Now that we have discussed the interpretation of the coefficients, we can turn to the problem at hand. The natural estimator of $\Delta(x)$ is simply $\hat{\Delta}(x) = \hat{\beta}_2 + \hat{\beta}_3 x$, and an example of estimated strategy is the plug-in estimator of the optimal strategy, say $\hat{\texttt{pol}}_1$, sometimes called the optimal treatment regime (OTR, Brinkley et al., 2010, Zhang et al., 2012, Janes et al., 2014). It is defined by $\hat{\texttt{pol}}_1(x) = \mathbb{1}_{\hat{\Delta}(x) \geq 0}$, and recommends to each patient the treatment maximizing the predicted outcome given the patient's characteristics X. For a general estimator $\hat{\texttt{pol}}(x)$, eventually different from $\hat{\texttt{pol}}_1(x)$, the associated improvement in population averaged outcome is

$$\Theta(\hat{\texttt{pol}}) \stackrel{\text{def}}{=} \mathbb{E}_X[-\Delta(X)\mathbb{1}_{\hat{\texttt{pol}}(X)=0}],$$

which is unknown and has to be estimated. More importantly, we should also provide a lower confidence bound for this quantity at a predefined level. We will assume that $\mathbb{P}_X$ is known throughout our theoretical discussions, but in practice, we use $\mathbb{P}_n$, the empirical distribution of $X_1, \ldots, X_n$. The distribution $\mathbb{P}_X$ used in the estimation of $\Theta(\hat{\texttt{pol}})$ could be different from the one generating the sample if, for example, we want to estimate the average improvement in a population with a different distribution than the one of the RCT.

We insist that $\Theta(\hat{\texttt{pol}})$ is a random variable, because it depends upon $\hat{\texttt{pol}}$. This dependency complicates the study of $\Theta(\hat{\texttt{pol}})$ under $\mathbb{P}_\beta$. To circumvent this complication, we have chosen to adopt a Bayesian estimation point of view and, conditionally on the sample $\mathcal{D}_n$, put a distribution on the regression parameter β. This allows to separate the uncertainty on β (random under the posterior distribution) from the uncertainty on the strategy (fixed under the posterior distribution). We will however be interested in the frequentist properties of our estimation, and therefore the use of a prior would be counter-productive. In other words, we use a constant

prior on $\mathbb{R}^4$ for β. Therefore, the posterior distribution of β given the sample $\mathcal{D}_n$ is simply $\Pi \overset{\text{def}}{=} \mathcal{N}(\hat{\beta}, \Sigma)$, where Σ is the covariance matrix defined above.

To distinguish the true fixed β, we denote by $\beta^\star \overset{\text{def}}{=} (\beta_1^\star, \beta_2^\star, \beta_3^\star, \beta_4^\star)$ a random variable whose distribution is Π. Similarly, we add a $\star$ superscript on the quantities depending on $\beta^\star$ instead of β. Thus, we set

$$\Theta^\star(\hat{\texttt{pol}}) = \mathbb{E}_X[-\Delta^\star(X)\mathbb{1}_{\hat{\texttt{pol}}(X)=0}],$$

where $\Delta^\star(X) = \beta_2^\star + \beta_3^\star X$. This quantity is easy to study, as shown in the next proposition. We let

$$\hat{\Theta}(\hat{\texttt{pol}}) = \mathbb{E}_X[-\hat{\Delta}(X)\mathbb{1}_{\hat{\texttt{pol}}(X)=0}].$$

Proposition 1. *Under* Π, $\Theta^\star(\hat{\texttt{pol}})$ *follows a Gaussian distribution with* $\mathbb{E}_\Pi[\Theta^\star(\hat{\texttt{pol}})] = \hat{\Theta}(\hat{\texttt{pol}})$ *and*

$$\mathrm{Var}_\Pi[\Theta^\star(\hat{\texttt{pol}})] = \mathbb{E}_\Pi\big[\mathbb{E}_X^2[(\Delta^\star(X) - \hat{\Delta}(X))\mathbb{1}_{\hat{\texttt{pol}}(X)=0}]\big].$$

Proof. We have

$$\Theta^\star(\hat{\texttt{pol}}) = \mathbb{E}_X[\Delta^\star(X)\mathbb{1}_{\hat{\texttt{pol}}(X)=0}] = \beta_2^\star \mathbb{P}_X(\hat{\texttt{pol}}(X)=0) + \beta_3^\star \mathbb{E}_X[X\mathbb{1}_{\hat{\texttt{pol}}(X)=0}],$$

which is a linear combination of a Gaussian vector and is therefore Gaussian. Using Fubini's theorem, we may write

$$\mathbb{E}_\Pi[\Theta^\star(\hat{\texttt{pol}})] = \mathbb{E}_X\big[\mathbb{E}_\Pi[-\Delta^\star(X)]\mathbb{1}_{\hat{\texttt{pol}}(X)=0}\big] = \mathbb{E}_X[-\hat{\Delta}(X)\mathbb{1}_{\hat{\texttt{pol}}(X)=0}] = \hat{\Theta}(\hat{\texttt{pol}}).$$

Besides,

$$\mathrm{Var}_\Pi[\Theta^\star(\hat{\texttt{pol}})] = \mathbb{E}_\Pi\big[(\Theta^\star(\hat{\texttt{pol}}) - \hat{\Theta}(\hat{\texttt{pol}}))^2\big] = \mathbb{E}_\Pi\big[\mathbb{E}_X^2[(\Delta^\star(X) - \hat{\Delta}(X))\mathbb{1}_{\hat{\texttt{pol}}(X)=0}]\big].$$

□

The credible α-quantile of $\Theta(\hat{\texttt{pol}})$—that is, the α-quantile of the posterior distribution—is then

$$\hat{q}_{n,\alpha}(\hat{\texttt{pol}}) = \hat{\Theta}(\hat{\texttt{pol}}) + q_\alpha \mathrm{sd}_\Pi(\Theta^\star(\hat{\texttt{pol}})), \tag{6.3}$$

where sd denotes the standard deviation. This quantity will be studied thoroughly in the next section, and we will see later that these credible quantiles are in fact confidence bounds with valid frequentist coverage for $\Theta(\hat{\texttt{pol}})$ under $\mathbb{P}_\beta$. In practice, the use of formula (6.3) requires to compute the standard deviation. It is however simpler to sample $\beta^\star$ from Π J times and compute $\Theta^\star(\hat{\texttt{pol}})$ each time, which gives us an empirical distribution $\{\Theta_j^\star(\hat{\texttt{pol}})\}_{1\le j\le J}$, from which we can then retrieve the quantiles of the posterior distribution of $\Theta(\hat{\texttt{pol}})$.

In addition to estimating the benefit of personalization, it is essential in clinical studies to be able to test the null hypothesis of no gain from personalization, that is, $H_0 : \Theta(\hat{\texttt{pol}}) \le 0$ against the alternative hypothesis of gain from personalization $H_1 : \Theta(\hat{\texttt{pol}}) > 0$. This can naturally be done using our credible α-quantile. Indeed, the natural rejection region of the null hypothesis at level α is simply $\{\hat{q}_{n,\alpha}(\hat{\texttt{pol}}) > 0\}$, as this means that the posterior distribution of $\Theta(\hat{\texttt{pol}})$ has more than $1-\alpha$ of its weight on the positive line. It is important to note that this test is not a proper test because our null hypothesis depends on a random variable instead of a fixed

quantity. This means that even when there is an improvement to be found, i.e., $\Theta(\texttt{opt}) > 0$, the null hypothesis $H_0 : \Theta(\hat{\texttt{pol}}) \leq 0$ can occur, either because of poor estimation of the strategy or because the estimated strategy defines an empty personalized set. This issue will depend on the choice of the strategy, as illustrated in the next section. However, while this is not a proper test, it is precisely the decision we care about. In fact, we will show in Section 6.4 that this test needs to be combined with the test for interaction presented in Definition 1. Furthermore, we will see that for the right choice of strategy, the resulting joint test is a valid test for presence of improvement.

We close this section by emphasizing that we have three probability measures of interest: $\mathbb{P}_X$ (the distribution of X under which we want to compute the expected gains), $\mathbb{P}_\beta$ (the frequentist probability of ϵ) and Π (the posterior probability of β given $\mathcal{D}_n$). The subscript β in the second one underlines the fact that β is a constant under this probability, whereas it is random under Π. To avoid confusion, we introduced a $\star$ superscript for the random variables under Π.

6.3 Choice of a treatment strategy

6.3.1 Problems of estimation under $\hat{\texttt{pol}}_1$

Our first example of strategy was the plug-in estimator of the optimal treatment strategy, that is $\hat{\texttt{pol}}_1(x) = \mathbb{1}_{\hat{\Delta}(x)\geq 0}$. We show here that, despite its simplicity, this policy behaves poorly for estimating the benefit of personalization. In particular, it can be overall detrimental even tough our estimation predicts that it will be beneficial, i.e., $\mathbb{P}_\beta(\Theta(\hat{\texttt{pol}}_1) < 0 | \hat{q}_{n,\alpha}(\hat{\texttt{pol}}_1) > 0)$ is large.

We illustrate this using simulations based on a linear model of the form (6.2), whose specifics are described in the legend of Figure 6.1. Briefly, under our simulation settings, 10% of patients should benefit more from the alternative than from the reference treatment (i.e., $P_{\text{neg}} = 0.1$), and under the optimal treatment strategy, the average outcome is improved by 0.013 (i.e., $\Theta(\texttt{opt}) = 0.013$). We consider the properties of the quantile $\hat{q}_{n,0.05}(\hat{\texttt{pol}}_1)$, which can be used to provide a test, and the mean of the distribution, $\hat{\Theta}(\hat{\texttt{pol}}_1)$, which provides an estimation of the benefit of personalizing treatment. As the true parameters of the linear model are known, we can compute the true value of $\Theta(\hat{\texttt{pol}}_1)$.

The quantile $\hat{q}_{n,0.05}(\hat{\texttt{pol}}_1)$ is positive, i.e., we are detecting the presence of a benefit of personalization, in 4.7% of the simulations. This is suboptimal as with a different strategy detailed later on, we will be able to detect a benefit of personalization in 34% of cases.

More worrying yet is the behavior of the true parameter $\Theta(\hat{\texttt{pol}}_1)$ against its estimator $\hat{\Theta}(\hat{\texttt{pol}}_1)$. This is shown in the upper panel of Figure 6.1. It should be noted that most points are close to the optimal $(0.013, 0.013)$, where the strategy is well estimated and the estimated gain is close to its real value. However, in a certain number of simulations the strategy is poorly estimated, and as a consequence the real gain is negative while the estimated gain appears very large. These situations are problematic because a decision based on such estimations will negatively impact health while being claimed to have a large positive impact.

Manual inspection of the problematic points shows that this behavior happens when the ratio $\hat{\beta}_2/\hat{\beta}_3$ is underestimating β_2/β_3. As $x = -\hat{\beta}_2/\hat{\beta}_3$ is the value at which $\hat{\Delta}$ changes sign, this means that $\hat{\texttt{pol}}_1$ will attribute the alternative treatment to too many participants compared to the optimal treatment strategy, and therefore many participants who would have benefited

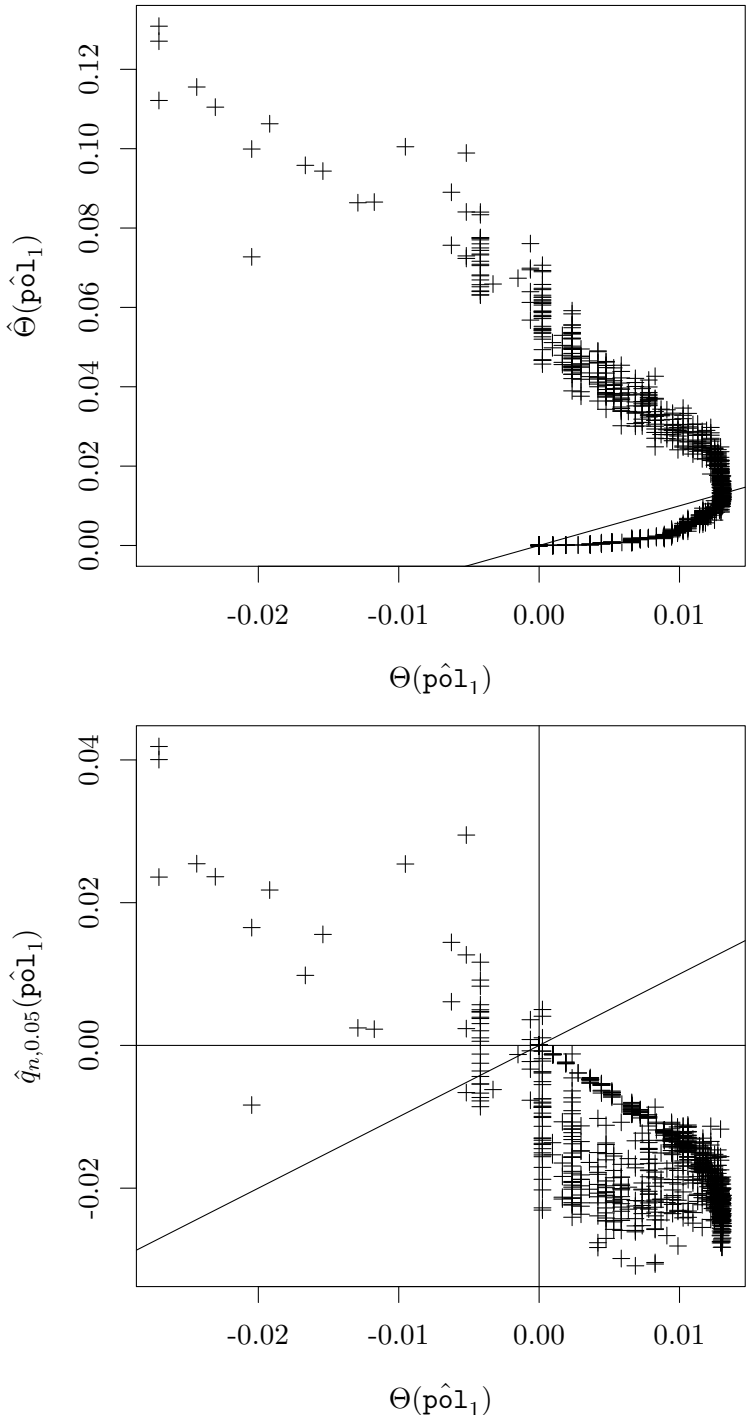

Figure 6.1: Graph of $\Theta(\hat{\mathtt{pol}}_1)$ against $\hat{\Theta}(\hat{\mathtt{pol}}_1)$ and $\hat{q}_{n,0.05}(\hat{\mathtt{pol}}_1)$. We simulated 10 000 datasets based on the linear model (6.2), i.e., we sampled ϵ for each simulation. The variable X is sampled once from a uniform distribution between -1 and 1. We take $n = 300$, $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1.3$, and $\sigma^2 = 1$. Under these simulation settings, optimal personalization affects 10% of the patients, $P_{\text{neg}} = 0.1$, and $\Theta(\mathtt{opt}) = 0.013$. We sampled from Π 10 000 times in order to compute the quantiles. To enhance readability, we have plotted only 1 000 points. The line $y = x$ is also plotted.

from the reference treatment will receive the alternative treatment. At the same time, since the posterior distribution we use is centered on poorly estimated coefficients, the estimation of benefit will be optimistic. The inverse situation can also be seen under the first bisector: when the ratio $\hat{\beta}_2/\hat{\beta}_3$ is overestimated, not enough participants will be attributed the alternative treatment, leading to a decreased population averaged outcome as compared to what could have been obtained by using the optimum.

As announced at the beginning of this subsection, the dramatic consequence of these problematic situations is even more poignant when we consider the lower quantile, i.e., the one we use to test the presence of an improvement. As we can see in the lower panel of Figure 6.1, $\hat{q}_{n,0.05}(\hat{\texttt{pol}}_1)$ seems to be positive mostly when $\Theta(\hat{\texttt{pol}}_1)$ is negative. Although $\hat{q}_{n,0.05}(\hat{\texttt{pol}}_1)$ is positive 4.7% of the time, in 90% of those cases $\Theta(\hat{\texttt{pol}}_1)$ is negative. This is the same phenomenon as above: an underestimated ratio leads to a large personalized set and at the same time overestimated bounds on the gain. While the quantile has the right coverage probability, the mistakes it makes are the one we care the most about: claiming a large improvement when personalizing treatment would be detrimental on average.

Fortunately, all these issues can be dealt with by considering a different strategy. The key idea is that if we can identify the patients that bring the most uncertainty to our estimation, then we can choose to exclude them from the personalization. This means that in case of uncertainty, we prefer to give the reference treatment. This asymmetry is analogous to the asymmetry created by defining a hypothesis as the null hypothesis in statistical test theory. This asymmetry is present in our quantity of interest $\Theta(\hat{\texttt{pol}})$ as no uncertainty comes from the patients receiving the reference treatment. Furthermore, this asymmetry is desirable, as a clinician implementing a personalization strategy will want to make sure that the change will be beneficial even if some patients who would have benefited from the alternative treatment are missed by this strategy.

In order to define a better strategy, we need to understand the uncertainty that patients bring to our aggregated quantities. To reach this goal, it is necessary to make a detour through an analysis of uncertainty at the individual level. This is the topic of the next subsection.

6.3.2 Uncertainty at the individual level

A patient with covariate $X = x$ has an estimated improvement of $\hat{\Delta}(x) = \hat{\beta}_2 + \hat{\beta}_3 x$. The distribution of improvement under Π is

$$\Delta^\star(x) = \beta_2^\star + \beta_3^\star x \sim \mathcal{N}\big(\hat{\Delta}(x), \mathrm{Var}_\Pi(\Delta^\star(x))\big),$$

where

$$\mathrm{Var}_\Pi(\Delta^\star(x)) = \Sigma_{2,2} + 2\Sigma_{2,3}x + \Sigma_{3,3}x^2.$$

The certainty with which such a patient will benefit from personalization is naturally measured by the quantity $\Pi(\Delta^\star(x) < 0)$. As $\Delta^\star(x)$ is Gaussian, this probability depends only on

$$z_\Delta(x) \stackrel{\text{def}}{=} \frac{\hat{\Delta}(x)}{\mathrm{sd}_\Pi(\Delta^\star(x))} = \frac{\hat{\beta}_2 + \hat{\beta}_3 x}{\sqrt{\Sigma_{2,2} + 2\Sigma_{2,3}x + \Sigma_{3,3}x^2}},$$

as we have $\Pi(\Delta^\star(x) < 0) = \Phi(-z_\Delta(x))$, with Φ the cumulative distribution function of the standard normal distribution.

In this context, we can derive a test for personalization at the individual level. Namely, we consider the null hypothesis $H_0 : \Delta(x) \geq 0$, i.e., the reference treatment is better than the alternative for a patient with covariate x. The alternative hypothesis is then $H_1 : \Delta(x) < 0$, and the rejection region of this test is $\{z_\Delta(x) < q_\alpha\}$. This rejection region is the region of superiority for the alternative treatment and was already defined in Shuster and van Eys [1983]. Observe that, for each individual, $\Phi(z_\Delta(x)) = \Pi(\Delta^\star(x) \geq 0)$ is the level at which we would reject the null.

This state of affairs suggests a new, natural, strategy, which personalizes treatment for patients in the region of superiority of the alternative treatment, i.e., patients such that $z_\Delta(x) < q_\alpha$. We call this strategy the individual strategy $\hat{\texttt{ind}}_\alpha$, and notice that it depends on α, the chosen confidence level. Thus, we have

$$\hat{\texttt{ind}}_\alpha : x \mapsto \mathbb{1}_{z_\Delta(x) \geq q_\alpha}. \tag{6.4}$$

The fundamental difference with the naïve plug-in $\hat{\texttt{pol}}_1$ strategy is that patients with $q_\alpha \leq z_\Delta(x) < 0$, who are predicted to benefit from the alternative treatment under the strategy $\hat{\texttt{pol}}_1$, would still receive the reference one under $\hat{\texttt{ind}}_\alpha$. The function $x \mapsto z_\Delta(x)$ is therefore sufficient to quantify the uncertainty with which a patient would benefit from the alternative treatment, and we need to study it in depth. It is important to keep in mind that we are mainly interested by the behavior of z_Δ when it is negative, as this corresponds to patients who are expected to benefit from personalization.

Note first that $z_\Delta(x) = 0 \Leftrightarrow x = -\hat{\beta}_2/\hat{\beta}_3$. The limits of z_Δ in $-\infty$ and $+\infty$ are, respectively, $\lim_{-\infty} z_\Delta = -\hat{\beta}_3/\sqrt{\Sigma_{3,3}}$ and $\lim_{+\infty} z_\Delta = \hat{\beta}_3/\sqrt{\Sigma_{3,3}}$. This is the z-statistic of the test for presence of an interaction seen in Definition 1. In addition,

$$\frac{dz_\Delta}{dx}(x) = \frac{Ax + B}{\mathrm{sd}_\Pi(\Delta^\star(x))^{3/2}},$$

where $A = -\hat{\beta}_2\Sigma_{3,3} + \hat{\beta}_3\Sigma_{2,3}$ and $B = -\Sigma_{2,3}\hat{\beta}_2 + \hat{\beta}_3\Sigma_{2,2}$.

The signs of A and B decide the shape of z_Δ. Since X and T are independent, $\Sigma_{2,3}$ will be approximately 0. If we set it equal to 0 in the previous expressions, we obtain $A = -\hat{\beta}_2\Sigma_{3,3}$ and $B = \hat{\beta}_3\Sigma_{2,2}$. For simplicity, we will assume that $\Sigma_{2,3} = 0$ throughout, as this allows to discuss more easily interpretable cases. For example, instead of $-\hat{\beta}_2\Sigma_{3,3} + \hat{\beta}_3\Sigma_{2,3} < 0$, we will have $\hat{\beta}_2 > 0$ and instead of $-\Sigma_{2,3}\hat{\beta}_2 + \hat{\beta}_3\Sigma_{2,2} > 0$, we will have $\hat{\beta}_3 > 0$. As $\Sigma_{2,3}$ is close to 0, the probability of the sign of A (respectively, of the sign of B) being different than minus the sign of $\hat{\beta}_2$ (respectively, than the sign of $\hat{\beta}_3$) is small. Note however that if we want to apply our work to observational data, where X and T are not independent, then the more complicated inequalities should be used.

Figure 6.2 shows how z_Δ typically varies, depending on the sign of $\hat{\beta}_2$ when $\hat{\beta}_3 > 0$. The sign of $\hat{\beta}_3$ does not affect the behavior: if we change its sign, the curve is simply the symmetric of the original curve with respect to the ordinate axis. If $\hat{\beta}_2 > 0$, then z_Δ is increasing on $(-\infty, -\hat{\beta}_2/\hat{\beta}_3]$ and therefore defines a bijection on its image set $(-\hat{\beta}_3/\sqrt{\Sigma_{3,3}}, 0]$. As $-\hat{\beta}_3/\sqrt{\Sigma_{3,3}}$ is the z-statistic for the test of presence of an interaction, for no individual can the alternative

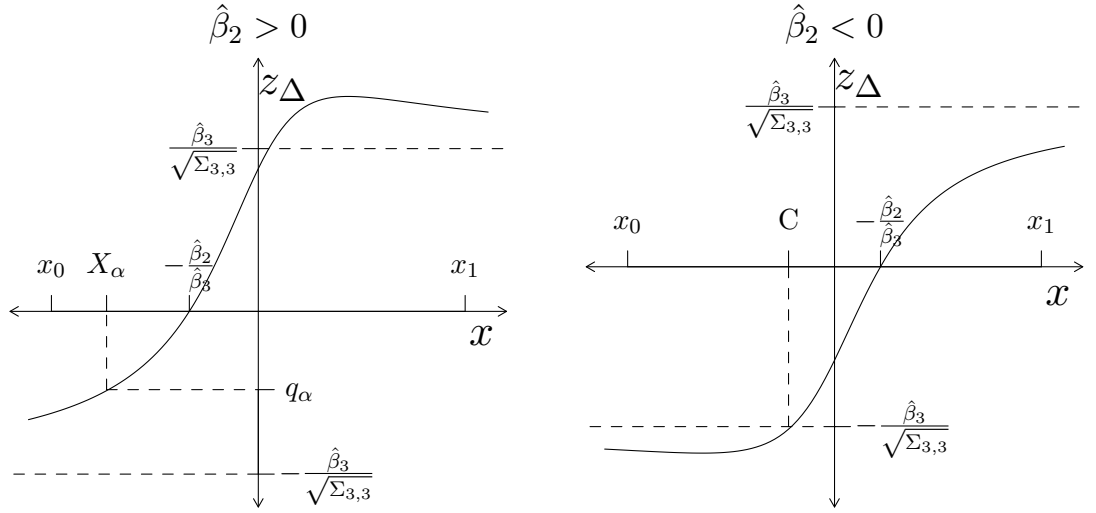

Figure 6.2: **Example of variation of z_Δ depending on the sign of $\hat{\beta}_2$.** The interaction coefficient $\hat{\beta}_3$ is positive in both plots. If it were negative, the curve would the symmetric of the plotted curve with respect to the ordinate axis.

treatment be recommended at a level α smaller than the p-value of the test for presence of interaction.

However, if $\hat{\beta}_2 < 0$, then z_Δ decreases before increasing. This means that for C characterized by $z_\Delta(C) = -\hat{\beta}_3/\sqrt{\Sigma_{3,3}}$, we have $\forall x < C$, $z_\Delta(x) < -\hat{\beta}_3/\sqrt{\Sigma_{3,3}}$, i.e., we can have more evidence for the use of the alternative treatment in some people than for the presence of an interaction. The condition $\hat{\beta}_2 < 0$ means that the observed treatment effect is negative. If we refer back to our definition of the reference and the alternative treatment in the introduction, $\hat{\beta}_2 < 0$ can happen only when the new treatment was superior to the old treatment but not significantly. It is therefore not desirable to recommend the alternative treatment to patients with $z_\Delta(x) < q_\alpha < -\hat{\beta}_3/\sqrt{\Sigma_{3,3}}$ as this means that we recommend an alternative treatment when neither the treatment effect nor the interaction effect are significant. In Section 6.4, we advocate to combine the interaction test with the test for improvement in order to exclude this situation.

Nevertheless, as long as we consider levels α larger than the p-value attained by the interaction, we are in a similar position to the one we had in the case $\hat{\beta}_2 > 0$. Indeed, in this case z_Δ defines a bijection from $(C, -\hat{\beta}_2/\hat{\beta}_3]$ to $(-\hat{\beta}_3/\sqrt{\Sigma_{3,3}}, 0]$. This means that we consider only levels α larger than the p-value for interaction.

The following proposition summarizes the results.

Proposition 2. *Assume, for simplicity, that $\Sigma_{2,3} = 0$. Then the function z_Δ defines a bijection*

to $(-|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}}, 0]$. *The domain of the bijection is*

$$\begin{cases} (-\infty, -\hat{\beta}_2/\hat{\beta}_3] & \text{if } \hat{\beta}_2 > 0 \text{ and } \hat{\beta}_3 > 0, \\ [-\hat{\beta}_2/\hat{\beta}_3, +\infty) & \text{if } \hat{\beta}_2 > 0 \text{ and } \hat{\beta}_3 < 0, \\ (C, -\hat{\beta}_2/\hat{\beta}_3] & \text{if } \hat{\beta}_2 < 0 \text{ and } \hat{\beta}_3 > 0, \text{ with } z_\Delta(C) = -|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}}, \\ [-\hat{\beta}_2/\hat{\beta}_3, C) & \text{if } \hat{\beta}_2 < 0 \text{ and } \hat{\beta}_3 < 0. \end{cases}$$

Thus, for all $\alpha < 0.5$ such that $q_\alpha > -|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}}$, we can define $X_\alpha \overset{\text{def}}{=} z_\Delta^{-1}(q_\alpha)$, as shown in Figure 6.2. We have capitalized the X to underline that this is a random variable under $\mathbb{P}_\beta$. This allows to rewrite the rejection region to test superiority of the alternative treatment for a patient with covariate x as $\{z_\Delta(x) < q_\alpha\} = \{x < X_\alpha\}$ if $\hat{\beta}_3 > 0$ and $\{z_\Delta(x) < q_\alpha\} = \{x > X_\alpha\}$ if $\hat{\beta}_3 < 0$. The point of Proposition 2 is to be able to parametrize on the scale of x instead of the scale of z_Δ—this new parameterization will play a key role in the next subsection.

Our detour through the analysis of the uncertainty at the individual level is now over. On the way, we have found another strategy of interest, $\hat{\texttt{ind}}_\alpha$, and a function sufficient to quantify the uncertainty, z_Δ. We are now prepared to study the uncertainty at the aggregated level, and use it to define a strategy with maximal guarantee on its gain.

6.3.3 Max lower bound strategy

Our goal is to find a strategy that defines a non-empty personalized set as often as possible, while having the most confidence that personalization will be beneficial. To this aim, we first need to restrict the range of possible policies we will be looking at. The two policies we have seen up to now, $\hat{\texttt{pol}}_1$ and $\hat{\texttt{ind}}_\alpha$, are both of the form $x \mapsto \mathbb{1}_{z_\Delta(x) \geq \eta}$, with $\eta = 0$ for $\hat{\texttt{pol}}_1$ and $\eta = q_\alpha$ for $\hat{\texttt{ind}}_\alpha$. We have also seen that z_Δ is sufficient to quantify uncertainty at the individual level. It is therefore reasonable to limit our search to policies of the same form as our two previously defined policies.

Conditionally on $\mathcal{D}_n$, z_Δ is a fixed function. Assume for now, without loss of generality, that $\hat{\beta}_3 > 0$, and, to fix ideas, that x_0 (the left extremity of the support of X) is in the domain of the bijection defined in Proposition 2. In this case, we can parametrize our class of policies on the scale of x instead of the scale of z_Δ. This suggests to look for our strategy in functions of the form $x \mapsto \mathbb{1}_{x \geq \rho}$, with $\rho \in [x_0, -\hat{\beta}_2/\hat{\beta}_3]$. (Of course, in the opposite situation where $\hat{\beta}_3 < 0$, the corresponding form is $x \mapsto \mathbb{1}_{x \geq \rho}$ for $\rho \in [-\hat{\beta}_2/\hat{\beta}_3, x_1]$.)

In order to achieve our goal, we propose to maximize the test statistic for positive impact of personalization, i.e., $\hat{q}_{n,\alpha}(\mathbb{1}_{x \geq \rho})$, over all possible choices of $\rho \in [x_0, -\hat{\beta}_2/\hat{\beta}_3]$. More precisely, let $A_\alpha = \arg\max_\rho \hat{q}_{n,\alpha}(\mathbb{1}_{x \geq \rho})$ be the set of maximizers. To define the strategy unambiguously, we choose the largest such threshold, i.e., we set

$$\rho_{\max,\alpha} = \max A_\alpha.$$

We call the resulting strategy the max lower bound strategy and denote it by $\hat{\texttt{mlb}}_\alpha$. Thus, by definition,

$$\hat{\texttt{mlb}}_\alpha(x) = \mathbb{1}_{x \geq \rho_{\max,\alpha}}.$$

To study the behavior of the random variable $\rho_{\max,\alpha}$, we simply use equation (6.3), which in this context takes the form

$$\hat{q}_{n,\alpha}(\mathbb{1}_{x\geq\rho}) = \hat{\Theta}(\mathbb{1}_{x\geq\rho}) + q_\alpha \mathrm{sd}_\Pi(\Theta^\star(\mathbb{1}_{x\geq\rho})).$$

As we want to maximize this quantity, we are going to derivate with respect to ρ. Assume, to simplify, that X has a bounded density with respect to the Lebesgue measure on $[x_0, x_1]$, i.e., $\mathbb{P}_X(dx) = f_X(x)dx$. In that case, the variance becomes

$$\mathrm{Var}_\Pi[\Theta^\star(\mathbb{1}_{x\geq\rho})] = \mathbb{E}_\Pi\Big[\Big(\int_{x_0}^{\rho}(\Delta^\star(x) - \hat{\Delta}(x))f_X(x)dx\Big)^2\Big] \stackrel{\text{def}}{=} v(\rho).$$

Thus, using the Lebesgue dominated convergence theorem, we have

$$\begin{aligned}\frac{dv}{d\rho}(\rho) &= \mathbb{E}_\Pi\Big[2f_X(\rho)(\Delta^\star(\rho) - \hat{\Delta}(\rho))\Big(\int_{x_0}^{\rho}(\Delta^\star(x) - \hat{\Delta}(x))f_X(x)dx\Big)\Big] \\ &= 2f_X(\rho)\mathrm{Cov}_\Pi(-\Delta^\star(\rho), \Theta^\star(\mathbb{1}_{x\geq\rho})).\end{aligned}$$

We conclude that

$$\frac{d\hat{q}_{n,\alpha}(\mathbb{1}_{x\geq\rho})}{d\rho}(\rho) = f_X(\rho)\left(-\hat{\Delta}(\rho) + q_\alpha \mathrm{sd}_\Pi(\Delta^\star(\rho))\mathrm{Cor}_\Pi(-\Delta^\star(\rho), \Theta^\star(\mathbb{1}_{x\geq\rho}))\right). \tag{6.5}$$

We are now ready to state our main theorem, which summarizes the connection between what happens at the individual level and at the aggregated level. For the sake of clarity, its proof is postponed to the Appendix. Observe that the result does not assume that x_0 is in the domain of the bijection.

Theorem 1. *Assume that X is either discrete or has a bounded density with respect to the Lebesgue on $[x_0, x_1]$. If $-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha$, then:*

(i) If $x_0 < X_\alpha$, the strategy $\hat{\mathtt{mlb}}_\alpha = \mathbb{1}_{x\geq\rho_{\max,\alpha}}$ defines a non-empty personalized set with $\rho_{\max,\alpha} \geq X_\alpha$. In this case, $\hat{q}_{n,\alpha}(\hat{\mathtt{mlb}}_\alpha) > 0$.

(ii) If, on the contrary, $X_\alpha \leq x_0$, then the personalized set is empty, i.e, $\rho_{\max,\alpha} = x_0$ and $\hat{\mathtt{mlb}}_\alpha = \mathtt{ref}$. In this case, $\hat{q}_{n,\alpha}(\hat{\mathtt{mlb}}_\alpha) = 0$.

On the other hand, if $\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha$, then:

(i) If $X_\alpha < x_1$, the strategy $\hat{\mathtt{mlb}}_\alpha = \mathbb{1}_{x\leq\rho_{\max,\alpha}}$ defines a non-empty personalized set with $\rho_{\max,\alpha} \leq X_\alpha$. In this case, $\hat{q}_{n,\alpha}(\hat{\mathtt{mlb}}_\alpha) > 0$.

(ii) If, on the contrary, $x_1 \leq X_\alpha$, then the personalized set is empty, i.e, $\rho_{\max,\alpha} = x_1$ and $\hat{\mathtt{mlb}}_\alpha = \mathtt{ref}$. In this case, $\hat{q}_{n,\alpha}(\hat{\mathtt{mlb}}_\alpha) = 0$.

This theorem is important insofar as it connects the behavior of the max lower bound strategy to the presence or absence of patients who benefit individually from the alternative treatment at level α. If there are some such patients, then we can personalize treatment for a slightly larger

set of patients while maximizing the confidence that personalization will be beneficial overall. If there are none, then the strategy is but the `ref` strategy, which does not personalize anyone and recommends the reference treatment to everyone.

Theorem 1 relies on the assumption that the unilateral test for interaction is significant (that is, $-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha$). If this assumption is not respected, i.e., $-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} \geq q_\alpha$, and if we also have $\hat{\beta}_2 < 0$, then $\hat{q}_{n,\alpha}(\hat{\texttt{mlb}}_\alpha)$ can be positive. As discussed in Subsection 6.3.2, this would mean an undesirable recommendation of the alternative treatment when neither the treatment effect nor the interaction effect are significant. To circumvent this problem, we propose to simply replace the $\hat{\texttt{mlb}}_\alpha$ by the restricted max lower bound strategy $\hat{\texttt{Mlb}}_\alpha$, which reduces to $\hat{\texttt{mlb}}_\alpha$ under the assumptions of the theorem and does not personalize otherwise. This will allow to control the type I error of the companion test, as we will see in Section 6.4. The restriction depends on the alternative hypothesis for interaction:

$$\hat{\texttt{Mlb}}_\alpha = \mathbb{1}_{-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} \geq q_\alpha} \texttt{ref} + \mathbb{1}_{-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha} \hat{\texttt{mlb}}_\alpha$$

when the test for interaction is unilateral with $H_1^{\text{interact}} : \beta_3 > 0$, and

$$\hat{\texttt{Mlb}}_\alpha = \mathbb{1}_{\hat{\beta}_3/\sqrt{\Sigma_{3,3}} \geq q_\alpha} \texttt{ref} + \mathbb{1}_{\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha} \hat{\texttt{mlb}}_\alpha$$

when the test for interaction is unilateral with $H_1^{\text{interact}} : \beta_3 < 0$. If the test is bilateral, i.e., $H_1^{\text{interact}} : \beta_3 \neq 0$, then we let

$$\hat{\texttt{Mlb}}_\alpha = \mathbb{1}_{-|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}} \geq q_{\alpha/2}} \texttt{ref} + \mathbb{1}_{-|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}} < q_{\alpha/2}} \hat{\texttt{mlb}}_\alpha.$$

As $q_{\alpha/2} < q_\alpha$, when $-|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}} < q_{\alpha/2}$ we are under one of the assumptions of Theorem 1.

We conclude this subsection by studying the gap between X_α and $\rho_{\max,\alpha}$. If we go back to equation (6.5), we see that if $\text{Cor}_\Pi(-\Delta^\star(\rho), \Theta^\star(\mathbb{1}_{x \geq \rho})) = 1$, then $\frac{d\hat{q}_{n,\alpha}(\mathbb{1}_{x \geq \rho})}{d\rho}(\rho) = 0$ if and only if $\rho_{\max,\alpha} = X_\alpha$ defined above. In practice, this will be approximately true as long as ρ is close to x_0, as shown in the next proposition.

Proposition 3. *We have, almost surely,* $\lim_{\rho \to x_0^+} \text{Cor}_\Pi(-\Delta^\star(\rho), \Theta^\star(\mathbb{1}_{x \geq \rho})) = 1$.

Proof. By Lemma 1 in the Appendix,

$$\begin{aligned}\text{Cor}_\Pi(-\Delta^\star(\rho), \Theta^\star(\mathbb{1}_{x \geq \rho})) &= \text{Cor}_\Pi(-\Delta^\star(\rho), -\Delta^\star(g(\rho))) \\ &\to_{\rho \to x_0^+} \text{Cor}_\Pi(-\Delta^\star(x_0), -\Delta^\star(x_0)) = 1.\end{aligned}$$

□

This means that in many cases the max lower bound strategy will be very close to the individual strategy defined in (6.4), i.e.,

$$\hat{\texttt{Mlb}}_\alpha \approx \hat{\texttt{ind}}_\alpha.$$

Furthermore, the proximity of $\text{Cor}_\Pi(-\Delta^\star(X_\alpha), \Theta^\star(\mathbb{1}_{x \geq X_\alpha}))$ with 1 is a diagnostic tool to evaluate the quality of this approximation. If $\text{Cor}_\Pi(-\Delta^\star(X_\alpha), \Theta^\star(\mathbb{1}_{x \geq X_\alpha}))$ is not close to 1, then we can personalize more patients than if we were considering uncertainty at the individual level, while controlling type I error. This is a consequence of subadditivity of standard deviation: if we sum variables that have correlation smaller than 1, the standard deviation of the sum is smaller than the sum of the standard deviations. The two possibilities are illustrated in the next subsection.

6.3.4 Illustration of the max lower bound strategy

In this subsection, we illustrate the behavior of the bilateral $\hat{\texttt{Mlb}}_\alpha$ strategy. We simulate two scenarios: the first scenario is the one we used for Figure 6.1, to exemplify the problems caused by considering $\hat{\texttt{pol}}_1$, the plug-in estimator of the optimal treatment strategy. Our second scenario illustrates the case where there is substantial distance between X_α and $\rho_{\max,\alpha}$.

In both scenarios, we simulate datasets based on the linear model (6.2). The specifics of the first scenario are detailed in Figure 6.1. To show the influence of the choice of strategy, we compute our quantities of interest for a grid of policies depending on a threshold η for z_Δ, with η ranging from $q_{0.02}$ to $q_{0.5}$. For each simulation, we select $\eta_{\max,0.05}$ empirically, i.e., we select the threshold for which $\hat{q}_{n,0.05}(\mathbb{1}_{z_\Delta(x)\geq\eta})$ is maximal. This is shown in Figure 6.3. As we can see, the problematic points in the upper left quadrant under $\hat{\texttt{pol}}_1$ are brought to the upper right quadrant by decreasing η. Only 0.2% of the simulations where $\hat{q}_{n,0.05}(\hat{\texttt{Mlb}}_{0.05}) > 0$ identify a strategy that is not beneficial, i.e., $\Theta(\hat{\texttt{Mlb}}_{0.05}) \leq 0$. Manual inspection of the instances where this is true show that they correspond to extreme underestimation of the ratio β_2/β_3.

As noted at the end of Section 6.2, the hypothesis we want to test $H_0 : \Theta(\hat{\texttt{Mlb}}_\alpha) \leq 0$ depends on a random variable. We have seen in Subsection 6.3.1 that for a poor choice of strategy such as $\hat{\texttt{pol}}_1$, $\mathbb{P}_\beta(H_0|\hat{q}_{n,\alpha}(\hat{\texttt{pol}}_1) > 0)$ can be substantial even though $\Theta(\texttt{opt}) > 0$. However, for the $\hat{\texttt{Mlb}}_\alpha$ strategy, the simulations are reassuring as this same probability is quite small and the strategy identified is not beneficial only if the personalized set is empty.

Power for detection of interaction in this case is 1 and we are therefore always under the assumption of our theorem. We detect a positive improvement due to personalization in 34% of simulations. There is a small difference between the probability of $\min(z_\Delta) < q_{0.05}$ at 34.7% and $\hat{q}_{n,0.05}(\hat{\texttt{Mlb}}_{0.05}) > 0$ at 34.3% that is likely due to Monte-Carlo noise. The proportion personalized P_{neg} is 5% when the personalized set is not empty instead of the 10.6% of the $\texttt{opt}$ strategy.

The factor identified in the derivative of $\rho \mapsto \hat{q}_{n,0.05}(\mathbb{1}_{x\geq\rho})$ as influencing the quality of the approximation of $\rho_{\max,0.05}$ by $X_{0.05}$, i.e., of $\eta_{\max,0.05}$ by $q_{0.05}$, $\text{Cor}_\Pi(-\Delta^\star(X_{0.05}), \Theta^\star(\mathbb{1}_{z_\Delta(x)<q_{0.05}}))$ is always larger than 0.98 in this scenario. This implies that $q_{0.05}$ is a good approximation of $\eta_{\max,0.05}$ and the $\hat{\texttt{Mlb}}_{0.05}$ strategy is almost equal to the $\hat{\texttt{ind}}_{0.05}$ strategy. Since we sample only 10 000 times from Π, there is still some noise and $q_{0.05}$ is selected only 94% of the time when $\hat{q}_{n,0.05}(\hat{\texttt{Mlb}}_{0.05}) > 0$, the remaining occurrences select either $q_{0.04}$ or $q_{0.07}$, the closest values on the grid.

To conclude on this scenario, the use of the $\hat{\texttt{Mlb}}_{0.05}$ strategy has allowed to deal with the problems that were due to the use of $\hat{\texttt{pol}}_1$, i.e., the low probability of identifying a significant improvement and, when a significant improvement was identified, the large probability of the strategy being detrimental.

With our second scenario, we want to illustrate the interest of using $\eta_{\max,0.05}$ instead of $q_{0.05}$. We simulate 100 datasets with parameters $\beta_1 = 1$, $\beta_2 = 0$, $\beta_3 = 0.5$, $n = 500$, and $\sigma^2 = 1$. As we still use a uniform distribution on $[-1, 1]$ for X, the optimal treatment strategy would be to personalize half the patients. We sampled 100 000 times from Π in order to minimize Monte Carlo noise and pinpoint precisely $\eta_{\max,0.05}$. We searched for η on a grid with 0.1 increments that contained $q_{0.05}$.

Out of the 70 occurrences where the test for interaction is significant, i.e., $-|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}} >$

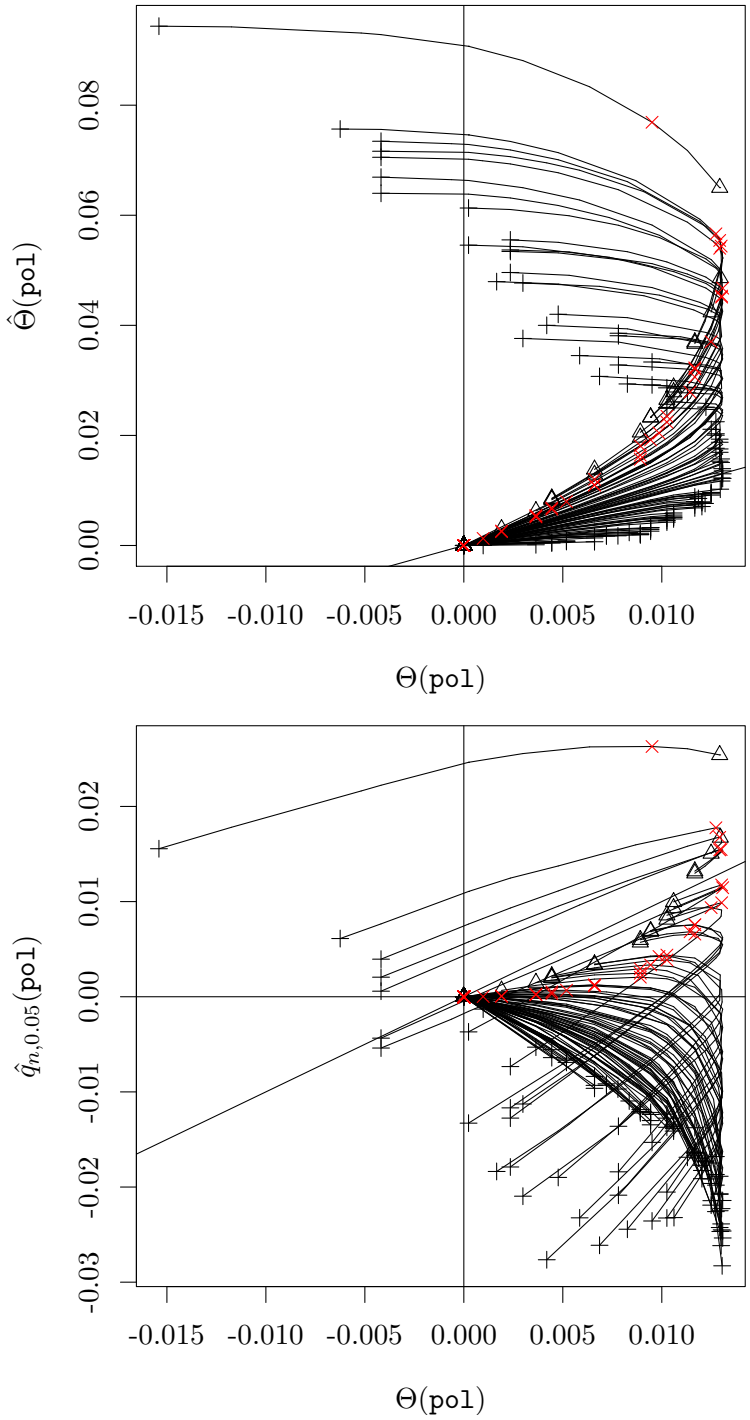

Figure 6.3: **Max lower bound strategy.** Graph of $\Theta(\texttt{pol})$ against $\hat{\Theta}(\texttt{pol})$ and $\hat{q}_{n,0.05}(\texttt{pol})$. Each trajectory represents the change in those quantities when η varies. The cross corresponds to $\eta = 0$, i.e., $\texttt{pol} = \hat{\texttt{pol}}_1$. The triangles correspond to the end of the trajectory, i.e., $\rho = q_{0.02} = -2.05$. The red X marks the quantities under the $\hat{\texttt{Mlb}}_{0.05}$ strategy. We plotted only the first 100 trajectories in order to have a readable output. In both plots, the first bisector, i.e., the $y = x$ line, is drawn.

$q_{0.025}$, and $\hat{q}_{n,0.05}(\widehat{\texttt{Mlb}}_{0.05}) > 0$, the best choice for η was $q_{0.05}$ only 30 times. All other selected thresholds were larger than $q_{0.05}$, as expected. In these 70 simulations, the average $\text{Cor}_\Pi(\Delta^\star(X_{0.05}), \Theta^\star(\mathbb{1}_{z_\Delta(x)<q_{0.05}}))$ was 0.91. This quantity was 0.99 when $q_{0.05}$ had been selected and 0.86 in the 40 other instances. In the 40 simulations where a larger threshold had been selected, the average P_{neg} was 42% for the threshold $q_{0.05}$ while it was 46% for $\eta_{\max,0.05}$. This shows that the gain from aggregating uncertainty instead of controlling it at the individual level can be substantial.

6.4 Testing for benefit of personalization

In this section, we focus on how to test for benefit of personalization using the new strategy $\widehat{\texttt{Mlb}}_\alpha$. As announced above, we show the need to combine the rejection region we proposed at the end of Section 6.2, $\{\hat{q}_{n,\alpha}(\widehat{\texttt{mlb}}_\alpha) > 0\}$, with the test for interaction presented in Definition 1. In fact, this is precisely what motivated the definition of the restricted max lower bound strategy $\widehat{\texttt{Mlb}}_\alpha$ that was introduced in the previous section. We then see that the combined test controls the type 1 error when there is no improvement under the optimal treatment strategy, i.e., $\Theta(\texttt{opt}) = 0$.

Let us recall that we want to test the null hypothesis of no benefit of personalization under the $\widehat{\texttt{mlb}}_\alpha$ strategy, i.e., $H_0 : \Theta(\widehat{\texttt{mlb}}_\alpha) \leq 0$, against the alternative hypothesis that there is a gain from personalization, i.e., $H_1 : \Theta(\widehat{\texttt{mlb}}_\alpha) > 0$. As noted above, these hypotheses depend on random variables. However, when the optimal strategy defines an empty personalized set, i.e., $\Theta(\texttt{opt}) = 0$, the estimated strategy can never be beneficial, since $\Theta(\widehat{\texttt{mlb}}_\alpha) \leq \Theta(\texttt{opt}) = 0$. Therefore, when $\Theta(\texttt{opt}) = 0$, $\mathbb{P}_\beta(H_0) = 1$. Thus, letting $H_0^{\texttt{opt}} : \Theta(\texttt{opt}) = 0$, we have $H_0^{\texttt{opt}} \subset H_0$, with the added advantage that $H_0^{\texttt{opt}}$ is a traditional null hypothesis that depends only on fixed quantities. In the sequel, we therefore study the well-defined type I error under the null hypothesis $H_0^{\texttt{opt}}$.

We now provide motivation for the combination of the interaction test with the test we proposed based on $\hat{q}_{n,\alpha}(\widehat{\texttt{mlb}}_\alpha)$. We will show that $\{\hat{q}_{n,\alpha}(\widehat{\texttt{mlb}}_\alpha) > 0\}$ does not control type I error under $H_0^{\texttt{opt}}$ and that this is linked with the pathological behavior of z_Δ when $\hat{\beta}_2 < 0$. As we have seen in Subsection 6.3.2, when $\hat{\beta}_2 < 0$ the function z_Δ has a downward bump and can therefore attain values smaller than $-|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}}$. This situation can occur quite frequently. Assume, for example, that $\beta_3 = \beta_2 = 0$, and let us illustrate this scenario with simulations following the protocol in the legend of Figure 6.4. As $\hat{\beta}_2$ is centered and Gaussian, we have $\mathbb{P}_\beta(\hat{\beta}_2 < 0) = 1/2$. Furthermore, the scenario $\beta_3 = \beta_2 = 0$ falls under $H_0^{\texttt{opt}}$. The simulations show that the quantile we have focused on, $\hat{q}_{n,0.05}(\widehat{\texttt{mlb}}_{0.05})$, is positive 1 time out of 4 when $\hat{\beta}_2 < 0$. As a consequence, if we use $\{\hat{q}_{n,0.05}(\widehat{\texttt{mlb}}_{0.05}) > 0\}$ as the rejection region of our test, the type I error rate will be 14%. The excess in type I error rate above its nominal level comes exclusively from the simulations where $\hat{\beta}_2 < 0$. As we see next, combining this rejection region with the one of the interaction test of Proposition 1 allows to control type I error under the null hypothesis $H_0^{\texttt{opt}}$.

Depending on the clinical context, the test for interaction can be unilateral or bilateral, and we advocate for the use of the corresponding (unilateral or bilateral) $\widehat{\texttt{Mlb}}_\alpha$. When the alternative hypothesis for the interaction test is $H_1^{\text{interact}} : \beta_3 > 0$, i.e., the test is unilateral, then the border

between $H_0^{\texttt{opt}}$ and the alternative hypothesis $H_1^{\texttt{opt}} : \Theta(\texttt{opt}) > 0$ and $\beta_3 > 0$ is

$$B = \{(\beta_2, \beta_3), \beta_3 \geq 0, \Delta(x_0) = 0\} = \{(\beta_2, \beta_3), \beta_3 \geq 0, \beta_2 = -x_0\beta_3\}.$$

The rejection region of the joint test is

$$R = \{-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha, \hat{q}_{n,\alpha}(\hat{\texttt{mlb}}_\alpha) > 0\}.$$

Combining the test based on our quantile with the interaction test means that we always respect the assumption of Theorem 1 in the rejection region, and we can therefore apply the theorem to write $R = \{-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha, x_0 < X_\alpha\}$. As we have $x_0 < X_\alpha \Leftrightarrow z_\Delta(x_0) < q_\alpha$, the rejection region then becomes:

$$R = \{-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha, z_\Delta(x_0) < q_\alpha\}.$$

On the border B, $\Delta(x_0) = 0$, and therefore, under $\mathbb{P}_\beta$, $z_\Delta(x_0) \sim \mathcal{N}(0,1)$. As $\mathbb{P}_\beta(z_\Delta(x_0) < q_\alpha) = \alpha$, we have control of the type I error rate:

Proposition 4. *If we consider a unilateral test for interaction in the joint test for presence of improvement, then for all* $(\beta_2, \beta_3) \in B$ *we have* $\mathbb{P}_\beta\,(\text{Rejection}) \leq \alpha$.

The situation is not so straightforward in the bilateral case and we will not formally prove that we have control over type I error. Indeed, when the alternative hypothesis for the interaction is $H_1^{\text{interact}} : \beta_3 \neq 0$, the rejection region changes with the sign of $\hat{\beta}_3$. We have $R = \{-|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}} < q_{\alpha/2}, z_\Delta(x_0) < q_\alpha\}$ if $\hat{\beta}_3 > 0$ and $R = \{-|\hat{\beta}_3|/\sqrt{\Sigma_{3,3}} < q_{\alpha/2}, z_\Delta(x_1) > q_\alpha\}$ if $\hat{\beta}_3 < 0$. As $\hat{\beta}_3$ is not always of the sign of β_3, a formal proof would be tedious. We can nevertheless expect type I error to be controlled, and we give some arguments to support this statement. Indeed, the border region between the null hypothesis and the alternative becomes $B = \{(\beta_2, \beta_3), \beta_3 \geq 0, \beta_2 = -x_0\beta_3\} \cup \{(\beta_2, \beta_3), \beta_3 \leq 0, \beta_2 = -x_1\beta_3\}$. In the case where $\beta_3 = 0$, the test for interaction is clearly sufficient to control the type I error. If β_3 is large and positive, we have with great probability that $\hat{\beta}_2 > 0$ and $\hat{\beta}_3 > 0$. In this case, $z_\Delta(x_0) \sim \mathcal{N}(0,1)$, and

$$\mathbb{P}_\beta\,(\text{Rejection}) \approx \mathbb{P}_\beta(z_\Delta(x_0) < q_\alpha) = \alpha.$$

A similar reasoning applies when β_3 is large and negative. Using simulations, we show in Figure 6.4 that this control is still attained for intermediate values between $\beta_3 = 0$ and β_3 positive and large.

To sum up the findings of this section, we have defined a strategy, $\hat{\texttt{mlb}}_\alpha$, and a companion test such that our strategy personalizes treatment when the null hypothesis of no improvement is rejected and recommends the reference treatment to everyone in the opposite case. This test is also a valid test for presence of improvement under the optimal treatment strategy. Therefore, $\hat{\texttt{mlb}}_\alpha$ personalizes treatment whenever we can detect the presence of a theoretical improvement.

6.5 Additional properties of the $\hat{\texttt{mlb}}_\alpha$ strategy

We study in this section the coverage probabilities of the strategy $\hat{\texttt{mlb}}_\alpha$, as well as the influence of parameters.

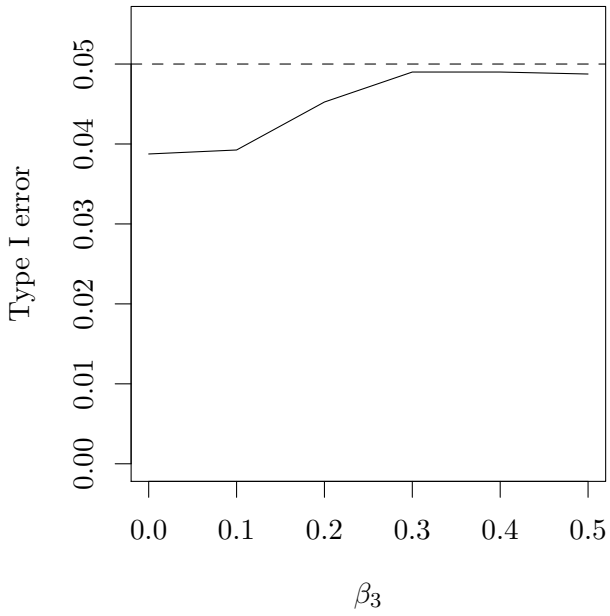

Figure 6.4: **Control of type I error for the joint test with bilateral test for interaction.** We simulated 4 000 datasets based on the linear model (6.2), i.e., we sampled ϵ for each simulation. The variable X is sampled once from a uniform distribution between -1 and 1. We take $n = 300$, $\beta_0 = 0$, $\beta_1 = 1$, and $\sigma^2 = 1$. We sampled 4 000 times from Π. The interaction coefficient β_3 varies between 0 and 0.5 by 0.1 increments with $\beta_2 = -\beta_3 \min_i(X_i)$. The dotted line shows $\alpha = 0.05$.

6.5.1 Coverage probabilities

We have defined our estimation procedure using Bayesian arguments, and we have used this estimation to select the strategy $\widehat{\texttt{Mlb}}_\alpha$. Now, we want to check that the resulting quantiles respect the expected frequentist coverage probabilities, i.e., $\mathbb{P}_\beta(\Theta(\widehat{\texttt{Mlb}}_\alpha) \leq \hat{q}_{n,\gamma}(\widehat{\texttt{Mlb}}_\alpha)) = \gamma$. Here $\gamma \in [0,1]$ is any confidence level, which does not have to be equal to α, the confidence level used in the definition of $\widehat{\texttt{Mlb}}_\alpha$. This subsection will use simulations to check this, keeping in mind that the question of frequentist validity of Bayesian credible bounds is an active field of theoretical research [e.g., van der Vaart, 1998].

Coverage probabilities are complicated in our setting by the fact that with positive probability, $\widehat{\texttt{Mlb}}_\alpha = \texttt{ref}$. In this case, the set of patients for whom the recommended treatment is the alternative treatment (i.e., the personalized set) is empty. As $\Theta^\star$ is an integral on the personalized set, we have $\Theta^\star(\texttt{ref}) = 0$, and the distribution of $\Theta^\star$ is a Dirac mass at 0. As $\Theta(\texttt{ref}) = 0$ as well, all quantiles are therefore correct in this case. This means that $\mathbb{P}_\beta(\Theta(\widehat{\texttt{Mlb}}_\alpha) = \hat{q}_{n,\gamma}(\widehat{\texttt{Mlb}}_\alpha)) > 0$, and the best we can hope for is to have control over strict left coverage $\mathbb{P}_\beta(\Theta(\widehat{\texttt{Mlb}}_\alpha) < \hat{q}_{n,\gamma}(\widehat{\texttt{Mlb}}_\alpha))$ and over strict right coverage $\mathbb{P}_\beta(\Theta(\widehat{\texttt{Mlb}}_\alpha) > \hat{q}_{n,\gamma}(\widehat{\texttt{Mlb}}_\alpha))$. In the frequentist world, valid coverage translates to

$$\mathbb{P}_\beta\big(\Theta(\widehat{\texttt{Mlb}}_\alpha) < \hat{q}_{n,\gamma}(\widehat{\texttt{Mlb}}_\alpha)\big) \leq \gamma$$

and

$$\mathbb{P}_\beta\big(\Theta(\widehat{\texttt{Mlb}}_\alpha) > \hat{q}_{n,\gamma}(\widehat{\texttt{Mlb}}_\alpha)\big) \leq 1-\gamma.$$

To investigate if coverage is respected by our estimation, we simulated data in the same fashion as above. We simulated 10 000 datasets based on the linear model (6.2), i.e., we sampled ϵ for each simulation. The variable X is sampled once from a uniform distribution between -1 and 1. We take $n = 300$, $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 0.3$, $\sigma^2 = 1$, and $\beta_3 \in \{0.5, 0.8, 1\}$. We select 0.05 for α and we use $\widehat{\texttt{Mlb}}_{0.05}$. We sampled $\beta^\star$ 10 000 times from Π.

Figure 6.5 shows the results for four thresholds, including $\alpha = 0.05$. Each bar corresponds to a set of simulations with fixed parameters and a choice of γ. It is divided in 3 with the length of each color corresponding (from bottom to top) to the probability of left coverage (blue), equality (green), or right coverage (red). Naturally, the total length of the bar is 1 as those three probabilities sum to 1. The line γ is drawn and the coverage inequalities hold if it is in the green. The main message of Figure 6.5 is that coverage is approximately respected. However, there is a clear asymmetry between left and right coverages. When β_3 decreases, the green bar corresponding to the probability of an empty personalized set increases. As that probability increases, the probability of right coverage (red) decreases while the probability of left coverage (blue) stays constant. It is only after the probability of right coverage is 0 that the probability of left coverage decreases with β_3.

6.5.2 Influence of parameters

To have some grasp on the situation, we propose the visualization in Figure 6.6 in the parameters' plane. The confidence ellipse around $(\hat{\beta}_2, \hat{\beta}_3)$ represents Π, the posterior distribution of (β_2, β_3) from which we sample $(\beta_2^\star, \beta_3^\star)$. To each X_i corresponds a line $\Delta^\star(X_i) = \beta_2^\star + \beta_3^\star X_i = 0$. The lines that are under the true value of (β_2, β_3) correspond to participants for whom personalization is beneficial.

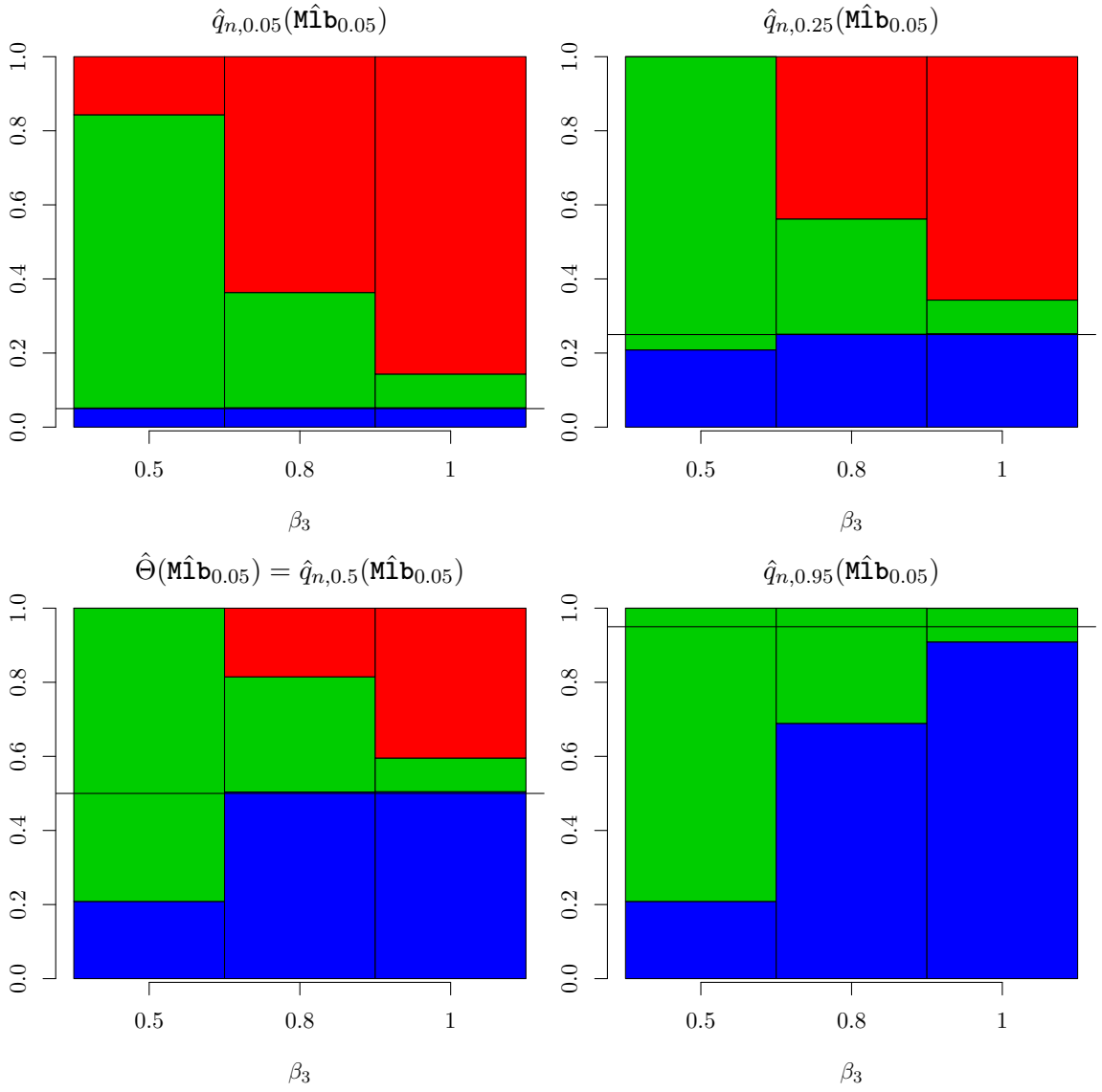

Figure 6.5: **Coverage of $\Theta(\hat{\texttt{Mlb}}_{0.05})$ by $\hat{q}_{n,\gamma}(\hat{\texttt{Mlb}}_{0.05})$ for $\gamma \in \{0.05, 0.25, 0.5, 0.95\}$.** The bar length corresponds to the probabilities of different events depending on the value of β_3. In each bar, from bottom to top, the length of the blue bar is the empirical probability of left coverage $\mathbb{P}_\beta(\Theta(\hat{\texttt{Mlb}}_{0.05}) < \hat{q}_{n,\gamma}(\hat{\texttt{Mlb}}_{0.05}))$. The green bar corresponds to $\mathbb{P}_\beta(\Theta(\hat{\texttt{Mlb}}_{0.05}) = \hat{q}_{n,\gamma}(\hat{\texttt{Mlb}}_{0.05}))$, i.e., $\mathbb{P}_\beta(\hat{\texttt{Mlb}}_{0.05} = \texttt{ref})$. The red bar corresponds to the probability of right coverage $\mathbb{P}_\beta(\Theta(\hat{\texttt{Mlb}}_{0.05}) > \hat{q}_{n,\gamma}(\hat{\texttt{Mlb}}_{0.05}))$. The γ line is drawn. As long as the line is in the green, coverage is respected. Note that the median $\hat{q}_{n,0.5}(\hat{\texttt{Mlb}}_{0.05})$ corresponds to $\hat{\Theta}(\hat{\texttt{Mlb}}_{0.05})$.

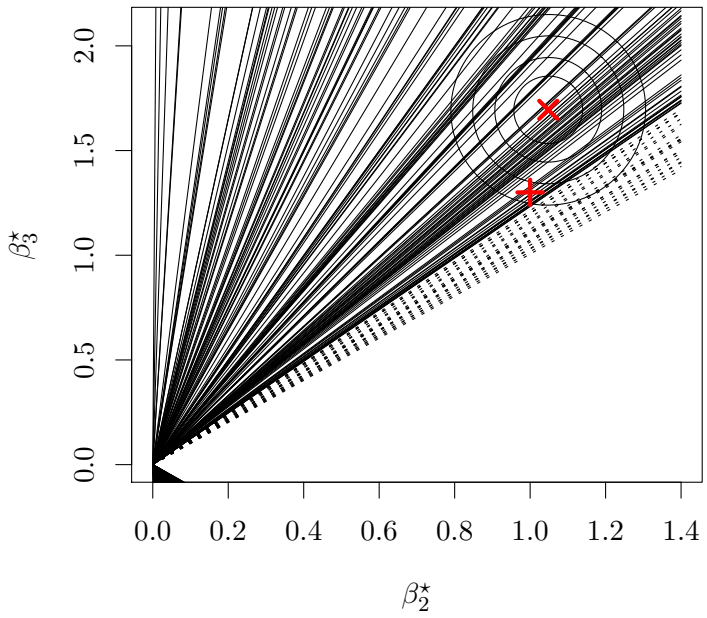

Figure 6.6: Confidence ellipse around $(\hat{\beta}_2, \hat{\beta}_3)$ (marked by a red x) at confidence level 25%, 50%, 75%, and 90%. The red cross marks the true value of the parameters (β_2, β_3). The lines have equation $x + yX_i = 0$ for all X_i in the sample. The dotted lines correspond to the observations that verify $z_\Delta(X_i) < \eta_{\max,0.05}$, i.e., the set of patients for which the alternative treatment is recommended under strategy $\widehat{\texttt{ind}}_{0.05}$. The axis of the ellipse are parallel to the x-axis and y-axis because of the independence between X and T.

The ellipse controls uncertainty in two dimensions instead of one and therefore leads to wider confidence region when projected in one dimension. This is why some of the dotted lines intersect the 90% confidence ellipse despite corresponding to points for which $z_\Delta(X) < q_{0.05}$.

With this representation, it is quite easy to infer the influence of parameters on the discovery rate. As the sample size n increases, the ellipse will shrink and the estimated parameters will converge towards the true parameters at rate $1/\sqrt{n}$, which will lead to more frequent identification of a non-empty personalized set. When the treatment effect β_2 increases, there are less patients who benefit from personalization and therefore it becomes harder to identify them. If β_3 increases, the opposite happens. If both parameters increase while their ratio remains fixed, it will become easier to identify a personalized set because the lines will be more spread out.

6.6 Extensions of the method

6.6.1 Extension to other outcomes

If our outcome Y is binary and not continuous, we would like to be able to use logistic regression. To be consistent, we assume that $Y = 1$ is a desirable outcome, such as healing or absence of negative outcome. In this context, the natural extension of our approach is to consider the logistic regression model

$$\text{logit}(\mathbb{P}(Y = 1|X, T)) = \beta_0 + \beta_1 X + \beta_2 T + \beta_3 XT,$$

where, for all $p \in [0, 1]$, $\text{logit}(p) = \log(\frac{p}{1-p})$. As the maximum likelihood estimator of β is asymptotically normal, we can apply the procedures described in the previous sections. However, the definition of $\Delta(X)$ as $\beta_2 + \beta_3 X$ is not natural as it is unclear what the corresponding Θ would mean. It is more logical to follow Janes et al. [2014], and let

$$\Delta(X) = \mathbb{P}(Y = 1|X, T = 1) - \mathbb{P}(Y = 1|X, T = 0),$$

in which case $\Theta(\texttt{pol})$ will be the difference in event rate between `pol` and `ref`. Thus,

$$\Delta(X) = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 + \beta_3 X)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 + \beta_3 X)} - \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)},$$

an expression that depends on the whole of β instead of just β_2 and β_3.

We can then, as before, sample $\beta^\star$ from Π, the asymptotic posterior distribution of β, and obtain quantiles of our quantity of interest. Instead of using z_Δ as we did under a Gaussian assumption, a natural idea is to work with $\Pi(\Delta^\star(x) > 0)$. Indeed, in the Gaussian case, we have $z_\Delta(x) < q_\alpha \Leftrightarrow \Pi(\Delta^\star(x) > 0) < \alpha$. We then look for policies of the form $\mathbb{1}_{\Pi(\Delta^\star(x)>0)<\gamma}$, with $\gamma \in [0, 1]$. To rephrase it, we personalize treatment for the patients most likely to benefit from the alternative treatment. Our threshold is now on the probability scale between 0 and 1. We apply this extension to real data in Section 6.7.

Similarly, our approach can be extended to a censored outcome. In that case, one should first select a measure of treatment contrast, which could be either the survival probability at a prespecified time τ, or the restricted mean survival up to τ [Irwin, 1949], for instance. A Cox proportional hazards model $\lambda(t|X, T) = \lambda_0(t) \exp(\beta_1 X + \beta_2 T + \beta_3 XT)$ can be used for regression, where $\lambda(t)$ denotes the hazard function and the baseline hazard $\lambda_0(t)$ is an unspecified non-negative function [Cox, 1972]. It is then possible to define Δ as the difference in predicted survival at time τ as

$$\Delta(X) = \hat{S}_0(\tau)^{\exp(\beta_1 X + \beta_2 + \beta_3 X)} - \hat{S}_0(\tau)^{\exp(\beta_1 X)},$$

by plugging the Breslow estimator of Λ_0 to obtain $\hat{S}_0(\tau) = \exp(-\hat{\Lambda}_0(\tau))$, and sampling $\beta^\star$ from the asymptotic posterior distribution of β obtained from the partial likelihood estimator $\hat{\beta}$. For restricted mean survival up to τ, Δ is defined as

$$\Delta(X) = \int_0^\tau \left\{ \hat{S}_0(\tau)^{\exp(\beta_1 X + \beta_2 + \beta_3 X)} - \hat{S}_0(\tau)^{\exp(\beta_1 X)} \right\} dt.$$

Both were used in Zhao et al. [2013], and the second one in Li et al. [2016], for instance. Otherwise, it is also possible to use a time-dependent logistic model estimated with inverse probability weighting for the survival at τ [Zheng et al., 2006], or a model directly estimating the restricted mean survival time to the covariates, also estimated through inverse probability of censoring weighting [Tian et al., 2014].

6.6.2 Extension to the multivariate case

In most RCTs, a large number of covariates describe each patient instead of just one, as we previously assumed. In order for personalized medicine to fulfill its promises, it is crucial that the information contained in this data be harnessed.

Let us then assume that $X = (X_1, \dots, X_d)$ is now d-dimensional. Using a machine learning algorithm, we can predict Y given X and T. In the present chapter, we use the random forests algorithm [Breiman, 2001a], and denote by $\hat{h}(X, T)$ such a prediction. We can then define, as before,

$$\hat{\Delta}(X) = \hat{h}(X, 1) - \hat{h}(X, 0).$$

The procedure to estimate $\Theta(\texttt{pol})$ we proposed in Section 6.2 was based on our knowledge of the joint distribution of $(\Delta^\star(x))_x$, with $\Delta^\star(x)$ following the posterior distribution of $\Delta(X)$. In the case of a linear model, this distribution was quite simple, thanks to the straightforward relation between Δ and the coefficients, namely $\Delta^\star(x) = \beta_2^\star + \beta_3^\star x$. Unfortunately, such a simple technique does not immediately transpose to machine learning algorithms. Of course, there are ways to estimate the uncertainty around a prediction, for example using bootstrap methods [Tibshirani, 1996b, Wager et al., 2014], but they only concern single predictions. Besides, studying the joint distribution of prediction errors of machine learning algorithms exceeds the scope of this chapter. Instead, we suggest in this section a simpler approach, which nicely extends what we have developed in the univariate case. The idea is to use $\hat{h}$ to build features that will be plugged in a linear model, for which we will be able to apply the procedure detailed in Section 6.2.

Our univariate linear model has three parts besides the intercept: a prognostic term $\beta_1 X$, a reference treatment average effect term β_2, and an interaction term $\beta_3 X$, with the last two multiplied by treatment assignment T. The prognostic term is the expected value given X when a patient receives the alternative treatment. Therefore, we define our first feature as $Z_1(X) = \hat{h}(X, 0)$ in order to capture the prognostic. Besides, the treatment effect $\Delta(X)$ is estimated by $\hat{h}(X, 1) - \hat{h}(X, 0)$. Since $\Delta(X) = \beta_2 + \beta_3 X$ in the linear case, and $\mathbb{E}_X X = 0$, it is logical to define the interaction term by $Z_3(X) = \hat{h}(X, 1) - \hat{h}(X, 0) - \bar{\Delta}$, where $\bar{\Delta}$ is the empirical mean of $\hat{\Delta}(X)$. All in all, we assume that our outcome follows a linear model, of the form

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 T + \beta_3 Z_3 T + \epsilon,$$

and we are all set to apply the procedure described in the first part of the chapter. We can, in particular, test for presence of an improvement and estimate the improvement under the max lower bound strategy.

Observe that, because of the way we have built our features, we expect the coefficients in the linear regression to be close to natural values: $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = \bar{\Delta}$, and $\beta_3 = 1$. Indeed, in that case we have

$$Y = 0 + \hat{h}(X, 0) + \bar{\Delta} T + (\hat{h}(X, 1) - \hat{h}(X, 0) - \bar{\Delta}) T = \hat{h}(X, 0) + (\hat{h}(X, 1) - \hat{h}(X, 0)) T,$$

i.e., $Y = \hat{h}(X, 0)$ if $T = 0$ and $Y = \hat{h}(X, 1)$ if $T = 1$. If the coefficients are far from their expected values, if for example β_3 is negative, then we should be suspicious of the results. This motivates the use of a unilateral test for interaction in the test described in Section 6.4.

Overfitting is also a concern in our setting, because an overfit prediction will always find an interaction term Z_3 by fitting the noise, and the same noise will be present when doing the linear regression. We will therefore find a significant interaction even if it is not in fact present. To avoid this unpleasant state of affairs, a possible route is to use the out-of-bag estimators of random forests. This is what we did in the application described in the next section, and we recommend to take careful consideration of this issue.

Let us finally mention that the crux of the approach is to replace the joint distribution of prediction errors by the joint distribution of $(\beta_2^\star + \beta_3^\star Z_3(x))_x$. A reason to expect the procedure to be somewhat conservative is that very different x will lead to similar values of Z_3. If we have two distant points x_1 and x_2 such that $Z_3(x_1) \approx Z_3(x_2)$, then the prediction errors of $Z_3(x_1)$ and $Z_3(x_2)$ might be less correlated than in our procedure, where the correlation is one. Given the subadditivity of standard deviation, taking this into account would lead the aggregate quantity to carry less uncertainty. A promising research perspective is therefore to study the joint distribution of prediction errors of machine learning algorithms to be able to provide tighter confidence bounds and therefore extend personalization to as many patients as possible.

6.7 Illustration on real data

We illustrate our approach on the data of the International Stroke Trial (IST) that are openly available [Sandercock et al., 2011]. IST is a randomized open treatment blinded outcome trial evaluating heparin and aspirin for acute ischemic stroke, which was conducted between 1991 and 1996 [Sandercock et al., 1997]. With a factorial design, half patients were allocated to receive unfractionated heparin (with balanced allocation between two dosages), and the other half "no heparin", and half patients were allocated to receive aspirin 300 mg daily and the other half "no aspirin". The original IST protocol specified two separate main analyses: heparin vs. no heparin and aspirin vs.no aspirin. We here focused on the aspirin vs. no aspirin analysis, and therefore considered the 19 435 patients as randomly allocated to aspirin for 9 720 of them and no aspirin for 9 715. The primary outcome was either death or dependence at 6 months after randomization. To be in-line with our previous notations, where larger values of the outcome Y were considered as beneficial, we reversed the binary outcome so that $Y = 1$ denoted patients alive and not dependent at 6 months. Overall, the proportion of patients alive and not dependent at 6 months was 37.8% with aspirin, vs. 36.5% ($p = 0.03$ after adjustment on baseline prognostic variables).

187 patients had missing outcome data in the IST database ($< 1\%$ of randomized patients), and we omitted those patients from the analyses, which were then carried out on 9 618 patients in the aspirin arm ($T = 1$) and 9 630 in the no aspirin arm ($T = 0$). We used 21 baseline covariates to model the outcome (see Table 6.1 in the Appendix). Missing covariate data were imputed once using an iterative Factorial Analysis for Mixed Data (FAMD) algorithm [Audigier et al., 2016] in order to obtain a dataset with no missing value.

Using this completed dataset, we trained random forests in classification to predict the

outcome. Each tree is grown with a bootstrap version of the sample. Each observation therefore has probability around 1/3 to not be used to grow a tree, and such an observation is called out-of-bag in the forest vocabulary. To avoid overfitting, we used the out-of bag estimate, i.e., only the trees for which the observation is out-of-bag are used to predict. This means that the observation was not used in the model that predicts its outcome. We grew 1 500 trees in order to have around 500 trees for each out-of-bag prediction. For each observation (X, T), we used the out-of-bag forest to predict the treatment effect $Z_1 = \hat{h}(X, 0)$ and $\hat{\Delta}(X) = \hat{h}(X, 1) - \hat{h}(X, 0)$, where $\hat{h}(X, T)$ is the mean vote of the forest.

We then computed $Z_3 = \hat{\Delta}(X) - \bar{\Delta}$, and did the logistic regression of Y on Z_1, T, and Z_3T. As Z_1 and Z_3 are not on the logit scale, we do not expect for their coefficients to be close to 1 but the sign of the coefficients should be positive. This is indeed the case. The coefficient for Z_1 was highly significant with $p < 10^{-16}$ and the coefficient for Z_3 was also significant with $p = 2 \times 10^{-6}$. The mean treatment effect term attains significance with $p = 0.01$ consistent with the result in the original study.

We then performed our procedure adapted to logistic regression by sampling 5 000 times from Π. We find a significant improvement under personalization with a p-value of $p = \Pi(\Theta^\star(\hat{\mathtt{Mlb}}_{0.05}) \leq 0) = 0.01$. The optimal threshold is $\gamma = 0.07$, slightly above $\alpha = 0.05$,the threshold corresponding to $\hat{\mathtt{ind}}_\alpha$. The proportion of patient with personalized treatment is $P_{\text{neg}}(\hat{\mathtt{Mlb}}_{0.05}) = 6.8\%$ and the estimated improvement in the event rate is 1.8 for 1 000, with a lower 95% confidence bound of 0.5 for 1 000. This is small but should be compared with the estimated 14 for 1 000 overall effect of aspirin. Furthermore this benefit is concentrated on the 6.8% of the population for whom aspirin should not be given (personalized set). For those patients, the probability of death or dependence at 6 months is 26 for 1 000 higher with aspirin than without aspirin, with a lower confidence bound of 8 for 1 000, which is sensibly larger than the average effect of aspirin. Therefore, it might be considered far from negligible to refrain from giving them aspirin (what we termed personalizing treatment, as compared to a standard strategy where all patients would be given aspirin).

6.8 Discussion

In this chapter, we have shown the importance of considering the uncertainty in the estimated strategy when estimating the benefit of personalization. Indeed, the optimal strategy is never known, and we have to estimate a strategy and its benefit simultaneously. In order to deal with this uncertainty, we chose to prioritize one treatment over the other, just as we prioritize the null hypothesis in significance testing. This has led us to advocate for the max lower bound strategy. It is the strategy for which the α-credible quantile of the improvement under personalization will be maximal. The personalized set it defines is often close to the region of superiority of the alternative treatment in Shuster and van Eys [1983], but can be substantially larger.

The asymmetry between treatments is induced by privileging the reference treatment that would usually be received if no treatment personalization was implemented. This implies in particular that the max lower bound strategy still assigns the reference treatment to some patients for whom it is predicted that the alternative treatment would lead to a more favorable outcome. This could seem unnatural, but we argue it is however necessary to control the risk

of implementing a strategy that leads to a worse average outcome than the reference non-personalized strategy. It is therefore crucial to ensure that a change in policy is beneficial to patients on average, in the spirit of the *primum non nocere* principle in medicine.

Our estimation of benefit is based on Bayesian arguments. The interest of considering Bayesian estimation is that it allows us to consider the estimated strategy as fixed. Nevertheless, we are interested in frequentist properties of our estimation such as type I error and coverage probabilities. Using theoretical arguments and simulations, we saw that these statistical properties appear to be respected. Unfortunately, the positivity of the credible quantile $\hat{q}_{n,\alpha}(\hat{\mathtt{mlb}}_\alpha)$ does not offer a stand-alone test because of somewhat pathological behavior when the estimated treatment effect $\hat{\beta}_2$ is negative. However, it can be combined with the test for presence of interaction, which allows to deal with this issue while not affecting the result of the test when $\hat{\beta}_2$ is positive. The resulting test is a valid test for presence of improvement under the optimal strategy. Another test for presence of improvement under the optimal strategy was proposed in Janes et al. [2014] but not studied.

In a more speculative part of our chapter, we proposed to extend our approach to the multivariate case, by creating features from the prediction of a machine learning algorithm, and applying what we have developed in the univariate case. An important consideration underlying this strategy was to avoid overfitting. A more formal extension of our approach to the multivariate case may involve the derivation of the joint distribution of prediction errors for machine learning algorithms, which needs further work. Also, it may be interesting to study the gain of using recently proposed X-learners [Künzel et al., 2017] to estimate the predictions $\hat{\mathbb{E}}[Y^1 - Y^0|X]$ (also termed conditional average treatment effects) instead of random forests, for instance. The issue of estimating credible quantiles with X-learner however also remains unsolved.

In the multivariate case, other authors have proposed different approaches to control for overfitting and uncertainty in model-based predictions of individual treatment effects $\hat{\mathbb{E}}[Y^1 - Y^0|X]$. For instance Li et al. [2016] have proposed a two- or three-step procedure. If there are sufficient data, the dataset is divided into three independent subsets. The first subset serves for estimating candidate models. The second subset is used to define the personalized set, for instance by using the lower one-sided 95% confidence bound of predictions. Then, the properties of the resulting strategy are estimated in the third subset. If data are not sufficient, the first two stages are replaced by a cross-training stage, where one subset is iteratively randomly divided into a training and an evaluation set, final predictions being then averaged over the repetitions. While this approach has been shown to have good properties in simulation studies, in particular in controlling the type I error rate, it necessitates a fair amount of data for splitting. It would be interesting to investigate how it compares with the max lower bound strategy.

It could also be interesting to study how the max lower bound strategy compares to (or could be used in conjunction with) outcome or rather residual weighted learning approaches, where the issue of determining the optimal strategy is directly handled as a weighted classification problem [Zhao et al., 2012, Zhou et al., 2017].

To approach the optimal treatment strategy and maximize outcomes in the long term, it could be recommended to conduct randomized trials focusing on patients in the region of no superiority of any treatment. Such trials may provide adequate data to refine treatment policies.

Last, all aforementioned methods that have been developed to identify policies are making

use of already collected data. Once an "optimal" strategy has been determined, it can—and should—also be evaluated in a randomized trial.

Code. The code used for the simulations and the figures is available at github.com/FelBalazard/PersMed

Acknowledgment. This chapter contains information from the IST dataset which is made available under the ODC Attribution License (`https://datashare.is.ed.ac.uk/bitstream/handle/10283/128/license_text?sequence=12&isAllowed=y`). We acknowledge the collaborative group that collected the data and funding that allowed making these data available.

6.9 Appendix

6.9.1 Proof of Theorem 1

For simplicity, it is assumed throughout that X has a bounded density with respect to the Lebesgue measure on $[x_0, x_1]$. Proofs are similar in the discrete case and left to the reader. We only consider the case where $-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha$. The proof is analogous in the other case.

We first begin by stating a lemma.

Lemma 1. *There exists a non-decreasing function g from $[x_0, x_1]$ such that $g(x_0) = x_0$, for all $\rho > x_0$, $x_0 < g(\rho) < \rho$ and*

$$\Theta^\star(\mathbb{1}_{x\geq\rho}) = -\mathbb{P}_X(X < \rho)\Delta^\star(g(\rho)).$$

Proof. Observe that

$$\Theta^\star(\mathbb{1}_{x\geq\rho}) = \mathbb{P}_X(X < \rho)(-\beta_2^\star - \beta_3^\star \mathbb{E}_X[X|X < \rho])$$

and let, for $\rho > x_0$, $g(\rho) = \mathbb{E}_X[X|X < \rho]$.

To see that g is non-decreasing, we derivate for $\rho > x_0$ and obtain

$$g'(\rho) = \frac{f_X(\rho)}{\mathbb{P}_X^2(X < \rho)} \int_{x_0}^{\rho} (\rho - x) f_X(x) dx \geq 0.$$

The inequality is strict except when $f_X(\rho) = 0$.

For all $\rho > x_0$, the inequalities $x_0 < g(\rho) < \rho$ are straightforward from the definition of g. We then have $\lim_{\rho \to x_0^+} g(\rho) = x_0$ and we prolongate by continuity so that $g(x_0) = x_0$. □

We are now ready to prove the theorem.

Proof of Theorem 1. The assumption $-\hat{\beta}_3/\sqrt{\Sigma_{3,3}} < q_\alpha$ guarantees that q_α is in the image set of the bijection defined in Proposition 2 and that we can safely define $X_\alpha = z_\Delta^{-1}(q_\alpha)$. The function $\rho \mapsto \hat{q}_{n,\alpha}(\mathbb{1}_{x\geq\rho})$ attains a maximum in $[x_0, x_1]$ as a continuous function in a compact set.

In the case $X_\alpha \leq x_0$, we use Lemma 1. Then, for all $\rho \in (x_0, -\hat{\beta}_2/\hat{\beta}_3)$, we may write

$$\Theta^\star(\mathbb{1}_{x\geq\rho}) = -\mathbb{P}_X(X < \rho)\Delta^\star(g(\rho))$$

and

$$\hat{q}_{n,\alpha}(\mathbb{1}_{x\geq\rho}) = \mathbb{P}_X(X < \rho)\left(-\hat{\Delta}(g(\rho)) + q_\alpha \mathrm{sd}_\Pi(\Delta^\star(g(\rho)))\right).$$

As $g(\rho) > x_0 \geq X_\alpha$, we have

$$-\hat{\Delta}(g(\rho)) + q_\alpha \mathrm{sd}_\Pi(\Delta^\star(g(\rho))) < 0,$$

and therefore $\hat{q}_{n,\alpha}(\mathbb{1}_{x\geq\rho}) < 0$, $\forall \rho \in (x_0, -\hat{\beta}_2/\hat{\beta}_3)$, and the maximum is 0 attained in x_0.

If $x_0 < X_\alpha$ and x_0 is in the domain of the bijection defined in Proposition 2, we can write using equation (6.5), for all $\rho \in [x_0, X_\alpha]$,

$$\begin{aligned}\frac{d\hat{q}_{n,\alpha}(\mathbb{1}_{x\geq\rho})}{d\rho}(\rho) = f_X(\rho)\big(&- \hat{\Delta}(\rho) + q_\alpha \mathrm{sd}_\Pi(\Delta^\star(\rho)) \\ &+ q_\alpha \mathrm{sd}_\Pi(\Delta^\star(\rho))(\mathrm{Cor}_\Pi(-\Delta^\star(\rho), \Theta^\star(\mathbb{1}_{x\geq\rho})) - 1)\big).\end{aligned}$$

The sum of the first two terms in the parenthesis is positive, except in X_α where it is 0, as $z_\Delta(\rho) > q_\alpha$ $\forall \rho \in [x_0, X_\alpha)$, thanks to Proposition 2. The last term is positive, except in x_0 where it is 0, as the product of two negative factors ($\alpha < 0.5$ and the correlation is smaller than 1). The parenthesis is therefore strictly positive. As $f_X(\rho)$ can be 0, we have $\frac{d\hat{q}_{n,\alpha}(\mathbb{1}_{x\geq\rho})}{d\rho}(\rho) \geq 0$, $\forall \rho \in [x_0, X_\alpha]$, and $\rho \mapsto \hat{q}_{n,\alpha}(\mathbb{1}_{x\geq\rho})$ increases on that interval. It follows that the maximum is attained for $\rho_{\max,\alpha} \geq X_\alpha$ and is positive, since $\hat{q}_{n,\alpha}(\mathbb{1}_{x\geq x_0}) = 0$ and $f_X(\rho)$ puts mass in a neighborhood of x_0.

If $x_0 < X_\alpha$ and x_0 is not in the domain of the bijection defined in Proposition 2, then the parametrization of our class of policies on the x scale is not valid. However, once parametrization is dealt with, the same arguments lead to the same result. We now detail the parametrization in this case.

If $x_0 > x_2 \stackrel{\text{def}}{=} \arg\min z_\Delta$, then we can still define a bijection with domain $[x_0, -\hat{\beta}_2/\hat{\beta}_3]$ and the calculations above apply.

However, if $x_0 < x_2$, we cannot replace $x \mapsto \mathbb{1}_{z_\Delta(x)\geq\eta}$ by $x \mapsto \mathbb{1}_{x\geq\rho}$. We have to define two bijections g_1 and g_2 with image set $[\eta_0 \stackrel{\text{def}}{=} \min(z_\Delta), z_\Delta(x_0)]$, as shown in Figure 6.7. The function g_1 is decreasing while g_2 is increasing, and we have, for all $\eta \leq z_\Delta(x_0)$, $z_\Delta(x) < \eta \Leftrightarrow g_1(\eta) < x < g_2(\eta)$.

The mean improvement of this strategy is then

$$\hat{\Theta}(\mathbb{1}_{z_\Delta(x)\geq\eta}) = -\int_{g_1(\eta)}^{g_2(\eta)} \hat{\Delta}(x) f_X(x) dx,$$

and its derivative

$$\frac{d\hat{\Theta}}{d\eta}(\eta) = -g_2'(\eta)\hat{\Delta}(g_2(\eta)) f_X(g_2(\eta)) + g_1'(\eta)\hat{\Delta}(g_1(\eta)) f_X(g_1(\eta)).$$

The variance under Π is

$$\mathrm{Var}_\Pi[\Theta^\star(\mathbb{1}_{z_\Delta(x)\geq\eta})] = \mathbb{E}_\Pi\Big[\Big(\int_{g_1(\eta)}^{g_2(\eta)} (\Delta^\star(x) - \hat{\Delta}(x)) f_X(x) dx\Big)^2\Big] \stackrel{\text{def}}{=} v(\eta).$$

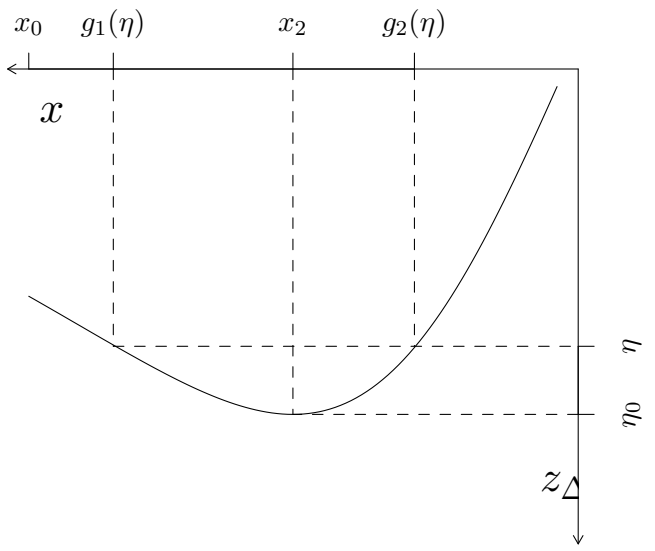

Figure 6.7: Double parametrization of z_Δ.

Thus,

$$\begin{aligned}\frac{dv}{d\eta}(\eta) &= 2g_2'(\eta)f_X(g_2(\eta))\mathrm{Cov}_\Pi(-\Delta^\star(g_2(\eta)), \Theta^\star(\mathbb{1}_{z_\Delta(x)\geq\eta})) \\ &\quad - 2g_1'(\eta)f_X(g_1(\eta))\mathrm{Cov}_\Pi(-\Delta^\star(g_1(\eta)), \Theta^\star(\mathbb{1}_{z_\Delta(x)\geq\eta})).\end{aligned}$$

From this, we obtain the derivative of our quantile with respect to η, that is,

$$\begin{aligned}&\frac{d\hat{q}_{n,\alpha}(\mathbb{1}_{z_\Delta(x)\geq\eta})}{d\eta}(\eta) \\ &\quad = g_2'(\eta)f_X(g_2(\eta))\Big(- \hat{\Delta}(g_2(\eta)) \\ &\qquad\qquad + q_\alpha \mathrm{sd}_\beta(\hat{\Delta}(g_2(\eta)))\mathrm{Cor}_\Pi(-\Delta^\star(g_2(\eta)), \Theta^\star(\mathbb{1}_{z_\Delta(x)\geq\eta}))\Big) \\ &\quad - g_1'(\eta)f_X(g_1(\eta))\Big(- \hat{\Delta}(g_1(\eta)) \\ &\qquad\qquad + q_\alpha \mathrm{sd}_\beta(\hat{\Delta}(g_2(\eta)))\mathrm{Cor}_\Pi(-\Delta^\star(g_1(\eta)), \Theta^\star(\mathbb{1}_{z_\Delta(x)\geq\eta}))\Big).\end{aligned}$$

As g_2 is increasing and g_1 is decreasing, both terms have the same sign, and the same argument as above shows that $\frac{d\hat{q}_{n,\alpha}(\mathbb{1}_{z_\Delta(x)\geq\eta})}{d\eta}(\eta) > 0$ for $\eta \leq q_\alpha$. □

6.9.2 Covariates used for the analysis of the IST data

Table 6.1: Covariates used for analysis of the IST data.

Variable
Delay between stroke and randomization in hours
Conscious state at randomization
Patient sex
Age in years
Symptoms noted on waking
Atrial fibrillation
CT before randomization
Infarct visible on CT
Heparin within 24 hours prior to randomization
Aspirin within 3 days prior to randomization
Systolic blood pressure at randomization (mmHg)
Face deficit
Arm/hand deficit
Leg/foot deficit
Dysphasia
Hemianopia
Visuospatial disorder
Brainstem/cerebellar signs
Other deficit
Stroke subtype
Local time - hours of randomization

Conclusion

Dans cette thèse, nous avons développé deux approches différentes de la médecine personnalisée. La première pourrait être qualifiée plus justement de médecine préventive. L'idée est que la génétique permet d'identifier une population à risque à laquelle on pourra proposer une intervention pour limiter le risque de développer la maladie. Nous avons illustré cette idée autour des données de l'étude Isis-Diab, consacrée au diabète de type 1.

Dans le chapitre 2, nous avons étudié la première étape : la prédiction du risque génétique. En particulier, nous avons proposé de prendre en compte l'information de phase afin d'améliorer les prédictions. Nous avons également répliqué une estimation du risque génétique du diabète de type 1 sur les patients d'Isis-Diab.

Afin de pouvoir proposer une intervention à la population à risque, il faut identifier les facteurs environnementaux qui sont liés à la maladie. Pour cela, nous avons analysé dans le chapitre 4, les questionnaires environnementaux de l'étude Isis-Diab. Cela nous a permis d'identifier une liste de facteurs à étudier plus avant.

L'analyse des études observationnelles doit viser à limiter autant que possible la confusion. Un moyen d'y parvenir est d'apparier les patients avec des témoins similaires. Cet appariement doit par la suite être pris en compte en utilisant des méthodes d'analyse appropriées. Dans le chapitre 3, nous avons prouvé que deux de ces méthodes, le test de Hotelling apparié et la régression logistique conditionnelle sont asymptotiquement équivalents.

Comme nous avions à notre disposition les données de patients avec un risque génétique et des facteurs environnementaux potentiellement liés à la maladie, nous nous sommes demandés s'il était possible d'utiliser ces données afin de confirmer les associations environnementales. Dans ce but, nous avons proposé, au chapitre 5, une nouvelle méthodologie : DAC pour Disease As Collider qui se base sur le biais de collision. Cela nous a amené à étudier plus précisément le modèle d'étude cas-seulement et l'influence du biais de collision dans celle-ci. Comme DAC ne se sert que des données des patients, l'information qu'elle apporte est indépendante du choix des témoins qui peut avoir influencé l'étude cas-témoins. Malheureusement, notre méthodologie n'avait pas de puissance statistique dans le cadre d'Isis-Diab.

La deuxième approche de la médecine personnalisée que nous avons envisagée est la personnalisation des traitements. Dans le chapitre 6, nous avons étudié comment cette personnalisation pouvait être décidée à partir d'essais cliniques. Nos contributions y sont méthodologiques : nous avons proposé une méthode pour estimer le bénéfice d'une politique de personnalisation ainsi qu'une politique qui maximise la confiance qu'on peut avoir en elle. Ces nouvelles méthodes ont permis de montrer que l'aspirine était néfaste pour une partie des patients traités pour une crise cardiaque.

Un axe de lecture de cette thèse est l'utilité des méthodes d'apprentissage statistique pour l'épidémiologie. Dans le chapitre 6, les forêts aléatoires nous ont permis de généraliser au cas multivarié ce que nous avons fait pour le cas univarié. Dans ce cadre, un atout essentiel des forêts aléatoires est d'offrir pour chaque observation une prédiction hors-du-sac (out-of-bag) qui ne se sert pas de l'observation. Cela permet d'avoir des prédictions qui ne sont pas affectées par le problème du surapprentissage.

Bien sûr, au chapitre 2, l'apprentissage statistique est dans son élément puisque qu'il s'agit de prédire un risque génétique. Malheureusement, la prise en compte de structure bio-logique afin d'adapter les algorithmes aux données génétiques n'a pas encore fait ses preuves. L'accès imparfait aux plus grands jeux de données fait que d'autres méthodes qui peuvent se contenter des résultats SNP par SNP des études GWAS, comme les scores de risque polygéniques, restent d'actualité.

Dans le cas de l'identification des facteurs environnementaux, un contrôle strict des faux positifs, et donc une approche basée sur les corrections pour tests multiples, reste primordial. Le principal problème des études observationnelles est la possibilité de biais dus à des variables de confusion et il s'agit donc de contrôler ces biais le mieux possible. Ainsi, utiliser des algorithmes de machine learning pour prédire le label cas ou témoin n'a que peu de sens car cela ne ferait que capturer ces biais. Toutefois, l'apprentissage statistique peut être mis au service du contrôle et de l'étude de ces biais. Ainsi, dans le chapitre 4, nous avons utilisé les forêts aléatoires afin d'estimer les scores de propension dans une des analyses. Un autre emploi, plus original, des forêts aléatoires a été de prédire l'âge des patients et témoins à partir du questionnaire. Cela a permis d'étudier si les témoins avaient bien pris en compte l'âge de référence : l'âge à partir duquel ils ne devaient plus prendre en compte leurs expériences pour remplir le questionnaire et qui correspond à l'âge au diagnostic du patient apparié. Cela a permis de montrer qu'il y a effectivement des témoins qui n'ont pas tenu compte de cet âge de référence.

Bibliography

1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye. Sparsnp: Fast and memory-efficient analysis of all snps for phenotype prediction. *BMC Bioinformatics*, 13(1):88, 2012.

G. Abraham, J. Tye-Din, O. Bhalala, A. Kowalczyk, J. Zobel, and M. Inouye. Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genetics*, 10(2): e1004137, 2014.

P. Achenbach, E. Bonifacio, K. Koczwara, and A.-G. Ziegler. Natural history of type 1 diabetes. *Diabetes*, 54(suppl 2):S25–S31, 2005.

D.E. Adkins, S.L. Clark, W.E. Copeland, M. Kennedy, K. Conway, A. Angold, H. Maes, Y. Liu, G. Kumar, A. Erkanli, et al. Genome-wide meta-analysis of longitudinal alcohol consumption across youth and early adulthood. *Twin Research and Human Genetics*, 18(4):335–347, 2015.

A. Agresti and M. Kateri. *Categorical Data Analysis*. Springer, Berlin, Heidelberg, 2011.

H.K. Åkerblom, TRIGR Study Group, et al. The trial to reduce iddm in the genetically at risk (trigr) study: recruitment, intervention and follow-up. *Diabetologia*, 54(3):627–633, 2011.

P.S. Albert, D. Ratnasinghe, J. Tangrea, and S. Wacholder. Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology*, 154(8): 687–693, 2001.

N.E. Allen, C. Sudlow, T. Peakman, R. Collins, et al. Uk biobank data: come and get it. *Science Translational Medicine*, 2014.

M.A. Atkinson, G.S. Eisenbarth, and A.W. Michels. Type 1 diabetes. *The Lancet*, 383(9911): 69–82, 2014.

V. Audigier, F. Husson, and J. Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10:5–26, 2016.

P.C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

J.-F. Bach. The effect of infections on susceptibility to autoimmune and allergic diseases. *New England Journal of Medicine*, 347(12):911–920, 2002.

F. Balazard, S. Le Fur, S. Valtat, Isis Diab collaborative group, A.-J. Valleron, and P. Bougnères. Association of environmental markers with childhood type 1 diabetes mellitus revealed by a long questionnaire on early life exposures and lifestyle in a case-control study. *BMC Public Health*, 2016.

Félix Balazard. Haplotype based genetic risk estimation for complex diseases. *PeerJ PrePrints*, 2016.

Jennifer M. Barker, Katherine J. Barriga, Liping Yu, Dongmei Miao, Henry A. Erlich, Jill M. Norris, George S. Eisenbarth, and Marian Rewers. Prediction of autoantibody positivity and progression to type 1 diabetes: Diabetes autoimmunity study in the young (daisy). *The Journal of Clinical Endocrinology & Metabolism*, 89(8):3896–3902, 2004. doi: 10.1210/jc.2003-031887.

P.J. Barter, M. Caulfield, M. Eriksson, S.M. Grundy, J. Kastelein, M. Komajda, J. Lopez-Sendon, L. Mosca, J.-C. Tardif, D.D. Waters, et al. Effects of torcetrapib in patients at high risk for coronary events. *New England Journal of Medicine*, 357(21):2109–2122, 2007.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

A. Beyerlein, F. Wehweck, A.-G. Ziegler, and M. Pflueger. Respiratory infections in early life and the development of islet autoimmunity in children at increased type 1 diabetes risk: evidence from the babydiet study. *JAMA Pediatrics*, 167(9):800–807, 2013.

E. Bonifacio, A. Beyerlein, M. Hippich, C. Winkler, K. Vehik, M. Weedon, M. Laimighofer, et al. Genetic scores to stratify risk of developing multiple islet autoantibodies and type 1 diabetes: A prospective study in children. *PLoS Medicine*, 15(4):1–18, 04 2018.

V. Botta, G. Louppe, P. Geurts, and L. Wehenkel. Exploiting snp correlations within random forest for genome-wide association studies. *PLOS ONE*, 9(4), 2014.

A.-L. Boulesteix, S. Janitza, J. Kruppa, and I.R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507, 2012.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001a.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001b.

N.E. Breslow and N.E. Day. *Statistical methods in cancer research. The analysis of case-control studies*. International Agency for Research on Cancer, Lyon, 1981.

N.E. Breslow, N.E. Day, K.T. Halvorsen, R.L. Prentice, and C. Sabai. Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, 108:299–307, 1978.

J. Brinkley, A.A. Tsiatis, and K.J. Anstrom. A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics*, 66:512–522, 2010.

G. Broberg and N. Roll-Hansen. *Eugenics and the welfare state: Norway, Sweden, Denmark, and Finland*. Michigan State University Press, 2005.

R. Bumgarner. Overview of dna microarrays: types, applications, and their future. *Current Protocols in Molecular Biology*, pages 22–1, 2013.

P. Burton, D. Clayton, L. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. Kwiatkowski, M. McCarthy, W. Ouwehand, N. Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

T. Cai, L. Tian, P.H. Wong, and L.J. Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12:270–282, 2011.

S. Caillat-Zucman, H.-J. Garchon, J. Timsit, Roger. Assan, C. Boitard, I. Djilali-Saiah, P. Bougnères, and J.F. Bach. Age-dependent hla genetic heterogeneity of type 1 insulin-dependent diabetes mellitus. *The Journal of Clinical Investigation*, 90(6):2242–2250, 1992.

R. Capdeville, E. Buchdunger, J. Zimmermann, and A. Matter. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nature Reviews Drug Discovery*, 1:493–502, 2002.

C.R. Cardwell, L.C. Stene, J. Ludvigsson, J. Rosenbauer, O. Cinek, J. Svensson, F. Perez-Bravo, et al. Breast-feeding and childhood-onset type 1 diabetes: a pooled analysis of individual participant data from 43 observational studies. *Diabetes Care*, page DC_120438, 2012.

N.M. Chapman, K. Coppieters, M. Von Herrath, and S. Tracy. The microbiology of human hygiene and its impact on type 1 diabetes. *Islets*, 4(4):253–261, 2012.

P.B. Chapman, A. Hauschild, C. Robert, J.B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, D. Hogg, P. Lorigan, C. Lebbe, T. Jouary, D. Schadendorf, A. Ribas, S.J. O'Day, J.A. Sosman, J.M. Kirkwood, A.M.M. Eggermont, B. Dreno, K. Nolop, J. Li, B. Nelson, J. Hou, R.J. Lee, K.T. Flaherty, G.A. McArthur, and BRIM-3 Study Group. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *New England Journal of Medicine*, 364:2507–2516, 2011.

H.P. Chase, D. Boulware, H. Rodriguez, D. Donaldson, S. Chritton, L. Rafkin-Mervis, J. Krischer, J.S. Skyler, M. Clare-Salzler, and Type 1 Diabetes TrialNet Nutritional Intervention to Prevent (NIP) Type 1 Diabetes Study Group. Effect of docosahexaenoic acid supplementation on inflammatory cytokine levels in infants at high genetic risk for type 1 diabetes. *Pediatric Diabetes*, 16(4):271–279, 2015.

N. Chatterjee, J. Shi, and M. García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17(7):392, 2016.

S. Chen and G. Parmigiani. Meta-analysis of brca1 and brca2 penetrance. *Journal of Clinical Oncology*, 25(11):1329, 2007.

Collaborative Group on Hormonal Factors in Breast Cancer et al. Alcohol, tobacco and breast cancer–collaborative reanalysis of individual data from 53 epidemiological studies, including 58 515 women with breast cancer and 95 067 women without the disease. *British Journal of Cancer*, 87(11):1234, 2002.

Collaborative Group on Hormonal Factors in Breast Cancer et al. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The Lancet Oncology*, 13(11):1141–1151, 2012.

S.R. Cole, R.W. Platt, E.F. Schisterman, H. Chu, D. Westreich, D. Richardson, and C. Poole. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39 (2):417–420, 2009.

M. Cornelis, E. Byrne, T. Esko, M. Nalls, A. Ganna, N. Paynter, K.L. Monda, N. Amin, et al. Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Molecular Psychiatry*, 20(5):647, 2015.

S.S. Coughlin. Recall bias in epidemiologic studies. *Journal of Clinical Epidemiology*, 43(1): 87–91, 1990.

D.R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B*, 34:187–220, 1972.

J. Craig et al. Complex diseases: research and applications. *Nature Education*, 1(1), 2008.

G. Dahlquist. Can we slow the rising incidence of childhood-onset autoimmune diabetes? the overload hypothesis. *Diabetologia*, 49(1):20–24, 2006.

G. Davey Smith and G. Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–R98, 2014.

N.E. Day and D.P. Byar. Testing hypotheses in case-control studies-equivalence of mantel-haenszel statistics and logit score tests. *Biometrics*, 35:623–630, 1979.

O. Delaneau, J.-F. Zagury, and J. Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10(1):5–6, 2013.

DIAMOND Project Group. Incidence and trends of childhood type 1 diabetes worldwide 1990–1999. *Diabetic Medicine*, 23(8):857–866, 2006.

J.-Y. Dong, W. Zhang, J.J. Chen, Z.-L. Zhang, S.-F. Han, and L.-Q. Qin. Vitamin d intake and risk of type 1 diabetes: a meta-analysis of observational studies. *Nutrients*, 5(9):3551–3562, 2013.

M.A. D'angeli, E. Merzon, L.F. Valbuena, D. Tirschwell, C.A. Paris, and B.A. Mueller. Environmental factors associated with childhood-onset type 1 diabetes mellitus: an exploration of the hygiene and overload hypotheses. *Archives of Pediatrics & Adolescent Medicine*, 164(8): 732–738, 2010.

F.M. Egro. Why is type 1 diabetes increasing? *Journal of Molecular Endocrinology*, 51(1): R1–R13, 2013.

R.F. Engle. Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of Econometrics*, 2:775–826, 1984.

J. Evans, I. Goldfine, B. Maddux, and G. Grodsky. Are oxidative stress- activated signaling pathways mediators of insulin resistance and β-cell dysfunction? *Diabetes*, 52(1):1–8, 2003.

M. Ewertz, S. Duffy, H.-O. Adami, G. Kvåle, E. Lund, O. Meirik, A. Mellemgaard, I. Soini, and H. Tulinius. Age at first birth, parity and risk of breast cancer: A meta-analysis of 8 studies from the nordic countries. *International Journal of Cancer*, 46(4):597–603, 1990.

Global Alliance for Genomics and Health. A federated ecosystem for sharing genomic, clinical data. *Science*, 352(6291):1278–1280, 2016.

D.S. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76, 1965.

K. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. Housley, S. Beik, N. Shoresh, H. Whitton, R. Ryan, A. Shishkin, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337, 2015.

Jose A. Fernandes F. and B.E. Shapiro. Tay-sachs disease. *Archives of Neurology*, 61(9):1466–1468, 2004.

G.P. Forlenza and M. Rewers. The epidemic of type 1 diabetes: what is it telling us? *Current Opinion in Endocrinology, Diabetes and Obesity*, 18(4):248–251, 2011.

J.C. Foster, J.M.G. Taylor, and S.J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30:2867–2880, 2011.

B. Freidlin and E.L. Korn. Biomarker enrichment strategies: Matching trial design to biomarker credentials. *Nature Reviews Clinical Oncology*, 11:81–90, 2014.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

H. Furberg, Y. Kim, J. Dackor, E. Boerwinkle, N. Franceschini, D. Ardissino, L. Bernardinelli, P. Mannucci, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*, 42(5):441, 2010.

S.H. Gage, G. Davey-Smith, J.J. Ware, J. Flint, and M.R. Munafò. G= e: What gwas can tell us about the environment. *PLoS Genetics*, 12(2):e1005765, 2016.

E. Gale, European Nicotinamide Diabetes Intervention Trial (ENDIT) Group, et al. European nicotinamide diabetes intervention trial (endit): a randomised controlled trial of intervention before the onset of type 1 diabetes. *The Lancet*, 363(9413):925–931, 2004.

E.A. Gale. A missing link in the hygiene hypothesis? *Diabetologia*, 45(4):588–594, 2002.

F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

N. Gatto, U. Campbell, A. Rundle, and H. Ahsan. Further development of the case-only design for assessing gene–environment interaction: evaluation of and adjustment for bias. *International Journal of Epidemiology*, 33(5):1014–1024, 2004.

N.W. Gillham. *A life of Sir Francis Galton: From African exploration to the birth of eugenics*. Oxford University Press, 2001.

J. Green, D. Casabonne, and R. Newton. Coxsackie b virus serology and type 1 diabetes mellitus: a systematic review of published case-control studies. *Diabetic Medicine*, 21(6):507–514, 2004.

S. Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, 2003.

S. Greenland, J. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46, 1999.

N. Gujral, H.J. Freeman, and A. Thomson. Celiac disease: prevalence, diagnosis, pathogenesis and treatment. *World Journal of Gastroenterology*, 18(42):6036, 2012.

J. Guo and Z. Geng. Collapsibility of logistic regression coefficients. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 263–267, 1995.

N.S. Hall. Ra fisher and his advocacy of randomization. *Journal of the History of Biology*, 40 (2):295–325, 2007.

P. Hall and C.C. Heyde. *Martingale limit theory and its application*. Academic Press., New York, 1980.

M.A. Hamburg and F.S. Collins. The path to personalized medicine. *New England Journal of Medicine*, 363:301–304, 2010.

T. Harder, K. Roepke, N. Diller, Y. Stechling, J.W. Dudenhausen, and A. Plagemann. Birth weight, early weight gain, and subsequent risk of type 1 diabetes: systematic review and meta-analysis. *American Journal of Epidemiology*, 169(12):1428–1436, 2009.

V. Harjutsalo, R. Sund, M. Knip, and P.-H. Groop. Incidence of type 1 diabetes in finland. *JAMA*, 310(4):427–428, 2013.

D.L. Hartl and A.G. Clark. *Principles of Population Genetics*, volume 116. Sinauer associates Sunderland, 1997.

B.N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, 2009.

J. Howson, J. Cooper, D. Smyth, N. Walker, H. Stevens, J.-X. She, G.S. Eisenbarth, M. Rewers, J.A. Todd, B. Akolkar, et al. Evidence of gene-gene interaction and age-at-diagnosis effects in type 1 diabetes. *Diabetes*, 61(11):3012–3017, 2012.

C. Hu, H. Ding, Y. Li, J. Pearson, X. Zhang, R. Flavell, F. Wong, and L. Wen. Nlrp3 deficiency protects from type 1 diabetes through the regulation of chemotaxis into the pancreatic islets. *Proceedings of the National Academy of Sciences*, 112(36):11318–11323, 2015.

S. Hummel, M. Pflüger, M. Hummel, E. Bonifacio, and A.-G. Ziegler. Primary dietary intervention study to reduce the risk of islet autoimmunity in children at increased risk for type 1 diabetes: the babydiet study. *Diabetes Care*, 34(6):1301–1305, 2011.

V. Hyttinen, J. Kaprio, L. Kinnunen, M. Koskenvuo, and J. Tuomilehto. Genetic liability of type 1 diabetes and the onset age among 22,650 young finnish twin pairs a nationwide follow-up study. *Diabetes*, 52(4):1052–1055, 2003.

M. Inouye, G. Abraham, C. Nelson, A. Wood, M. Sweeting, F. Dudbridge, et al. Genomic risk prediction of coronary artery disease in nearly 500,000 adults: implications for early screening and primary prevention. *bioRxiv*, 2018.

J.O. Irwin. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene*, 47:188–189, 1949.

S.S. Jamuar and E.-C. Tan. Clinical application of next-generation sequencing for mendelian diseases. *Human Genomics*, 9(1):10, 2015.

H. Janes, M.D. Brown, Y. Huang, and M.S. Pepe. An approach to evaluating and comparing biomarkers for patient treatment selection. *The International Journal of Biostatistics*, 10: 99–121, 2014.

C. Kang, H. Janes, and Y. Huang. Combining biomarkers to optimize patient treatment recommendations. *Biometrics*, 70:695–707, 2014.

J. Kang, S. Kugathasan, M. Georges, H. Zhao, and J. Cho. Improved risk prediction for crohn's disease with a multi-locus approach. *Human Molecular Genetics*, 20(12):2435–2442, 2011.

M.J. Khoury. Dealing with the evidence dilemma in genomics and personalized medicine. *Clinical Pharmacology and Therapeutics*, 87:635–638, 2010.

M.J. Khoury and W.D. Flanders. Nontraditional epidemiologic approaches in the analysis of gene environment interaction: Case-control studies with no controls! *American Journal of Epidemiology*, 144(3):207–213, 1996.

Y. Kim, W. Wang, M. Okla, I. Kang, R. Moreau, and S. Chung. Suppression of nlrp3 inflammasome by γ-tocotrienol ameliorates type 2 diabetes. *Journal of Lipid Research*, 57(1):66–76, 2016.

M. Knip and O. Simell. Environmental triggers of type 1 diabetes. *Cold Spring Harbor Perspectives in Medicine*, 2(7):a007690, 2012.

M. Knip, H.K. Åkerblom, D. Becker, et al. Hydrolyzed infant formula and early β-cell autoimmunity: a randomized clinical trial. *JAMA*, 311(22):2279–2287, 2014.

M. Knip, H.K. Åkerblom, E. Al Taji, D. Becker, et al. Effect of hydrolyzed infant formula vs conventional formula on risk of type 1 diabetes: The trigr randomized clinical trial. *JAMA*, 319(1):38–48, 2018.

A. Kuhad, M. Bishnoi, V. Tiwari, and K. Chopra. Suppression of nf-$\kappa\beta$ signaling pathway by tocotrienol can prevent diabetes associated cognitive deficits. *Pharmacology Biochemistry and Behavior*, 92(2):251–259, 2009.

Sören Künzel, Jasjeet Sekhon, Peter Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv*, 2017.

S. Le Fur, P. Bougnères, and A.-J. Valleron. Comparison of a french pediatric type 1 diabetes cohort's responders and non-responders to an environmental questionnaire. *BMC Public Health*, 14(1):1241, 2014.

L. Lello, S.G. Avery, L. Tellier, A. Vazquez, G. de los Campos, and S. Hsu. Accurate genomic prediction of human height. *bioRxiv*, 2017.

J. Li, L. Zhao, L. Tian, T. Cai, B. Claggett, A. Callegaro, B. Dizier, B. Spiessens, F. Ulloa-Montoya, and L.J. Wei. A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies. *Biometrics*, 72: 877–887, 2016.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

A. Like, D. Guberski, and L. Butler. Influence of environmental viral agents on frequency and tempo of diabetes mellitus in bb/wor rats. *Diabetes*, 40(2):259–262, 1991.

I. Lipkovich, A. Dmitrienko, J. Denne, and G. Enas. Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30:2601–2621, 2011.

P.A. Lombardo. *A century of eugenics in America: from the Indiana experiment to the human genome era*. Indiana University Press, 2011.

M.H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247, 2010.

K.E. Malone, J.R. Daling, D.R. Doody, L. Hsu, L. Bernstein, R.J. Coates, P.A. Marchbanks, M.S. Simon, J.A. McDonald, S.A. Norman, et al. Prevalence and predictors of brca1 and brca2 mutations in a population-based study of breast cancer in white and black american women ages 35 to 64 years. *Cancer Research*, 66(16):8297–8308, 2006.

S.J. Mandrekar and D.J. Sargent. Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. *Journal of Clinical Oncology*, 27:4027–4034, 2009.

T. Manolio, F. Collins, N. Cox, D. Goldstein, L. Hindorff, D. Hunter, M. I McCarthy, E. Ramos, L. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute*, 22:719–748, 1959.

J.H. McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, 2009.

Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947.

N.E. Miller, D.S. Thelle, O.H. Forde, and O.D. Mjos. The tromso heart study: high-density lipoportein and coronary heart disease: a prospective case-control study. *The Lancet*, 309 (8019):965–968, 1977.

B. Modan, P. Hartge, G. Hirsh-Yechezkel, A. Chetrit, F. Lubin, et al. Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a brca1 or brca2 mutation. *New England Journal of Medicine*, 345(4):235–240, 2001.

P.G. Moorman, E.S. Iversen, P.K. Marcom, J.R. Marks, F. Wang, E. Lee, G. Ursin, T.R. Rebbeck, S.M. Domchek, B. Arun, et al. Evaluation of established breast cancer risk factors as modifiers of brca1 or brca2: a multi-center case-only analysis. *Breast Cancer Research and Treatment*, 124(2):441–451, 2010.

J.M. Norris, X. Yin, M.M. Lamb, K. Barriga, J. Seifert, M. Hoffman, et al. Omega-3 polyunsaturated fatty acid intake and islet autoimmunity in children at increased risk for type 1 diabetes. *JAMA*, 298(12):1420–1428, 2007.

R.A. Oram, K. Patel, A. Hill, B. Shields, T.J. McDonald, A. Jones, A.T. Hattersley, and M.N. Weedon. A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. *Diabetes Care*, 39(3):337–344, 2016.

C. Patel, J. Bhattacharya, and A.J. Butte. An environment-wide association study (ewas) on type 2 diabetes mellitus. *PLOS ONE*, 5(5):1–10, 05 2010.

K.A. Patel, R.A. Oram, S.E. Flanagan, E. De Franco, K. Colclough, S. Ellard, M.N. Weedon, A.T. Hattersley, et al. Type 1 diabetes genetic risk score: a novel tool to discriminate monogenic and type 1 diabetes. *Diabetes*, page db151690, 2016.

J. Pearl. *Causality*. Cambridge university press, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

W. Piegorsch, C. Weinberg, and J. Taylor. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13 (2):153–162, 1994.

C. Piffaretti, S. Fosse Edorh, R. Coutant, C. Choleau, S. Guilmin Crepon, and L. Mandereau Bruno. Incidence du diabète de type 1 chez l'enfant en France en 2013-2015, à partir du système national des données de santé (SNDS). Variations régionales. *Numéro thématique. Journée mondiale du diabète 2017.*, (27-28):571–578, November 2017.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. De Bakker, M. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

M. Qian and S.A. Murphy. Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39:1180–1210, 2011.

S.M. Rappaport, D.K. Barupal, D. Wishart, P. Vineis, and A. Scalbert. The blood exposome and its role in discovering causes of disease. *Environmental Health Perspectives*, 122(8):769, 2014.

T. Rasmussen, E. Witsø, G. Tapia, L.C. Stene, and K.S. Rønningen. Self-reported lower respiratory tract infections and development of islet autoimmunity in children with the type 1 diabetes high-risk hla genotype: the midia study. *Diabetes Metabolism Research and Reviews*, 27(8):834–837, 2011.

M.J. Redondo, J. Jeffrey, P.R. Fain, G.S. Eisenbarth, and T. Orban. Concordance for islet autoimmunity among monozygotic twins. *New England Journal of Medicine*, 359(26):2849–2850, 2008.

P. Rosenbaum. *Observational Studies*. Springer, New York, 2002.

P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983a.

P.R. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983b.

D. Rubin. Matching to remove bias in observational studies. *Biometrics*, 29:159–183, 1973.

P.A.G. Sandercock, R. Collins, C. Counsell, B. Farrell, R. Peto, J. Slattery, and C. Warlow. The International Stroke Trial (IST): A randomized trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *The Lancet*, 349:1569–1581, 1997.

P.A.G. Sandercock, M. Niewada, and A. Członkowska. The International Stroke Trial database. *Trials*, 12:101, 2011.

S. Schmidt and D.J. Schaid. Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology*, 150(8):878–885, 1999.

D. Schneider and M. von Herrath. Potential viral pathogenic mechanism in human type 1 diabetes. *Diabetologia*, 57(10):2009–2018, 2014.

G.G. Schwartz, A.G. Olsson, M. Abt, C.M. Ballantyne, P.J. Barter, J. Brumm, B.R. Chaitman, I.M. Holme, D. Kallend, L.A. Leiter, et al. Effects of dalcetrapib in patients with a recent acute coronary syndrome. *New England Journal of Medicine*, 367(22):2089–2099, 2012.

J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

J. Shen, L. Wang, S. Daignault, D.E. Spratt, T.M. Morgan, and J.M.G. Taylor. Estimating the optimal personalized treatment strategy based on selected variables to prolong survival via random survival forest with weighted bootstrap. *Journal of Biopharmaceutical Statistics*, 21: 1–20, 2017.

Y. Shen and T. Cai. Identifying predictive markers for personalized treatment selection. *Biometrics*, 72:1017–1025, 2016.

J. Shuster and J. van Eys. Interaction between prognostic factors and treatment. *Controlled Clinical Trials*, 4:209–214, 1983.

R.M. Simon, S. Paik, and D.F. Hayes. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *Journal of the National Cancer Institute*, 101:1446–1452, 2009.

M. Simpson, H. Brady, X. Yin, J. Seifert, K. Barriga, M. Hoffman, T. Bugawan, A.E. Barón, R.J. Sokol, G. Eisenbarth, et al. No association of vitamin d intake or 25-hydroxyvitamin d levels in childhood with risk of islet autoimmunity and type 1 diabetes: the diabetes autoimmunity study in the young (daisy). *Diabetologia*, 54(11):2779, 2011.

D.P. Singal and M.A. Blajchman. Histocompatibility (hl-a) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus. *Diabetes*, 22(6):429–432, 1973.

J.S. Skyler. Primary and secondary prevention of type 1 diabetes. *Diabetic Medicine*, 30(2): 161–169, 2013.

M. Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477, 2008.

J.K. Snell-Bergeon, J. Smith, F. Dong, A. Barón, K. Barriga, J. M Norris, and M. Rewers. Early childhood infections and the risk of islet autoimmunity: the diabetes autoimmunity study in the young (daisy). *Diabetes Care*, 35(12):2553–2558, 2012.

J. Snoep, A. Morabia, S. Hernández-Díaz, M.A. Hernán, and J.P. Vandenbroucke. Commentary: a structural approach to berkson's fallacy and a guide to a history of opinions about it. *International Journal of Epidemiology*, 43(2):515–521, 2014.

D. Speed, N. Cai, M.R. Johnson, S. Nejentsev, D.J. Balding, UCLEB Consortium, et al. Reevaluation of snp heritability in complex human traits. *Nature Genetics*, 49(7):986, 2017.

J. Stamatoyannopoulos. Connecting the regulatory genome. *Nature Genetics*, 48(5):479, 2016.

R. Tattersall. *Diabetes: the biography*. Oxford University Press, 2009.

TEDDY Study Group. The environmental determinants of diabetes in the young (teddy) study: study design. *Pediatric Diabetes*, 8(5):286–298, 2007.

P.I. Terasaki. A brief history of hla. *Immunologic Research*, 38(1-3):139–148, 2007.

R. Tewhey, V. Bansal, A. Torkamani, E. Topol, and N. Schork. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223, 2011.

L. Tian, L. Zhao, and L.J. Wei. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15:222–233, 2014.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996a.

R. Tibshirani. A comparison of some error estimates for neural network models. *Neural Computation*, 8:152–163, 1996b.

A. Torkamani, N.E. Wineinger, and E.J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, page 1, 2018.

J. Tyrrell, A.R. Wood, R.M. Ames, H. Yaghootkar, et al. Gene–obesogenic environment interactions in the uk biobank study. *International Journal of Epidemiology*, 46(2):559–575, 2017.

US Food and Drug Administration. Paving the way for personalized medicine: FDA's role in a new era of medical product development. Technical report, US Food and Drug Administration, Silver Spring, 2013. URL `https://www.fda.gov/downloads/scienceresearch/specialtopics/personalizedmedicine/ucm372421.pdf`.

A.W. van der Vaart. *Bayes Procedures*, pages 138–152. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

T. Vatanen, A.D. Kostic, E. d'Hennezel, H. Siljander, et al. Variation in microbiome lps immunogenicity contributes to autoimmunity in humans. *Cell*, 165(4):842–853, 2016.

K.C. Verbeeten, C.E. Elks, D. Daneman, and K.K. Ong. Association between childhood obesity and subsequent type 1 diabetes: a systematic review and meta-analysis. *Diabetic Medicine*, 28(1):10–18, 2011.

P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, 2008.

P.M. Visscher and J.B. Walsh. Commentary: Fisher 1918: the foundation of the genetics and analysis of complex traits. *International Journal of Epidemiology*, 2017.

A. Vivot, I. Boutron, P. Ravaud, and R. Porcher. Guidance for pharmacogenomic biomarker testing in labels of FDA-approved drugs. *Genetics in Medicine*, 17:733–738, 2015.

A. Vivot, J. Li, J.D. Zeitoun, S. Mourah, P. Crequit, P. Ravaud, and R. Porcher. Pharmacogenomic biomarkers as inclusion criteria in clinical trials of oncology-targeted drugs: A mapping of ClinicalTrials.gov. *Genetics in Medicine*, 18:796–805, 2016.

B.F. Voight, G.M. Peloso, M. Orho-Melander, R. Frikke-Schmidt, M. Barbalic, M.K. Jensen, G. Hindy, H. Hólm, E.L. Ding, T. Johnson, et al. Plasma hdl cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*, 380(9841):572–580, 2012.

S. Wager, T. Hastie, and B. Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651, 2014.

B. Wang, W.J. Canestaro, and N.K. Choudhry. Clinical evidence supporting pharmacogenomic biomarker testing provided in US Food and Drug Administration drug labels. *JAMA Internal Medicine*, 174:1938–1944, 2014.

Z. Wei, K. Wang, H. Qu, H. Zhang, J. Bradfield, C. Kim, E. Frackleton, C. Hou, J. T. Glessner, R. Chiavacci, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genetics*, 5(10):e1000678, 2009.

Z. Wei, W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, C. Kim, F. Mentch, K. Van Steen, P. Visscher, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *The American Journal of Human Genetics*, 92(6):1008–1012, 2013.

D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, 2014.

S. Wood. mgcv: mixed gam computation vehicle with gcv/aic/reml smoothness estimation. r package version 1.7-22, 2012.

N. Wray, J. Yang, M. Goddard, and P. Visscher. The genetic interpretation of area under the roc curve in genomic profiling. *PLoS Genetics*, 6(2):e1000864, 2010.

J. Yang, S.H. Lee, M.E. Goddard, and P.M. Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

W.-C. Yeung, W.D. Rawlinson, and M.E. Craig. Enterovirus infection and type 1 diabetes mellitus: systematic review and meta-analysis of observational molecular studies. *British Medical Journal*, 342:d35, 2011.

B. Zhang, A.A. Tsiatis, E.B. Laber, and M. Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68:1010–1018, 2012.

L. Zhao, L. Tian, T. Cai, B. Claggett, and L.J. Wei. Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, 108:527–539, 2013.

Y. Zhao, D. Zeng, A.J. Rush, and M.R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107:1106–1118, 2012.

Y.Q. Zhao, D. Zeng, E.B. Laber, R. Song, M. Yuan, and M.R. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102:151–168, 2015.

Y. Zheng, T. Cai, and Z. Feng. Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics*, 62:279–287, 2006.

X. Zhou, N. Mayer-Hamblett, U. Khan, and M.R. Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112:169–187, 2017.

R.C. Ziegelstein. Personomics. *JAMA Internal Medicine*, 175:888–889, 2015.

A.-G. Ziegler, M. Hummel, M. Schenker, and E. Bonifacio. Autoantibody appearance and risk for development of childhood diabetes in offspring of parents with type 1 diabetes: the 2-year analysis of the german babydiab study. *Diabetes*, 48(3):460–468, 1999. doi: 10.2337/diabetes.48.3.460.

Résumé

Cette thèse porte sur l'élucidation des causes génétiques et environnementales des maladies complexes. L'application centrale est l'étude Isis-Diab sur le diabète de type 1. Nous proposons une méthode afin d'utiliser l'information de phase pour améliorer les prédictions de risque génétique. Nous répliquons une estimation du risque génétique sur les patients d'Isis-Diab. Nous prouvons un résultat d'équivalence asymptotique entre deux méthodes d'analyse des données appariées. Nous analysons ensuite l'étude cas-témoins d'Isis-Diab basée sur des questionnaires environnementaux. Nous essayons de confirmer les résultats de cette étude en croisant le risque génétique et les facteurs environnementaux chez les patients d'Isis-Diab. Cela nous amène à proposer une nouvelle méthodologie basée sur le biais de collision et à étudier l'influence de ce dernier dans les études cas-seulement pour les interactions gène-environnement. Finalement, nous étudions la possibilité d'utiliser des essais thérapeutiques randomisés pour personnaliser les traitements. Nous proposons une nouvelle méthodologie pour estimer le bénéfice de la personnalisation et nous recommandons un choix de stratégie de personnalisation.

Mots-clés : génétique, épidémiologie, apprentissage statistique, médecine personnalisée.

Abstract

This thesis concerns the identification of the genetic and environmental causes of complex diseases. Our central application is Isis-Diab a study of type 1 diabetes. We propose a method to use phase information to improve genetic risk predictions. We replicate an estimate of genetic risk on Isis-Diab patients. We prove an asymptotic equivalence result between two paired data analysis methods. We then analyze the Isis-Diab case-control study based on environmental questionnaires. We try to confirm the results of this study by crossing the genetic risk and environemental factors in Isis-Diab patients. This leads us to propose a new methodology based on collider bias and a study of its influence in case-only studies for gene-environment interactions. Finally, we study the possibility of using randomized clinical trials to personalize treatments. We propose a new methodology to evaluate the benefit of personalization and we recommend a choice of personalization strategy.

Key-words: genetics, epidemiology, statistical learning, personalized medecine.