Cartographie globale des essais cliniques

Ignacio Atal

UNIVERSITÉ PARIS DESCARTES

École doctorale Pierre Louis de Santé Publique :
Épidémiologie et Sciences de l'Information Biomédicale

Centre de Recherche Épidémiologie et Statistique Sorbonne Paris Cité,
INSERM UMR1153, équipe METHODS

Cartographie globale des essais cliniques

Ignacio Atal

Thèse de doctorat de Santé Publique
Spécialité Épidémiologie

Dirigée par Raphaël Porcher
Co-encadrée par Ludovic Trinquart

Présentée et soutenue publiquement le 23 Novembre 2017

Devant un jury composé de :

Rapporteurs :		
Jacques DEMOTES	Professeur	Université Bordeaux 2
Bruno FALISSARD	Professeur	Université Paris-Sud
Examinateurs :		
Marie-Christine JAULENT	Directrice de recherche	Université Paris Descartes
Aurélie NÉVÉOL	Chargée de recherche	Université Paris-Sud
Philippe RAVAUD	Professeur	Université Paris Descartes
Directeur de thèse :		
Raphaël PORCHER	Maître de conférences	Université Paris Descartes

Résumé : Pour comprendre comment se construisent les connaissances sur l'effet des interventions en médecine, il est nécessaire de savoir où est faite la recherche clinique dans le monde, quelles maladies sont étudiées, et quels acteurs la mettent en place. Une vision globale du système de recherche peut aider à identifier des lacunes dans la production de connaissances et à orienter l'activité de recherche vers les priorités de santé, en particulier dans les régions où les ressources sont limitées. Dans ce travail nous avons construit des cartographies de la recherche clinique, c'est-à-dire des analyses agrégées de ce système complexe visant à extraire de l'information sur l'activité globale de recherche. Nous avons utilisé les registres d'essais cliniques inclus dans l'International Clinical Trials Registry Platform de l'Organisation Mondiale de la Santé pour cartographier l'activité de recherche.

Dans un premier travail nous avons évalué pour 7 régions l'alignement entre l'effort local de recherche sur 10 ans et le fardeau de 27 groupes de maladies. Ce travail a nécessité le développement d'un algorithme de classification automatique des maladies étudiées dans les essais clinique basé sur des méthodes de traitement automatique du langage. À partir des données de 117,180 essais randomisés, nous avons montré que la recherche faite dans les pays riches était bien alignée avec leurs besoins. Dans toutes les autres régions nous avons identifié des lacunes dans l'effort de recherche. En particulier, en Afrique Subsaharienne, même si des causes majeures de fardeau comme le VIH et le paludisme reçoivent un effort de recherche important, d'autres priorités locales, les maladies infectieuses communes et les pathologies du nouveau né, ont été négligées par l'effort de recherche.

Dans un deuxième travail nous avons évalué l'influence du type de promoteur (industriel ou non-industriel) dans l'utilisation de réseaux de pays pour recruter des patients dans des essais cliniques multi-pays. Nous avons montré que 30% contre 3% des essais à promoteur industriel et non-industriel sont multi-pays, respectivement. Les pays d'Europe de l'Est participent dans leur ensemble de façon surreprésentée dans la recherche multi-pays industrielle. Ceci suggère les grandes capacités des industriels à globaliser leur recherche en s'appuyant sur des réseaux de pays bien définis.

À l'échelle de tous les essais clinique enregistrés, nos travaux ont mis en évidence des lacunes majeures dans l'effort de recherche mondial, et montré l'influence des différents acteurs dans la globalisation de celle-ci. Ces travaux forment une brique pour le développement d'un observatoire global de la recherche médicale.

Mots-clés : cartographies, essais cliniques, meta-recherche, fardeau mondial des maladies, systèmes complexes

Title : World mapping of clinical trials

Abstract : By knowing what clinical research is undertaken worldwide, where it is conducted, which diseases are studied, and who is supporting it, we could have a better understanding on how is created the knowledge concerning health interventions. A global landscape of health research may inform policy makers on knowledge gaps and on how to reallocate resources to address health needs, in particular in low-resource settings. In this thesis we mapped clinical research, i.e. we analyzed at a macro-level the complex system of health research to bring information on the global landscape of health research effort. We based our analyses on clinical trial registries included in the International Clinical Trials Registry Platform from the World Health Organization.

In a first project, we evaluated within 7 regions the local alignment between the effort of research and the burden for 27 groups of diseases. This work needed the development of a knowledge-based classifier of clinical trial registries according to diseases studied based on natural language processing methods. We mapped 117,180 randomized controlled trials. For high-income countries, the research effort was well aligned with the needs. In all other regions we identified reseach gaps. In particular, for Sub-Saharan Africa, where major causes of burden such as HIV and malaria received a high research attention, research was lacking for major causes of burden, especially for common infectious diseases and neonatal disorders.

In a second project, we compared the mappings of multi-country trials for industry- and non-industry–sponsored clinical trials, and analyzed the networks of collaboration of countries participating together to the same multi-country trials. We showed that among industry- and non-industry–sponsored trials, 30% and 3% were multi-country, respectively. The collaboration within Eastern European countries was particularly over-represented for industry-sponsored research. Industry sponsors may thus have a greater capacity to conduct globalized research, using well-defined networks of countries.

Our large-scale mappings of all registered clinical trials shed light on major gaps in the effort of health research as compared to health needs. In addition, we showed the influence of different sponsors in the globalization of clinical research. These projects are in-line with the development of a global observatory for health research.

Keywords : mapping, clinical trials, meta-research, global burden of diseases, complex systems

"Les acteurs de l'an 1812 ont depuis longtemps quitté la scène. Les intérêts personnels qu'ils poursuivaient ont disparu sans laisser trace et seuls subsistent pour nous les résultats historiques de cette époque. Mais si nous admettons que les habitants de l'Europe DEVAIENT *s'enfoncer, sous la conduite de Napoléon, au cœur de la Russie et y périr, toute la conduite contradictoire, absurde et cruelle de ceux qui ont participé à cette guerre nous devient compréhensible"*

Léon Tolstoï, *La Guerre et la Paix*

Remerciements

Tout d'abord, je tiens à exprimer ma sincère gratitude envers mes encadrants, Dr Raphaël Porcher et Dr Ludovic Trinquart. Leurs conseils et soutient m'ont permis de prendre goût au métier de recherche.

Je tiens aussi à remercier Pr Philippe Ravaud, pour nos discussions toujours enrichissantes et grâce à qui ce travail de thèse s'est déroulé dans les meilleures conditions que l'on puisse imaginer.

Je remercie mes rapporteurs, Pr Jacques Demotes et Pr Bruno Falissard, et mes examinatrices, Dr Marie-Christine Jaulent et Dr Aurélie Névéol, d'avoir accepté d'évaluer ce travail.

Je remercie toutes les personnes avec qui j'ai eu l'occasion de collaborer ou tout simplement discuter pendant ce travail. Je remercie spécialement l'équipe du Centre d'Épidémiologie Clinique de l'Hôpital Hôtel-Dieu, au sein de laquelle travailler a toujours été agréable et stimulant.

Je remercie du fond du cœur tous mes amis. Je remercie en particulier toutes les personnes avec qui j'ai partagé, et celles qui m'ont accueilli, nourri et logé pendant cette thèse nomade, à Santiago, Cutipay, Rio de Janeiro, Barcelone, Paris et ailleurs.

Gracias a mis padres, mis hermanos Juan Pablo, Raimundo y Vicente, mi hermana Gabriela, e Ita, por saberlos siempre presentes.

Affiliations

1. Centre d'Épidémiologie Clinique, Hôpital Hôtel-Dieu, Paris, France

2. INSERM U1153, Paris, France

3. Université Paris Descartes - Sorbonne Paris Cité, Paris, France

Financement

1. Programme Équipe Espoirs de la Recherche, Fondation pour la Recherche Médicale, France (DEQ20101221475).

Publications scientifiques liées à la thèse

Atal I, Trinquart L, Porcher R, Ravaud P (2015) Differential Globalization of Industry- and Non-Industry– Sponsored Clinical Trials. *PLoS ONE* 10(12) :e0145122. doi :10.1371/journal.pone.0145122

Atal I, Zeitoun JD, Névéol A, Ravaud P, Porcher R, Trinquart L (2016) Automatic classification of registered clinical trials towards the Global Burden of Diseases taxonomy of diseases and injuries. *BMC Bioinformatics* 17(1) :392. doi :10.1186/s12859-016-1247-7

Atal I, Trinquart L, Ravaud P, Porcher R (2017) Does clinical research effort match public health needs ? A large-scale mapping of 115,000 randomized trials and 2.2 billion disability-adjusted life years. *(en révision)*

Autres publications scientifiques

Dechartres A, Ravaud P, **Atal I**, Riveros C, Boutron I (2016) Association between trial registration and treatment effect estimates : a meta-epidemiological study. *BMC Medicine* 14(1) :100. doi :10.1186/s12916-016-0639-x

Dechartres A, Bond EG, Scheer J, Riveros C, **Atal I**, Ravaud P (2016) Reporting of statistically significant results at ClinicalTrials.gov for completed superiority randomized controlled trials. *BMC Medicine* 14(1) :192. doi :10.1186/s12916-016-0740-1

Dechartres A, Trinquart L, **Atal I**, Moher D, Dickersin K, Boutron, I, Perrodeau E, Altman D, Ravaud P (2017) Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews : research on research study. *BMJ* 357 :j2490. doi :10.1136/bmj.j2490

Zeitoun JD, Baron G, Vivot A, **Atal I**, Downing NS, Ross J, Ravaud P (2017) Post-marketing research and its outcome for novel anticancer agents approved by both the FDA and EMA between 2005 and 2010 : a cross-sectional study. *International Journal of Cancer* (in press)

Liste d'abréviations

AACT *Aggregate Analysis of ClinicalTrials.gov*

AFD *Agence Française de Développement*

CIM-10 Classification Internationale des Maladies, 10^{e} révision

DALY *Disability adjusted life-years* (Années de vie corrigées de l'incapacité)

EBM *Evidence-based medicine* (Médecine basée sur les faits)

ECRIN *European Clinical Research Infrastructure Network*

EDCTP *European and Developing Countries Clinical Trials Partnership*

GBD *Global Burden of Diseases* (Charge mondiale de morbidité)

ICMJE *International Committee of Medical Journal Editors*

ICTRP *International Clinical Trials Registry Platform* (Système d'enregistrement international des essais cliniques)

IHME *Institue for Health Metrics and Evaluation*

LCNMA *Live cumulative network meta-analyses* (Meta-analyses en réseau cumulatives et dynamiques

MeSH Medical Subject Headings

NIH *U.S. National Institutes of Health*

NLM *U.S. National Library of Medicine*

OMS Organisation Mondiale de la Santé

PIB Produit Interne Brut

TAL Traitement Automatique du Langage

UMLS *Unified Medical Language System*

USAID *U.S. Agency for International Development*

Table des matières

Chapitre 1

Introduction

La recherche clinique est toute recherche conduite chez l'être humain visant au développement des connaissances biologiques ou médicales. C'est un système complexe, globale, impliquant un grand nombre d'acteurs, et prenant un grand nombre de formes. Cartographier la recherche clinique, c'est-à-dire savoir où se fait la recherche dans le monde, quelles maladies sont étudiées, et qui fait cette recherche, pourrait aider à comprendre comment se construit la connaissance sur l'effet des interventions en médecine. Des cartographies de la recherche clinique pourraient être notamment une source importante d'information pour aider à orienter les programmes de recherche vers les domaines où existent des manques.

1.1 La recherche clinique : un système complexe

1.1.1 Formes de la recherche clinique

Tout type de connaissance biologique ou médicale chez l'homme peut être l'objet d'étude de la recherche clinique. La recherche peut être centrée sur le patient, le médecin, l'hôpital, le système de soins, ou la population générale. Elle peut étudier des maladies, mais aussi des états de santé comme la douleur, ou la qualité des soins. Ainsi, le spectre des questions étudiées est large, allant par exemple

de l'étude de l'effet d'un traitement pharmacologique sur la survie du patient à l'étude de l'effet biologique d'une molécule dans un organe humain spécifique, en passant par l'étude de l'effet d'une formation de médecins sur la qualité de prise en charge des patients. Pour chaque type de question, il existe des types d'études adaptés, notamment séparés en études interventionnelles et observationnelles.

Les études interventionnelles, appelées aussi *essais cliniques* en santé, ont pour objectif d'évaluer l'efficacité et la sécurité de nouvelles interventions. Des interventions peuvent correspondre à des traitements pharmacologiques donnés aux patients tels que des médicaments ou des vaccins, ou bien des traitements non-pharmacologiques tels que des techniques chirurgicales, des procédures de prise en charge, des méthodes de diagnostic, des méthodes de changement du comportement, l'utilisation des plateformes technologiques dans les hôpitaux ou des méthodes d'éducation pour des mesures de prévention, par exemple.

En particulier, les nouveaux traitements pharmacologiques doivent être mis à l'épreuve chez des patients dans des essais cliniques pour être autorisés par des agences de régulation. Il est habituel de diviser les essais cliniques médicamenteux en différentes phases représentant les stades du développement du traitement. Les essais de phase I visent principalement à évaluer la tolérance du médicament chez un nombre relativement restreint de patients. Les essais de phase II ont pour but d'évaluer l'activité biologique du nouveau traitement et de compléter l'évaluation de la tolérance sur un échantillon un peu plus large. Les essais de phase III ont pour objectif d'évaluer l'efficacité des traitements chez les patients, en les comparant à un placebo ou aux traitements déjà existants. C'est généralement à l'issue d'un ou plusieurs essais de phase III qu'une nouvelle molécule peut postuler à une autorisation de mise sur le marché par des agences de régulation. On ajoute parfois des essais dits de phase IV, qui sont utilisés pour évaluer l'efficacité et la sécurité des traitements déjà autorisés dans des conditions de prise en charge habituelle. Les études interventionnelles peuvent avoir plusieurs plans d'expérience, le plus connu étant les essais randomisés, pierre angulaire de l'évaluation thérapeutique.

Dans les études observationnelles, l'étude ne prescrit pas d'intervention sur son objet, les investigateurs organisant simplement un recueil de données et une stratégie d'analyse de ces données. Ces études peuvent être utilisées par exemple en épidémiologie ou en pharmacovigilance, mais aussi pour comparer l'effet de différentes stratégies de prise en charge existantes dans le cadre de la comparaison de l'efficacité de traitements (*comparative effectiveness research*).

Toutes ces formes de recherche clinique sont à la faveur de la *médecine basée sur les faits* (EBM, de l'anglais *evidence-based medicine*), c'est-à-dire la pratique médicale telle que les décisions d'interventions sont fondées sur le plus grand nombre de données accessibles et pertinentes concernant l'efficacité et la sécurité.

1.1.2 Un système global

La recherche clinique est faite partout dans le monde. C'est en particulier le cas des essais cliniques internationaux, recrutant des patients simultanément dans des centres de plusieurs pays. Un intérêt majeur de ces essais est d'accélérer le recrutement des participants à l'essai, par exemple pour des maladies rares, mais aussi de produire une information ayant une meilleure validité externe, c'est-à-dire qui s'applique à la plus vaste population et au plus grand nombre de contextes.

En effet, les résultats de la recherche clinique conduite dans un pays ne sont pas toujours directement transposables à d'autres (Rothwell, 2005; Mahaffey et al., 2011). D'une part, différentes caractéristiques génétiques peuvent donner lieu à une variation dans la réponse biologique à certains traitements (Mok et al., 2009). D'autre part, des différences culturelles peuvent influencer l'effet d'un traitement. Par exemple, l'observance des traitements, c'est-à-dire la régularité de suivi par le patient du traitement qui lui est prescrit, peut varier selon les pays (Kotseva et al., 2009). Le système de soins dans lequel est évalué un traitement peut varier d'un pays à l'autre, et influencer de façon très importante l'effet du traitement. Par exemple, dans un essai clinique international, la prise en charge complémentaire au traitement expérimental peut varier entre les pays (Rothwell, 2007). Des facteurs

socio-économiques et environnementaux peuvent aussi affecter l'applicabilité de résultats de recherche, en particulier dans les pays en voie de développement. Par exemple, l'oxytocine, utilisé pour prévenir et traiter l'hémorragie du post-partum, nécessite d'être conservé à froid et doit être administré par des professionnels de santé qualifiés. Or, dans le contexte de régions tels que l'Afrique subsaharienne ou l'Asie du Sud, où l'hémorragie du post-partum entraîne 30% des décès maternels, les ressources manquent pour une conservation et administration correcte de l'oxytocine (Say et al., 2014; Torloni et al., 2016). De même, les recherches étudiant des interventions visant des changements du comportement ou les systèmes de soins sont particulièrement sensibles au contexte dans lequel elles sont conduites (GESICA Investigators, 2005). Par ailleurs, les interventions testées chez l'adulte ne sont pas toujours transposables chez les populations pédiatriques ou gériatriques, alors qu'elles sont généralement sous-représentées dans les essais cliniques par leur vulnérabilité (Herrera et al., 2010; Joseph et al., 2015).

Ainsi, dans de nombreux cas il est nécessaire de chercher des solutions locales là où les solutions connues ne peuvent pas être appliquées ou ne sont pas efficaces. Par exemple, des solutions recommandées internationalement pour la prise en charge d'enfants en état de choc suite à une maladie infectieuse se sont avérées peu efficaces dans le contexte de prise de soins en Afrique subsaharienne (Maitland et al., 2011).

Pour assurer l'applicabilité des résultats de recherche dans leurs pays, plusieurs agences locales régulant la mise sur le marché des médicaments exigent que les interventions soient évalués localement par des essais cliniques. C'est le cas en particulier des États-Unis, de l'Union Européenne, de la Chine, de l'Inde, du Japon, de la Corée du Sud, du Brésil, du Nigeria et des Philippines.

1.1.3 Un système dirigé par de multiples acteurs

Il existe plusieurs acteurs décidant quelle recherche est faite, quelles interventions sont évaluées, quelles maladies sont étudiées, et où est conduite la recherche.

Ces acteurs peuvent être par exemple des industries du médicament, des universités, des organismes gouvernementaux ou non-gouvernementaux, des associations caritatives, des agences publiques, ou des investigateurs. Ils peuvent jouer des rôles différents lors d'une recherche : décider quelle question est étudiée, financer la recherche, ou bien en être le promoteur, c'est-à-dire le responsable de la conduite de l'étude. Plusieurs acteurs peuvent participer simultanément à une même étude en jouant des rôles différents. Par exemple, un industriel du médicament peut planifier, financer et assurer la conduite d'un essai clinique pour évaluer une de ses molécules, mais peut aussi aider au financement d'une étude académique en fournissant sa molécule pour un essai conduit par un groupe d'investigateurs et promu par un établissement hospitalier ou universitaire, par exemple.

Il existe des préoccupations concernant l'influence de l'industrie pharmaceutique dans la décision des programmes de recherche et dans la construction des connaissances médicales sur l'effet des traitements pharmacologiques (Stamatakis et al., 2013).

Les différents acteurs ont des intérêts et contraintes différents pour planifier leur programme de recherche, en particulier pour décider les maladies étudiées et la localisation de la recherche. Par exemple, des industries pharmaceutiques peuvent décider d'étudier certains types de molécules ayant des retours sur investissement plus intéressants que d'autres (Carter et al., 2016), ou au contraire ne pas chercher à développer des molécules pour des maladies affectant principalement les populations à revenu faible (Pedrique et al., 2013). D'autres organisations à but non lucratif comme l'AFM-Théléton ou la Foundation for AIDS Research concentrent leurs programmes de recherche sur des maladies rares ou affectant des populations négligées. Des organismes comme l'Agence Française de Développement (AFD), la United States Agency for International Development (USAID) ou Wellcome Trust décident leur programme de recherche basés sur le priorité de santé publique au niveau global (Moran, 2016).

De même, les industries pharmaceutiques chercheraient à externaliser leurs es-

sais cliniques vers des régions où il est moins coûteux de faire de la recherche comme les pays d'Europe de l'Est (Caldron et al., 2012). Des initiatives comme le European Clinical Research Infrastructure Network (ECRIN) cherchent à faciliter la recherche internationale au sein de l'Europe, en particulier entre acteurs académiques (Demotes-Mainard and Ohmann, 2005). De même, le European and Developing Countries Clinical Trials Partnership (EDCTP) a été créé pour faciliter les collaborations entre pays d'Europe et les pays d'Afrique Subsaharienne en matière de recherche clinique sur le VIH, le paludisme et la tuberculose (Matee et al., 2009).

Ainsi, des facteurs économiques, opérationnels et politiques influencent le système complexe de la recherche clinique, et définissent la cartographie de celle-ci.

1.2 Cartographier la recherche clinique

La recherche clinique est donc un système global dans lequel un grand nombre d'acteurs décide à son échelle individuelle quelle recherche est faite, où elle est faite et quelles sont les maladies étudiées. Pour élucider les forces qui meuvent ce système complexe et identifier des déséquilibres dans la production de connaissances médicales, nous allons créer des *cartographies de la recherche clinique*.

Définition. *Une cartographie de la recherche clinique est toute analyse agrégée du système de recherche clinique construit à partir des interactions complexes entre acteurs, pays et maladies, visant à donner de l'information sur la production de connaissances médicales.*

Pour construire des cartographies il existe de nombreuses bases de données traçant l'activité de recherche dans le monde. La complexité du système et la nature des bases de données nécessitent l'utilisation de méthodes spécifiques pour extraire des connaissances au niveau macroscopique et créer de l'information utile à la prise de décision sur les programmes de recherche.

1.2.1 Objectifs

Savoir quelle recherche est faite dans le monde, quelles maladies sont étudiées, et qui met en place cette recherche est indispensable pour comprendre comment se développent les connaissances sur l'effet des interventions en médecine. Une meilleure compréhension du système pourrait aider à élucider les forces qui dessinent le programme global de recherche, identifier des lacunes dans celui-ci, et informer les décideurs du programme de recherche pour les combler (Terry et al., 2014). En particulier, dans les régions où les ressources sont limitées, cartographier la recherche clinique permettrait de coordonner les programmes de recherche vers les priorités de santé publique.

Savoir de façon systématique quelle recherche a été faite et quelle recherche est en cours est ainsi le premier pas pour diminuer le gâchis de la recherche (Chalmers et al., 2014). En effet, un panorama global de l'état de la recherche clinique permettrait d'atteindre une couverture de santé universelle en cherchant (1) à implémenter universellement les interventions déjà existantes, (2) à évaluer et mettre en place des interventions plus efficaces et moins coûteuses que celles déjà existantes, et (3) à développer de nouvelles interventions là où il existe des besoins (Røttingen et al., 2013; World Health Organization, 2010).

L'information donnée aux acteurs de l'activité de recherche clinique doit par ailleurs être modulables à leurs besoins. La complexité du système nécessite des cartographies multi-échelle et adaptatives aux besoins spécifiques de chacun (Terry et al., 2014).

1.2.2 Modèle pour construire des cartographies

Pour construire des cartographies, nous allons nous baser sur un modèle générique basé sur trois types d'entités interagissant entre elles : les acteurs décidant, finançant et conduisant la recherche, les pays où est faite la recherche, et les maladies ou états de santé étudiés (Figure 1.1). La recherche clinique est par ailleurs

divisée "temporellement" en trois étapes : les entrées (investissements), les processus (essais cliniques), et les résultats (publications scientifiques et nouveaux produits).

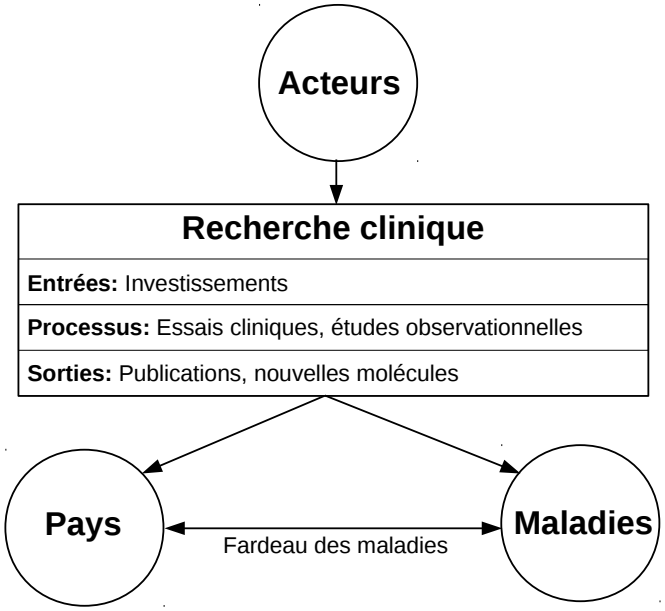


FIGURE 1.1 – Modélisation du système de recherche clinique

En effet, les acteurs décident dans quels pays et sur quelles maladies est faite la recherche. Ils peuvent avoir des intérêts ou contraintes spécifiques pour conduire de la recherche dans un pays ou sur une maladie donnés. Ces intérêts ou contraintes vont dépendre des spécificités des pays ou des maladies et des capacités opérationnelles des acteurs à mettre en place la recherche. Par ailleurs, la recherche a aussi différentes formes impliquant des différents avancements dans la production des connaissances. L'ampleur des problèmes de santé locaux, mesuré par le fardeau des maladies, peut aussi influencer la localisation de la recherche.

Le système de recherche clinique peut ainsi être modélisé par une réunion de *graphes bipartis* entre acteurs, pays et maladies. Un graphe biparti est un réseau dans lequel les nœuds sont séparés en deux groupes de nature différente, et les arêtes connectent uniquement des nœuds de nature différente (Figure 1.2).

Dans notre système d'étude, trois graphes bipartis apparaissent :

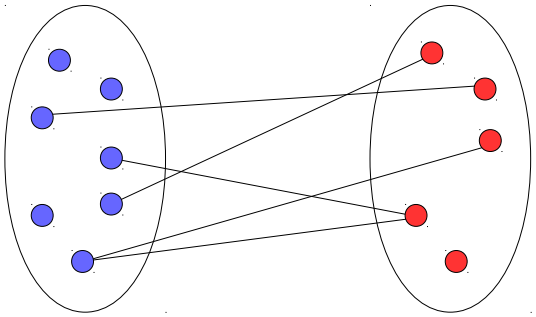


FIGURE 1.2 – Graphe biparti

Dans un graphe biparti, deux types entités, par exemple acteurs et pays, sont connectés par des activités de la recherche clinique, par exemple un lien pourrait représenter les essais financés par un acteur dans un pays donné.

— Le graphe où les nœuds sont d'une part les acteurs de la recherche clinique, et d'autre part les pays où se fait la recherche. Les arêtes sont les études de recherche clinique menées par des acteurs dans des pays donnés.
— Le graphe où les nœuds sont d'une part les acteurs de la recherche clinique, et d'autre part les maladies sur lesquelles porte la recherche. Les arêtes sont les études de recherche clinique menées par des acteurs sur des maladies données.
— Le graphe où les nœuds sont d'une part les pays où se fait la recherche clinique, et d'autre part les maladies sur lesquelles porte la recherche. Les arêtes sont les études de recherche clinique menées dans des pays sur des maladies données.

Dans ces graphes, les nœuds et les arêtes sont caractérisés par des variables spcifiques. Par exemple, les acteurs peuvent être catégorisés par type (public ou privé), par taille et moyens économiques, ou par leurs rôles dans la mise en place de la recherche (décideur, investisseur ou responsable de la conduite). De même, les pays peuvent être catégorisés par niveau de revenu, régions géographiques ou caractéristiques démographiques et culturelles. Les maladies peuvent être catégorisées par types (e.g. transmissibles ou non transmissibles), par la population généralement affectée (e.g. enfants, adultes, femmes) ou par leur complexité. Fi-

nalement, les arêtes, correspondant aux études de recherche clinique, peuvent être catégorisés par type d'étude, par taille ou par étape dans la production de recherche (investissements, essais cliniques, publications).

Une cartographie de la recherche clinique correspond donc à toute analyse de ce réseau complexe d'interactions entre acteurs, pays et maladies visant à extraire de l'information agrégée sur la création des connaissances médicales (Figure 1.3).

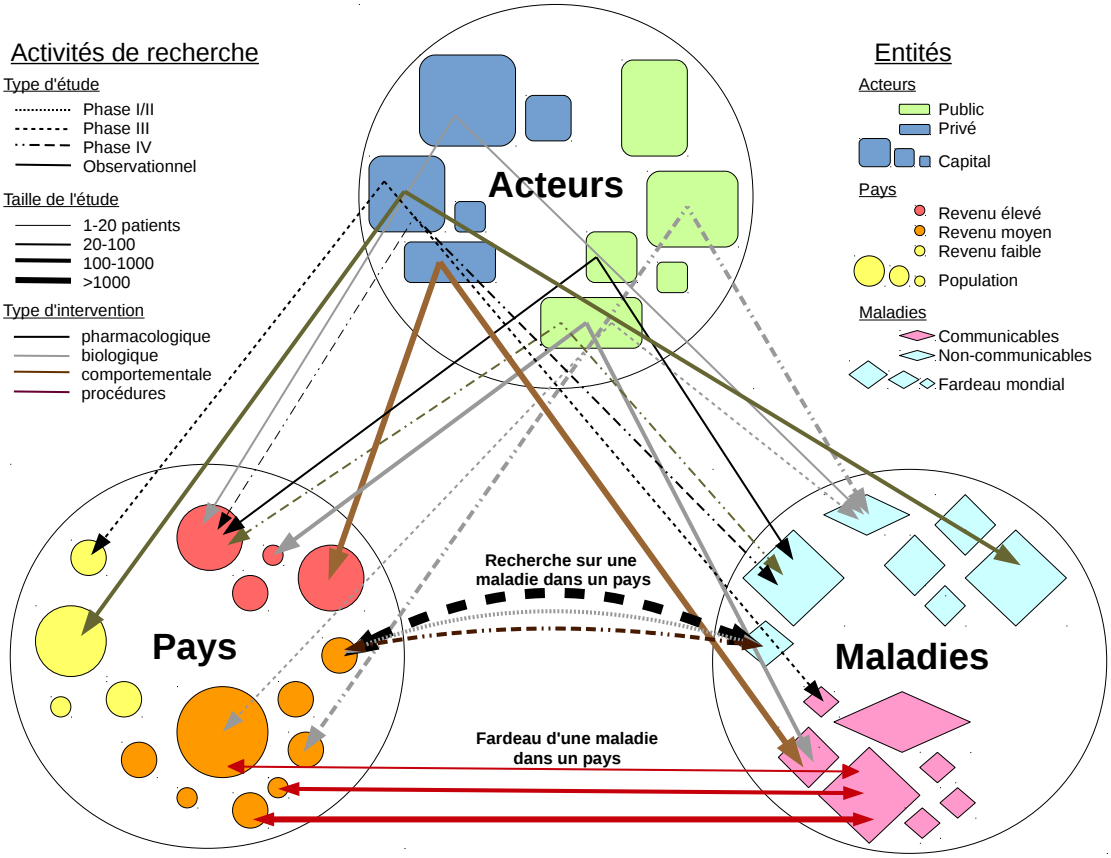


FIGURE 1.3 – Exemple de modèle complexe pour construire des cartographies

1.2.3 Méthodes d'analyse

En partant de ce modèle, on peut déduire des typologies d'analyse qui s'adaptent à la grande variété d'objectifs possibles des cartographies. Nous distinguons trois grands axes présentés par degré de complexité.

Un premier axe d'analyse correspond à des analyses descriptives centrées sur un type d'entité. Par exemple, les analyses centrées sur les pays correspondent à décrire pour un pays (ou région) donnée les acteurs y faisant la recherche et les maladies étudiées. De la même façon, les analyses centrées sur les maladies correspondent à décrire pour une maladie donnée quels sont les acteurs faisant la recherche sur cette maladie et dans quels pays elle est faite. Il est ainsi possible de décrire en détail l'état de la recherche dans un pays ou dans un domaine, quels sont les principaux acteurs y participant, et identifier des maladies ou pays en manque de recherche, respectivement.

Un deuxième axe d'analyse correspond à comparer la production de recherche avec la prévalence des maladies pour élucider des lacunes dans la connaissance vis-à-vis des besoins de santé. Conceptuellement, ceci correspond à analyser l'alignement entre le graphe biparti de la recherche entre les pays et les maladies à un graphe biparti parallèle dans lequel les arêtes entre pays et maladies seraient le nombre de malades par maladie dans chaque pays. Ainsi, il est possible d'identifier des maladies ou pays où l'effort de recherche est trop bas par rapport au fardeau des maladies, et identifier l'influence des différents acteurs dans l'obtention des connaissances manquantes.

Un troisième axe correspond à analyser des réseaux d'interaction entre les éléments d'un même type d'entité, interactions qui découlent des graphes bipartis. Par exemple, on peut dire que deux acteurs de la recherche sont proches s'ils font de la recherche dans les mêmes pays ou sur les mêmes maladies. Ces réseaux peuvent être ensuite analysés pour faire ressortir des informations comme l'influence de certains acteurs, leur centralité, la co-occurrence de paires d'acteurs, ou l'existence de communautés d'acteurs. La théorie des graphes ouvre un grand nombre de possibilités d'analyses de ces réseaux pour élucider forces dessinant les programmes de recherche (Albert and Barabasi, 2002).

Les cartographies ont un caractère multi-échelle, correspondant aux niveaux de détail de la description de chaque type d'entité. Par exemple, des analyses

peuvent être à l'échelle globale, à l'échelle d'une région, ou à l'échelle d'un pays. De la même façon, des cartographies peuvent être faites à l'échelle des spécialités médicales comme l'ophtalmologie, ou à l'échelle d'un problème de santé précis.

La complexité de ces types d'analyses, leur caractère multi-échelle, le nombre de variables mises en jeu, nécessitent aussi de méthodes avancées de visualisation, adaptatives aux spécificités de chaque question, mais simples pour aider à la prise de décision.

1.2.4 Bases de données existantes

Pour construire des cartographies il est nécessaire d'avoir accès à des données donnant de l'information sur le système de recherche médicale. Il existe une grande quantité de sources de données concernant les entrées (investissements), les processus (essais cliniques) et les résultats (publications scientifiques) de la recherche clinique, mais aussi les entités (acteurs, pays et maladies) et le fardeau des maladies. Ces ressources varient dans leur accessibilité, échelle, exhaustivité, opérabilité et leur qualité et fiabilité.

Résultats de la recherche clinique

Dans la recherche clinique, ses résultat ont été disséminés bien avant d'autres informations concernant l'activité de recherche comme les investissements ou les essais cliniques conduits. Cette dissémination se fait principalement par le biais des publications dans des journaux scientifiques. Il existe de nombreuses bases électroniques réunissant les publications d'un grand nombre de journaux scientifiques dans le domaine biomédical, les principales étant MEDLINE, EMBASE, CENTRAL et Web of Science. Ces bases donnent accès généralement aux titres et résumés pour chaque publication, et quelques fois au texte intégral. Par ailleurs, d'autre meta-données telles que les affiliations des auteurs peuvent être accessibles. Ces bases ne sont pas exhaustives. Notamment, la plupart des publications

présentes sont écrites en anglais, et les publications en d'autres langues sont sous-représentées (Larsen and Ins, 2010).

Il apparaît cependant qu'une part non négligeable de la recherche clinique n'est jamais publiée dans des journaux scientifiques, soit parce qu'elle n'a pas réussi à arriver à terme, soit parce que les résultats n'étaient pas satisfaisants pour le promoteur, les investigateurs ou les journaux scientifiques (Jones et al., 2013; Ross et al., 2009). Dans le cas des essais cliniques, l'absence de publication des résultats pose des problèmes éthiques majeurs : la participation volontaire de tout patient essayant des traitements expérimentaux doit apporter de l'information à la communauté scientifique sur leur effet. Ceci est particulièrement préoccupant lorsque les essais non publiés sont ceux ayant eu des résultats défavorables ou simplement non favorisant le traitement expérimental. Or, c'est justement le cas : les essais avec des résultats positifs ont davantage tendance à êtres publiés que ceux avec des résultats négatifs (Hopewell et al., 2009). La non connaissance des résultats négatifs biaise ainsi l'état des connaissances sur l'effet des interventions en favorisant les traitements expérimentaux, c'est ce qu'on appelle le *biais de publication* (Dickersin, 1990). En effet, la non publication de résultats de recherche clinique peut avoir des conséquences sur la prise en charge des patients. Par exemple, un essai ne montrant pas de différence entre deux interventions peut avoir des conséquences sur les pratiques médicales (Curley et al., 2005). La non publication de résultats de recherche est une des principales sources de gâchis dans la recherche médicale (Chan et al., 2014).

D'autres sources peuvent renseigner sur les résultats de la recherche clinique. En particulier, on peut trouver de l'information non publiée dans les résumés de conférences de spécialités. De l'information supplémentaire peut aussi être trouvée dans les demandes de brevets ou dans les demandes d'autorisation aux agences de régulation.

Processus de la recherche clinique

Pour éviter les problèmes de biais de publication, le International Committee of Medical Journal Editors (ICMJE) requiert depuis Septembre 2005 que les essais cliniques de phase II ou plus soient enregistrés dans des registres de façon prospective, c'est-à-dire avant que le premier patient ne reçoive le traitement expérimental (De Angelis et al., 2004). Ces registres d'essais cliniques sont une source de données permettant de surveiller les processus de la recherche clinique (Dickersin and Rennie, 2012). Le registre d'essais cliniques le plus connu est ClinicalTrials.gov (U.S. National Institutes of Health, 2000). Mis en ligne en 2000 par le National Institutes of Health (NIH), il compte aujourd'hui presque 250,000 essais cliniques enregistrés. D'autres registres d'essais cliniques au niveau national, régional ou international ont ouvert partout dans le monde. L'Organisation Mondiale de la Santé (OMS) a mis en place en 2005 le International Clinical Trials Registry Platform (ICTRP), méta-registre dans lequel sont accessibles publiquement les données de 16 registres répondant aux exigences de complétude et qualité défini par l'OMS (World Health Organization, 2005, 2012). Cette base de données sera utilisée dans le Chapitre 2 pour cartographier les processus de la recherche clinique, et sera décrite plus en détail dans la Section 2.1.

Les registres d'essais cliniques ne sont pas non plus libres de biais. Premièrement, tous les essais cliniques ne sont pas concernés par l'enregistrement prospectif, c'est en particulier le cas des essais cliniques de phase I. Pour les essais de phase I et II à promoteur industriel, il existe des rapports comme le CenterWatch Weekly (accessible sur souscription) dans lesquels il est possible de trouver des essais non présents dans les registres d'essais cliniques (Cottingham et al., 2014). Deuxièmement, dans les registres d'essais cliniques on ne trouve pas uniquement des essais cliniques, mais aussi des études observationnelles. Ces derniers ne sont pas non plus concernés par l'enregistrement prospectif, même s'il existe des efforts pour que les exigences s'étendes à cette forme de recherche (Boccia et al., 2015; The PLOS Medicine Editors, 2014; Dal-Ré et al., 2014). Troisièmement, tous les

journaux du domaine biomédical ne promeuvent pas les exigences de l'ICMJE. Même des journaux faisant partie de l'ICMJE n'ont adhéré que très récemment à l'exigence d'enregistrement prospectif pour publier des essais randomisés (Leopold et al., 2016). Le taux d'enregistrement des essais cliniques publiés reste néanmoins bas : parmi un échantillon de d'essais randomisés publiés en 2010, seulement 55% étaient avaient été enregistrés, 58% parmi les publications dans des journaux ICMJE versus 45% pour les journaux non-ICMJE (van de Wetering et al., 2012). Par ailleurs, le taux d'enregistrement peut varier selon les disciplines (30% en kinésithérapie (Babu et al., 2014) versus 60% en urologie (Kunath et al., 2011)), les régions (seulement 17% des publications avec un premier auteur latino-américain (Reveiz et al., 2012)) ou le type d'intervention (15% des essais évaluant des test diagnostiques (Korevaar et al., 2014)).

À ce jour, plus de 10 ans après le début de l'exigence de l'ICMJE, des grandes revues scientifiques continuent à recevoir des soumissions d'essais cliniques non enregistrés de façon prospective (Weber et al., 2015). Le manque d'enregistrement prospectif est particulièrement préoccupant lorsque celui-ci pourrait être associé à des effets de traitement plus élevés (Dechartres et al., 2016).

Entrées de la recherche clinique

Les entrées du système de production de recherche clinique, c'est-à-dire les investissements, sont plus difficiles à tracer par des bases de données que les processus et les résultats.

Des enquêtes annuelles menées par l'UNESCO et l'OCDE permettent de connaître l'investissement total des pays en recherche et développement en matière de santé. Ces données ont très peu de granularité, et sont manquantes pour 63% des pays (Røttingen et al., 2013). Par ailleurs, les données sont principalement manquante dans les pays à revenu faible, moyen faible, et moyen élevé : 86%, 81% et 63% des pays n'ont pas de données disponibles sur l'investissement total en recherche médicale, respectivement. Au contraire, les données manquent pour 28% des pays

à revenu élevé, correspondant principalement à des micro-États.

Des efforts sont mis en place entre les pays d'Amérique Latine, l'Espagne et le Portugal (RICYT, 1990) et en Afrique (NEPAD, 2005) pour créer des indicateurs locaux sur les investissements en recherche médicale. D'autres enquêtes spécifiques à des maladies existent pour le monitorage des investissements en recherche, en particulier pour les maladies négligées (Policy Cures, 2007), la tuberculose (Treatment Action Group, 2005) et le VIH (Resource Tracking Working Group, 2004).

Entités

De l'information spécifique aux entités du système de recherche clinique (pays, maladies et acteurs) est aussi nécessaire pour la création des cartographies.

Pour les pays, il est habituel de comparer l'activité de recherche par niveau économique et de développement. Pour cela, les indicateurs généralement utilisés sont les groupes de niveau de revenu fournis par la Banque Mondiale (pays à revenu élevé, moyen-élevé, moyen-faible et faible), le Produit Interne Brut (PIB) fourni par l'OCDE, et des indicateurs de développement comme l'Indice de Développement Humain des Nations Unies. Par ailleurs, des efforts récents cherchent à créer des bases de données internationales spécifiques aux systèmes de soins par pays, comme les dépenses en santé par personne et les sources de ces dépenses (Global Burden of Disease Health Financing Collaborator Network, 2017). Au sein de chaque pays il existe un grand nombre de données pertinentes pour être inclues dans des cartographies de la recherche, mais elles sont rarement comparables entre pays, et généralement sont peu accessibles ou peu exploitables.

Des données spécifiques aux maladies peuvent aussi être informatives pour les cartographies. Par exemple les tranches d'âge des populations concernées par la maladie, les différents types de maladies (chroniques vs aiguës, transmissibles vs non-transmissibles), les voies de transmission pour les maladies contagieuses, ou les groupes de maladies telles que les co-morbidités sont communes, par exemple.

De façon similaire, des caractéristiques générales des acteurs participant et appuyant la recherche peuvent être inclus dans les cartographies, comme leurs statut (privé, public, non-gouvernemental par exemple), et leurs moyens économiques. Des données plus détaillées sur leurs mode de fonctionnement sont plus difficilement accessibles à grande échelle.

Fardeau des maladies

Depuis les années 1990, le U.S. Institute for Health Metrics and Evaluation (IHME) conduit le projet Global Burden of Diseases (GBD), dont l'objectif est de d'estimer le fardeau des maladies dans le monde et son évolution dans le temps. Financé par la Bill & Mellinda Gates Foundation depuis l'édition GBD 2010, l'IHME estime à ce jour le fardeau de 315 maladies et blessures, 79 facteurs de risque, et plus de 2,600 séquelles dans 195 pays entre 1990 et 2015 (GBD 2015 DALYs and HALE Collaborators, 2016). Le fardeau est évalué par sexe et groupes d'âge. Ces données sont de plus accessibles publiquement.

Le fardeau des maladies et des blessures est mesuré avec plusieurs indicateurs, notamment la mortalité, le nombre d'années de vie perdues due à une mort précoce, le nombre d'années vécues avec une incapacité pour des maladies ou blessures non fatales, et finalement le nombre d'années de vie corrigées de l'incapacité (DALY, de l'anglais *disability-adjusted life-years*). Un DALY correspond à une année de vie saine perdue en raison soit d'une mort précoce, soit d'une incapacité provoquée par des maladies ou blessures (Devleesschauwer et al., 2014). Cette mesure unique a pour objectif d'aider les décideurs en matière de santé à comparer le fardeau pour des maladies de nature différente dans des populations différentes, et ainsi définir les priorités de santé publique. Par exemple, les DALY permettent de comparer le fardeau de maladies mortelles comme le cancer, à celui de maladies non mortelles comme la dépression. Cette mesure donne aussi un poids plus élevé aux maladies affectant les nouveaux nés et les enfants. Les DALY sont aussi utilisés pour des analyses de coût-efficacité des interventions en termes de coût par DALY prévenu

(Neumann et al., 2016).

Les implications de l'utilisation d'une mesure unique pour comparer le fardeau et mesurer l'efficacité des interventions ont été débattues depuis sa création (Anand and Hanson, 1997). Toutefois, dans le cadre de la création de cartographies de la recherche clinique, cette unique mesure nous permet de comparer l'activité de recherche au fardeau pour toutes les maladies, et ainsi identifier des lacunes dans la production de connaissances médicales (Section 2.2).

1.3 Cartographies existantes de la recherche clinique

Il existe un grand nombre de travaux publiés entrant dans notre définition de cartographie. Nous résumons ici les principaux résultats de cartographies existantes dans la littérature concernant les entrées (investissements), les processus (essais cliniques) et les sorties (publications ou nouveaux traitements).

1.3.1 Investissements

Il manque de bases de données compréhensibles pour cartographier précisément les investissements en recherche dans le monde, principalement dans les pays en voie de développement (Section 1.2.4). En utilisant un grand nombre de sources de données nationales et internationales, Røttingen et al. ont estimé l'investissement en recherche biomédicale en 2009 à 240 milliards USD : 89.5% vient des pays à revenu élevé, et 0.1% vient des pays à revenu faible (Røttingen et al., 2013). Dans les pays à revenu élevé, 60% de l'investissement vient du secteur commercial, 30% du secteur public, et 10% d'autres secteurs (dont organisations privées à but non lucratif). Entre pays, les parts relatives du privé et du public dans les sources d'investissement sont très variables. Ce travail a aussi permis de mettre en avant

que seulement 1% de l'investissement mondial en recherche biomédical était dédié aux maladies négligées.

Chakma et al. ont comparé l'investissement en recherche biomédicale des pays à revenu élevé à ceux des pays d'Asie (Chakma et al., 2014). Ils ont comparé les différents types d'investissement (public ou privé) et étudié l'évolution entre 2007 et 2012. Ils montrent que, d'une part les investissements publiques et privés augmentent considérablement en Asie, donnant plus de poids à cette région dans le panorama global de la recherche, alors que d'autre part tous les investissements ont diminué dans les pays riches, en particulier l'investissement de l'industrie en Europe.

Viergever and Hendriks ont recensé les plus grands investisseurs publics et philanthropiques dans le monde (Viergever and Hendriks, 2016). Le premier investisseur est le U.S. National Institutes of Health (NIH), contribuant 26.1 milliards USD en 2013, suivi de la Commission Européenne avec 3.7 milliards USD. Le plus grand investisseur philanthropique était Wellcome Trust avec 909 millions USD. Les mécanismes de distribution des investissements entre régions et maladies variait de façon substantielle d'une institution à l'autre.

Plusieurs travaux ont analysé en détail la distribution des investissements du NIH par maladie. En utilisant principalement des données nationales d'hospitalisation et mortalité, trois études ont montré que le taux d'investissement en recherche du NIH par maladie est correctement corrélé avec le fardeau de celles-ci aux États-Unis (Gross et al., 1999; Gillum et al., 2011; Sampat et al., 2013). Des analyses spécifiques au cancer montrent qu'il existe un déséquilibre entre l'investissement du NIH et le fardeau par type de cancer (Brower, 2005; Carter and Nguyen, 2012). Similairement, Hazo et al. ont analysé le financement de la Commission Européenne en matière de santé mentale, par pays et par maladie, et ont mis en évidence des pays où les financements sont absents (Hazo et al., 2016). De même, Guegan et al. ont étudié les projets de recherche en santé publique financés en Angleterre par le U.K. National Institute for Health Research (Guegan et al., 2016). Nous

tenons à souligner que ces analyses détaillées ont été conduites uniquement sur les données d'organismes de pays à revenu élevé, probablement puisqu'elles sont difficilement accessibles pour des organismes de pays en voie de développement (Røttingen et al., 2013).

Une étude a analysé les mécanismes d'investissement des organismes la recherche en santé globale (Moran, 2016). Il existe deux types d'acteurs publiques finançant ce type de recherche : les agences de développement comme l'USAID et l'AFD, ayant des stratégies de financement basées sur des systèmes de priorisation à échelle internationale, et les organismes locaux de recherche scientifique comme le NIH ou l'INSERM, dont les projets en santé globale naissent à l'initiative des chercheurs et non pas d'une stratégie à grande échelle. Or, seulement 10% du financement en santé globale provient des agences de développement, versus 83% venant des organismes de recherche.

Finalement, d'autres études ont montré que l'investissement en recherche médicale par maladie et par région est corrélé avec le nombre d'essais cliniques et le nombre de publications scientifiques (Røttingen et al., 2013; Vanderelst and Speybroeck, 2013).

1.3.2 Processus de la recherche clinique

Plusieurs études ont cartographié les processus de la recherche clinique à partir des registres d'essais cliniques (Section 2.1).

Analyses descriptives des essais enregistrés

Les travaux présentés ci-dessus ont fait des analyses descriptives des essais enregistrés (Section 1.2.3, premier axe d'analyse). Dans la littérature, quelques travaux ont décrit la totalité des essais enregistrés. Ces analyses concernent uniquement des caractéristiques des essais permettant un traitement automatique des registres

à grande échelle, comme par exemple leur localisation et des caractéristiques du plan d'expérience.

Deux études ont analysé la localisation des essais cliniques et son évolution dans le temps pour la totalité des essais enregistrés dans ClinicalTrials.gov et l'ICTRP (Thiers et al., 2008; Drain et al., 2014). Ces deux études ont montré les grandes inégalités entre pays en terme de nombre d'essais cliniques faits par habitant. Les différences vont de plus de 100 essais par million d'habitants dans certains pays à revenu élevé comme les États-Unis et le Danemark, à moins de 0.3 essais par million d'habitants en moyenne dans les pays à revenu faible. Par ailleurs, ces deux études mettent en évidence la croissance du nombre d'essais conduits dans les pays en voie de développement, en particulier en Asie et en Amérique Latine.

D'autres études ont décrit les caractéristiques de tous les essais enregistrés dans ClinicalTrials.gov comme la taille, les plans d'expérience et les phases. Deux travaux ont étudié l'influence du type de promoteur (industriel ou non-industriel) dans ces caractéristiques (Califf et al., 2012; Roumiantseva et al., 2013). Les études évaluant des traitements pharmacologiques ou biologiques ont plus tendance à être promus par les industriels, contrairement à ceux évaluant des interventions comportementales et des procédures de prise en charge. La quantité de données manquantes concernant les caractéristiques des études reste élevée, elle peut varier selon le type de sponsor, mais a tendance à s'améliorer globalement dans le temps. Murthy et al. ont comparé, pour les essais à promoteur industriel, les caractéristiques de ceux conduits exclusivement dans les pays à revenu élevé, et ceux recrutant des patients dans d'autres pays (Murthy et al., 2015). Les essais recrutant en dehors des pays à revenu élevé ont tendance à être de plus grande taille, plus longs, et correspondent à des phases plus avancées.

Nous avons recensé un grand nombre d'études décrivant les caractéristiques des essais enregistrés à plus petite échelle. Ces études se concentrent soit sur une spécialité médicale ou état de santé précis comme par exemple les maladies cardiovasculaires (Alexander et al., 2013; Hill et al., 2014), l'oto-rhino-laryngologie

(Witsell et al., 2013), l'Alzheimer (Cummings et al., 2014), la néphrologie (Inrig et al., 2014), les pathologies respiratoires du sommeil (Todd et al., 2013), les artériopathies des membres inférieurs (Subherwal et al., 2014), l'oncologie (Dear et al., 2011, 2012), le diabètes (Lakey et al., 2013), les maladies rares (Bell and Tudur Smith, 2014) ou la pédiatrie (Pasquali et al., 2012). Le changement d'échelle vers un plus petit nombre d'essais à analyse rend possible un traitement manuel des données, et permet ainsi des descriptions plus fines. Les résultats sont très dépendants du sujet traité, mais montrent que des analyses spécifiques permettent d'identifier des lacunes dans l'activité de recherche. Par exemple, sur 2,484 essais enregistrés étudiant le diabètes, seulement 1.4% rapporte des critères de jugement liés à la mortalité ou à des complications cardiovasculaires (Lakey et al., 2013).

Comparaison des essais enregistrés au fardeau des maladies

Nous présentons ici des travaux comparant l'activité de recherche en termes de nombre d'essais cliniques enregistrés, et le fardeau des maladie (Section 1.2.3, deuxième axe d'analyse).

Deux travaux ont évalué l'alignement entre les études interventionnelles en cours et le fardeau mondial des maladies. Le premier, utilisant tout ClinicalTrials.gov, a comparé pour 16 grandes maladies leur part relative dans la recherche à leur part relative dans le fardeau mondial en termes de DALY (Dal-Ré, 2011). Il a mis en évidence que le paludisme, la tuberculose et les maladies liées à la diarrhée étaient sous-étudiés relativement à leur part dans le fardeau, alors que le diabètes, le cancer du poumon et le VIH étaient sur-étudiés. Le deuxième travail, utilisant un échantillon de 5% de l'ICTRP, a comparé pour un plus grand nombre de maladies leur part relative dans la recherche à leur part dans le fardeau global (Viergever et al., 2013). Ils ont aussi comparé, pour chaque région géographique et chaque groupe de niveau de revenu, leur part relative dans la recherche et leur part relative dans le fardeau global. Alors que 40% du fardeau global est dû aux maladies communicables, maternes, périnatales ou nutritionnelles, seulement 10%

des essaies les étudiaient. Par ailleurs, dans les pays à revenu élevé, le nombre d'essais par millions de DALY est de 292.7 vs 0.8 dans les pays à revenu faible. Ainsi, ces études montrent que, à l'échelle globale, les essais ne se concentrent ni dans les régions ni sur les maladies avec le plus de fardeau.

Deux autres travaux de recherche on effectué des analyses similaires sur la population pédiatrique : ils ont comparé les essais en pédiatrie au fardeau chez l'enfant à partir de ClinicalTrials.gov et l'ICTRP (Bourgeois et al., 2014; Joseph et al., 2017). Leurs méthodes d'analyse et leurs résultats sont similaires. Alors que les enfants représentent 34% du fardeau global en termes de DALY, seulement 15% des essais enregistrés recrutaient des enfants. Il existe aussi de grandes inégalités dans la distribution de la recherche entre niveaux de revenu : 98% du fardeau chez l'enfant vient des régions à revenu moyen ou faible, mais seulement 22% des essais pédiatriques sont faits dans ces régions. Pour toutes les régions, les maladies pédiatriques les plus étudiées localement ne correspondent pas à celles avec le plus grand fardeau. En particulier, les pathologies du nouveau né sont à l'origine d'un quart du fardeau chez l'enfant dans les régions à revenu moyen ou faible, mais sont étudiés uniquement par 4% des essais conduits dans ces régions. Ainsi, pour le cas de la pédiatrie, d'une part l'effort de recherche est inégal entre les régions, et d'autre part, au sein de chaque région l'effort de recherche n'est pas aligné avec les besoins locaux.

Quelques travaux ont comparé l'effort de recherche au fardeau pour des problématiques de santés spécifiques. Ahmad et al. ont comparé la répartition géographique des essais randomisés publiés et en cours à la prévalence et la mortalité pour deux priorités de santé publique : le tabagisme et le VIH (Ahmad et al., 2011). Les Figures 1.4 et 1.5 extraites de ce travail montrent la différence de localisation entre la recherche et les besoins. Il est intéressant de noter que les interventions visant à diminuer ou à arrêter la consommation du tabac, et les interventions visant à prévenir l'infection du VIH sont particulièrement susceptibles d'être dépendantes du contexte dans lequel elles sont évaluées. Hirsch et al. ont comparé le nombre

d'essais enregistrés pour chaque type de cancer à leur mortalité et incidence mondiale (Hirsch et al., 2013). La corrélation entre l'effort de recherche et les besoins est modeste, et l'effort reste relativement bas pour le cancer du poumon, cause de 27.6% de la mortalité due au cancer et correspondant à 14.5% des nouveaux cancers diagnostiqués, mais seulement étudié par 9.2% des essais en cancer.

Finalement, quelques travaux ont utilisé des bases de données différentes aux registres d'essais cliniques pour comparer les processus de la recherche clinique au fardeau des maladies dans des contextes spécifiques. Premièrement, Cottingham et al. montre le non-alignement entre les essais industriels de phases précoces (I et II) et le fardeau par maladie en utilisant des données issues de rapports industriels (Cottingham et al., 2014). Deuxièmement, Canario et al. utilise des bases nationales décrivant les projets de recherche en République Dominicaine pour évaluer l'effort de recherche local (Canario et al., 2016). Ils ont montré le manque de corrélation entre recherche et fardeau local, et ont mis en évidence l'influence des industriels dans le programme de recherche local.

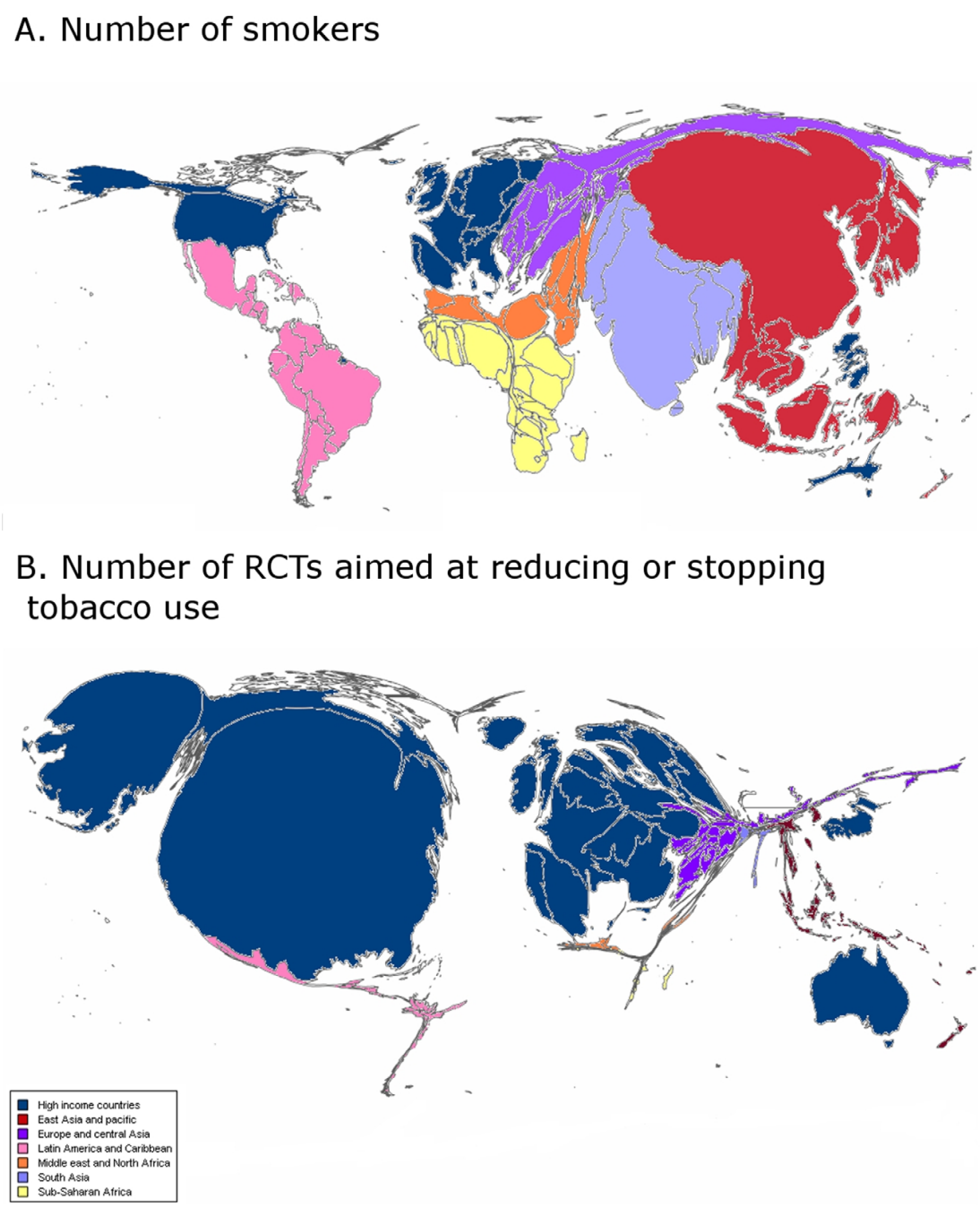


FIGURE 1.4 – Comparaison entre le nombre de fumeurs et la quantité de recherche visant à diminuer ou arrêter la consommation de tabac par pays

La taille des pays est proportionnelle A) au nombre de fumeurs et B) au nombre d'essais publiés ou en cours visant à réduire ou arrêter la consommation de tabac. Alors que 70% de la mortalité liée à la consommation de tabac se trouve dans les pays à revenu moyen ou faible, seulement 2% des essais randomisés sont faits dans ces régions. Source : (Ahmad et al., 2011)

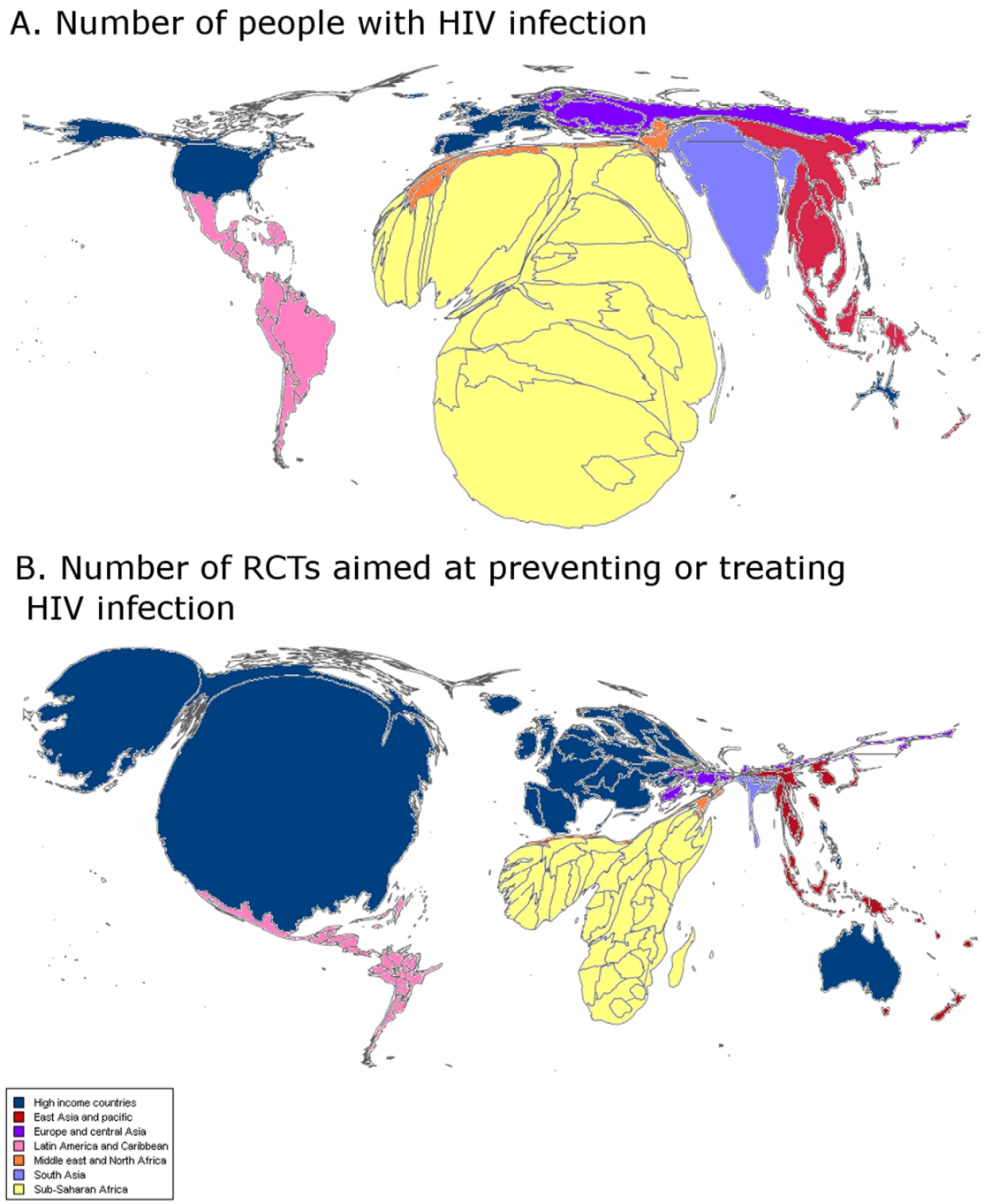


FIGURE 1.5 – Comparaison entre le nombre porteurs du VIH et la quantité de recherche visant à prévenir ou à traiter le VIH par pays

La taille des pays est proportionnelle A) au nombre de porteurs du VIH et B) au nombre d'essais publiés ou en cours visant à prévenir ou à traiter le VIH. Alors que 99% de la mortalité liée au VIH advient dans les pays à revenu moyen ou faible, seulement 31% des essais sont faits dans ces régions. Source : (Ahmad et al., 2011)

1.3.3 Résultats de la recherche

Nous présentons ici des travaux ayant cartographié les résultats de l'activité de recherche clinique (publications, revues systématiques et nouveaux traitements) pour étudier la production des connaissances médicales. La plupart de ces travaux comparent les résultats de l'activité de recherche au fardeau des maladies (Section 1.2.3, deuxième axe d'analyse).

Publications scientifiques

Deux études ont comparé à grande échelle le nombre d'articles publiés au fardeau des maladies, par pays et par maladie. À partir de toutes les publications biomédicales indexées dans PubMed, Vanderelst et Speybroeck ont montré que le poids d'une maladie dans le nombre total d'articles publiés était plus aligné avec le fardeau des pays riches qu'avec celui des pays à revenu moyen ou faible (Vanderelst and Speybroeck, 2013). Ils ont aussi identifié que les maladies tropicales négligées étaient sous-étudiées et les cancers étaient sur-étudiés relativement au fardeau global. Evans et al. ont conduit une analyse plus fine à partir de tous les articles inclus dans MEDLINE (Evans et al., 2014). En premier lieu, ils ont comparé pour chaque maladie la production scientifique au *marché global pour les traitements* associé à une maladie m (MGT_m), égal à la somme par pays du produit entre le fardeau lié à la maladie dans le pays p ($\mathrm{DALY}_{m,p}$) et le PIB du pays :

$$\mathrm{MGT}_m = \sum_{p \in Pays} \mathrm{DALY}_{m,p} \times \mathrm{PIB}_p$$

Cet indicateur cherche à mesurer la somme des pouvoirs d'achat des malades pour une maladie donnée, relativement au poids des maladies. Elle donne ainsi plus de poids aux maladies affectant les pays riches, mais aussi aux maladies à fardeau élevé. À partir de 300,000 liens article–maladie, l'étude montre que le nombre d'articles publiés sur une maladie données n'est pas corrélé au fardeau global de la maladie, mais au marché global pour les traitements associés à celle-ci. Ils ont

ensuite comparé au sein de chaque pays la production scientifique au fardeau des maladies. Pour assigner un pays à une publication, ils ont utilisé l'affiliation des auteurs. Ils ont montré que localement, le nombre d'articles publiés sur une maladie est correctement corrélé avec le fardeau local de la maladie. Or, la quantité d'articles scientifiques publiés dans les régions à revenu faible reste très bas par rapport aux autres régions. Une analyse similaire faite sur un échantillon de 1097 publications d'essais randomisés confirme ces résultats, mais montre que l'étendu de la corrélation locale entre production scientifique et fardeau est assez pauvre en Afrique Subsaharienne, Asie du Sud et l'Europe Centrale et de l'Est (Emdin et al., 2015).

D'autres études ont fait des analyses similaires portant sur des maladies, des régions, ou des types de recherche spécifiques. Par exemple, pour le cas de la cardiopathie ischémique, il a été montré que 19% des publications sont produites dans des régions à revenu moyen ou faible, alors que 75% de la mortalité advient dans ces régions (Okhovati et al., 2015). Trois études ont évalué l'alignement entre les essais publiés conduits en Afrique Subsaharienne et le fardeau local (Isaakidis et al., 2002; Swingler et al., 2005; Ndounga Diakou et al., 2017). Ils suggèrent que la recherche locale est assez bien corrélée aux besoins, même si des grandes causes de fardeau local comme les anomalies congénitales, les infections respiratoires et les pathologies du nouveau né restent sous-étudiées. Par ailleurs, ces études montrent que la plus part de la recherche faite dans cette région est financée par des organismes publiques, que les auteurs des publications sont affiliés principalement à des organismes de pays riches, et que le peu de recherche financée par des organismes privés n'étudie pas les maladies affectant les populations locales. Une étude récemment publiée a analysé de façon très détaillée 2,292 publications d'essais cliniques portant sur la santé maternelle conduits dans les pays en voie de développement (Chersich and Martin, 2017). Ils ont mis en évidence des états de santé négligés par la recherche relativement à leur fardeau, et des régions participant peu à la recherche relativement à leur développement économique. Ils ont aussi souligné le

manque de coordination entre les différents acteurs finançant ces recherches pour combler ces lacunes, en particulier le NIH, l'USAID et la Bill & Mellinda Gates Foundation. En effet, tous ces acteurs financent de la recherche portant sur les mêmes maladies et dans les mêmes régions. Par ailleurs, sur ce même jeu de données Chersich et al. ont montré que le taux d'articles publiés par un premier auteur venant du pays en voie de développement concerné par l'étude varie considérablement selon la source de financement (Chersich et al., 2016). Finalement, d'autres travaux ont montré que les grands journaux médicaux (comme *The BMJ*, *The New England Journal of Medecine* ou *The Lancet*) ne publiaient pas des travaux portant sur les principales causes des fardeau (Rochon et al., 2004; Perel et al., 2008).

Revues systématiques

Plusieurs études ont comparé la couverture des thématiques étudiées par des revues systématiques aux besoins de santé. Globalement, il a été montré que la quantité de revues systématiques portant sur une maladie donnée est correctement aligné avec le poids de cette maladie dans le fardeau global (Swingler et al., 2003; Yoong et al., 2015). D'autres séries d'articles ont étudié la couverture des revues systématiques Cochrane spécifiquement à la dermatologie (Karimkhani et al., 2014), l'ophtalmologie (Boyers et al., 2015), l'oto-rhino-laryngologie (Pederson et al., 2015) et aux blessures (Karimkhani et al., 2016). Dans tous les cas, des maladies spécifiques ont été identifiées comme étant sous-étudiées par les revues systématiques.

Nouveaux traitements

L'analyse des nouveaux traitements disponibles sur le marché permet d'identifier des lacunes dans les résultats de recherche. Par exemple, parmi 1,393 médicaments mis sur le marché entre 1975 et 1999, seulement 1% visait à traiter les maladies tropicales négligées ou la tuberculose (Trouiller et al., 2002). La même

analyse faite 10 ans plus tard montre que parmi 850 médicaments mis sur le marché entre 2000 et 2011, le pourcentage était monté à 4% (Pedrique et al., 2013). Or, les maladies tropicales correspondent à 12% du fardeau global. De même, en 2008 il a été mis en avant qu'une seule nouvelle classe de drogues a été mise sur le marché pour les soins obstétricaux en deux décennies (Fisk and Atun, 2008). Par ailleurs, l'activité de recherche dans ce domaine représente seulement 3% de celle portant sur les maladies cardiovasculaires.

1.3.4 Analyses d'interactions

Dans la littérature, quelques travaux ont utilisé les données traçant l'activité de recherche pour étudier les interactions entre éléments du système de recherchee (Section 1.2.3, troisième axe d'analyse).

Tsalatsanis et al. ont construit des réseaux d'essais randomisés en oncologie tels que chaque nœud correspond à un essai randomisé, et deux nœuds sont reliés par une arête si les essais partagent des caractéristiques comme le plan expérimental ou le type d'intervention évaluée (Tsalatsanis et al., 2011). L'objectif était d'évaluer la relation entre le succès d'un essai et la position de cet essai dans le réseau, mais aucune relation n'a été trouvée. En partant aussi des essais en oncologie, Cheng et al. ont construit des réseaux de sites de recrutement tels que chaque nœud correspond à un site de recrutement, et deux sites sont reliés s'ils on y a fait des essais sur des types de cancer similaires (Cheng et al., 2012). Ils ont montré que les sites étaient plus interconnectés dans les essais de phase avancée (III, IV) que dans les phases précoce (I, II).

Les publications scientifiques ont été largement analysées sous forme de réseaux. Par exemple, des travaux ont étudié les collaborations entre universités dans la recherche scientifique en utilisant les affiliations des co-auteurs des d'articles de recherche (Jones et al., 2008). De même, plusieurs travaux ont travaillé sur des réseaux de citations, dans lesquels les nœuds peuvent correspondre à des articles de recherche, des chercheurs, ou des équipes de recherche, et deux nœuds sont reliés

s'ils se citent les uns les autres. Par exemple, à partir de réseaux de citations il a été montré l'influence de la distance géographique dans la collaboration scientifique (Pan et al., 2012). Dans le domaine médical, des réseaux de citation ont été utilisés, par exemple, pour montrer la distance existante entre communautés académiques ayant des positionnements opposés vis-à-vis de l'influence de la consommation de sel dans les maladies cerebro-cardiovasculaires (Trinquart et al., 2016), ou bien pour étudier l'évolution dans le temps de l'interdisciplinarité dans la recherche concernant le VIH (Adams and Light, 2014), ou pour montrer la centralité des chercheurs financés par l'industrie pharmaceutique dans la production de recherche médicale (Dunn et al., 2012).

Chapitre 2

Cartographies des essais cliniques enregistrés

J'ai réalisé deux cartographies de la recherche clinique donnant lieu à la production de deux articles scientifiques. Ces deux cartographies ont été faites à grande échelle à partir des registres d'essais cliniques. La première avait pour but d'analyser les maladies étudiées par les essais cliniques randomisés dans chaque région du monde, et étudier pour chaque région l'alignement entre les maladies étudiées par la recherche et le fardeau des maladies. La deuxième cartographie avait pour objectif d'étudier l'influence du type de promoteur, industriel ou non-industriel, dans la globalisation des essais cliniques, et étudier les différentes capacités des promoteurs à conduire des essais cliniques internationaux. Ces travaux ne sont pas présentés par ordre chronologique, mais suivant l'ordre des axes d'analyse présentés dans la Section 1.2.3.

2.1 Registres d'essais cliniques

Ces deux travaux ont été fondés sur les registres d'essais cliniques publiquement accessibles par le portail de l'OMS, l'ICTRP (World Health Organization,

2005). Ce portail réunit à ce jour les essais cliniques enregistrés dans 16 registres nationaux, régionaux ou internationaux (Tableau 2.1).

Registre	Nombre d'essais cliniques enregistrés	Pourcentage d'essais cliniques enregistrés (%)
ClinicalTrials.gov	119 046	68.6
EU Clinical Trials Register	21 268	12.3
Japan Primary Registries Network	11 251	6.5
International Standard Randomized Controlled Trial Number Register	9174	5.3
Australian New Zealand Clinical Trials Registry	7900	4.6
Iranian Registry of Clinical Trials	5266	3.0
Clinical Trials Registry-India	3731	2.2
The Netherlands National Trial Register	3481	2.0
Chinese Clinical Trial Register	2924	1.7
German Clinical Trials Register	1431	0.8
Clinical Research Information Service, Republic of Korea	679	0.4
Pan African Clinical Trial Registry	289	0.2
Brazilian Clinical Trials Registry	272	0.2
Cuban Public Registry of Clinical Trials	169	0.1
Sri Lanka Clinical Trials Registry	116	0.1
Thai Clinical Trials Registry	95	0.1

TABLE 2.1 – Caractéristiques des essais cliniques enregistrés dans le International Clinical Trials Registry Platform entre 2005 et 2013
N=173 532. Source : (Viergever and Li, 2015)

Il est important de remarquer qu'un même essai peut être enregistré dans plusieurs registres à la fois, et un des rôles de l'ICTRP est d'identifier et de regrouper ces doublons. Aussi, un essai clinique peut être enregistré dans le registre d'un pays différent à celui dans lequel a lieu l'essai. Par exemple, 57% des essais cliniques enregistrés dans ClinicalTrials.gov, le registre des États-Unis, ne recrutent leurs patients qu'en dehors des États-Unis (U.S. National Institutes of Health, 2000).

Pour être inclus dans l'ICTRP, les essais cliniques enregistrés dans ces registres

doivent rapporter en anglais un ensemble minimum d'information. En particulier, les informations suivantes doivent être renseignées :

— Date d'enregistrement et date planifiée de début de l'essai ;
— Promoteur principal (responsable légal de la conduite de l'étude) et promoteurs secondaires ;
— Intitulés succinct et détaillé de l'étude ;
— Pays de recrutement des patients ;
— Maladie(s) ou état(s) de santé étudié(s) ;
— Intervention(s) étudiée(s) ;
— Type d'étude (interventionnel ou observationnel) et plan d'expérience ;
— Taille planifiée de l'étude ;
— État d'avancement du recrutement (en attente, en cours, suspendu, complété) ;
— Critères de jugement principal et secondaire(s).

En plus de ces informations, chaque registre peut demander des informations supplémentaires pour l'enregistrement d'un essai. Par exemple, dans ClinicalTrials.gov on trouve les coordonnées de chaque centre de recrutement par pays. On y trouve aussi de façon standardisé le type des promoteurs principal et secondaires, notamment s'ils correspondent à des industriels médicamenteux.

L'ICTRP est actuellement la seule base de données publiquement accessible et prête à l'utilisation avec laquelle il est possible de faire des cartographies de la recherche clinique à grande échelle et en amont du biais de publication.

2.2 Alignement entre l'effort de recherche et les besoins de santé

2.2.1 Résumé

Ce travail avait pour objectif d'évaluer l'alignement entre l'effort de recherche clinique et les besoins de santé publique pour toutes les régions du monde et un grand nombre de maladies. Nous avons considéré le nombre d'essais randomisés effectués dans chaque région comme mesure de l'effort de recherche clinique, et le fardeau des maladies comme mesure des besoins de santé (Section 1.2.3, deuxième axe d'analyse).

Nous avons considéré 7 régions définies par proximité épidémiologique dans l'étude GBD 2010 (correspondant aux super-régions présentées dans la Figure 2.1) (Murray et al., 2012). Nous avons considéré 27 grands groupes de maladies correspondant à 98.9% du fardeau global des maladies mesurées en années de vie corrigées de l'incapacité (DALY). Ces groupes ont été définis à partir des catégories de maladies définies dans l'étude GBD 2010 pour évaluer le fardeau de maladies. Le regroupement en 27 groupes a été défini par un panel d'experts comme étant suffisamment informatif pour créer des cartographies globales des essais cliniques par grandes thématiques de santé publique (Section 3.2).

Pour chaque région et chaque groupe de maladies nous avons compté le nombre d'essais randomisés enregistrés commencés entre 2006 et 2015 et inclus dans l'IC-TRP. Nous avons classifié les essais cliniques par groupe de maladies en utilisant une méthode de classification automatique que nous avons développée et qui est présentée dans la Section 3.2. La classification n'étant pas parfaite, nous avons incorporé de l'incertitude au nombre d'essais randomisés par groupe de maladies en tenant en compte les valeurs prédictives positives et négatives de classification spécifique à chaque groupe de maladies (voir Annexe page 165). Ceci nous a permis de dériver des intervalles d'incertitude à 95% du nombre d'essais par groupe

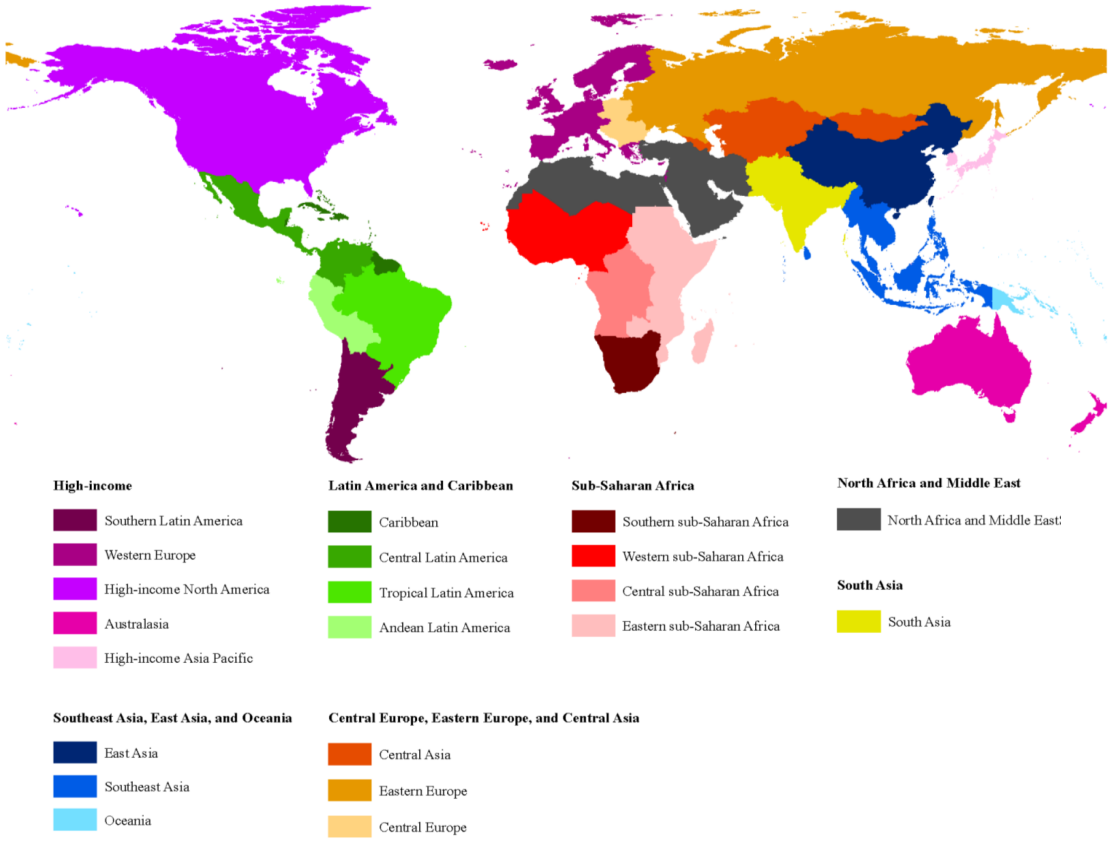


FIGURE 2.1 – Régions et super-régions GBD
Source : (Murray et al., 2012)

de maladies dans chaque région. Pour chaque région et chaque groupe de maladies nous avons évalué le fardeau en 2005 en DALY (Lozano et al., 2012).

Nous avons comparé le nombre d'essais initiés dans la période 2006-2015 au fardeau de 2005, en supposant que les besoins en 2005 pourraient influencer le programme de recherche des 10 années suivantes. Nous avons évalué le nombre d'essais randomisés par million de DALY par région et par groupe de maladies.

Au sein de chaque région, nous avons identifié des groupes de maladies pour lesquelles il existait des lacunes dans l'effort de recherche local. Pour cela, nous avons défini *a priori* une lacune dans l'effort de recherche comme étant un groupe de maladies tel que sa part dans la totalité de la recherche locale est inférieure à la moitié de sa part dans le fardeau local. Par exemple, si 40% du fardeau de l'Afrique

Subsaharienne était dû au VIH, et moins de 20% des essais cliniques en Afrique Subsaharienne étudiaient le VIH, nous avons considéré qu'il existait une lacune dans l'effort de recherche sur le VIH en Afrique Subsaharienne. Cette définition permet d'identifier des grandes disparités dans l'effort de recherche relativement au fardeau de chaque maladie. Par ailleurs, les lacunes de recherche ont été rapportées seulement quand elles étaient robustes vis-à-vis de l'incertitude du nombre d'essais par groupe de maladies.

Similairement, pour chaque groupe de maladies, nous avons identifié des disparités dans la distribution de l'effort de recherche à travers les régions à revenu moyen ou faible relativement au fardeau de chaque région. Nous avons identifié les régions telles que leur participation dans l'effort de recherche sur une maladie était inférieure à la moitié de leur part dans le fardeau global de la maladie. Par exemple, si 20% du fardeau du paludisme affectait l'Asie du Sud, mais moins de 10% des essais cliniques sur le paludisme se sont fait en Asie du Sud, nous avons considéré qu'il y avait une lacune dans la distribution de l'effort de recherche sur le paludisme entre les régions. Cette analyse est faite uniquement à travers les régions autres que celle à revenu élevé, puisque la part de cette dernière dans l'effort de recherche global est écrasante vis-à-vis des autres régions (Section 1.3).

Par ailleurs, nous avons conduit des analyses secondaires avec d'autres mesures de fardeau, en particulier le nombre de morts, le nombre d'années de vie perdues due à une mort précoce et le nombre d'années vécues avec une incapacité pour des maladies non fatales. Nous avons aussi mesuré l'activité de recherche en tenant en compte le nombre de patients planifiés d'être recrutés dans chaque essai clinique.

Nous avons cartographié 117,180 essais cliniques randomisés ayant planifié de recruter 44.0 millions de patients, et 2,220 million de DALY. Il y a eu 200 fois plus d'essais randomisés par million de DALY dans les pays à revenu élevé que dans les autres régions (130.9 vs 6.9). L'effort de recherche dans les pays à revenu élevé était bien aligné avec leurs besoins : on n'a pas identifié de maladie présentant une lacune locale dans l'effort de recherche. Par contre, dans toutes les autres régions

nous avons identifié des lacunes dans l'effort de recherche local. Par exemple, en Afrique Subsaharienne et Asie du Sud nous avons identifié des lacunes dans l'effort de recherche sur les maladies infectieuses communes (5.8% [intervalle d'incertitude à 95% 4.7–6.9] des essais vs 22.9% des DALY en Afrique Subsaharienne, et 7.0% [5.8–8.3] des essais vs 21.4% des DALY en Asie du Sud) et sur les pathologies du nouveau né (2.0% [0.9-4.5] des essais vs 11.6% des DALY en Afrique Subsaharienne, et 2.3% [1.1-4.9] des essais vs 16.5% des DALY en Asie du Sud). Nous n'avons pas identifié de lacune de l'effort de recherche en Afrique Subsaharienne pour le VIH ni pour le paludisme, correspondant tous les deux à 30.1% du fardeau de la région. Par rapport aux autres régions à revenu moyen ou faible, la part de l'Afrique Subsaharienne dans l'effort global de recherche sur les maladies infectieuses était faible.

Les comparaisons entre effort de recherche et fardeau pour les autres régions et maladies sont accessibles via une visualisation interactive en ligne sur https ://clinicalepidemio.fr/RCTvsBurden (Figure 2.2). Dans ce site il est possible de mesurer l'effort de recherche en tenant en compte du nombre de patients inclus dans les essais, et d'utiliser d'autres mesures de fardeau. Ainsi, les décideurs des programmes de recherche pourront évaluer les différences entre effort de recherche et fardeau pour leurs maladies d'intérêt, leur région d'intérêt, en utilisant différentes mesures.

Cette analyse à grande échelle de l'alignement entre l'effort de recherche clinique et les besoins de santé publique a confirmé à l'échelle de la totalité des essais cliniques enregistrés que la plupart des essais randomisés sont faits dans les pays à revenu élevé sur des maladies les affectant. Il existe des lacunes majeures dans l'effort de recherche dans des régions à revenu faible, en particulier en Afrique subsaharienne et en Asie du Sud sur des maladies prédominantes dans ces régions, notamment les maladies infectieuses et les pathologies du nouveau né.

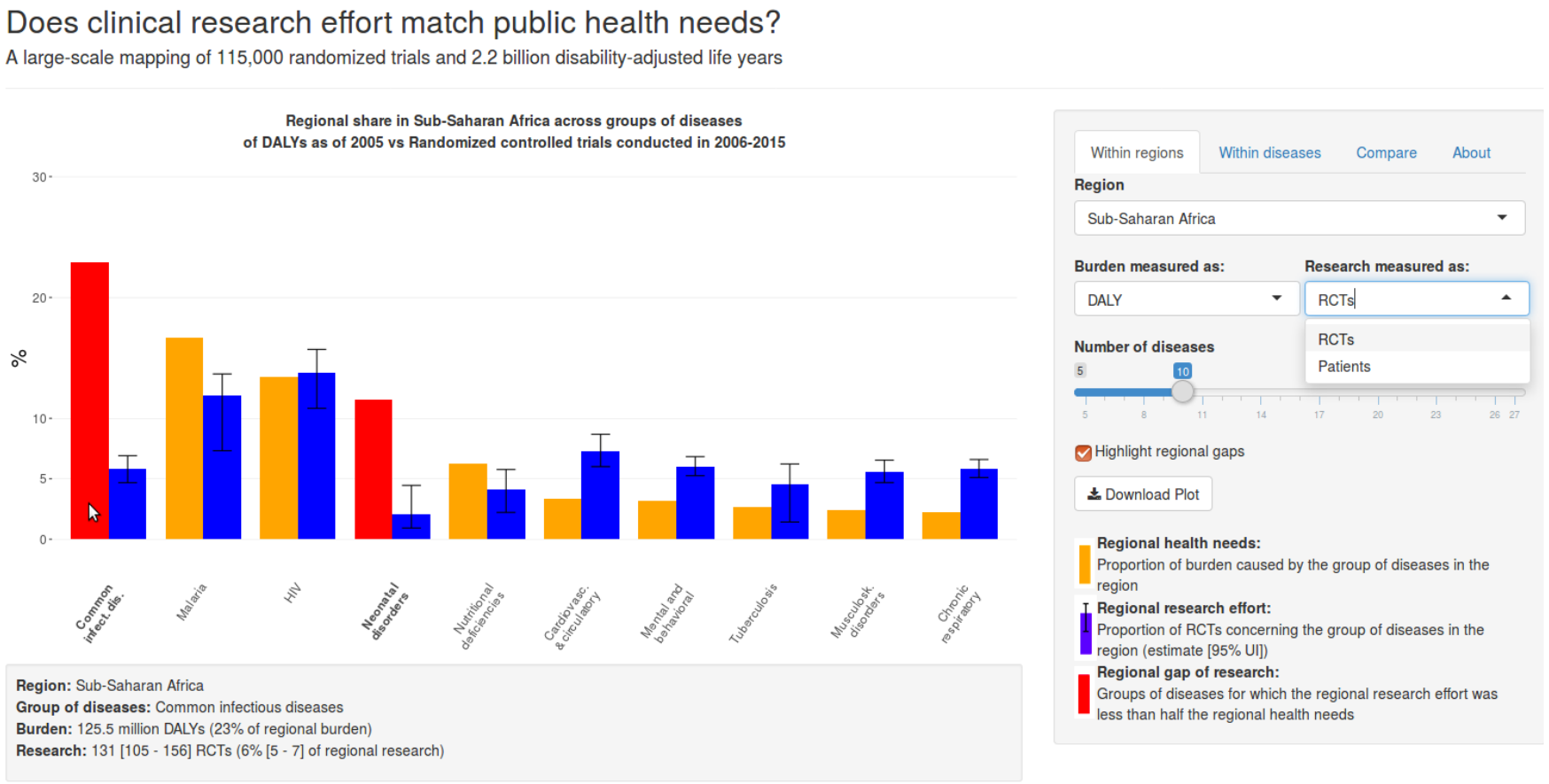


FIGURE 2.2 – Visualisation interactive pour comparer l'effort de recherche au fardeau des maladies à grande échelle https ://clinicalepidemio.fr/RCTvsBurden/

Does clinical research effort match public health needs? A large-scale mapping of 115,000 randomized trials and 2.2 billion disability-adjusted life years

Authors: Ignacio ATAL (MSc), Ludovic TRINQUART (PhD), Philippe RAVAUD (MD, PhD), Raphaël PORCHER (PhD)

Authors affiliations:

Centre d'Épidémiologie Clinique, Hôpital Hôtel-Dieu, Paris, France (I Atal, R Porcher, P Ravaud, L Trinquart)

INSERM U1153, Paris, France (I Atal, R Porcher, P Ravaud, L Trinquart)

Columbia University, Mailman School of Public Health, Epidemiology Department, New York, NY, USA (P Ravaud, L Trinquart)

Université Paris Descartes, Paris, France (I Atal, R Porcher, P Ravaud)

Boston University School of Public Health, MA, USA (L Trinquart)

Word count: Abstract: 273; Manuscript: 3641

Corresponding author: Ignacio Atal, MSc
Centre d'Epidémiologie Clinique, Hôpital Hôtel-Dieu,
1, place du parvis Notre Dame, 75004 Paris, France
Tel: (+33) 1 42 34 79 02
E-mail: ignacio.atal-ext@aphp.fr

ABSTRACT

Background: Concerns exist as to whether the allocation of resources in clinical research is aligned with public health needs. We evaluated the alignment between the effort of clinical research through the conduct of randomized controlled trials (RCTs) and health needs measured as the burden of diseases for all regions and a broad range of diseases.

Methods: We grouped countries into 7 regions and diseases into 27 groups. We mapped all RCTs initiated between 2006 and 2015 that were registered at the WHO International Clinical Trials Registry Platform to regions and diseases. The burden of diseases in 2005 was mapped as disability-adjusted life years (DALYs), based on the 2010 Global Burden of Diseases study. Within regions, we defined a research gap when the proportion of RCTs concerning a disease in the region was less than half the relative burden of the disease.

Results: We mapped 117 180 RCTs planning to enroll 44.0 million patients, and 2220 million DALYs. In high- versus non–high-income countries, 130.9 versus 6.9 RCTs per million DALYs were conducted. We did not identify any research gap in high-income countries. For Sub-Saharan Africa, we identified research gaps for common infectious diseases (CID) and neonatal disorders (ND): 5.8% [95% uncertainty interval 4.7–6.9] and 2.0% [0.9-4.5] of RCTs in Sub-Saharan Africa concerned CID and ND, although these diseases represented 22.9% and 11.6% of the burden in the region, respectively. For South Asia we identified research gaps for the same two groups of diseases.

Conclusions: For Sub-Saharan Africa and South Asia, the distribution of the research effort was misaligned with the distribution of the burden for two major causes of burden, CID and

ND.

Keywords: clinical trials, burden of diseases, mapping, research priorities, research gaps

What was already known about the topic

Health research efforts are sparse in low- and middle-income countries as compared to high-income countries. Studies have shown that health efforts are globally aligned with high-income countries needs, and that diseases prevalent in low-income countries are neglected. On the contrary, other studies have suggested that local research effort in Sub-Saharan Africa is aligned with the local burden.

What new knowledge the manuscript contributes

Based on all randomized controlled trials registered in the WHO ICTRP, we give a detailed analysis on the local alignment between a 10 years health research effort and health needs for seven epidemiological regions. We defined a criteria for misalignment between the conduct of randomized controlled trials and the burden of diseases to highlight major research gaps within regions. We support at a large scale previous knowledge on the unequal distribution of health research efforts across regions and diseases. We showed that in high-income countries health research efforts were aligned with health needs. In all other regions we identified local research gaps relatively to the burden. For Sub-Saharan Africa, highly prevalent diseases as HIV and malaria were indeed receiving high research effort, but other major causes of burden remain neglected by research, in particular common infectious diseases and neonatal disorders. We highlighted other research gaps not stated elsewhere, in particular concerning common infectious diseases and neonatal disorders in South Asia, and cardiovascular and circulatory diseases in Eastern Europe and Central Asia.

INTRODUCTION

The global landscape of epidemiology is complex. Infectious diseases are predominant in low-income countries but high-income countries are mainly affected by non-communicable diseases.[1] Nevertheless, the incidence of diseases predominant in poor countries is increasing among poorer people in rich countries, when the incidence of non-communicable diseases is increasing in low- and low- to middle-income countries.[2, 3]

In this changing landscape, concerns have been raised regarding the alignment of the allocation of clinical research and public health needs.[4, 5] Clinical research activities, and in particular the conduct of randomized controlled trials (RCTs), may be driven by specific interests or constraints that may differ from local health priorities.[6] A comprehensive mapping of RCTs used as a proxy for clinical research effort may be helpful to understand the processes guiding clinical research, and to steer limited resources toward local health priorities, particularly in low resource settings.[4, 7]

Several studies have shown that research is lacking in low-income countries[8, 9] and that diseases receiving the most research attention are those that are predominant in high-income countries.[4, 10] Other studies have suggested that in low-income regions such as Sub-Saharan Africa, the conduct of RCTs is aligned with the burden across diseases.[11] However, previous studies focused on specific regions or specific diseases, and a global-scale analysis may bring novel insights.

We evaluated the alignment between the research effort (measured as the number of RCTs conducted) and the burden of disease across all world regions and a broad range of diseases. Within each region, we estimated the research effort across diseases, and

identified the diseases for which the research effort was too low as compared with the burden they cause. At a global level, for each disease, we estimated the research effort across non–high-income regions, and identified the regions for which the research effort was too low as compared with the regional disease burden.

METHODS

We compared the effort in clinical research to the health needs across regions and diseases. The number of RCTs was used to the measure the research effort, and the burden of diseases to measure health needs. By using clinical trial registries, we mapped the RCTs initiated between 2006 and 2015 to 7 regions and 27 groups of diseases. By using the 2010 Global Burden of Diseases (GBD) study,[1] we mapped the burden in 2005. For each region, we analyzed the distribution of the research effort across groups of diseases and identified diseases for which the regional effort of research was lacking as compared to the regional burden. For each group of diseases, we analyzed the distribution of the research effort across regions, excluding high-income countries, and identified regions for which the disease-specific effort of research was lacking as compared with the disease burden.

Mapping the effort of clinical research

We downloaded all records of clinical trials registered in the World Health Organization International Clinical Trials Registry Platform (WHO ICTRP) by January 1, 2016.[12] We identified RCTs according to the study type and study design fields of the trial records (e.g., by excluding observational and non-randomized trials, see Supplementary Information). Because the International Committee of Medical Journal Editors recommended registration before considering interventional trials for publication since September 2005, we restricted our analyses to RCTs enrolling the first patient after January 2006.[13]

Country locations were extracted from the clinical trial records. We categorized countries into

7 epidemiological regions defined in the 2010 GBD study: high-income countries, Latin America and Caribbean, Eastern Europe and Central Asia, South Asia, Southeast and East Asia and Oceania, North Africa and Middle East, and Sub-Saharan Africa. Countries not included in the 2010 GBD study were excluded.

We classified RCTs in terms of 27 predefined groups of diseases.[14] RCTs were classified automatically by using a validated knowledge-based classifier. Trials classified for none of these disease groups may have studied non-disease contributors to morbidity (injuries), health conditions considered not relevant for burden estimation by the 2010 GBD study (e.g., pain management), or residual causes of burden excluded from the 27-class grouping.[1]

We then calculated the number of RCTs in each region for each group of diseases as a measure of clinical research effort. Multi-regional RCTs and RCTs concerning more than one group of diseases were counted in each region and each group of diseases.

We accounted for potential misclassification of an RCT by the knowledge-based classifier by using a probabilistic analysis (see Supplementary Information).[15] For each group of diseases, the true sensitivity and specificity are not known with certainty; thus we first simulated a sensitivity and specificity of the classifier for each group of diseases by using the observed true positive and negative rates and their uncertainty (Table S1).[14] Then we simulated a corrected number of RCTs for this group of diseases by using the simulated sensitivity and specificity of the classifier.[15] This process was repeated 10 000 times. We reported the median and the 2.5^{th}–97.5^{th} percentiles as estimates and 95% uncertainty interval (UI), respectively.

Mapping health needs

We measured health needs by using disability-adjusted life years (DALYs), a unique metric summarizing the years of life lost due to premature death and the years lived with disability. We used the estimation of DALYs in 2005 from the 2010 GBD study.[1] The DALYs for each region and each of the 27 groups of diseases was estimated as the sum of DALYs across the component countries and causes of disability or death reported in the 2010 GBD study, respectively.

Comparison between research effort and health needs

With the hypothesis that health needs in 2005 should drive the research agenda in 2006-2015, we compared the number of RCTs initiated in 2006-2015 to the burden in 2005.

First, we compared regions according to the total number of RCTs per DALYs. Then, we analyzed the distribution of the research effort across groups of diseases to identify global research gaps. At a world level, we considered a global research gap for a group of diseases when the proportion of global research effort concerning that disease was less than half the proportion of the global burden caused by that disease. A research gap was claimed only if the two-fold difference was maintained over the 95% UI of the proportion of RCTs. For instance, if 30% of the global burden was caused by neoplasms, we considered a global research gap for this disease if the upper limit for the 95% UI of the proportion of RCTs concerning neoplasms was less than 15%.

We conducted the same analysis within regions to identify regional research gaps. For instance, if 30% of the total burden in Sub-Saharan Africa was caused by HIV/AIDS, we

considered a regional research gap for this disease if the upper limit for the 95% UI of the proportion of RCTs conducted in Sub-Saharan Africa and concerning HIV/AIDS was less than 15%. In addition, we estimated the proportion of RCTs that should be reallocated between groups of diseases within a region so as to eliminate research gaps. In the case presented above, if the proportion of RCTs conducted in Sub-Saharan Africa and concerning HIV/AIDS was 10%, then reallocating 5% of trials would eliminate this gap.

Finally, for each group of diseases, we analyzed the distribution of the research effort concerning the diseases across regions to identify disease-specific research gaps. We excluded high-income countries from these analyses to focus on the allocation of research effort across non–high-income countries. We considered a disease-specific research gap in a region when the upper limit for the 95% UI of the proportion of the disease-specific research effort in the region among all non–high-income regions was less than half the proportion of the burden affecting that region. We also estimated a proportion of RCTs that should be reallocated across non–high-income regions to eliminate these gaps. For instance, if 20% of the burden caused by malaria affected South Asia but only 4% of RCTs conducted in non–high-income countries were conducted in South Asia, then reallocating 6% of RCTs would eliminate this gap.

Secondary analyses

We conducted similar analyses by considering the sample size of the RCTs instead of the number of RCTs as a measure of clinical research effort. We extracted the target sample size of each RCT from the registry record (see Supplementary Information). When the sample

size was not available or when the sample size was < 10 and > 200 000, we excluded RCTs from the analyses by sample size.

In addition, we conducted sensitivity analyses by using 3 other measures of the burden, namely years of life lost (YLL), years lived with disability (YLD) and number of deaths, as provided by the 2010 GBD study.

Research reproducibility

All data analyses involved use of R 3.3.1 (R Development Core Team, Vienna, Austria). Data underlying the results are available within the article, and the code used is available as a Jupyter notebook (github.com/iatal/RCTvsBurden). It is shared as open source under the MIT license. An interactive visualization tool allowing for comparison between research effort (in RCTs or patients planned to be enrolled) and disease burden (in DALY, YLL, YLD or number of deaths) for the 7 regions and 27 groups of diseases is available at https://clinicalepidemio.fr/RCTvsBurden/.

RESULTS

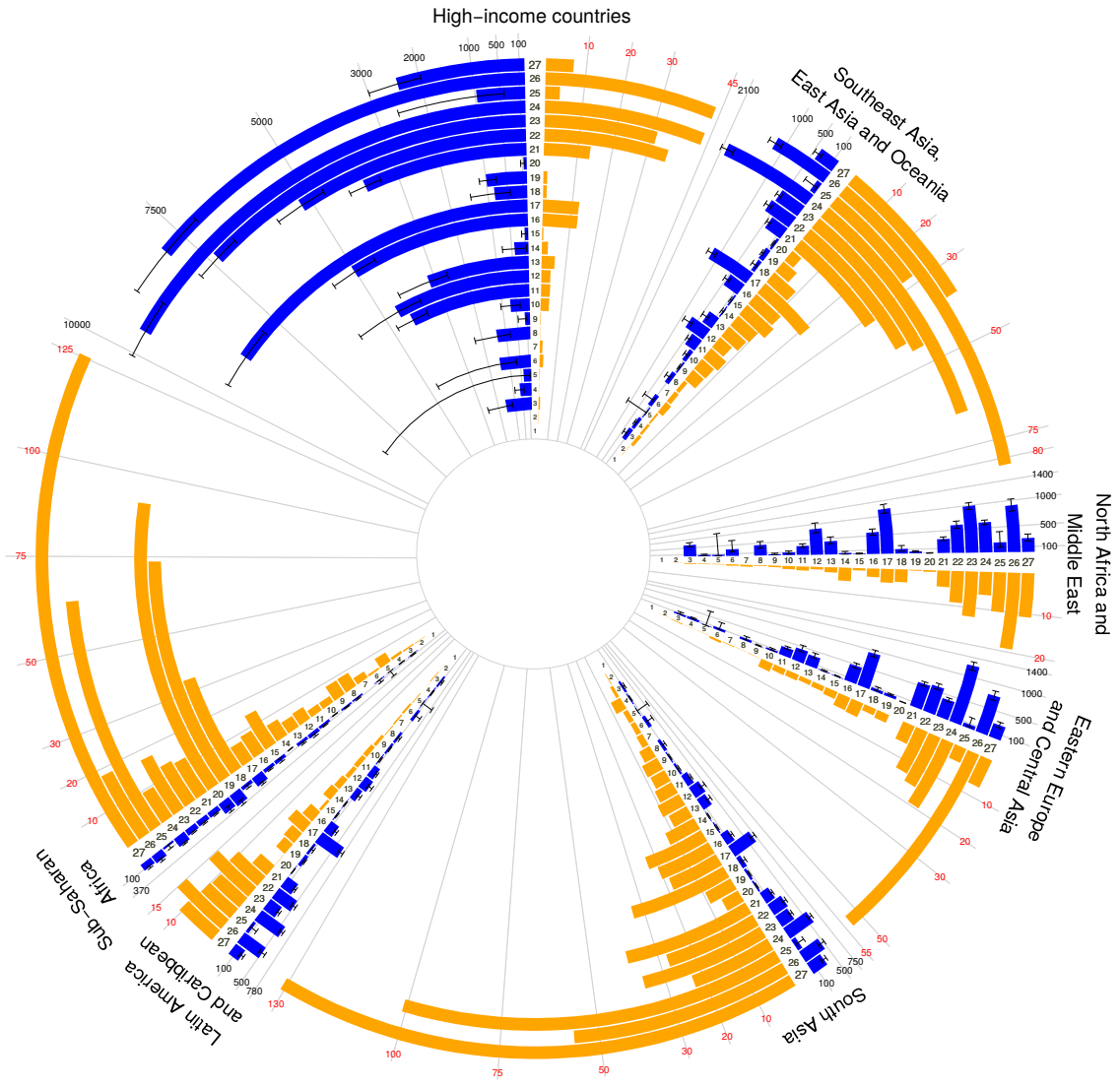
Mapping the effort of clinical research

We analyzed 117 180 registered RCTs initiated between 2006 and 2015: 107 558 planned to enroll 44.0 million patients (Figure S1). Overall, an estimated 82 179 RCTs [95% UI 78 662–85 358] were relevant to the burden of diseases. For high- versus non–high-income countries, 60 631 [58 035–62 973] versus 27 564 [26 405–28 597] RCTs were relevant to the burden of diseases (Figure 1), and 19.0 [18.0–19.9] versus 11.1 [10.4–11.8] million patients were planned to be enrolled in these RCTs (Figure S2). Groups of diseases concerned by the highest number of RCTs were neoplasms (12 024 [11 148–12 728]), diabetes, urinary diseases and male infertility (11 700 [10 614–12 854]) and cardiovascular and circulatory diseases (10 676 [9 527–11 943]).

Mapping health needs

We mapped 2220 million DALYs across the 7 regions and the 27 groups of diseases. The region with the highest disease burden in 2005 was South Asia (610 million DALYs), followed by Sub-Saharan Africa (548 million DALYs) (Figure 1). The groups of diseases causing the highest burden in 2005 were common infectious diseases (329 million DALYs), cardiovascular and circulatory diseases (287 million DALYs) and neonatal disorders (220 million DALYs).

Figure 1: Number of randomized controlled trials (RCTs) versus number of disability-adjusted life years (DALYs) for the 7 regions and the 27 groups of diseases

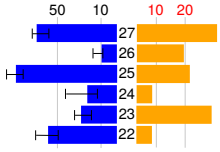


Groups of diseases

27 Common infectious diseases
26 Cardiovascular and circulatory diseases
25 Neonatal disorders
24 Neoplasms
23 Mental and behavioral disorders
22 Musculoskeletal disorders
21 Chronic respiratory diseases
20 Malaria
19 HIV/AIDS
18 Nutritional de ciencies
17 Diabetes, urinary diseases and male infertility
16 Neurological disorders
15 Tuberculosis
14 Congenital anomalies

Regional research e ort
(Number of RCTs [95% UI])

Regional health needs
(Million DALYs)



Groups of diseases

Skin and subcutaneous diseases 13
Digestive diseases (except cirrhosis) 12
Sense organ diseases 11
Cirrhosis of the liver 10
Neglected tropical diseases excluding malaria 9
Maternal disorders 8
Hemoglobinopathies and hemolytic anemias 7
Oral disorders 6
Sexually transmitted diseases excluding HIV 5
Hepatitis 4
Gynecological diseases 3
Sudden infant death syndrome 2
Leprosy 1

Figure 1: Number of randomized controlled trials (RCTs) versus number of disability-adjusted life years (DALYs) for the 7 regions and the 27 groups of diseases: *Number of registered RCTs initiated in 2006-2015 and number of DALYs (per million) in 2005, for the 7 regions across the 27 groups of diseases. For number of RCTs, shows the estimates and 95% uncertainty intervals (UIs) after incorporating classification uncertainty of RCTs across groups of diseases. Regions are ordered clockwise by the total number of RCTs relevant to the burden of diseases. Groups of diseases are ordered according by the global burden.*

Comparison between research effort and health needs

During 2006-2015, 37.0 RCTs [95% UI 35.4–38.4] per million DALYs in 2005 worldwide were conducted, 130.9 [125.3–136.0] in high-income countries versus 6.9 [6.6–7.2] in non–high-income countries. We identified global research gaps for common infectious diseases (4.2% [3.4–5.2] of RCTs vs 14.8% of DALYs), neonatal disorders (1.7% [0.8–4.3] vs 9.9%), malaria (0.5% [0.3–0.6] vs 4.6%), HIV (1.7% [1.3–2.1] vs 4.3%) and tuberculosis (0.4% [0.1–0.5] vs 2.5%).

Regional research gaps

All numbers of RCTs and DALYs within regions per groups of diseases are presented in Figure 1. Global and regional research gaps differed (Figure 2). We did not identify any research gaps in high-income countries. For South Asia, we identified research gaps for common infectious diseases (7.0% [95% UI 5.8–8.3] of RCTs vs 21.4% of DALYs) and Neonatal disorders (2.3% [1.1-4.9] vs 16.5%). We also identified research gaps for these 2

causes of burden in Sub-Saharan Africa (5.8% [4.7–6.9] vs 22.9% and 2.0% [0.9-4.5] vs 11.6%, respectively), with no research gaps for malaria and HIV (11.9% [7.3–13.7] vs 16.7% and 13.8% [10.8–15.7] vs 13.4%, respectively). For North Africa and Middle East we observed a research gap for common infectious diseases (3.3% [2.6-4.2] vs 10.5%), and in Latin America and Caribbean for neonatal disorders (1.6% [0.7-4.0] vs 8.7%). For Eastern Europe and Central Asia we identified a research gap for cardiovascular and circulatory diseases (12.5% [11.1–14.0] vs 35.1%). For Southeast Asia, East Asia and Oceania we identified a research gap for tuberculosis (0.6% [0.2-0.9] vs 2.8%). For all regions but high-income countries and Eastern Europe and Central Asia, we observed research gaps in congenital anomalies. For Southeast Asia, East Asia and Oceania, as well as Latin America and Caribbean, less than 5% of RCTs would need to be reallocated across diseases to eliminate regional gaps. For South Asia and Sub-Saharan Africa, the proportion of reallocation needed was higher, 12% and 10% respectively.

Figure 2: Regional proportion of randomized controlled trials (RCTs) and disability-adjusted life years (DALYs) within regions across groups of diseases and regional research gaps

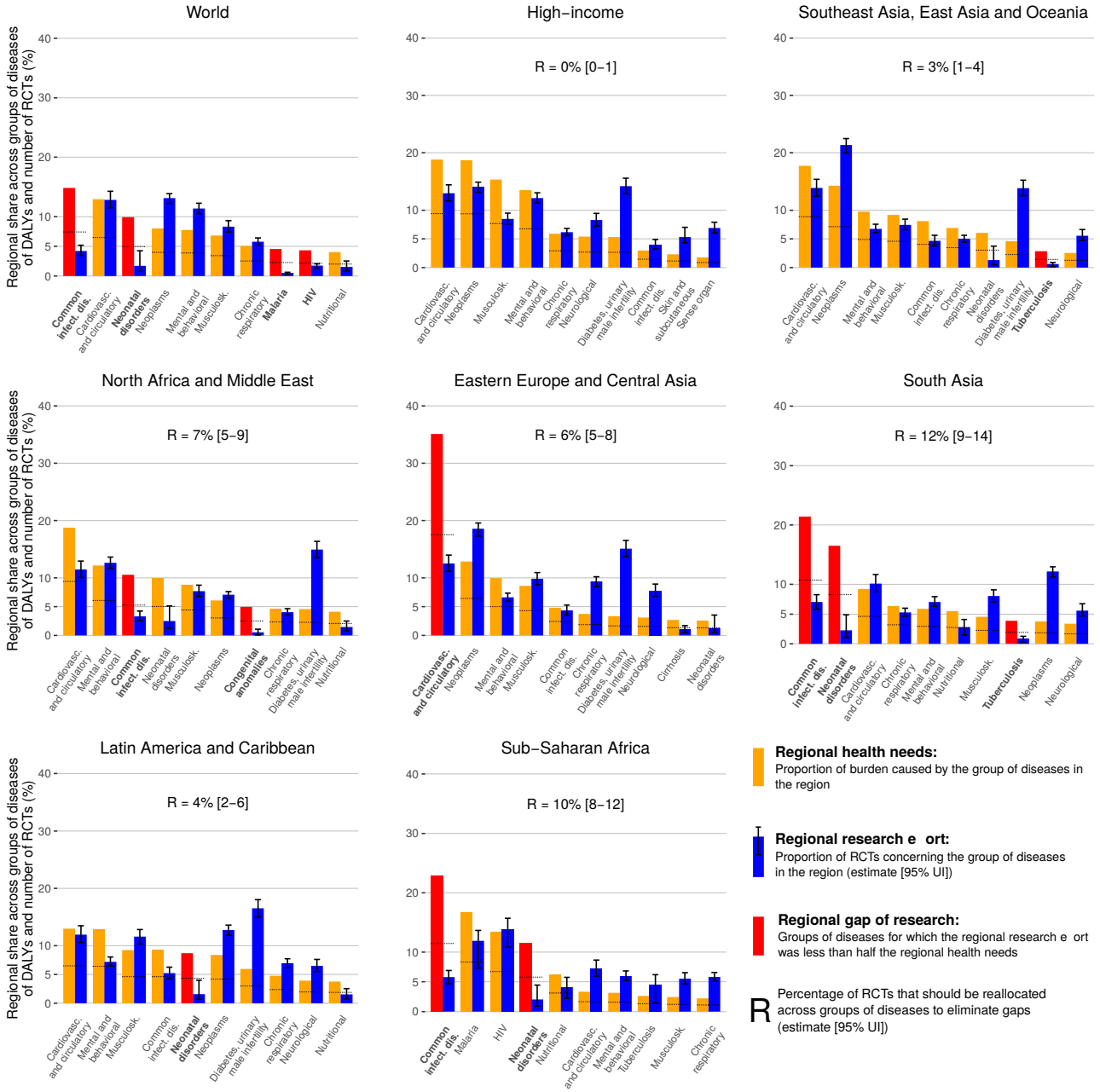


Figure 2: Regional proportion of randomized controlled trials (RCTs) and disability-adjusted life years (DALYs) within regions across groups of diseases and regional research gaps

Worldwide, and for each region, compares the regional research effort (i.e., the proportion of registered RCTs initiated in 2006-2015) to the regional health needs (i.e., the proportion of DALYs in 2005) for the 10 groups of diseases causing the highest burden in the region. For regional research effort, shows the estimates and 95% uncertainty intervals (UIs) after incorporating classification uncertainty of RCTs across groups of diseases. Regions are by the total number of RCTs relevant to the burden of diseases. Groups of diseases are ordered by the regional burden. For each region, highlights groups of diseases showing regional research gaps (i.e., for which the regional research effort was less than half the regional health needs). For each region, annotates the proportion of RCTs that should be reallocated (R) across groups of diseases to eliminate regional research gaps. For other groups of diseases and other measures of burden see https://clinicalepidemio.fr/RCTvsBurden/.

Disease-specific research gaps

When studying the allocation of RCTs across non–high-income regions, disease-specific gaps appeared for Sub-Saharan Africa and South Asia (Figure 3). For common infectious diseases, 9.8% [95% UI 8.8–10.7] of RCTs were conducted in Sub-Saharan Africa, although the region contributed to 38.9% of the disease burden; a reallocation of 10% of these RCTs across non–high-income countries would be needed to eliminate this gap. For neonatal disorders and neurological disorders we documented gaps for Sub-Saharan Africa and South Asia, and the reallocation of RCTs across non–high-income regions needed to eliminate these gaps was 11% and 9%, respectively (Figure 3 and interactive visualization figure at https://clinicalepidemio.fr/RCTvsBurden/).

Figure 3: Disease-specific proportion of randomized controlled trials (RCTs) and disability-adjusted life years (DALYs) across non–high-income regions and disease-specific research gaps

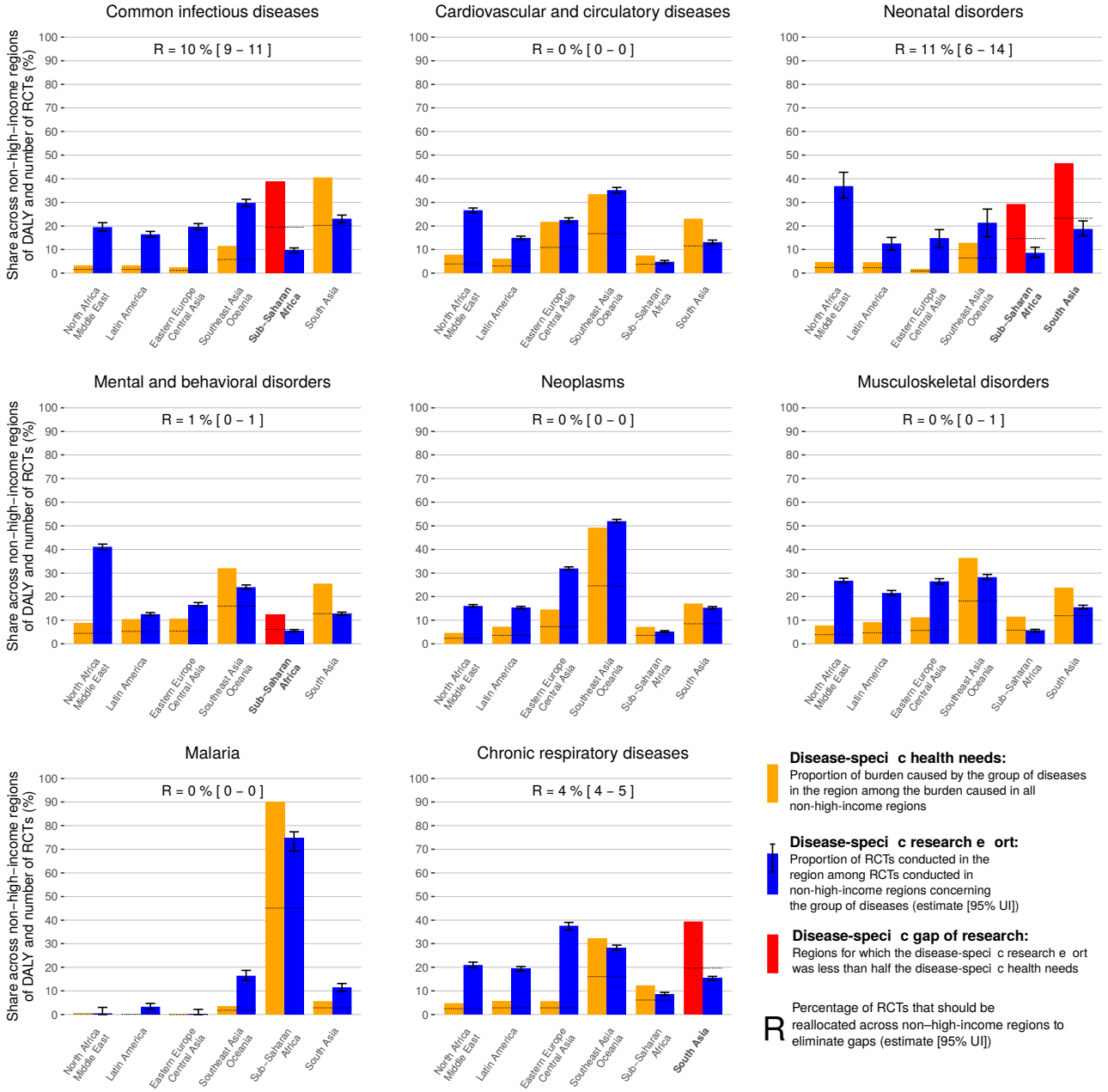


Figure 3: Disease-specific proportion of randomized controlled trials (RCTs) and disability-adjusted life years (DALYs) across non–high-income regions and disease-specific research gaps

For the 8 groups of diseases causing the highest number of DALYs in 2005, compares for each region, excluding high-income countries, the disease-specific research efforts (i.e., the proportion of registered RCTs initiated in 2006-2015) to the disease-specific health needs (i.e., the proportion of DALYs in 2005) among all non–high-income regions. For disease-specific research effort in each region, shows the estimates and 95% uncertainty intervals (UIs) after incorporating classification uncertainty of RCTs across groups of diseases. Groups of diseases are ordered by the total burden in non–high-income regions. Regions are ordered by the total burden. For each group of diseases, highlights regions showing disease-specific research gaps (i.e., for which the disease-specific research effort in the region was less than half the disease-specific health needs of the region). For each group of diseases, annotates the proportion of RCTs that should be reallocated (R) across non–high-income regions to eliminate disease-specific research gaps. Visualization for other groups of diseases and other measures of burden are available at https://clinicalepidemio.fr/RCTvsBurden/.

Secondary analyses

When considering the number of patients planned to be enrolled instead of the number of RCTs as a measure of research effort, we observed fewer gaps in research. For high- vs non–high-income countries, 81.9 [95% UI 77.6-86.0] versus 5.6 [5.2-5.9] thousand patients per million DALYs were planned to be enrolled. All numbers of patients planned to be enrolled and DALYs within regions per groups of diseases are presented in Figure S2. We did not observe any global research gap, nor regional research gap in high-income countries (Figure S3). A regional research gap for common infectious diseases was observed for Sub-Saharan

Africa and North Africa and Middle East. We did not observe any regional research gap for neonatal disorders, nor a research gap for Eastern Europe and Central Asia for cardiovascular and circulatory diseases. Regional gaps for congenital anomalies were maintained in all regions but Sub-Saharan Africa (interactive visualization at https://clinicalepidemio.fr/RCTvsBurden/).

When comparing the share of patients planned to be enrolled to the share of burden across non–high-income regions, only South Asia presented disease-specific gaps in research for malaria, chronic respiratory diseases, and tuberculosis (Figure S4 and interactive visualization figure at https://clinicalepidemio.fr/RCTvsBurden/). For Sub-Saharan Africa, we did not observe any disease-specific regional gap, because RCTs in Sub-Saharan Africa planned to enroll more patients as compared with other regions (Figure S5).

Almost all these regional research gaps identified were stable in sensitivity analyses when comparing the conduct of RCTs to the burden measured as the number of YLL or number of deaths (see interactive figure at https://clinicalepidemio.fr/RCTvsBurden/). The only regional research gap that was not confirmed over these analyses was the conduct of RCTs for neonatal disorders versus the number of deaths in Latin America and Caribbean. However, when comparing the conduct of RCTs to the number of YLD, none of these regional research gaps were stable. In fact, RCTs were lacking in all regions as compared to YLD caused by musculoskeletal disorders, mental and behavioral deficiencies, and nutritional disorders.

DISCUSSION

In our study, we performed a worldwide large-scale comparison between the conduct of RCTs and the burden of diseases. Most RCTs were conducted in high-income countries, and their share across groups of diseases was aligned with the burden in those countries. Diseases mostly affecting low-income regions were understudied as compared to their global burden. Among non–high-income regions, South Asia and Sub-Saharan Africa presented the least balanced research efforts as compared to their regional burden, and regional research efforts were lacking for some diseases with high burden.

RCTs are needed globally to study the efficacy of health interventions in the most diverse population and to find local solutions when the efficacy of interventions depends on local settings. Our multi-scale approach allowed for the detection of research gaps by using a unified method both for regional and disease-specific research gaps. Regional gaps show how some health conditions might be under-studied as compared to the burden they cause. For instance, in Eastern Europe and Central Asia clinical trials did not cover the principal causes of burden. This may be because trials in this region are mainly sponsored by industry. [6] Regional gaps may thus be considered by local funders or health authorities to drive research towards local needs. Studying disease-specific gaps is more relevant in a global health perspective. When gaps exist, the evidence does not come from the regions with highest burden, which may lower the transposition of findings where they are needed the most due to lack of generalisability.

Previous studies evaluated the alignment between the research effort and burden by

conducting manual mappings over samples of RCTs, allowing for a fine-grained analysis within research topics or regions.[5, 8, 10, 11, 16–18] Our findings are in line with all these previous studies, and add region- and disease–specific-level results that have not been studied yet. Other studies have also compared the burden of diseases to other markers of research efforts such as publications,[19] systematic reviews,[20, 21] or funding.[22, 23] For instance *Evans et al.* classified all published articles in MEDLINE according to diseases and the principal investigator's country of affiliation and showed that within countries, the production of biomedical literature was correlated with the disease burden.[19] In our study we mapped the research effort according to the country of recruitment in RCTs, which may differ from the principal investigator's country of affiliation, particularly for RCTs conducted in low-income countries,[11, 24] and we were able to identify diseases for which RCTs were lacking as compared to the regional disease burden.

Our study has several strengths. First, we bring a global overview of the mapping of research as compared to the disease burden, which allowed for comparing with the same point of view the local production of research and the local burden in all regions for all diseases. We revealed large gaps of research by pre-specifying a binary definition of gap that was robust to possible misclassification of RCTs. The visualization tool might allow decision makers to look the data according to their own needs by focusing on specific regions or diseases. To be able to conduct such a large-scale mapping, we needed to rely on automatic classification algorithms.[14] These methods are not perfect, but we accounted for the classification uncertainty. Because of our fully automatic classification methods and our shared codes, our work is fully reproducible and updates can be done easily. Finally, this study is in line with the

development of the global observatory of health research called by the WHO, by linking clinical trial registries to the GBD database.[7]

Our study has also several limitations. First, our mapping is based on data from WHO ICTRP. We cannot exclude that some RCTs are not registered, particularly for some health conditions with low registration compliance,[25, 26] or in countries without clinical trial registries. Nevertheless the WHO ICTRP gathers registries from all regions of the world, and particularly the Pan African Clinical Trial Registry. Second, RCTs are not the unique measure of effort in clinical research. Observational studies could be included to complete these analyses. The amount of funding or investment could also be another way of quantifying the effort in clinical research. However, observational studies do not have the same level of requirement for registration as RCTs,[27] and we lack a database to monitor funding flows.[4] Third, we measured health needs based on data from the GBD study, which includes estimates that have been largely criticized.[28] As well, whether the clinical research effort should be aligned with the disease burden in DALYs is debated.[29] Indeed, it may be appropriate to over-study a particular condition because of its complexity (e.g diabetes), or it may be acceptable to under-study some medical conditions when conducting RCTs is difficult (e.g., congenital anomalies). Prioritizing clinical research according to the burden may maximize its potential impact, but research may be also prioritized according to the gaps in the knowledge we have on health intervention's efficacy.[30] For instance, we may already know solutions for preventing some common infectious diseases such as diarrhea but they may not be implemented in Sub-Saharan Africa or South Asia because of local settings. However, since they remain major causes of burden in those regions, there is still a need for further research.

[31] In addition, in our study we did not require perfect alignment, but we looked for gaps in research as defined by a proportion of research effort less than half the proportion of disease burden. We also showed that the existing gaps remained when using other metrics of burden such as the number of deaths and YLL. We did not specifically analyze health conditions presenting major global health challenges such as obesity or tobacco addiction. In the 2010 GBD study, the burden of these conditions was distributed across causes of disability or death associated with these conditions. Similarly, RCTs concerning, e.g., diabetes in obese patients, would be included in the mapping of diabetes. However, specific research mappings for these conditions need to be conducted to be compared to their burden.[32] We also pooled all kinds of treatments evaluated in RCTs. The efficacy of non-pharmacological treatments (NPT) may particularly show variations across regions. and separate mappings for NPTs would be useful. However, classifying RCTs in the WHO ICTRP by type of intervention remains challenging. Last, we mapped RCTs initiated during 2006-2015 to disease burden in 2005, considering that health needs in 2005 might drive the next 10-year research agenda. Further work would be needed to study the evolution of the mapping with time.

In conclusion, we compared a 10-year clinical research effort through the conduct of RCTs to the burden of diseases for 7 regions and 27 groups of diseases. We showed that the global research agenda has been driven by the interests and needs of high-income countries and that major causes of burden affecting non–high-income regions are locally under-studied.

AUTHORS' CONTRIBUTIONS

All authors conceived and designed the study. IA acquired and analyzed the data. All authors interpreted data. The initial manuscript was drafted by IA. All authors contributed to subsequent revisions and approved the final manuscript.

ACKNOWLEDGEMENTS

We would like to thank Elise Diard for help with the website hosting the interactive visualization tool, and Laura Smales for language revision of manuscript.

COMPLIANCE WITH ETHICAL STANDARDS

Funding: This work did not receive any specific grant.

Conflict of interest: All authors declare no competing interests.

Ethical approval: Not applicable for this study. The study only used data concerning the design and settings of clinical trials retrieved from publicly accessible clinical trial registries, and national-level aggregated database of the burden of diseases from publicly accessible databases.

Informed consent: Not applicable

REFERENCES

1. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380:2095–128.

2. Hotez PJ, Peiperl L. Noncommunicable diseases: A globalization of disparity? PLoS Med. 2015;12:e1001859.

3. Hotez PJ. Blue marble health redux: Neglected tropical diseases and human development in the Group of 20 (G20) nations and Nigeria. PLoS Negl Trop Dis. 2015;9:e0003672.

4. Røttingen JA, Regmi S, Eide M, Young AJ, Viergever RF, Ardal C, et al. Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? Lancet. 2013;382:1286–307.

5. Ahmad N, Boutron I, Dechartres A, Durieux P, Ravaud P. Geographical representativeness of published and ongoing randomized controlled trials. the example of: Tobacco consumption and HIV infection. PLoS One. 2011;6:e16878.

6. Atal I, Trinquart L, Porcher R, Ravaud P. Differential globalization of industry- and non-industry–sponsored clinical trials. PLoS One. 2015;10:e0145122.

7. Terry RF, Salm JF, Nannei C, Dye C. Creating a global observatory for health R&D. Science. 2014;345:1302–4.

8. Viergever RF, Terry RF, Karam G. Use of data from registered clinical trials to identify gaps in health research and development. Bull World Health Organ. 2013;91:416–425C.

9. Pedrique B, Strub-Wourgaft N, Some C, Olliaro P, Trouiller P, Ford N, et al. The drug and vaccine landscape for neglected diseases (2000–11): a systematic assessment. Lancet Glob Health. 2013;1:e371–e379.

10. Emdin CA, Odutayo A, Hsiao AJ, Shakir M, Hopewell S, Rahimi K, et al. Association between randomised trial evidence and global burden of disease : cross sectional study (Epidemiological Study of Randomized Trials — ESORT). BMJ. 2015;350:h117.

11. Ndounga Diakou LA, Ntoumi F, Ravaud P, Boutron I. Published randomized trials performed in Sub-Saharian Africa focus on high-burden diseases but are frequently funded and led by high-income countries. J Clin Epidemiol. 2017;82:29–36.

12. World Health Organization. International Clinical Trials Registry Platform. http://www.who.int/ictrp/. Accessed 1 Jan 2016.

13. de Angelis et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. Ann Intern Med. 2004;141:477–8.

14. Atal I, Zeitoun J-D, Névéol A, Ravaud P, Porcher R, Trinquart L. Automatic classification of registered clinical trials towards the Global Burden of Diseases taxonomy of diseases and injuries. BMC Bioinformatics. 2016;17.

15. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol. 2005;34:1370–1376.

16. Isaakidis P, Swingler GH, Pienaar E, Volmink J, Ioannidis JPA. Relation between burden of disease and randomised evidence in sub-Saharan Africa: survey of research. BMJ. 2002;324:702.

17. Rochon PA, Mashari A, Cohen A, Misra A, Laxer D, Streiner DL, et al. Relation between randomized controlled trials published in leading general medical journals and the global burden of disease. CMAJ. 2004;170:1673–1677.

18. Perel P, Miranda JJ, Ortiz Z, Casas JP. Relation between the Global Burden of Disease and randomized clinical trials conducted in Latin America published in the five leading medical journals. PLoS One. 2008;3:e1696.

19. Evans JA, Shim J-M, Ioannidis JPA. Attention to local health burden and the global disparity of health research. PLoS One. 2014;9:e90147.

20. Karimkhani C, Boyers LN, Prescott L, Welch V, Delamere FM, Nasser M, et al. Global Burden of Skin Disease as reflected in Cochrane Database of Systematic Reviews. JAMA Dermatol. 2014;150:945–51.

21. Yoong SL, Hall A, Williams CM, Skelton E, Oldmeadow C, Wiggers J, et al. Alignment of systematic reviews published in the Cochrane Database of Systematic Reviews and the Database of Abstracts and Reviews of Effectiveness with global burden-of-disease data: a bibliographic analysis. J Epidemiol Community Health. 2015;69:708–714.

22. Hazo JB, Gervaix J, Gandré C, Brunn M, Leboyer M, Chevreul K. European Union investment and countries' involvement in mental health research between 2007 and 2013. Acta Psychiatr Scand. 2016;134:138–49.

23. Gillum LA, Gouveia C, Dorsey ER, Pletcher M, Mathers C, McCulloch CE, et al. NIH disease funding levels and burden of disease. PLoS ONE. 2011;6:e16837.

24. Chersich MF, Blaauw D, Dumbaugh M, Penn-Kekana L, Dhana A, Thwala S, et al. Local and foreign authorship of maternal health interventional research in low- and middle-income countries: systematic mapping of publications 2000–2012. Glob Health. 2016;12.

25. Babu AS, Veluswamy SK, Rao PT, Maiya AG. Clinical Trial Registration in physical therapy journals: a cross-sectional study. Phys Ther. 2014;94:83–90.

26. Milette K, Roseman M, Thombs BD. Transparency of outcome reporting and trial registration of randomized controled trials in top psychosomatic and behavioral health journals: a systematic review. J Psychosom Res. 2011;70:205–17.

27. Boccia S, Rothman KJ, Panic N, Flacco ME, Rosso A, Pastorino R, et al. Registration practices for observational studies on clinicaltrials.gov indicated low adherence. J Clin Epidemiol. 2015;70:176–82.

28. Supervie V, Costagliola D. Time for a revolution in tracking the HIV epidemic. The Lancet. 2016;3:e337-9.

29. Prasad V. How should research be funded? Difficulties with the argument for proportionality to causes of death or years of life lost. J Natl Compr Canc Netw. 2016;14:365–366.

30. Viergever RF. The mismatch between the health research and development (R&D) that is needed and the R&D that is undertaken: an overview of the problem, the causes, and solutions. Glob Health Action. 2013;6:22450.

31. Greenland K, Chipungu J, Curtis V, Schmidt WP, Siwale Z, Mudenda M, et al. Multiple behaviour change intervention for diarrhoea control in Lusaka, Zambia: a cluster randomised trial. Lancet Glob Health. 2016;4:e988-77.

32. Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet. 2014;384:766–81.

2.3 Influence du promoteur dans la globalisation des essais cliniques

2.3.1 Résumé

Dans ce travail, nous avons étudié l'influence du type de promoteur dans la globalisation des essais cliniques (Atal et al., 2015). Nous avons en particulier analysé la capacité des promoteurs industriels et non-industriels (e.g. universités, organismes gouvernementaux et non-gouvernementaux) à mettre en place des essais cliniques internationaux, c'est-à-dire ceux recrutant des patients simultanément dans plusieurs pays.

Pour cela, nous avons analysé tous les essais cliniques commencés entre 2006 et 2013 et enregistrés dans l'ICTRP de l'OMS. Nous avons cartographié par pays et par groupes de revenu de la Banque Mondiale les essais cliniques internationaux et uni-pays pour les promoteurs industriels et non-industriels, et leur évolution entre 2006 et 2012. À partir des essais cliniques internationaux, nous avons construit des réseaux de collaboration entre pays, dans lesquels chaque nœud est un pays, et une arête entre deux pays correspond au nombre d'essais cliniques internationaux dans lesquels les deux pays participent simultanément. Nous avons analysé les réseaux de collaboration entre pays pour la recherche industrielle et non-industrielle. Pour chaque réseau nous avons identifié des groupes de pays participant de façon surreprésentée aux mêmes essais internationaux.

Nous avons travaillé sur 119,679 essais cliniques conduits dans 177 pays. Dans les pays à revenu élevé, les promoteurs industriels ont conduit trois fois plus d'essais cliniques par million d'habitant que les promoteurs non-industriels (75.0 vs. 24.5), alors que dans les pays à revenu faible ils ont conduit dix fois moins (0.08 vs. 1.08). Presque un tiers (30.3%) des essais cliniques industriels ont été conduits dans plusieurs pays simultanément, alors que seulement 3.2% des essais cliniques non-industriels étaient internationaux. Les essais cliniques internationaux industriels

sont devenu plus inter-continentaux entre 2006 et 2012 (de 54.8% à 67.3%), alors que pour ceux à promoteur non-industriel la tendance était inversées (de 42.4% à 37.2%). Les réseaux de collaboration entre pays participant ensemble aux essais cliniques internationaux avaient des propriétés différentes pour les deux types de promoteurs. En particulier, dans le réseau de collaboration dérivé des essais à promoteur industriel nous avons identifié une division de l'Europe en deux groupes : d'une part les pays d'Europe Occidentale collaborant de façon surreprésentée entre eux, et d'autre part les pays d'Europe de l'Est. Pour le réseau dérivé des essais non-industriels, les pays scandinaves forment un groupe de collaboration séparé du reste de l'Europe.

Ce travail nous a permis d'élucider l'influence du type de promoteur dans la globalisation des essais cliniques, et les différentes capacités des promoteurs industriels et non-industriels à mettre en place des essais cliniques internationaux. En particulier, les promoteurs industriels ont une grande capacité à globaliser leur recherche, en s'appuyant sur des réseaux de pays bien définis.

PLOS ONE

RESEARCH ARTICLE

Differential Globalization of Industry- and Non-Industry–Sponsored Clinical Trials

Ignacio Atal[1,2,4], Ludovic Trinquart[1,2,3]*, Raphaël Porcher[1,2,4], Philippe Ravaud[1,2,3,4]

1 Centre d'Épidémiologie Clinique, Hôpital Hôtel-Dieu, Paris, France, **2** INSERM U1153, Paris, France, **3** Columbia University, Mailman School of Public Health, Epidemiology Department, New York, New York, United States of America, **4** Université Paris Descartes, Paris, France

* ludovic.trinquart@htd.aphp.fr

Abstract

Background

Mapping the international landscape of clinical trials may inform global health research governance, but no large-scale data are available. Industry or non-industry sponsorship may have a major influence in this mapping. We aimed to map the global landscape of industry- and non-industry–sponsored clinical trials and its evolution over time.

Methods

We analyzed clinical trials initiated between 2006 and 2013 and registered in the WHO International Clinical Trials Registry Platform (ICTRP). We mapped single-country and international trials by World Bank's income groups and by sponsorship (industry- vs. non- industry), including its evolution over time from 2006 to 2012. We identified clusters of countries that collaborated significantly more than expected in industry- and non-industry–sponsored international trials.

Results

119,679 clinical trials conducted in 177 countries were analysed. The median number of trials per million inhabitants in high-income countries was 100 times that in low-income countries (116.0 vs. 1.1). Industry sponsors were involved in three times more trials per million inhabitants than non-industry sponsors in high-income countries (75.0 vs. 24.5) and in ten times fewer trials in low- income countries (0.08 vs. 1.08). Among industry- and non-industry–sponsored trials, 30.3% and 3.2% were international, respectively. In the industry-sponsored network of collaboration, Eastern European and South American countries collaborated more than expected; in the non-industry–sponsored network, collaboration among Scandinavian countries was overrepresented. Industry-sponsored international trials became more inter-continental with time between 2006 and 2012 (from 54.8% to 67.3%) as compared with non-industry–sponsored trials (from 42.4% to 37.2%).

OPEN ACCESS

Citation: Atal I, Trinquart L, Porcher R, Ravaud P (2015) Differential Globalization of Industry- and Non-Industry–Sponsored Clinical Trials. PLoS ONE 10 (12): e0145122. doi:10.1371/journal.pone.0145122

Editor: Joel Lexchin, York University, CANADA

Received: August 11, 2015

Accepted: November 28, 2015

Published: December 14, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by an academic grant (DEQ20101221475) for the programme Equipe Espoirs de la Recherche, from the Fondation pour la Recherche Médicale, France. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Conclusions

Based on trials registered in the WHO ICTRP we documented a substantial gap between the globalization of industry- and non-industry–sponsored clinical research. Only 3% of academic trials but 30% of industry trials are international. The latter appeared to be conducted in preferentially selected countries.

Introduction

Clinical trials are fundamental in advancing knowledge and improving health care globally.[1,2] By evaluating health interventions, clinical trials bring evidence about pharmacological and non-pharmacological therapies. International collaboration in clinical trials offers numerous advantages for the generation and interpretation of evidence.[3–6] Apart from accelerating the accrual of patients, especially for uncommon diseases, an important advantage in operating trials across countries is to increase the applicability of research findings. International collaboration in health research may also play an important role in reducing waste in health research.[7] Moreover, international clinical trials may strengthen health care systems in emerging economies because externally sponsored trials may increase the research capacity of sites in developing countries.[2,3]

As recently stated in *Science*, "the issue of knowing what research is currently being undertaken—where, by whom, and which organizations are supporting it—is a black hole in the public health landscape".[8] The international landscape of health research should be mapped to inform global governance and policy development.[9] In the last two decades, the number of clinical trials has expanded worldwide, and developing countries are increasingly involved, with a migration of trials from North America and Europe to Asia and Latin America.[10–12] Unravelling the forces that shape the research agenda may help steer it toward the most relevant health issues, to address the disparity between the local health burden and the production of health knowledge through clinical trials.[13,14]

A specific area of concern is the extent to which the clinical research landscape is dominated by industry sponsors.[15] In particular, international collaboration in clinical trials is constrained by scientific, ethical, economical, operational, and regulatory considerations. Different sponsors may have different capacities to address these constraints, and industry- and non-industry–sponsored research may thus show different collaborative patterns. Recent work suggest that private biomedical R&D expenditures in the United States have been reallocated to Asia and Oceania in the last five years.[16] Indeed, the pharmaceutical industry may be increasingly using global networks.[12] To our best knowledge, no quantitative large-scale data on this issue are available.

In 2006, the World Health Organization (WHO) established the International Clinical Trials Registry Platform (ICTRP), which gathers 16 worldwide registries of clinical trials meeting criteria of content, accessibility, quality and validity.[17] Based on clinical trials registered included in the WHO ICTRP, we aimed to map the global landscape of industry- and non-industry–sponsored clinical trials and its evolution over time.

Methods

To analyse the global landscape of clinical trials, we used data for all registered clinical trials that were included in the WHO ICTRP. We first mapped the trials and then studied the system of country–country collaboration for industry- and non-industry–sponsored clinical trials.

Collaboration between two countries was defined as the number of international clinical trials conducted simultaneously in at least these two countries. We analysed the system of country-country collaboration by deriving networks of collaboration for industry- and non-industry–sponsored clinical trials. In these networks, each node represents a country and an edge between two countries represents the number of international trials conducted simultaneously in at least these two countries. To describe the patterns of collaborations, we analysed these networks with a complex systems approach as detailed below.

Data

We retrieved records of clinical trials registered before February 2, 2014 in the ICTRP. After eliminating duplicates, we extracted the start date, the primary sponsor and the country locations for each trial (for details, see S1 Appendix). Because since September 2005, the International Committee of Medical Journal Editors has required registration before considering a trial for publication, we restricted the analysis to clinical trials with start dates between 2006 and 2013.[18]

Trials were classified by sponsor type (industry or non-industry) based on the primary sponsor, defined in the WHO ICTRP as the "organization which takes responsibility for the initiation, management, and/or financing of a clinical trial". The sponsor type was available for trials registered in ClinicalTrials.gov (90.7% of all included trials). For each of the remaining trials (9.3%), we determined whether the primary sponsor name matched that of the trial registered in ClinicalTrials.gov. If no match was found, we used a pre-specified list of keywords such as "Ltd.", "Inc.", and "University" to categorize the primary sponsor (for a detailed list, see S1 Appendix). We excluded 2.5% of all trials for which the sponsor type remained unclear.

The geographic classification of countries was based on the GeoNames and EuroVoc databases as well as the Organization for Economic Co-operation and Development (OECD) classification.[19,20] The country populations and income classifications were obtained from the World Bank database 2012.

Mapping of clinical trials

We first mapped the global distribution of clinical trials. Second, we mapped industry- and non- industry–sponsored clinical trials and the proportion of industry-sponsored trials. Finally, we mapped single-country and international clinical trials and the proportion of international trials for each sponsor type.

These mappings were performed both at the country-level and for groups of countries. At the country level, we mapped the density of clinical trials as the number of trials per million inhabitants. The density was considered only in countries with more than 250,000 inhabitants. At the country level, the share of sponsorship and of international trials was considered only in countries with at least more than 50 trials initiated (in total, industry- or non-industry–sponsored depending on the analysis) during the 2006–2013 period.

Collaboration network analysis

We analysed the industry- and non-industry–sponsored networks of collaboration using a null model analysis and a cluster analysis. In a network of collaboration, each node represents a country and an edge between two countries represents the country-country collaboration, corresponding to the number of international trials conducted simultaneously in at least these two countries.

To asses if some country-country collaborations were overrepresented as compared to what would be expected because of chance, we conducted a so-called null-model analysis, as

developed in the field of ecology.[21,22] For a given pair of countries, this method compares the observed number of collaborative trials between two given countries to the distribution of this number under the null hypothesis that all countries collaborate with each other purely at random. We derived the null distributions by generating 90,000 networks of collaboration through a permutation-based algorithm which preserved the numbers of trials initiated in each country and the numbers of countries involved in each trial (i.e., the margins of the collaboration matrix).[23] To test for overrepresentation, we compared each observed country-country collaboration to the 99.9th percentile of the corresponding null distribution. For each null distribution of country-country collaboration, 90,000 random networks of collaboration allowed for identifying the value below which 99.9% ± 0.01% of observations fall. To avoid sparse collaboration matrices, we had suppressed the countries participating in the lowest numbers of international trials. We removed countries successively until 95% and 90% of the total country-country collaborations remained for industry- and non-industry–sponsored networks, respectively.

For overrepresented country-country collaborations, we measured the extent to which both countries were involved more than expected by chance in the same international trials. This degree of overrepresentation was estimated as the ratio of the distance between the observed country-country collaboration and the mean of the null distribution to the distance between the 99.9th percentile and the mean of the null distribution. Then, we constructed co-occurrence networks where each node represents a country and an edge connects two countries if their collaboration is overrepresented, in which case its width corresponds to the degree of overrepresentation as previously.

To identify groups of countries where collaboration was higher than expected, we conducted a cluster analysis on the co-occurrence networks. Cluster analysis of networks is a data-driven approach allowing a network to be partitioned into groups to provide a simpler understanding of the network structure. The clustering algorithm we used partitioned the countries into clusters whereby the flow of collaboration is maximized within a cluster and minimized between clusters.[24] Countries in the same cluster were more likely to be involved together or with the same countries in clinical trials, and countries in different clusters had fewer chances of being involved together or with common countries in clinical trials.

Evolution over time

We studied the evolution of the mappings over time. Because of retrospective registration of trials, which may be more prevalent for trials that started in 2013, we restricted the time evolution analysis to the 2006–2012 period.[25] We computed the mappings for each year of the period and checked if trends existed.

All analyses involved use of R 3.0.2,[26] except for cluster analysis, which involved InfoMap code 0.13.5,[27] and co-occurrence network visualization, which involved NodeXL 1.0.1.251.[28]

Ethics statement

An ethics statement was not required for this work.

Results

We analysed 119, 679 clinical trials initiated during the 2006–2013 period (S1 Dataset). These trials were conducted in 177 countries, accounting for 99.3% of the worldwide population. In all, 30.1% of trials were industry-sponsored and 69.9% non-industry–sponsored (S1 Fig).

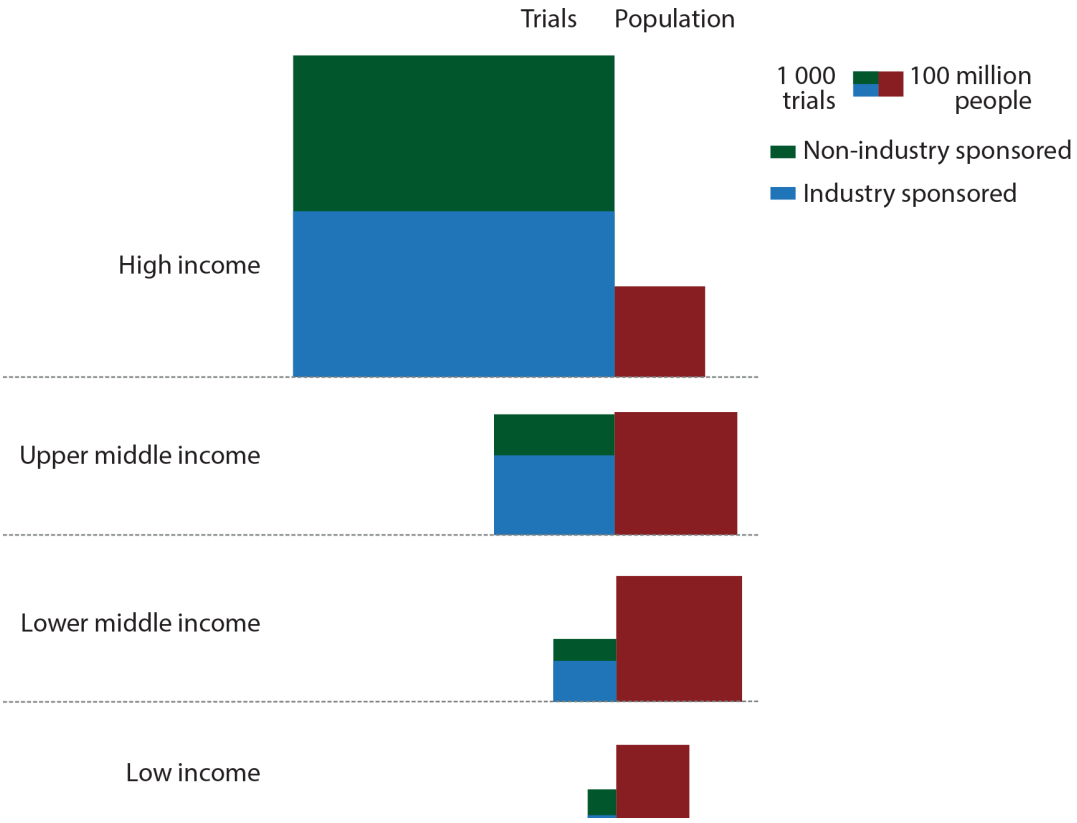


Fig 1. Distribution of clinical trials and population per income groups. For each income group, the size of the green (blue, respectively) area is proportional to the number of industry-sponsored (non-industry–sponsored) trials initiated during the 2006–2013 period, and the size of the red area is proportional to the population as of 2012. Equal-sized trial and population squares correspond to an overall density of 10 trials per million inhabitants. The proportion of industry-sponsored clinical trials was 51.6%, 66.0%, 65.4% and 9.3% in high-, upper-middle-, lower-middle- and low-income countries, respectively. In high-income countries, the density of trials ranged from 2.2 trials per million inhabitants in Trinidad and Tobago to 645.7 for Denmark. In upper-middle-income countries, it ranged from 0.05 to 225.9, with more than 50 trials per million inhabitants in four countries, all in Eastern Europe (Hungary, Bulgaria, Romania and Serbia). The variation was less pronounced in lower-middle-income countries (between 0.04 and 22.6) and low- income countries (between 0.13 and 14.0).

doi:10.1371/journal.pone.0145122.g001

Global mapping of clinical trials

Overall, the number of trials conducted in each country was extremely variable between income groups (Fig 1). The median number of trials per million inhabitants was 116.0 in high-income countries, 13.8 in upper-middle-income countries, but 1.8 and 1.1 in lower-middle- and low-income countries, respectively. In all, 1.65 billion people (23.4% of the world population) lived in countries where less than two trials per million inhabitants were initiated, most in low- or lower-middle- income countries (92.2%). The regions with the highest median density of clinical trials were Western Europe and Eastern Europe with 166.59 and 76.24 trials per million inhabitants, respectively (S2 Fig and S1 Table).

Sponsorship mapping

The proportion of industry-sponsored trials showed substantial variations across geographical regions and income groups. In particular, the proportion of industry-sponsored trials was 91.5% in Eastern Europe, 58.9% in Western Europe and 29.2% in Africa (Fig 2A). Similarly, the proportion of industry-sponsored trials was 67.0% and 76.4% in high- and upper-middle-

Fig 2. Sponsorship ratios of clinical trials. The radial barplot shows the proportion of industry- sponsored clinical trials (in dark blue) in the 87 countries where at least 50 trials were initiated during the 2006–2013 period. Countries were grouped by (a) geographical region and (b) income groups. For each group of countries, the red line represents the mean proportion of industry-sponsored clinical trials. The exact sponsorship ratios per country can be found on S4 Table. In high-income countries the proportion of industry-sponsored trials ranged from 33.6% in the United States to 90% or more in seven Eastern European countries. In upper-middle- income countries, the proportion ranged from 2.1% for Iran to more than 97% for countries such as Bulgaria and Romania. In lower-middle-income countries, the proportion ranged from less than 20% in three African countries to 97.2% in Ukraine.

doi:10.1371/journal.pone.0145122.g002

income countries, as compared to 11.7% in low-income countries (Fig 2B). In all income groups except the low-income group, the proportion of industry-sponsored trials varied, with the highest proportion consistently in Eastern European countries. In low-income countries, the proportion of industry-sponsored trials was homogeneously low as compared to the other income groups.

In high- and upper-middle-income countries, the median number of industry-sponsored trials per million inhabitants was three times that of non-industry–sponsored trials (75.0 vs. 24.5 and 7.3 vs. 2.5, respectively). In lower-middle-income countries, the three-fold difference was reversed (0.23 vs. 0.90), and in low-income countries the median number of industry-sponsored trials per million inhabitants was ten times less that of non-industry–sponsored trials (0.08 vs. 1.08). In fact, in all low-income countries but one, less than one industry-sponsored clinical trial per million inhabitants was initiated between 2006 and 2013.

Collaboration mapping

Global collaboration mapping. Most trials were conducted in a single country (88.6%). Single-country trials were mainly conducted in high-income countries (88.6%), particularly in the United States (42.3%), Western Europe (30.6%), and Asia (16.6%). Among international trials, 43.5% were conducted in a single continent. International single-continental trials were mainly conducted in Europe (65.8%) and North America (25.1%). Moreover, more than 90%

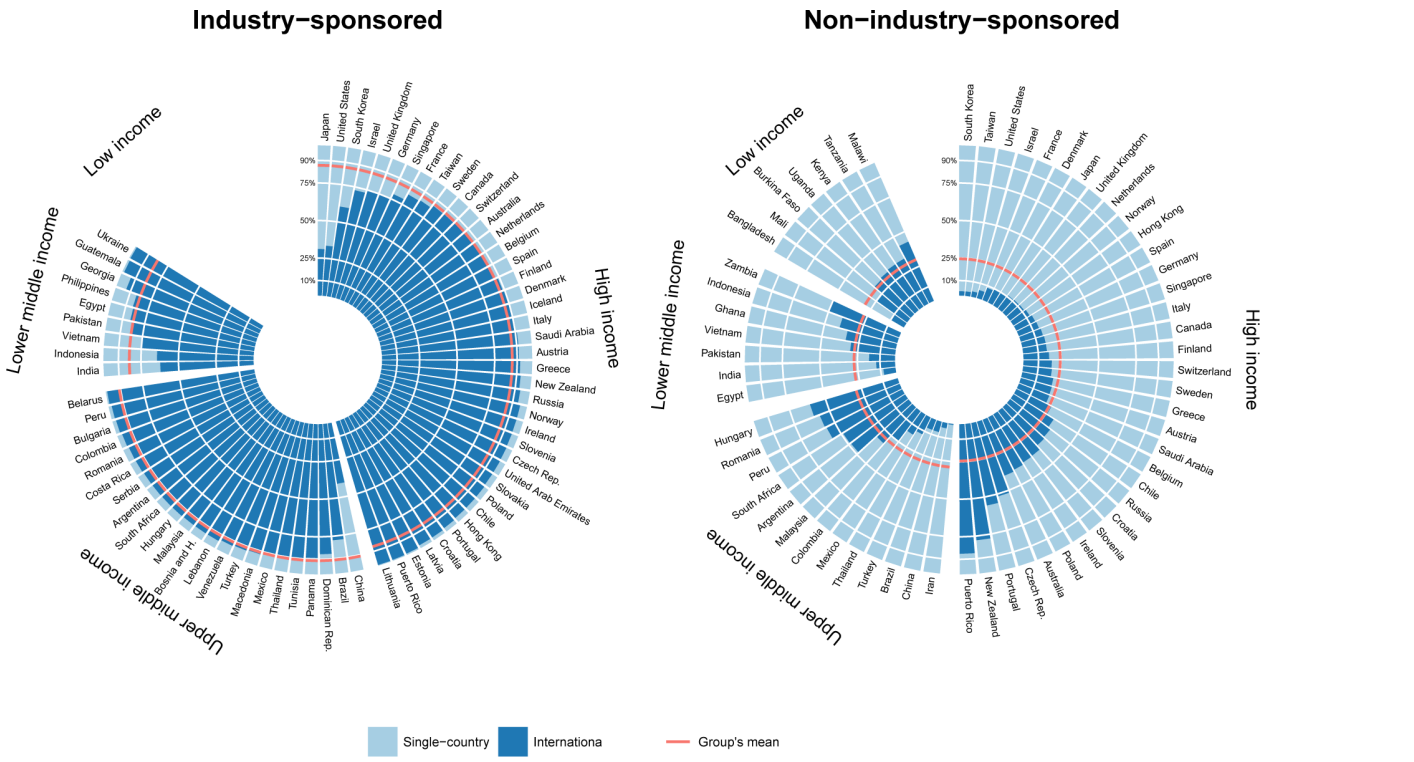


Fig 3. Collaboration ratios of industry- and non-industry–sponsored clinical trials. The radial barplot shows the proportion of international trials (in dark blue) per country for (left) industry- and (right) non-industry–sponsored trials in countries where at least 50 industry- or non-industry–sponsored trials were initiated during the 2006–2013 period. The 72 countries considered for industry- sponsored and the 62 countries considered for non-industry–sponsored trials were grouped by income groups. For each income group and sponsor type, the red line represents the mean proportion of international clinical trials. The exact collaboration ratios per country of industry- and non-industry-sponsored trials per country can be found on S5 and S6 Tables respectively. In all Eastern European and South American countries except Brazil, more than 90% of industry-sponsored research was international, whereas the proportion of international non-industry–sponsored research was lower and more variable, ranging from 25.0% in Colombia to 60.9% in Hungary.

doi:10.1371/journal.pone.0145122.g003

of all international trials were conducted in at least one North American or Western European country, and United-States–Canada collaborations represented 25.8% of international trials (3,523 trials).

Differences between industry- and non-industry–sponsored collaborations. Most of the single-country trials were non-industry sponsored (76.3%). The distribution of single-country trials by income groups and geographical regions was similar for both sponsor types (S3 and S5 Figs.). Most of the international trials were industry-sponsored (80.1%). The proportion of international trials was 30.3% for industry-sponsored and 3.2% for non-industry–sponsored trials. The proportion of international trials conducted in several continents was 60.6% for industry-sponsored and 40.1% for non-industry–sponsored trials. The median number of countries included in international trials was two for industry-sponsored and five for non-industry–sponsored trials.

For industry- and non-industry–sponsored trials, 46.2% and 13.0% of international trials were conducted in at least one Eastern European country, 18.3% and 7.9% in at least one South American country, and 2.4% and 8.9% in at least one African country other than South Africa, respectively (S4 and S5 Figs.).

Among industry-sponsored trials, the proportion of international trials in most high-income countries (37 of 40) was more than 70%, and in 21 of these 37 countries, it was greater than 90% (Fig 3). In contrast, among non-industry–sponsored trials, the proportion of

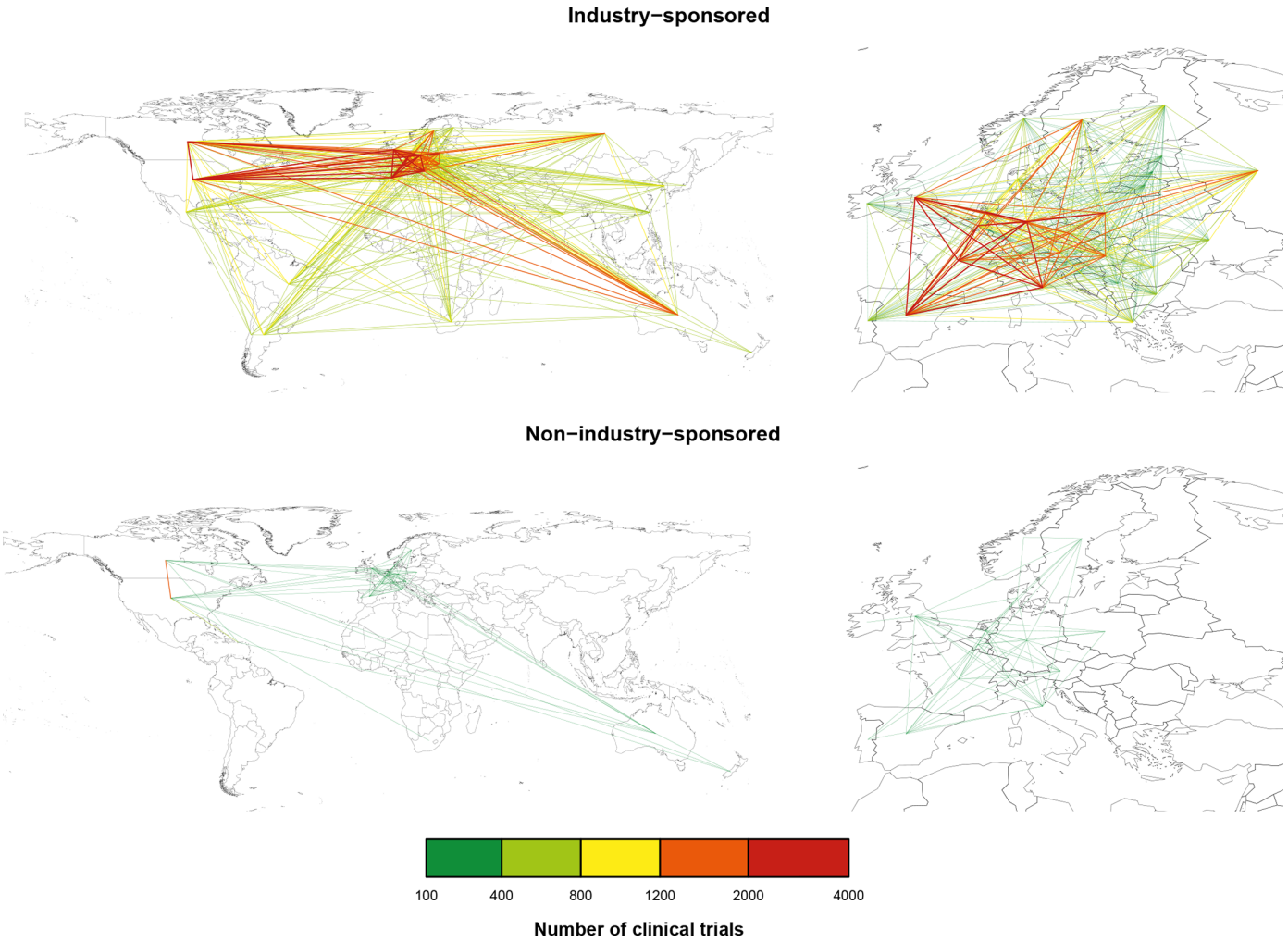


Fig 4. World and European collaboration networks in industry- and non-industry–sponsored clinical trials. Collaboration network of industry- (top) and non-industry–sponsored (bottom) clinical trials for registered trials initiated from 2006 to 2013; the color of a link between two countries corresponds to the number of clinical trials simultaneously conducted in both countries. For clarity, links between 100 and 400 clinical trials are not shown for the world's industry-sponsored network.

doi:10.1371/journal.pone.0145122.g004

international trials in half of the high-income countries was less than 20%. Similar discrepancies were observed for all other income groups.

In high- and upper-middle-income countries, the median number of industry-sponsored international trials per million inhabitants was 10 times that of non-industry–sponsored international trials (61.8 and 6.6 vs. 7.0 and 0.6, respectively). In low-income countries, the 10-fold difference was reversed (0.03 vs. 0.37, respectively).

Collaboration network analysis

The industry-sponsored network of collaboration included 138 countries and 4,711 country-country collaborations; 613 country-country collaborations accounted for more than 250 trials, with 2,870 trials for the United States–Canada collaboration (Fig 4). The non-industry–sponsored network of collaboration included 154 countries and 3,259 country-country collaborations. The United States–Canada collaboration was the unique collaboration, with more than

250 trials (653 trials). After trimming, the two networks comprised 60 countries (1,770 country-country collaborations) and 65 countries (1,736 country-country collaborations), respectively.

We found 440 (24.9%) and 316 (18.2%) overrepresented collaborations among industry- and non- industry–sponsored country-country collaborations, respectively. The 20 most overrepresented collaborations for industry- and non-industry–sponsored networks were between neighbor countries.

Cluster analysis of the co-occurrence networks identified 5 and 8 clusters for industry- and non-industry–sponsored trials, respectively (Fig 5). Most of the clusters corresponded to geographical regions. In the industry-sponsored network, the largest cluster corresponded to South American and Eastern European countries, which were apart from the Western European countries. In the non-industry–sponsored network, Scandinavian countries were clustered apart from the other European countries.

Evolution over time

Overall, the number of trials increased in all regions and income groups between 2006 and 2012. The distribution of trials over geographical regions and over income groups evolved differently when comparing sponsors (Fig 6). In particular, for industry-sponsored trials, the proportion of trials initiated in Western Europe was 42.3% in 2006 and 37.1% in 2012 and for non-industry–sponsored trials was 28.4% in 2006 and 35.3% in 2012. Conversely, the proportion of trials initiated in North America remained stable for industry-sponsored trials (23.1% on average) but was 53.1% in 2006 and 39.7% in 2012 for non-industry–sponsored trials.

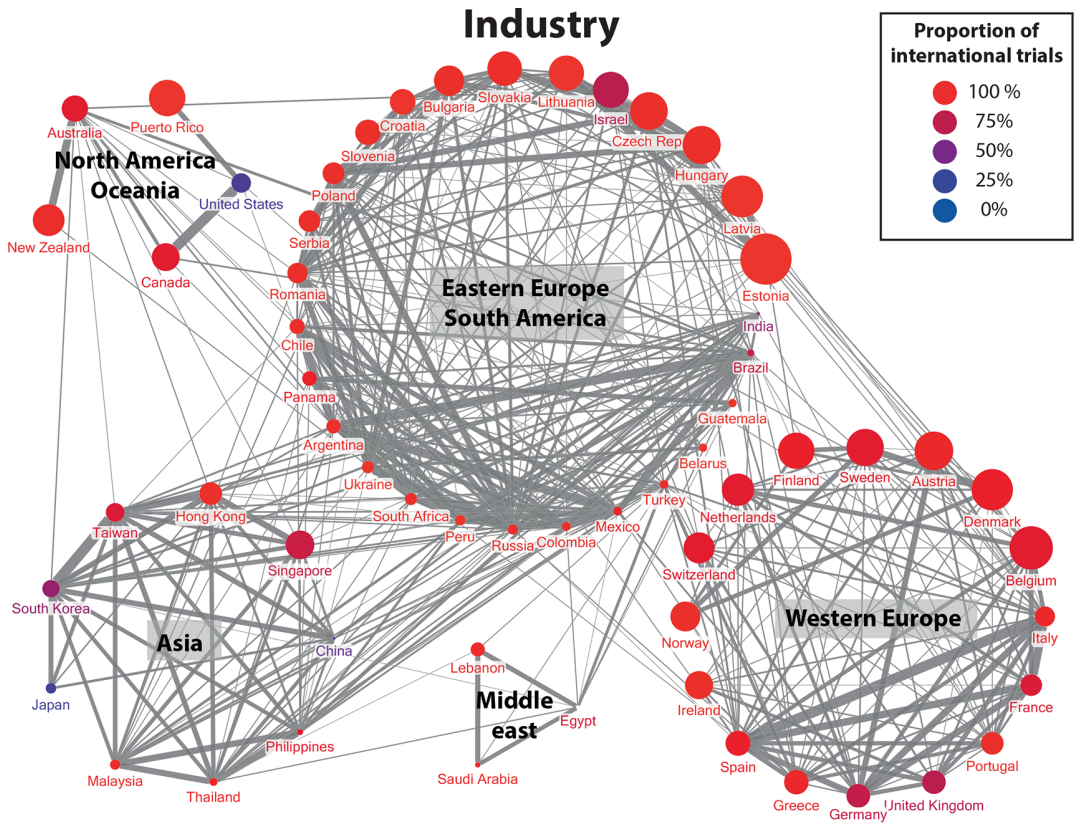
In total, the proportion of industry-sponsored trials was 32.9% in 2006 and 28.8% in 2012. This decrease was not equivalent in all geographical regions and income groups (S2 and S3 Tables). The proportion of industry-sponsored trials decreased by approximately 20% between 2006 and 2012 in Africa but remained stable in North America.

The number of international trials increased over the study period for both sponsor types. Meanwhile, the share of international trials among all trials decreased for both sponsor types (from 34.2% to 29.1% and from 3.6% to 2.9% for industry- and non-industry–sponsored trials, respectively). For industry-sponsored trials, the proportion of international trials conducted in several continents was 54.8% in 2006 and 67.3% in 2012 but for non-industry–sponsored trials was 42.4% in 2006 and 37.2% in 2012.

Discussion

In this bird's eye analysis of all registered clinical trials that were included in the WHO ICTRP, we found that clinical trials were unequally distributed in the world. Sponsorship has a major influence in this unequal mapping. International collaboration in clinical trials was mainly used by industry-sponsors, while non-industry–sponsored trials were mainly conducted in a single country.

Clinical trials were particularly prevalent in high-income countries and Eastern Europe and lacking in low-income countries. We documented substantial gaps in the global distribution of clinical trials between industry- and non-industry–sponsored research. Most of the clinical trials conducted in Eastern Europe were industry-sponsored but in Africa were non-industry–sponsored. International collaboration was sparse for academic sponsors, with 97% of academic-sponsored trials conducted in a single country. International collaboration was mainly used by industry sponsors in well-defined networks such as Eastern Europe and South America. In these regions, few single-country trials were conducted, so these countries may not conduct their own clinical trials. International trials were mainly conducted between neighboring


Industry
Proportion of international trials
100 %
75%
50%
25%
0%
North America
Oceania
Australia
Puerto Rico
United States
New Zealand
Canada
Eastern Europe
South America
Bulgaria
Slovakia
Lithuania
Israel
Croatia
Czech Rep
Slovenia
Hungary
Poland
Latvia
Serbia
Estonia
Romania
India
Chile
Brazil
Panama
Guatemala
Argentina
Belarus
Ukraine
Finland
Sweden
Austria
Turkey
Hong Kong
South Africa
Netherlands
Denmark
Taiwan
Peru
Russia
Colombia
Mexico
Switzerland
Belgium
Singapore
South Korea
Western Europe
Asia
China
Norway
Italy
Lebanon
Middle
east
Egypt
Ireland
France
Japan
Philippines
Spain
Portugal
Malaysia
Thailand
Saudi Arabia
Greece
Germany
United Kingdom

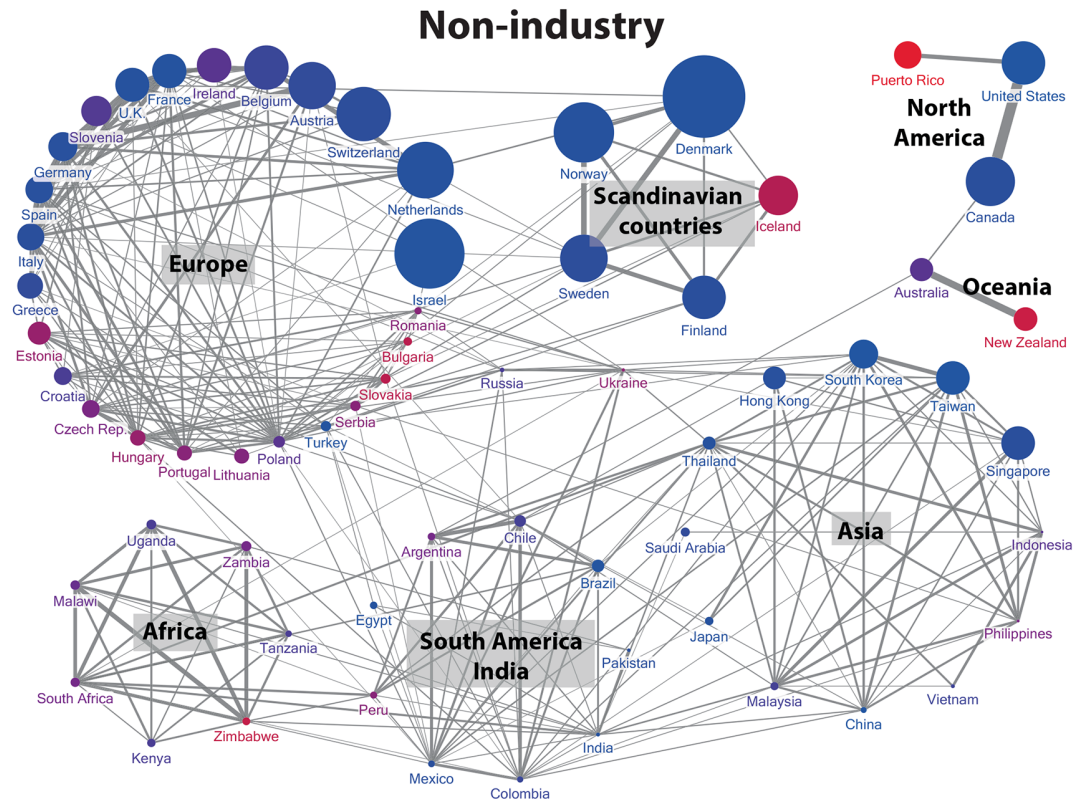

Non-industry
Puerto Rico
United States
North
America
France
Ireland
Belgium
U.K
Austria
Slovenia
Switzerland
Denmark
Germany
Norway
Canada
Netherlands
Scandinavian
countries
Spain
Iceland
Italy
Europe
Sweden
Australia
Oceania
Israel
Greece
Finland
New Zealand
Romania
Estonia
Bulgaria
Russia
Ukraine
South Korea
Croatia
Slovakia
Hong Kong
Taiwan
Czech Rep.
Serbia
Turkey
Thailand
Hungary
Poland
Portugal
Lithuania
Singapore
Uganda
Zambia
Argentina
Chile
Saudi Arabia
Asia
Indonesia
Brazil
Malawi
Egypt
Africa
Japan
Philippines
Tanzania
South America
India
Pakistan
South Africa
Peru
Malaysia
Vietnam
Zimbabwe
China
India
Kenya
Mexico
Colombia

Fig 5. Industry- and non-industry–sponsored co-occurrence networks. Country-country industry- (top) and non-industry–sponsored (bottom) networks for which links between countries are as wide as the estimated overrepresentation of the country-country collaboration. Size of nodes is proportional to the number of (top) industry- or (bottom) non-industry–sponsored clinical trials per million inhabitants. The color of the node represents the collaborative ratio of the country: the color corresponds to a gradient between blue, representing 100% of trials conducted in that country being single-country and red, 100% international trials. Among the 15 industry-sponsored most overrepresented collaborations, three were between Eastern European countries, four between South American countries, two between Asian countries, three between Western European countries (the France–Italy–Spain triangle), and the collaborations United States–Canada and Australia–New Zealand. The 15 most significantly overrepresented non-industry–sponsored collaborations were United States–Canada, United States–Puerto Rico, Australia–New Zealand, Malawi–Zimbabwe, South Korea–Taiwan, three collaborations between Northern European countries and six collaborations between other Western European countries. Among industry-sponsored collaborations, the trimming suppressed all African countries. Among non-industry–sponsored collaborations, African countries did not have overrepresented collaborations with European or North American countries.

doi:10.1371/journal.pone.0145122.g005

countries, but the groups of countries that collaborated differed for both sponsor types.[29] The locations of industry-sponsored trials remained stable since 2006, whereas non-industry–sponsored trial locations showed a shift from North America to Western European and Asian countries. More international trials were conducted over time, but the share of international trials among all clinical trials decreased for both sponsor types. In addition, industry-sponsored international trials became more inter-continental over time.

Clinical research is needed globally to validate treatment efficacy in the broadest population, to find local answers where universal questions may not be valid, and to improve health systems in emerging economies, but the location of clinical trials depends on sponsor's strategies and constraints.[5] Recently, Drain et al showed the unequal distribution and the global migration of clinical trials but did not study the impact of sponsorship.[12] Previous studies showed the unequal mapping of industry-sponsored clinical trials.[11,30] Our results are in line with these previous results and add to the substantial influence of sponsorship in the unequal global distribution of clinical trials and its evolution. The differential patterns of collaboration between the two sponsor types may underlie differentiated strategies and constraints in conducting international trials. In particular, academic sponsors may not have the operational and financial capacities to conduct trials worldwide, whereas industry sponsors may have more economical reasons to conduct international trials in specific regions such as Eastern Europe and South America. Initiatives have attempted to enhance academic collaboration networks. For instance, the European Clinical Research Infrastructures Network aims at promoting collaborative clinical research in Europe.[31] As well, the new European Union regulation on clinical research adopted last year modified the procedures for the authorization of clinical trials in order to stimulate international research.[32] Other initiatives attempt to favor trials between European and African countries.[33] Nevertheless, the number of international clinical trials conducted by non-industry sponsors still remains extremely low as compared to those conducted by the pharmaceutical industry. The future replication of these analyses would allow monitoring research agendas and assess the impact of initiatives or regulations aiming to stimulate international research.

The principal strengths of our study are the global overview of the system of clinical trials and the complex systems approach to analyse the system of international trials. However, the limitations are the self-reported nature of clinical trial registries data and the heterogeneity of what is considered a trial.[18,34] In particular, registries cannot verify the veracity of input trial information. There could be discrepancies between declared trial sites and sites that actually enrolled patients. However, the absence of verification concerns both industry-sponsored and non-industry-sponsored trials. In addition, our analysis is restricted to registered clinical trials. Not all registered clinical trials can be considered as means to increase clinical knowledge. Some clinical trials are conducted only for registration purposes, and several phase IV trials may be conducted for marketing purposes.[35] In addition, the vast majority of observational studies are not prospectively registered and so are consequently not covered by our analyses.

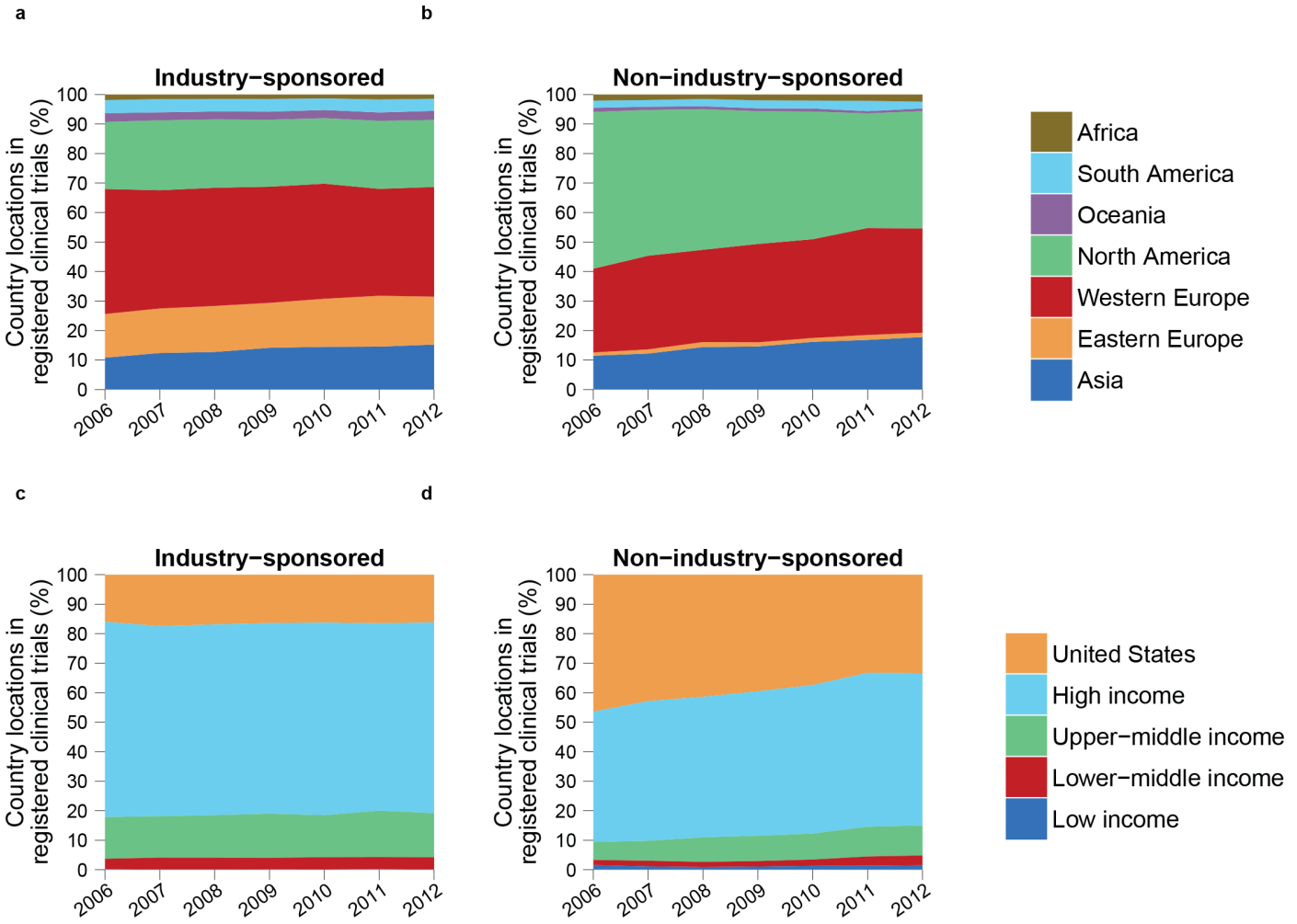


Fig 6. A differentiated global migration of clinical trials on sponsorship. The annual distribution of country trial locations of clinical trials initiated between 2006 and 2012. Countries were grouped by (top) geographical regions and (bottom) income groups, and trials were classified on sponsor type: (left) industry- and (right) non-industry–sponsored. Country-locations of international trials were considered individually: a trial conducted simultaneously in two South American countries would count twice when calculating the share of South America. For income groups, data for the United States and other high-income countries are shown separately. The proportion of trials initiated in Asia increased by a similar amount for both sponsor types (from 10.8% to 15.2% for industry and from 11.5% to 17.8% for non-industry). The proportion of trials initiated in Africa, South America, Oceania, and Eastern Europe remained stable for both sponsor types. The distribution of trials by income groups remained stable for industry-sponsored trials. For non-industry–sponsored trials, the proportion of trials initiated in high-income countries was 90.6% in 2006 and 85.0% in 2012, and the proportion of trials initiated in upper-middle-income countries was 6.1% in 2006 and 10.2% in 2012. The proportion of lower-middle- and low-income groups remained stable for non- industry–sponsored trials. The exact share of country trial locations of industry- and non-industry-sponsored trials per geographic region can be found on S7 and S8 Tables respectively. The exact share of country trial locations of industry- and non-industry-sponsored trials per income group can be found on S9 and S10 Tables respectively.

doi:10.1371/journal.pone.0145122.g006

[36–38] Moreover, the compliance with clinical trial registration may vary across countries and may be lower in low- and middle-income countries. Clinical trials sponsored by local sources (predominantly non-industry sponsors) and conducted in a single country in these regions may be less likely to be registered. In such a case, our findings regarding the share of single-country non-industry-sponsored clinical trials could be considered conservative. Other sources would allow us to recover data about unregistered trials, as publications registered in bibliographical databases, data from regional R&D hubs, or funders' databases. However, these additional resources are not readily usable nor accessible. In addition, unregistered trials are

unlikely to change the gap we found between the proportion of international clinical trials between industry- and non-industry-sponsored trials.

In this work, we chose to categorise sponsorship according to the primary sponsor, the primary organization that has the responsibility of the conduct of the clinical trial. We consider that the primary sponsor would be the most likely to enable or promote the conduct of international trials. One limitation is that we could not analyse the trial funding because such data are not reported in the WHO ICTRP. Sponsorship may not be a perfect proxy for funding research in that companies may influence steps of clinical research other than by sponsorship.[39] However, this situation would mainly concern non-industry-sponsored trials because industry-sponsored trials have high chance of being funded by the industry, but non-industry-sponsored trials may also be (partially) funded by industry.

Some of our choices of thresholds may have affected our results, such as the inclusion of countries for analyses. Nevertheless, these choices were unlikely to change our findings because of the magnitude of the discrepancies we found in analyses by sponsor type. Another limitation is the restriction to the 2006–2012 period for the time evolution analysis, but we considered that we did not have a reliable scope of global mapping outside that period. Another limitation that is not considered in our mapping is the country- or region-specific health needs. For instance, different areas of research may be more likely to motivate international collaboration, in particular in non-industry-sponsored settings.[40–43]. The next step will be to assess whether registered clinical trials correspond to health needs assessed locally.[44] Finally, we did not consider the number of patients included in each trial, which could result in more accurate measures of the amount of research in the population. The target sample size can be extracted from the trial registries but may not exactly correspond to the real sample size, and we have no information on the country-level sample size for international trials. Industry sponsors may have more capacity to conduct larger trials than academic researchers, which may increase the existing gaps between the mappings of both sponsor types.

The collaboration network analysis sheds light on the groups of countries that were more likely to be included together in international clinical trials. However, countries nearby in the collaboration network do not necessarily have scientific or logistic expertise to collaborate in international clinical trials. The weight and nature of the collaboration between countries participating in the same international trials may depend on the will of the primary sponsor to simply outsource the recruitment of patients or the entire conduct of the clinical trial. From the perspective of external validity, the physical location of trial sites does clearly mean that the trial is international. If a trial is performed in multiple geographical regions, one can assess whether the treatment effect is similar or heterogeneous across these settings.

This work is in-line with a series of works aiming to create a global observatory of health research.[8,9] The WHO ICTRP is the single source allowing a bird's-eye view of the mapping of clinical trials.[18,45] Acknowledging all the limitations of clinical trial data and the WHO ICTRP, the substantial gaps we show between the mappings of industry- and non-industry-sponsored trials and collaboration networks are unlikely to be changed. In conclusion, clinical trials are unequally distributed in the world. Substantial gaps exist between the mappings of industry- and non-industry–sponsored trials. International collaboration is lacking in academic-sponsored trials but is a predominant feature of industry-sponsored trials in well-defined networks of countries.

Supporting Information

S1 Appendix. Data extraction and Sponsor classification.
(DOCX)

S1 Dataset. Minimal dataset.
(CSV)

S1 Fig. Flowchart.
(PDF)

S2 Fig. Distribution of clinical trials and population per geographical regions. For each geographic region, the size of the green (blue, respectively) area is proportional to the number of industry- (non-industry–, respectively) sponsored trials initiated during the 2006–2013 period, and the size of the red area is proportional to the population as of 2012. Equal sized trial and population squares correspond to an overall density of 10 trials per million inhabitants. The proportion of industry-sponsored clinical trials was 57.0%, 37.5%, 51.0%, 92.5%, 65.4%, 77.2% and 45.8% in Western Europe, North America, Asia, Eastern Europe, South America, Oceania and Africa, respectively.
(PDF)

S3 Fig. Mapping of single-country for industry- and non-industry–sponsored clinical trials. The number of single-country clinical trials per million inhabitants for industry-sponsored (top) and non-industry–sponsored (bottom) research for registered trials initiated between 2006 and 2013.
(PDF)

S4 Fig. Mapping of international trials for industry- and non-industry–sponsored clinical trials. The number of international clinical trials per million inhabitants for industry-sponsored (top) and non-industry–sponsored (bottom) research for registered trials initiated between 2006 and 2013.
(PDF)

S5 Fig. Mapping of single-country and international clinical trials for industry- and non-industry–sponsored clinical trials in Europe. The number of single-country (top) and international (bottom) clinical trials per million inhabitants for industry-sponsored (left) and non-industry–sponsored (rigth) research for registered trials initi- ated between 2006 and 2013 in Europe.
(PDF)

S1 Table. Summary of the number of registered trials initiated in 2006–2013 per million inhabitants per geographical region.
(PDF)

S2 Table. Proportion of industry-sponsored trials per year for each geographical region.
(PDF)

S3 Table. Proportion of industry-sponsored trials per year for each income group.
(PDF)

S4 Table. Proportion of industry-sponsored trials for each country.
(PDF)

S5 Table. Proportion of international clinical trials among industry-sponsored trials for each country.
(PDF)

S6 Table. Proportion of international clinical trials among non-industry-sponsored trials for each country.
(PDF)

S7 Table. Distribution of country trial location of industry-sponsored trial over geographical regions per year.
(PDF)

S8 Table. Distribution of country trial location of non-industry-sponsored trial over geographical regions per year.
(PDF)

S9 Table. Distribution of country trial location of industry-sponsored trial over income groups per year.
(PDF)

S10 Table. Distribution of country trial location of non-industry-sponsored trial over income groups per year.
(PDF)

Acknowledgments

We would like to thank Elise Diard for help with the graphics and Laura Smales for language revision of manuscript.

Author Contributions

Conceived and designed the experiments: IA LT RP PR. Performed the experiments: IA LT RP. Analyzed the data: IA LT RP PR. Contributed reagents/materials/analysis tools: IA LT RP. Wrote the paper: IA LT RP PR.

References

1. Pang T, Terry RF, The PLoS Medicine Editors. Who/PLoS Collection "No Health Without Research": A Call for Papers. PLoS Med. 2011; 8: 1–2. doi: 10.1371/journal.pmed.1001008
2. Dye C, Boerma T, Evans D, Harries A, Lienhardt C, McManus J, et al. Research for universal health coverage. Sci Transl Med. 2013; 5: 1–146. doi: 10.1126/scitranslmed.3006971
3. Søreide K, Alderson D, Bergenfelz A, Beynon J, Connor S, Deckelbaum DL, et al. Strategies to improve clinical research in surgery through international collaboration. Lancet. 2013; 382: 1140–1151. doi: 10.1016/S0140-6736(13)61455-5 PMID: 24075054
4. OECD. Facilitating International Cooperation in Non-Commercial Clinical Trials. OECD Glob Sci Forum. 2011;
5. Lang T, Siribaddana S. Clinical trials have gone global: Is this a good thing? PLoS Med. 2012; 9: 6. doi: 10.1371/journal.pmed.1001228
6. Trimble EL, Abrams JS, Meyer RM, Calvo F, Cazap E, Deye J, et al. Improving Cancer Outcomes Through International Collaboration in Academic Cancer Treatment Trials. J Clin Oncol. 2009; 27: 5109–5114. doi: 10.1200/JCO.2009.22.5771 PMID: 19720905
7. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. Lancet. 2014; 383: 156–165. doi: 10.1016/S0140-6736(13)62229-1 PMID: 24411644
8. Terry RF, Salm JF, Nannei C, Dye C. Creating a global observatory for health R&D. Science. 2014; 345: 1302–1304. doi: 10.1126/science.1258737 PMID: 25214621
9. Røttingen JA, Regmi S, Eide M, Young AJ, Viergever RF, Ardal C, et al. Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? Lancet. 2013; 382: 1286–1307. doi: 10.1016/S0140-6736(13)61046-6 PMID: 23697824
10. Glickman SW., McHutchison JG., Peterson ED. et al. Ethical and Scientific Implications of the Globalization of Clinical Research. N Engl J Med. 2009; 360: 816–23. doi: 10.1056/NEJMsb0803929 PMID: 19228627
11. Thiers F a., Sinskey AJ, Berndt ER. Trends in the globalization of clinical trials. Nat Rev Drug Discov. 2008; 7: 13–14. doi: 10.1038/nrd2441

12. Drain PK, Robine M, Holmes KK, Bassett I V, Clinical I, Registry T, et al. Trail watch: global migration of clinical trials. Nat Rev Drug Discov. 2014; 13: 166–7. doi: 10.1038/nrd4260 PMID: 24577390
13. Evans J a, Shim J-M, Ioannidis JP a. Attention to local health burden and the global disparity of health research. PLoS One. 2014; 9: e90147. doi: 10.1371/journal.pone.0090147 PMID: 24691431
14. Viergever RF, Terry RF, Karam G. Use of data from registered clinical trials to identify gaps in health research and development. Bull World Heal Organ. 2013; 416–425C.
15. Patsopoulos N a, Ioannidis JP a, Analatos A a. Origin and funding of the most frequently cited papers in medicine: database analysis. BMJ. 2006; 332: 1061–1064. doi: 10.1136/bmj.38768.420139.80 PMID: 16547014
16. Justin Chakma, B. Sc., Gordon H. Sun, M.D., Jeffrey D. Steinberg, Ph.D., Stephen M. Sammut, M.A. MBA, and Reshma Jagsi, M.D. DP. Asia's Ascent—Global Trends in Biomedical R&D Expenditures. N Engl J Med. 2013; 3–6. doi: 10.1056/NEJMp1313927
17. WHO. International Clinical Trials Registry Platform [Internet]. Available: http://apps.who.int/trialsearch/
18. Dickersin K, Rennie D. The evolution of trial registries and their use to assess the clinical trial enterprise. JAMA. 2012; 307: 1861–4. doi: 10.1001/jama.2012.4230 PMID: 22550202
19. Wick M. GeoNames. Symposium on Space-Time Integration in Geography and GIScience. 2011.
20. Lazzeri V. Eurovoc. Terminol. 1983; 31–35.
21. Connor EF, Simberloff D. The assembly of species communities—chance or competition. Ecology. 1979; 60: 1132–1140. doi: 10.2307/1936961
22. Gotelli NJ. Null model analysis of species co-occurrence patterns. Ecology. 2000; 81: 2606–2621. doi: 10.1890/0012-9658(2000)081[2606:NMAOSC]2.0.CO;2
23. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. Package "vegan." R Packag ver 20–8. 2013; 254.
24. Rosvall M, Bergstrom C. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci. 2007; 105: 1118–23. doi: 10.1073/pnas.0706851105 PMID: 18216267
25. Viergever RF, Karam G, Reis A, Ghersi D. The Quality of Registration of Clinical Trials: Still a Problem. 2014; 9: 12–20. doi: 10.1371/journal.pone.0084727
26. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing Vienna Austria. 2013. pp. {ISBN} 3–900051–07–0. Available: http://www.r-project.org/
27. Edler D, Rosvall M. The MapEquation software package, available online at http://www.mapequation.org. 2013.
28. Hansen DL, Shneiderman B, Smith M a. Analyzing Social Media Networks with NodeXL. Anal Soc Media Networks with NodeXL. 2011; 11–29. doi: 10.1016/B978-0-12-382229-1.00002–3
29. Pan RK, Kaski K, Fortunato S. World citation and collaboration networks: uncovering the role of geography in science. Sci Rep. 2012; 2: 1–7. doi: 10.1038/srep00902
30. Karlberg JPE. Correspondence—Globalization of sponsored clinical trials. Nat Rev Drug Discov. 2007; 2007: 1–3.
31. Demotes-Mainars J, Ohmann C. European Clinical Research Infrastructures Network: promoting harmonisation and quality in European clinical research. Lancet. 2005; 365: 107–108. PMID: 15639280
32. European Parliament and Council of the European Union. Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC. Off J Eur Union. 2014; L: 1–76. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:JOL_2014_158_R_0001&from=EN. Accessed 14 July 2014.
33. Matee MI, Manyando C, Ndumbe PM, Corrah T, Jaoko WG, Kitua AY, et al. European and Developing Countries Clinical Trials Partnership (EDCTP): the path towards a true partnership. BMC Public Health. 2009; 9: 249. doi: 10.1186/1471-2458-9-249 PMID: 19619283
34. Viergever RF, Li K. Trends in global clinical trial registration: an analysis of numbers of registered clinical trials in different parts of the world from 2004 to 2013. 2015; 1–16. doi: 10.1136/bmjopen-2015-008932
35. Sox HC, Rennie D. Seeding trials: just say "no". Annals of internal medicine. 2008. pp. 279–280. VL—149
36. Editors TPM. Observational Studies: Getting Clear about Transparency. PLoS Med. 2014; 11: e1001711. doi: 10.1371/journal.pmed.1001711 PMID: 25158064

37. Dal-Re R, Ioannidis JP, Bracken MB, Buffler PA, Chan AW, Franco EL, et al. Making prospective registration of observational research a reality. Sci Transl Med. 2014; 6: 224cm1. doi: 10.1126/scitranslmed.3007513 PMID: 24553383

38. Boccia S, Rothman KJ, Panic N, Flacco ME, Rosso A, Pastorino R, et al. Registration practices for observational studies on clinicaltrials.gov indicated low adherence. J Clin Epidemiol. Elsevier Ltd; 2015; doi: 10.1016/j.jclinepi.2015.09.009

39. Stamatakis E, Weiler R, Ioannidis JPA. Undue industry influences that distort healthcare research, strategy, expenditure and practice: a review. 2010; 1–7. doi: 10.1111/eci.12074

40. Sista ND, Karim QA, Hinson K, Donnell D, Eshleman SH, Vermund SH. Experience in international clinical research: The HIV Prevention Trials Network. Clin Investig (Lond). 2011; 1: 1609–1618. http://dx.doi.org/10.4155/cli.11.156

41. International Maternal Pediatric Adolescent AIDS Clinical Trials Network (IMPAACT) [Internet]. Available: impaactnetwork.org/index.htm

42. HIV Vaccine Trials Network (HVTN) [Internet]. Available: http://www.hvtn.org/en.html

43. AIDS Clinical Trials Group (ACTG) [Internet]. Available: https://actgnetwork.org/

44. Emdin C a, Odutayo A, Hsiao AJ, Shakir M, Hopewell S, Rahimi K, et al. Association between randomised trial evidence and global burden of disease: cross sectional study (Epidemiological Study of Randomized Trials—ESORT). BMJ. 2015; 350. doi: 10.1136/bmj.h117

45. Weber WEJ, Merino JG, Loder E. Trial registration 10 years on. BMJ. 2015; 3572: h3572. doi: 10.1136/bmj.h3572

2.3.3 Analyse des réseaux de collaboration

Nous détaillons ici l'analyse des réseaux de collaboration conduite dans l'article précédemment présenté. Comme annoncé dans la Section 1.2.3 (troisième axe d'analyse), nous avons utilisé les données de la recherche clinique pour déduire des interactions entre les éléments d'une même entité, dans ce cas les pays. À partir des essais cliniques internationaux nous avons construit des *réseaux de collaboration* entre pays, dans lesquels chaque nœud est un pays, et une arête entre deux pays correspond au *nombre de collaborations* entre ces pays, défini par le nombre d'essais cliniques internationaux dans lesquels les deux pays participent simultanément. Pour les promoteurs industriels et non-industriels, nous avons travaillé sur 10,931 et 2,709 essais cliniques internationaux dans lesquels ont participé 138 et 154 pays, respectivement.

À partir de ces réseaux nous avons cherché à identifier des groupes de pays participant ensemble dans des essais cliniques internationaux de façon surreprésentée. En théorie des graphes, les *analyses de partitionnement* ou *analyses de communautés* permettent d'identifier automatiquement et sans *a priori* des groupement de nœuds tels que les nœuds appartenant à un même groupe soient le plus connectés possible, et les nœuds de groupes différents soient le moins connectés possible. Nous avons ainsi conduit une analyse de partitionnement sur les réseaux de collaboration pour identifier des groupes de pays tels que la collaboration entre pays au sein d'un groupe soit maximisée, et la collaboration entre pays appartenant à différents groupes soit minimisée.

Il existe un grand nombre d'algorithmes pour conduire des analyses de partitionnement. Au moment de faire notre analyse, la dernière revue systématique publiée avait identifié un algorithme récent ayant des performances supérieures aux autres (Lancichinetti and Fortunato, 2009). Cet algorithme utilise des marches aléatoires sur le réseau (Rosvall and Bergstrom, 2007). Un *agent* marche de façon aléatoire sur les nœuds d'un réseau, pouvant avancer uniquement entre des nœuds connectés par une arête, et ayant plus de probabilités de prendre des arêtes avec un

poids plus élevé (Figure 2.3). En faisant marcher aléatoirement un grand nombre d'agents, il est possible d'identifier un partitionnement des nœuds tels que le flux des trajectoires est maximisé au sein d'un groupe, et minimisé entre les groupes.

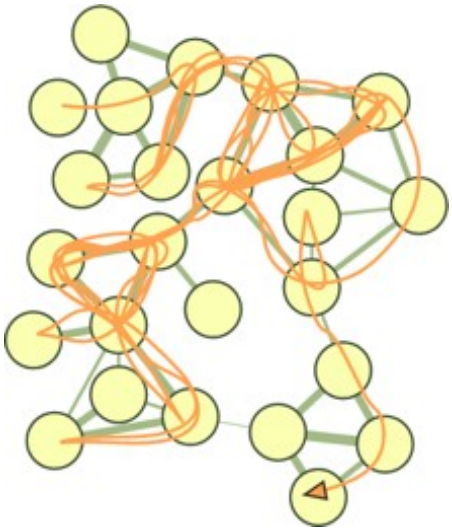


FIGURE 2.3 – Marche aléatoire sur un réseau
Source : (Rosvall and Bergstrom, 2007)

Un agent marche de façon aléatoire sur les nœuds d'un réseau. Il peut avancer uniquement entre nœuds connectés. Quand plusieurs possibilités de nœuds se présentent pour le prochain pas, l'agent a plus de probabilités de prendre le chemin ayant l'arête avec le plus de poids. Dans le cas des réseaux de collaboration, le poids des arêtes entre deux pays équivaut au nombre d'essais cliniques dans lesquels ces deux pays ont participé simultanément. Source : (Rosvall and Bergstrom, 2007)

Nous avons appliqué cet algorithme sur les réseaux de collaboration entre pays. Or, aucun partitionnement des pays n'a été identifié (Tables 2.2 et 2.3). En fait, ces réseaux étaient très *denses* : 49.8% et 28.0% des 9,453 et 11,781 hypothétiquement possibles collaborations entre paires de pays (arêtes) étaient non nulles, c'est-à-dire avaient participé simultanément à au moins un essai international. Ceci veut aussi dire que, en moyenne, tout pays était connecté par un essai international avec approximativement la moitié et un tiers des autres pays dans les réseaux à promoteur industriel et non-industriel, respectivement. Les marches aléatoires sur ces réseaux denses n'arrivaient ainsi pas à distinguer des partitionnement entre les pays.

Nous avons donc cherché à identifier des paires de pays ayant collaboré de façon surreprésentée comparées à ce qu'on aurait attendu seulement par le hasard. Ceci afin de créer des nouveaux réseaux dits de *co-occurrence* entre pays, dans lesquels sont uniquement mises en valeur les collaborations surreprésentées. Ainsi, nous avons été capables de diminuer la densité des réseaux en supprimant les collaborations non surreprésentées statistiquement. Nous avons finalement conduit une analyse de partitionnement sur ces réseaux de co-occurrence pour identifier des groupes de pays participant de façon surreprésentée ensemble à des essais cliniques internationaux.

Cohabitation surreprésentée chez les pinsons de Darwin

Pour identifier si deux pays ont participé ensemble à des essais cliniques internationaux de façon surreprésentée, nous nous sommes inspirés d'une méthode développée dans le domaine de l'écologie. Dans les années 1970, Connor et al. proposent une méthode pour analyser les motifs de partage de ressources chez les pinsons de Darwin, 13 espèces d'oiseaux habitant dans les 17 îles formant l'archipel de Galapagos en Équateur (Connor and Simberloff, 1979). La particularité des pinsons est que la taille et la forme de leurs becs sont très variables d'une espèce à l'autre, variabilité qui serait expliquée par la spécificité des ressources que chaque espèce consomme. Pour chaque espèce on connaît l'ensemble d'îles dans lesquelles elle habite, et pour chaque île on connaît l'ensemble d'espèces qu'elle reçoit. La méthode développée permet d'identifier quelles paires d'espèces s'évitent ou, au contraire, cohabitent, de façon significative. L'issu de ces recherches montre que les espèces qui s'évitent de façon significative sont celles ayant des becs similaires, alors que celles cohabitant de façon significative ont des types de becs différents. Leur répartition dans les îles serait ainsi reliée au partage des ressources alimentaires.

La méthode pour tester si deux espèces de pinsons s'évitent ou cohabitent de façon significative consiste à évaluer si le nombre observé d'îles dans lesquelles ces

deux espèces cohabitent significativement différent de la distribution de ce nombre sous l'*hypothèse nulle* que les espèces choisissent leurs îles au hasard, tout en préservant pour chaque espèce le nombre total d'îles dans laquelle elle est présente, et pour chaque île le nombre d'espèces qu'elle reçoit (Gotelli, 2000).

Conceptuellement, on commence par créer à partir des données observées une *matrice de présence-absence* des espèces dans les îles (Figure 2.4).

$$\begin{array}{c} \\ \text{espèce 1} \\ \text{espèce 2} \\ \vdots \\ \text{espèce n} \end{array} \begin{array}{c} \begin{array}{cccc} \text{île 1} & \text{île 2} & \cdots & \text{île m} \end{array} \\ \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 1 \end{pmatrix} \end{array}$$

FIGURE 2.4 – Matrice de présence-absence

Dans cette matrice, l'élément (i, j) vaut 1 si l'espèce i est présente dans l'île j, et 0 sinon. La somme par lignes et par colonnes de cette matrice (les *marges*) donne le nombre d'îles dans lesquelles habite chaque espèce, et le nombre d'espèce que reçoit chaque île, respectivement. L'hypothèse nulle correspond donc à tirer au hasard une matrice de présence-absence qui préserve les marges. Pour évaluer si deux espèces cohabitent de façon surreprésentée il est nécessaire de comparer le nombre observé d'îles où ces deux espèces cohabitent à la distribution de ce nombre sur les matrices ayant les mêmes marges que la matrice de présence-absence observée.

Énumérer toutes les matrices de présence-absence à marges fixes est computationellement très coûteux (Miller and Harrison, 2013). Il existe une méthode pour générer de façon non biaisée des matrices aléatoires à marges fixes basée sur un processus de Monte Carlo par chaînes de Markov (Miklós and Podani, 2004; Kannan et al., 1999). Elle consiste à modifier de manière itérative la matrice de présence-absence initiale par permutations préservant les marges comme suit (Figure 2.5) :

1. Choisir au hasard une sous-matrice de taille 2×2 ayant des 1 dans la diagonale et des 0 ailleurs, ou vice-versa

2. Inverser les 1 et les 0

$$\begin{array}{c} \\ i_1 \\ \vdots \\ i_2 \end{array} \begin{array}{c} \begin{array}{cccc} & j_1 & \cdots & j_2 \end{array} \\ \begin{pmatrix} & \cdots & & \\ 1 & \cdots & 0 & \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & \end{pmatrix} \end{array} \Rightarrow \begin{array}{c} \\ i_1 \\ \vdots \\ i_2 \end{array} \begin{array}{c} \begin{array}{cccc} & j_1 & \cdots & j_2 \end{array} \\ \begin{pmatrix} & \cdots & & \\ 0 & \cdots & 1 & \\ \vdots & \ddots & \vdots & \vdots \\ 1 & \cdots & 0 & \end{pmatrix} \end{array}$$

FIGURE 2.5 – Permutation d'une sous-matrice de taille 2×2 préservant les marges

Un grand nombre de répétitions de ces permutations donne lieu à une nouvelle matrice de présence-absence avec les mêmes marges que la matrice initiale, et tirée au hasard de façon uniforme. Finalement, un grand nombre de telles matrices permet de déduire des distributions de cohabitation sous l'hypothèse nulle pour chaque paire d'espèces de pinsons.

Collaborations surreprésentées entre pays et réseaux de co-occurrence

Nous avons utilisé cette méthode pour identifier des paires de pays participant de façon significative aux mêmes essais cliniques internationaux, mais aussi pour évaluer le *degré de surreprésentation* de leur collaboration. Ceci nous a permis de définir des *réseaux de co-occurrence*.

Pour chaque paire de pays, nous avons comparé le nombre de collaborations (i.e. le nombre observé d'essais cliniques dans lesquels ils participent ensemble) à la distribution de ce nombre sous l'*hypothèse nulle* que la localisation des essais est au hasard, tout en préservant pour chaque pays le nombre d'essais cliniques internationaux auxquels il participe, et pour chaque essai le nombre de pays y participant.

Les matrices de présence-absence étant de taille $1.5 \times 10e^6$ et $4 \times 10e^5$ pour les recherches industrielles et non-industrielle, et ayant seulement 5.2% et 2.2% de

cases non nulles respectivement, trouver aléatoirement des sous-matrices à permuter était computationnellement trop coûteux. Nous avons donc diminué la taille de ces matrices en supprimant les pays participant le moins à des essais internationaux jusqu'à garder au moins 95% et 90% des collaborations entre paires de pays pour les recherches industrielles et non-industrielles, respectivement. Pour les promoteurs industriels et non-industriels, nous avons ainsi travaillé sur un sous-ensemble de 10,832 et 2,569 essais internationaux conduits dans 60 et 65 pays, respectivement. Dans ces réseaux réduits, 100% et 83% des possibles collaborations entre paires de pays étaient non nulles, respectivement.

Nous avons généré 90,000 matrices de présence-absence pour la recherche industrielle et non-industrielle, à partir desquelles nous avons déduit pour chaque paire de pays des distributions sous l'hypothèse nulle du nombre de collaborations (Figure 2.6). Nous avons décidé que le nombre observé de collaborations entre deux pays était surreprésentée s'il était supérieur au 99.9^{e} percentile de la distribution sous l'hypothèse nulle. Nous avons ensuite défini le *degré de surreprésentation* d'une collaboration surreprésentée comme la distance entre le nombre de collaborations observée et la moyenne de la distribution de ce nombre sous l'hypothèse nulle, normalisé par la largeur moyenne–99.9^{e} percentile de la distribution sous l'hypothèse nulle (Figure 2.6).

Pour les réseaux de collaboration réduits issus de la recherche à promoteur industriel et non-industriel nous avons identifié 440 et 316 collaborations entre pays surreprésentées, correspondant à 25% et 15% de toutes les collaborations possibles entre paires de pays, respectivement.

À partir des analyses de surreprésentation nous avons ainsi construit des *réseaux de co-occurrence* entre pays pour la recherche à promoteur industriel et non-industriel, dans lesquels deux pays sont reliés pas une arête si leur nombre de collaborations est surreprésenté, et le poids de l'arête est égal au degré de surreprésentation de leur collaboration. Ainsi, dans ces réseaux de co-occurrence sont mises en valeur uniquement les collaborations entre pays surreprésentées.

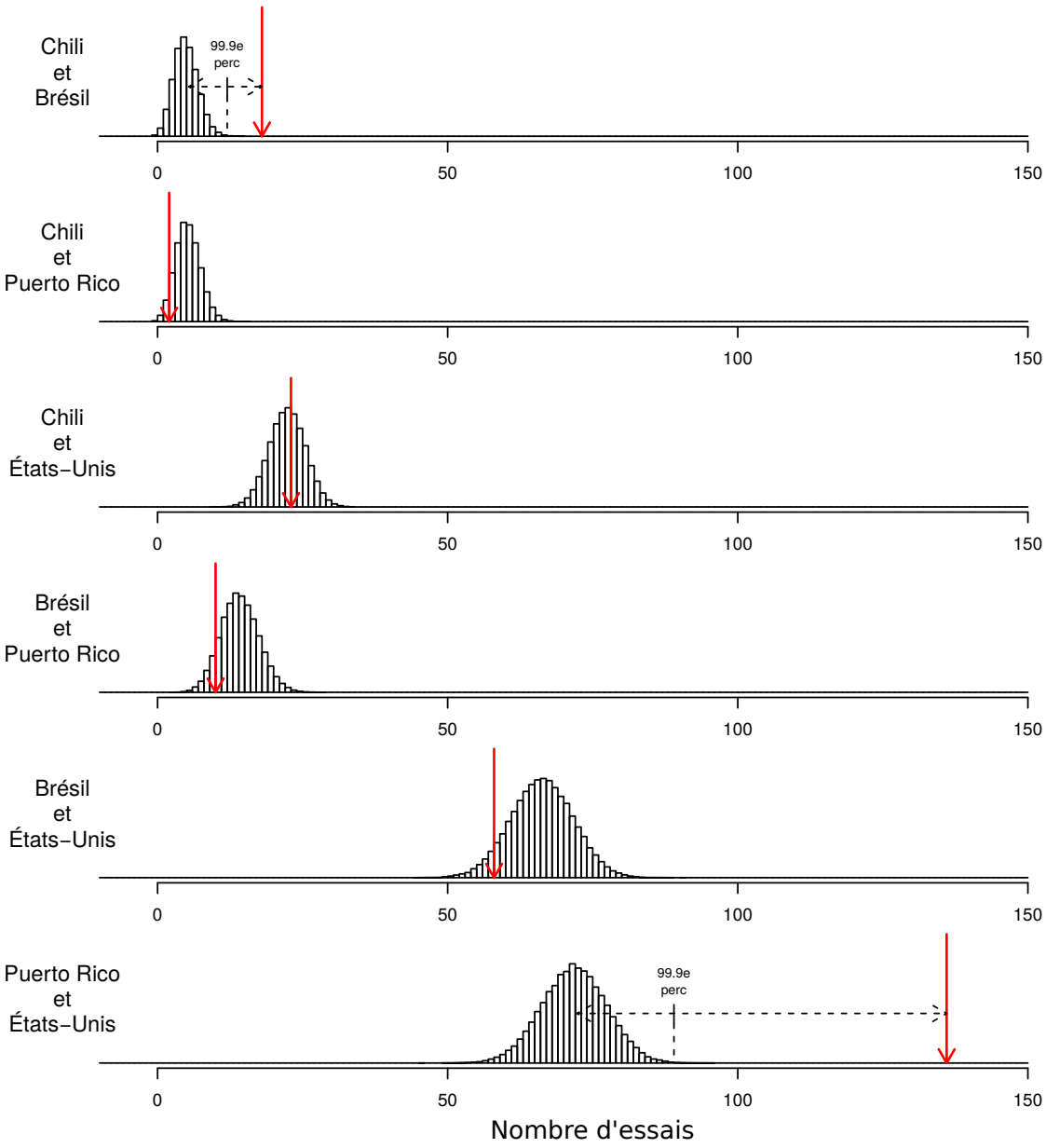


FIGURE 2.6 – Identification de collaborations surreprésentées dans les essais internationaux à promoteur non-industriel

Distributions sous l'hypothèse nulle et valeur observée (flèche rouge) du nombre d'essais à promoteur non-industriel conduits simultanément dans deux pays parmi le Chili, le Brésil, Puerto Rico et les États-Unis. Parmi les paires de pays présentés, les seuls nombres de collaboration surreprésentés sont Chili–Brésil et Puerto Rico–États-Unis, avec un degré de surreprésentation de 1.9 et 3.8, respectivement. Les autres nombres de collaborations ne sont pas significativement supérieurs à des collaborations sous l'hypothèse que la localisation des essais est faite seulement par le hasard.

Identification de groupes de collaboration surreprésentée

Comme spécifié au début de cette section, nous avons identifié des groupes de pays collaborant entre eux de façon surreprésentée en faisant une analyse de partitionnement sur les réseaux de co-occurrence en utilisant un algorithme de marches aléatoires (Rosvall and Bergstrom, 2007). Les résultats du partitionnement des réseaux de co-occurrence industriels et non-industriels sont présentés dans les Tables 2.2 et 2.3. Un des résultats les plus marquants de ces partitionnements était la différente division de l'Europe dans ces deux réseaux : pour le réseau industriel l'Europe été divisée en Europe de l'Est d'une part et Europe Occidentale d'autre part ; pour le réseaux non-industriel elle était divisée en pays scandinaves d'une part, et le reste de l'Europe d'autre part.

Depuis que nous avons conduit se travail, un autre algorithme de partitionnement a commencé à être très utilisé dans la communauté d'analyse de réseaux, et dont les performances se sont avérées être supérieures aux marches aléatoires sur un grand ensemble de réseaux et sur un grand nombre de métriques d'évaluation (Emmons et al., 2016). Cet algorithme, dit de *Louvain*, se base sur l'optimisation de la *modularité* du réseau (Blondel et al., 2008). La modularité est une métrique du réseau qui compare la densité d'arêtes entre nœuds d'un même groupe à la densité d'arêtes en nœuds de groupes différents.

Nous avons évalué si le partitionnement trouvé lors de notre travail restait stable en utilisant l'algorithme de Louvain. Par ailleurs, nous avons aussi évalué la capacité de Louvain à trouver un partitionnement sur les réseaux de collaboration directement sans avoir besoin de passer par la surreprésentation des liens et les réseaux de co-occurrence. Les résultats comparant les deux algorithmes de partitionnement, par marches aléatoires et Louvain, appliqués aux réseaux de collaboration complets et réduits (i.e. après enlever les pays participant le moins à des essais internationaux) d'une part, et aux réseaux de co-occurrence d'autre part, sont présentés dans les Tables 2.2 et 2.3.

Nous pouvons observer que l'algorithme par marches aléatoires a effectivement

besoin de passer par la création des réseaux de co-occurrence pour identifier un partitionnement des pays. Or, Louvain trouve des groupements de pays quand il est appliqué directement aux réseaux de collaboration. Les groupements trouvés par Louvain sur les réseaux de collaboration sont moins fins que ceux trouvés par les deux algorithmes sur les réseaux de co-occurrence, ce qui pourrait justifier le fait d'avoir conduit l'analyse des surreprésentations des collaborations.

On remarque que dans tous les réseaux à promoteur industriel, Louvain groupe séparément l'Europe de l'Est du reste de l'Europe, alors que dans les réseaux à promoteur non-industriel, Louvain laisse l'Europe unifiée dans les réseaux de collaboration, et la divise en trois groupes dans le réseau de co-occurrence : l'Europe occidentale, l'Europe de l'Est et les pays scandinaves. Toutefois, ceci confirme l'ampleur de l'existence d'un réseau de collaboration spécifique à l'Europe de l'Est dans la recherche industrielle. Par ailleurs, quand on applique Louvain au réseau de co-occurrence industriel, l'Amérique Latine n'est plus groupée avec l'Europe de l'Est, mais avec le Proche Orient, mettant en doute la robustesse du positionnement de l'Amérique Latine dans le groupement initialement trouvé. Mise à part les différences énoncées ci-dessus, pour les deux types de promoteurs les partitionnements des réseaux de co-occurrence par marches et aléatoires et Louvain donnent des résultats assez similaires.

	Marches aléatoires		Louvain	
	Partitionnement des pays	Régions	Partitionnement des pays	Régions
Réseau de collaboration complet	Benin, Burkina Faso, Cameroon, Congo DR, Congo, Côte d'Ivoire, Gabon, Gambia, Ghana, Kenya, Madagascar, Malawi, Mali, Mozambique, Nigeria, Rwanda, Senegal, Sudan, Tanzania, Togo, Uganda, Zambia	Afrique	Benin, Burkina Faso, Cameroon, Congo RD, Congo, Côte d'Ivoire, Gabon, Gambia, Ghana, Kenya, Madagascar, Malawi, Mali, Mozambique, Nigeria, Rwanda, Senegal, Sudan, Tanzania, Togo, Uganda, Zambia	Afrique
	Albania, Algeria, Andorra, Argentina, Armenia, Australia, Austria, Bahamas, Bahrain, Bangladesh, Belarus, Belgium, Belize, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Cambodia, Canada, Chile, China, Colombia, Costa Rica, Croatia, Cuba, Cyprus, Czech Republic, Denmark, Dominican Republic, Ecuador, Egypt, El Salvador, Estonia, Ethiopia, Finland, France, Georgia, Germany, Greece, Grenada, Guatemala, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, North Korea, South Korea, Kuwait, Kyrgyzstan, Latvia, Lebanon, Libya, Lithuania, Luxembourg, Macedonia, Malaysia, Malta, Mexico, Moldova, Monaco, Morocco, Netherlands, Netherlands Antilles, New Zealand, Nicaragua, Norway, Oman, Pakistan, Panama, Paraguay, Peru, Philippines, Poland, Portugal, Puerto Rico, Qatar, Romania, Russian Federation, Saint Kitts and Nevis, Saudi Arabia, Serbia, Singapore, Slovakia, Slovenia, South Africa, Spain, Sri Lanka, Sweden, Switzerland, Syria, Taiwan, Thailand, Tunisia, Turkey, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Venezuela, Vietnam, Yemen	Reste du monde	Armenia, Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Estonia, Georgia, Hungary, Iceland, Kazakhstan, Latvia, Lithuania, Macedonia, Malta, Moldova, Poland, Romania, Russian Federation, Serbia, Slovakia, Slovenia, South Africa, Ukraine, Uzbekistan	Europe de l'est
			Australia, Austria, Belgium, Belize, Botswana, Canada, Denmark, Finland, France, Germany, Greece, Ireland, Israel, Italy, North Korea, Luxembourg, Monaco, Netherlands, Netherlands Antilles, New Zealand, Norway, Paraguay, Portugal, Puerto Rico, Saint Kitts and Nevis, Spain, Sweden, Switzerland, United Kingdom, United States	Pays à revenu élevé
			Albania, Algeria, Andorra, Argentina, Bahamas, Bahrain, Bangladesh, Brazil, Brunei, Cambodia, Chile, China, Colombia, Costa Rica, Cuba, Cyprus, Dominican Republic, Ecuador, Egypt, El Salvador, Ethiopia, Grenada, Guatemala, Honduras, Hong Kong, India, Indonesia, Iran, Jamaica, Japan, Jordan, South Korea, Kuwait, Kyrgyzstan, Lebanon, Libya, Malaysia, Mexico, Morocco, Nicaragua, Oman, Pakistan, Panama, Peru, Philippines, Qatar, Saudi Arabia, Singapore, Sri Lanka, Syria, Taiwan, Thailand, Tunisia, Turkey, United Arab Emirates, Uruguay, Venezuela, Vietnam, Yemen	Asie + Am. Latine
Réseau de collaboration réduit	Argentina, Australia, Austria, Belarus, Belgium, Brazil, Bulgaria, Canada, Chile, China, Colombia, Croatia, Czech Republic, Denmark, Egypt, Estonia, Finland, France, Germany, Greece, Guatemala, Hong Kong, Hungary, India, Ireland, Israel, Italy, Japan, South Korea, Latvia, Lebanon, Lithuania, Malaysia, Mexico, Netherlands, New Zealand, Norway, Panama, Peru, Philippines, Poland, Portugal, Puerto Rico, Romania, Russian Federation, Saudi Arabia, Serbia, Singapore, Slovakia, Slovenia, South Africa, Spain, Sweden, Switzerland, Taiwan, Thailand, Turkey, Ukraine, United Kingdom, United States	Tous les pays	Belarus, Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Russian Federation, Serbia, Slovakia, Slovenia, Ukraine	Europe de l'est
			Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom	Europe occidentale
			Argentina, Australia, Brazil, Canada, Chile, China, Colombia, Egypt, Guatemala, Hong Kong, India, Israel, Japan, South Korea, Lebanon, Malaysia, Mexico, New Zealand, Panama, Peru, Philippines, Puerto Rico, Saudi Arabia, Singapore, South Africa, Taiwan, Thailand, Turkey, United States	Reste du monde
Réseau de co-occurrence	Argentina, Belarus, Brazil, Bulgaria, Chile, Colombia, Croatia, Czech Republic, Estonia, Guatemala, Hungary, Israel, Latvia, Lithuania, Mexico, Panama, Peru, Poland, Romania, Russian Federation, Serbia, Slovakia, Slovenia, South Africa, Turkey, Ukraine	Europe de l'est + Am Lat.	Belarus, Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Russian Federation, Serbia, Slovakia, Ukraine, Slovenia	Europe de l'est
	Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom	Europe occidentale	Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom	Europe occidentale
	China, Hong Kong, India, Japan, South Korea, Malaysia, Philippines, Singapore, Taiwan, Thailand	Asie	Argentina, Brazil, Chile, Colombia, Egypt, Guatemala, Israel, Lebanon, Mexico, Panama, Peru, Saudi Arabia, South Africa, Turkey	Am. Lat + Proche orient
	Australia, Canada, New Zealand, Puerto Rico, United States	US + Australie	China, Hong Kong, India, Japan, South Korea, Malaysia, Philippines, Singapore, Taiwan, Thailand	Asie
	Egypt, Lebanon, Saudi Arabia	Proche orient	Australia, Canada, New Zealand, Puerto Rico, United States	US + Australie

TABLE 2.2 – Partitions des réseaux de collaboration et de co-occurrence industriels avec deux algorithmes

Deux algorithmes de partitionnement ont été comparés : marches aléatoires (Rosvall and Bergstrom, 2007) et Louvain (Blondel et al., 2008). Chacun a été appliqué à trois réseaux : le réseau de collaboration total, le réseau de collaboration avec uniquement les pays participant le plus à des essais internationaux, et le réseau de co-occurrence.

	Marches aléatoires		Louvain	
	Partitionnement des pays	Régions	Partitionnement des pays	Régions
Réseau de collaboration complet	Albania, Algeria, American Samoa, Argentina, Armenia, Australia, Austria, Azerbaijan, Bangladesh, Belarus, Belgium, Benin, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Cayman Islands, Central African Republic, Chad, Chile, China, Colombia, Congo DR, Congo, Costa Rica, Côte d'Ivoire, Croatia, Cuba, Cyprus, Czech Republic, Denmark, Dominican Republic, Ecuador, Egypt, El Salvador, Estonia, Ethiopia, Finland, France, French Polynesia, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Guatemala, Guinea, Guinea Bissau, Haiti, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kenya, South Korea, Kuwait, Laos, Latvia, Lebanon, Liberia, Lithuania, Luxembourg, Macedonia, Malawi, Malaysia, Mali, Malta, Mauritania, Mauritius, Mexico, Moldova, Monaco, Mongolia, Morocco, Mozambique, Myanmar, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, Norway, Pakistan, Palestine, Panama, Paraguay, Peru, Philippines, Poland, Portugal, Puerto Rico, Qatar, Romania, Russian Federation, Rwanda, Saudi Arabia, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, South Africa, Spain, Sri Lanka, Sudan, Suriname, Swaziland, Sweden, Switzerland, Taiwan, Tanzania, Thailand, Tunisia, Turkey, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Venezuela, Vietnam, Zambia, Zimbabwe Solomon Islands, Vanuatu Afghanistan, Djibouti Barbados, Trinidad and Tobago	Tous les pays	Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, French Polynesia, Germany, Greece, Guinea Bissau, Hungary, Iceland, Iran, Iraq, Ireland, Israel, Italy, Latvia, Lebanon, Lithuania, Luxembourg, Macedonia, Malta, Moldova, Monaco, Netherlands, Norway, Palestine, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Suriname, Sweden, Switzerland, United Arab Emirates, United Kingdom American Samoa, Armenia, Australia, Azerbaijan, Barbados, Canada, Cayman Islands, Jordan, Mauritius, New Zealand, Puerto Rico, Qatar, Trinidad and Tobago, United States Afghanistan, Albania, Algeria, Argentina, Bangladesh, Belarus, Benin, Bolivia, Botswana, Brazil, Brunei, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Chile, China, Colombia, Congo DR, Congo, Costa Rica, Côte d'Ivoire, Cuba, Djibouti, Dominican Republic, Ecuador, Egypt, El Salvador, Ethiopia, Gabon, Gambia, Georgia, Ghana, Guatemala, Guinea, Haiti, Honduras, Hong Kong, India, Indonesia, Jamaica, Japan, Kenya, South Korea, Kuwait, Laos, Liberia, Malawi, Malaysia, Mali, Mauritania, Mexico, Mongolia, Morocco, Mozambique, Myanmar, Nepal, Nicaragua, Niger, Nigeria, Pakistan, Panama, Paraguay, Peru, Philippines, Russian Federation, Rwanda, Saudi Arabia, Senegal, Sierra Leone, Singapore, South Africa, Sri Lanka, Sudan, Swaziland, Taiwan, Tanzania, Thailand, Tunisia, Turkey, Uganda, Ukraine, Uruguay, Uzbekistan, Venezuela, Vietnam, Zambia, Zimbabwe' Solomon Islands, Vanuatu	Europe US + Australie Reste du monde
Réseau de collaboration réduit	Argentina, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, China, Colombia, Croatia, Czech Republic, Denmark, Egypt, Estonia, Finland, France, Germany, Greece, Hong Kong, Hungary, Iceland, India, Indonesia, Ireland, Israel, Italy, Japan, Kenya, South Korea, Lithuania, Malawi, Malaysia, Mexico, Netherlands, New Zealand, Norway, Pakistan, Peru, Philippines, Poland, Portugal, Puerto Rico, Romania, Russian Federation, Saudi Arabia, Serbia, Singapore, Slovakia, Slovenia, South Africa, Spain, Sweden, Switzerland, Taiwan, Tanzania, Thailand, Turkey, Uganda, Ukraine, United Kingdom, United States, Vietnam, Zambia, Zimbabwe	Tous les pays	Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Lithuania, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom Australia, Canada, New Zealand, Puerto Rico, United States Argentina, Brazil, Chile, China, Colombia, Egypt, Hong Kong, India, Indonesia, Japan, Kenya, South Korea, Malawi, Malaysia, Mexico, Pakistan, Peru, Philippines, Russian Federation, Saudi Arabia, Singapore, South Africa, Taiwan, Tanzania, Thailand, Turkey, Uganda, Ukraine, Vietnam, Zambia, Zimbabwe	Europe US + Australie Reste du monde
Réseau de co-occurrence	Austria, Belgium, Bulgaria, Croatia, Czech Republic, Estonia, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Lithuania, Netherlands, Poland, Portugal, Romania, Serbia, Spain, Slovakia, Switzerland, United Kingdom, Turkey, Slovenia Denmark, Finland, Iceland, Norway, Sweden Canada, Puerto Rico, United States Australia, New Zealand Kenya, Malawi, South Africa, Tanzania, Uganda, Zambia, Zimbabwe Argentina, Brazil, Chile, Colombia, Egypt, India, Mexico, Pakistan, Peru China, Hong Kong, Indonesia, Japan, South Korea, Malaysia, Philippines, Singapore, Taiwan, Thailand, Saudi Arabia, Vietnam Russian Federation, Ukraine	Europe Pays scand. US Australie Afrique Am Lat + Inde Asie Russia	Austria, Belgium, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal, Spain, Switzerland, United Kingdom Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Israel, Lithuania, Poland, Romania, Russian Federation, Serbia, Slovakia, Turkey, Ukraine, Slovenia Denmark, Finland, Iceland, Norway, Sweden Australia, Canada, Puerto Rico, New Zealand, United States Kenya, Malawi, South Africa, Tanzania, Uganda, Zambia, Zimbabwe Argentina, Brazil, Chile, China, Colombia, Egypt, Hong Kong, India, Indonesia, Japan, South Korea, Malaysia, Mexico, Pakistan, Peru, Philippines, Singapore, Taiwan, Thailand, Saudi Arabia, Vietnam	Europe occidentale Europe de l'est Pays scand. US + Australie Afrique Am Lat + Asie

TABLE 2.3 – Partitions des réseaux de collaboration et de co-occurrence non-industriels avec deux algorithmes

Chapitre 3

Automatisation des cartographies de la recherche

Pour mettre en place des cartographies de la recherche à grande échelle qui prennent en compte le plus grand nombre de sources, il est nécessaire de développer des méthodes automatiques d'indexation et d'extraction des connaissances pour rendre ces bases interopérables. En effet, les bases renseignant sur l'activité de recherche clinique sont hétérogènes et de grand volume et n'ont pas été faites initialement pour monitorer l'activité de recherche à grande échelle. L'OMS a appelé à la création d'un observatoire global de la recherche et du développement médical, dont l'objectif est de rendre interopérable les bases de données traçant l'activité de recherche afin de pour produire de l'information utile pour la prise de décision sur les programmes de recherche. Nous avons contribué au développement de cet observatoire en créant un algorithme de classification automatique des essais cliniques enregistrés dans l'ICTRP selon les catégories de morbi-mortalité de l'étude GBD. Cet algorithme nous a permis d'identifier à l'échelle de tous les essais enregistrés dans l'ICTRP des lacunes dans l'effort de recherche relativement aux besoins de santé (Section 2.2).

3.1 Observatoire global de la recherche médicale

3.1.1 Appel de l'Organisation Mondiale de la Santé

Il existe un grand nombre de sources de données informant sur l'état de la recherche médicale, les financements, la recherche en cours, et les produits issus de celle-ci (Section 1.2.4). De surcroît, la transition vers l'ouverture des données accroît le nombre de sources disponibles traçant les activités de la recherche clinique, à niveau international, régional et national. L'OMS a appelé à la création d'un *observatoire global de la recherche et développement médical* (Figure 3.1) (Terry et al., 2014).

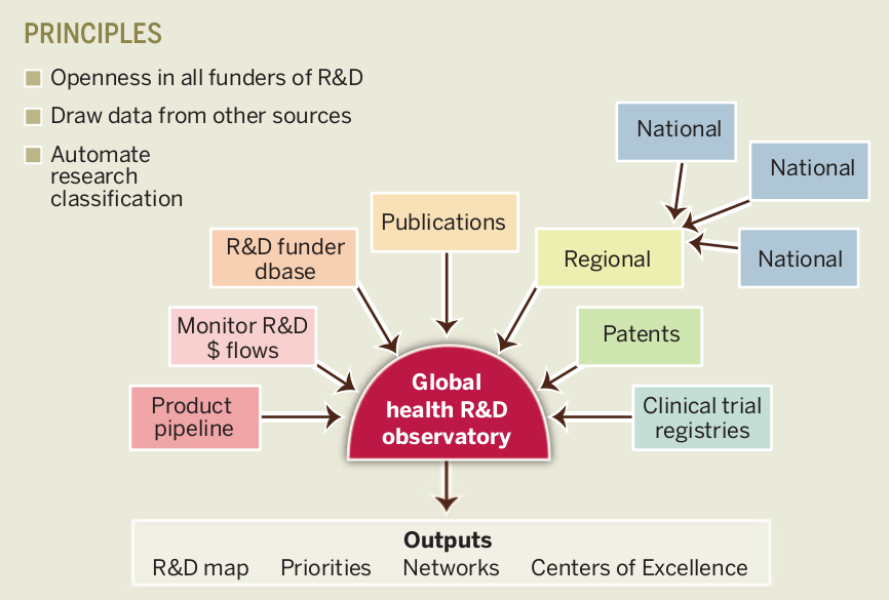


FIGURE 3.1 – Observatoire global de la recherche et le développement médical
Source : (Terry et al., 2014)

L'objectif est de créer un système unique qui recueille toutes les sources d'information sur les activités de recherche médicale pour fournir de l'information aux décideurs du programme de recherche sur les besoins dans la production de connaissances médicales, par l'identification de priorités de recherche, de domaines sous-étudiés et de centres d'excellence. Un tel observatoire est particulièrement né-

cessaire dans les régions et les domaines médicaux où les ressources sont limitées (Chersich and Martin, 2017). Cet observatoire doit en particulier assurer la qualité de l'information fournie, être modulable à la diversité des besoins de chaque décideur (diversité de niveau d'échelle, de spécificité de la maladie ou la région d'intérêt) et permettre un accès unique à la plus grande quantité d'information pour réanalyse spécifique.

Pour créer cet observatoire, il est nécessaire de normaliser les données concernant la recherche clinique en utilisant des nomenclatures standardisées (Terry et al., 2012). Par exemple, tous les investissements, essais cliniques et publications concernant une même maladie devraient être labellisés identiquement pour pouvoir facilement les réunir. Par ailleurs, le grand volume de données digitales non structurées (en particulier du texte libre) nécessite des moyens automatiques d'indexation et d'extraction des connaissances. Or, les données sur l'activité de recherche médicale n'ont pas été initialement recueillies en vue d'analyses agrégées, ni pour être inter-opérables avec d'autres bases de données.

3.1.2 Opérabilité et interopérabilité des sources de données

En effet, les données sur l'activité de recherche ne sont pas toujours collectées dans des formats exploitables pur des analyses automatiques à grande échelle, ni avec l'utilisation de nomenclatures standardisées permettant l'inter-opérabilité de ces bases d'origines très hétérogènes. Pour créer des cartographies de la recherche clinique, il est nécessaire d'une part de réussir à consulter chaque base de données de façon systématique et compréhensive, et d'autre part de croiser ces différentes sources de données. La grande variété de sources de données, leur hétérogénéité et leur volume pose la question sur le caractère opérationnel de celles-ci, et leur capacité à être interopérables.

Les analyses qui seront possibles à partir de ces bases de données vont dépendre

de l'échelle de l'analyse menée. Les bases de données accessibles dans des formats facilement exploitables comme XML ou JSON et utilisant des taxonomies standardisées pourront être traitées automatiquement et utilisées pour des analyses à grande échelle. D'autres bases de données, par exemple au format PDF, pourront être utilisées uniquement pour construire des cartographies à petite échelle moyennant du travail manuel pour extraire et standardiser les données afin de les rendre opérationnelles.

Les bases de données réunissant les publications scientifiques dans le domaine biomédical ont été largement indexées de façon standardisées pour faciliter le travail bibliographique. C'est le cas par exemple des publications inclues dans MEDLINE, indexées par des codes Medical Subject Headings (MeSH) via PubMed de façon totalement automatique ou semi-automatique avec vérification manuelle (Mork et al., 2013). De la même façon, EMBASE indexe les publications avec des codes SNOMED-CT. Ces indexations peuvent être réutilisées pour la création de cartographies à très grande échelle à partir des publications scientifiques (Evans et al., 2014).

Parmi les registres d'essais cliniques présentes dans l'ICTRP, ClinicalTrials.gov a été retravaillé par la U.S. National Library of Medicine (NLM) pour faciliter l'extraction d'information ou faire des analyses agrégées. De façon similaire à PubMed, les registres d'essais cliniques sont indexés par des codes MeSH de façon automatisée, mais l'indexation n'est pas vérifiée manuellement (Mork et al., 2013). Par ailleurs, le Aggregated Analysis of Clinical Trials (AACT) est un projet qui vise à retravailler la base ClinicalTrials.gov pour permettre des analyses agrégées de celle-ci (Tasneem et al., 2012). Cette base regroupe les essais enregistrés par spécialité médicales à partir de l'indexation MeSH de celle-ci, et permet des exports munis d'analyses de tous les essais d'une spécialité médicale. Le AACT est à la base de la plupart des analyses descriptives d'essais cliniques spécifiques à une spécialité médicale présentés dans la Section 1.3.2. Les autres registres inclus dans l'ICTRP ne bénéficient pas de ce travail de restructuration. En fait, les systèmes d'extrac-

tion d'information mis en place par ClinicalTrials.gov et l'ICTRP reposent sur des systèmes d'indexations différents, et peuvent ainsi donner des résultats différents pour une même requête (Chang et al., 2016).

L'interopérabilité entre les registres d'essais cliniques et les publications n'est par ailleurs pas assurée. On ne peut pas automatiquement relier le registre d'un essai clinique à sa publication correspondante, et vice-versa. Des travaux sont en cours pour identifier les essais n'ayant pas publié leurs résultats (Powell-Smith and Goldacre, 2016) et pour donner un seul point d'accès à toute l'information existante sur un même essai (Goldacre and Gray, 2016).

Aujourd'hui, les méthodes de traitement automatique du langage (TAL) dans le domaine médical commencent à devenir assez mûres pour être efficaces dans des contextes cliniques spécifiques (Demner-Fushman et al., 2009; Névéol and Zweigenbaum, 2015). Un grand avantage du domaine médical par rapport à d'autres est que le travail sur le traitement de l'information biomédicale et sa normalisation date de plusieurs décennies, donnant lieu à des systèmes ontologiques avancés tels que le Unified Medical Language System® (UMLS®) créé par la NLM (Bodenreider, 2004). L'UMLS® comporte un meta-thésaurus de 12 millions de concepts (appelés concepts UMLS) unifiant approximativement 200 systèmes de terminologie médicale, un réseau sémantique entre ces concepts, et un système lexical qui enrichi les concepts UMLS par des catégorisations syntaxiques et des variantes. À partir de l'UMLS® ont été développés des systèmes d'indexation de texte libre biomédical comme MetaMap® créé par la NLM (Aronson, 2001), à la base du système d'indexation MeSH de PubMed depuis plus de 10 ans (Mork et al., 2017). L'utilisation du TAL et autres techniques d'extraction de connaissances deviennent impératives pour maximiser la collecte, la synthèse et la réutilisation de données déjà existantes pour les rendre interopérables.

3.2 Classification automatique des essais cliniques

3.2.1 Résumé

Dans les registres d'essais cliniques, il n'existe pas de champ normalisé spécifiant quelle maladie est étudiée. Cette information peut être trouvée dans des champs non-structurés de texte libre. Pour cartographier à grande échelle les essais cliniques enregistrés par maladie, il est donc nécessaire d'analyser ce texte libre afin de classifier les essais selon une nomenclature standardisée des maladies. L'étude GBD définit une taxonomie des maladies utilisée pour estimer le fardeau des maladies par pays. Ce troisième travail présente le développement d'un algorithme de classification automatique des essais cliniques enregistrés vers les catégories de mortalité ou d'incapacité définies dans l'étude GBD (Atal et al., 2016). Autrement dit, notre objectif était de rendre les registres d'essais cliniques interopérables avec l'information fournie par l'étude GBD. Cet inter-opérabilité permet ainsi la création des cartographies de l'effort de recherche clinique par maladie comparables à leur fardeau (Section 2.2).

La taxonomie de causes de morbi-mortalité de l'étude GBD 2010 comporte plusieurs hiérarchies de détail, au niveau le plus détaillé comportant 291 causes de morbi-mortalité. Nous avons décidé de grouper ces causes en grandes thématiques de santé publique pour cartographier la recherche à grande échelle. Un panel d'expert a regroupé les causes en 27 groupes de maladies, et un groupe réunissant es accidents ou blessures non liés à des maladies (e.g. accidents de la route, suicide ou actes de guerre).

L'algorithme qu'on a développé classifie chaque essai clinique selon ces 28 groupes de causes de mortalité ou d'incapacité. Chaque groupe est défini dans l'étude GBD à partir de codes CIM-10 (classification internationale des maladies, 10^{e} révision) (World Health Organization, 2011). L'algorithme utilise dans un premier temps le système de connaissances UMLS® pour dériver des chemins entre des champs de texte du registre et des codes CIM-10. Dans un deuxième temps il

décide de la classification GBD finale à partir de règles de priorisation entre ces chemins.

L'algorithme utilise des méthodes de TAL implémentées dans MetaMap® pour identifier des concepts UMLS correspondant à des noms de maladies ou blessures dans des champs de texte du registre. Il utilise ensuite IntraMap (Fung and Bodenreider, 2005), système de projection inter-terminologique, pour trouver le(s) code(s) CIM-10 le(s) plus proche(s) des concepts identifiés. Ceci dessine ainsi des chemins entre un registre et une ou plusieurs des 28 catégories de mori-mortalité. Ces chemins peuvent se recouper ou au contraire être disjoints. Nous avons ainsi créé des règles de priorisation entre ces chemins en fonction de l'origine de ceux-ci et de leur recoupements.

Nous avons comparé la classification automatique à une classification manuelle de 2,763 essais cliniques provenant d'autres études pour estimer les valeurs prédictives positives et négatives spécifique à chaque groupe de causes de morbi-mortalité (Emdin et al., 2015; Viergever et al., 2013; Zeitoun et al., 2017). Pour construire un des jeux de données de validation utilisé (Zeitoun et al., 2017), nous avons développé une interface web pour classifier rapidement des essais enregistrés selon les causes de mortalité ou d'incapacité de l'étude GBD (Figure 3.2). L'outil permet de visualiser le registre et de naviguer dans la taxonomie GBD et les codes CIM-10 de chaque cause de morbi-mortalité.

L'algorithme de classification a identifié le groupe de morbi-mortalité exact pour 78% des essais cliniques manuellement classifiés. Les performances ont été particulièrement bonnes pour identifier les essais étudiant le cancer (sensibilité 97.4%, spécificité 97.5%). La sensibilité a été modérée (53 %) pour les essais ne correspondant à aucune catégorie de morbi-mortalité (portant par exemple sur le traitement de la douleur ou des symptômes transverses à plusieurs maladies). La sensibilité a été faible pour les essais portant sur les blessures (16 %). Cet algorithme peut être utilisé pour des analyses à grande échelle de l'épidémiologie de la recherche clinique comparée au fardeau des maladies.

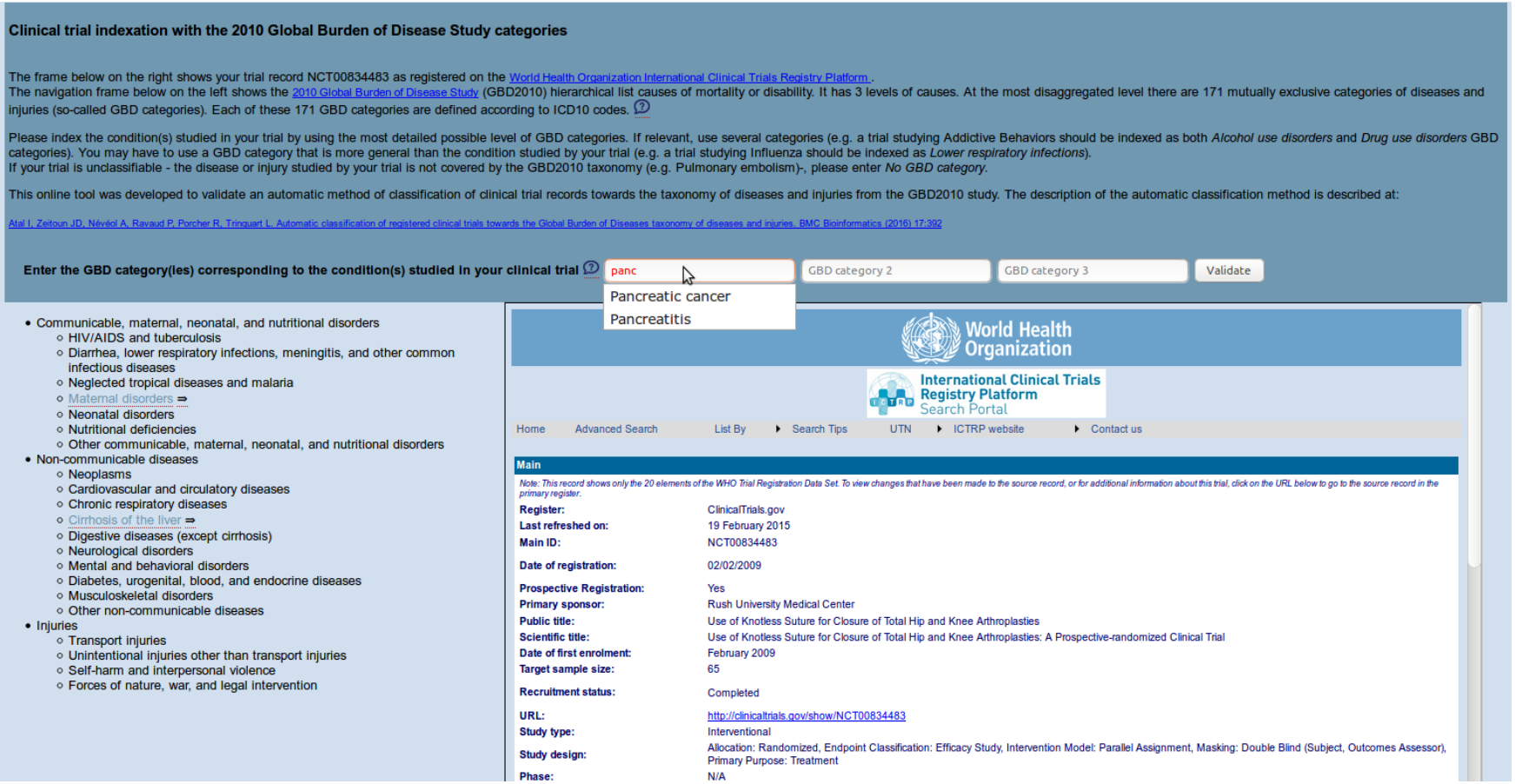


FIGURE 3.2 – Outil pour classification manuelle d'essais enregistrés dans l'ICTRP avec des catégories GBD
Un exemple dans : http ://www.clinicalepidemio.fr/gbd_study_who/

Nous avons utilisé l'algorithme pour classifier tous les essais enregistrés inclus dans l'ICTRP (N=109,603 lorsque nous avons conduit l'étude). La complexité de l'algorithme de classification, en particulier l'indexation automatique de texte libre par MetaMap®, faisait que les temps de calcul pour classifier tous les registres dépassait les 10 jours sur un ordinateur personnel. Nous avons ainsi développé un programme pour implémenter MetaMap® de façon distribuée sur un *cluster* de calcul (Benchoufi and Atal, 2016). Sur un *cluster* de 15 machines, le temps de calcul est passé à 3 heures.

L'algorithme de classification et la classification des 109,603 essais enregistrés sont accessibles librement pour la communauté scientifique. Fondé sur l'analyse de champs de texte libre, notre algorithme peut être facilement réutilisé pour la classification d'autres bases de données traçant l'activité de recherche pour lesquelles les maladies étudiées sont rapportées dans des champs non structurés (e.g. publications d'essais randomisés ou revues systématiques), mais aussi peut être adapté pour des classifications vers d'autres taxonomies des causes de morbi-mortalité (e.g. mises à jour de la taxonomie GBD, ou bien d'autres groupements à partir de codes CIM-10). Ce travail constitue ainsi une brique pour l'observatoire global de la recherche et du développement en matière de santé.

Atal et al. BMC Bioinformatics (2016) 17:392
DOI 10.1186/s12859-016-1247-7

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Automatic classification of registered clinical trials towards the Global Burden of Diseases taxonomy of diseases and injuries

Ignacio Atal[1,2,3*], Jean-David Zeitoun[1,2,3], Aurélie Névéol[4], Philippe Ravaud[1,2,3,5], Raphaël Porcher[1,2,3] and Ludovic Trinquart[1,2,5]

Abstract

Background: Clinical trial registries may allow for producing a global mapping of health research. However, health conditions are not described with standardized taxonomies in registries. Previous work analyzed clinical trial registries to improve the retrieval of relevant clinical trials for patients. However, no previous work has classified clinical trials across diseases using a standardized taxonomy allowing a comparison between global health research and global burden across diseases. We developed a knowledge-based classifier of health conditions studied in registered clinical trials towards categories of diseases and injuries from the Global Burden of Diseases (GBD) 2010 study.
The classifier relies on the UMLS® knowledge source (Unified Medical Language System®) and on heuristic algorithms for parsing data. It maps trial records to a 28-class grouping of the GBD categories by automatically extracting UMLS concepts from text fields and by projecting concepts between medical terminologies. The classifier allows deriving pathways between the clinical trial record and candidate GBD categories using natural language processing and links between knowledge sources, and selects the relevant GBD classification based on rules of prioritization across the pathways found. We compared automatic and manual classifications for an external test set of 2,763 trials. We automatically classified 109,603 interventional trials registered before February 2014 at WHO ICTRP.

Results: In the external test set, the classifier identified the exact GBD categories for 78 % of the trials. It had very good performance for most of the 28 categories, especially "Neoplasms" (sensitivity 97.4 %, specificity 97.5 %). The sensitivity was moderate for trials not relevant to any GBD category (53 %) and low for trials of injuries (16 %). For the 109,603 trials registered at WHO ICTRP, the classifier did not assign any GBD category to 20.5 % of trials while the most common GBD categories were "Neoplasms" (22.8 %) and "Diabetes" (8.9 %).

Conclusions: We developed and validated a knowledge-based classifier allowing for automatically identifying the diseases studied in registered trials by using the taxonomy from the GBD 2010 study. This tool is freely available to the research community and can be used for large-scale public health studies.

Keywords: Clinical trials, Global burden of diseases, Disease classification, Mapping

Background

The World Health Organization (WHO) has indicated the pressing need for a comprehensive monitoring of health research and development (R&D) to coordinate limited resources towards reducing the gaps between health research and health needs [1–3]. Mapping the global landscape of health R&D will allow for identifying diseases for which there is too much or too little research at a local level as compared to their burden at the same level [4]. The WHO is developing the Global Observatory on Health R&D and aims at analyzing multiple data sources to quantify the global state of health R&D, including clinical trial registries, publications, product pipelines, patents and grants [3, 5].

Although concerning a particular type of health R&D activity, one source of data, clinical trial registries, is

* Correspondence: ignacio.atal-ext@aphp.fr
[1]Centre d'Épidémiologie Clinique, Hôpital Hôtel-Dieu, Paris, France
[2]INSERM U1153, Paris, France
Full list of author information is available at the end of the article

readily available and could be used to rapidly achieve a global mapping [6]. Worldwide, clinical trials are registered in publicly accessible repositories with a common structure of data fields [7]. The WHO gathers 16 registries in the International Clinical Trials Registry Platform (ICTRP), now the largest repository of clinical trials worldwide [8].

However, the diseases studied by clinical trials registered in the WHO ICTRP are not described in trial records by using a standardized taxonomy but rather as free text with considerable heterogeneity. With more than 300,000 clinical trial records in the WHO ICTRP and more than 20,000 new records registered every year, the use of automatic methods for classification is imperative [8, 9]. Natural Language Processing (NLP) allows clinical knowledge representation in standardized formats and is becoming mature enough to be used efficiently for targeted applications [10, 11]. In particular, NLP methods have been developed to face the limitations of the retrieval systems of clinical trial registries such as clinicaltrials.gov. [12, 13] For instance, clinical trial records have been notably analyzed using NLP to provide formal representations of eligibility criteria, or to enrich eligibility criteria with meta-data to improve the retrieval of relevant clinical trials for patients [14–26]. However, none of these studies have analyzed the performance of retrieval of clinical trials across diseases, but rather across features of eligibility criteria (e.g. age, BMI [1] or more complex features) for specific diseases.

Moreover, the health conditions studied in registered clinical trials must be classified by using a taxonomy of diseases that allows for comparisons between the numbers of clinical trials and the actual burden of diseases. A consensual taxonomy over which the evolution of the burden is estimated regionally was developed by the US Institute for Health Metrics and Evaluation for the Global Burden of Diseases (GBD) 2010 study [27, 28]. Previous studies have developed NLP methods to index clinical trial records using Medical Subject Headings (MeSH) [29], and to regroup clinical trials across medical specialties [30]. However, to our knowledge no previous work has classified clinical trials using a taxonomy allowing a comparison between global health research and global burden across diseases.

Objective

We aimed to develop and validate a method that automatically maps the health conditions studied in registered clinical trials to the taxonomy from the GBD 2010 study. Towards that goal, we relied on Natural Language Processing to analyze the free-text description of health conditions found in clinical trial records, and a standardized knowledge representation of diseases to encode the information extracted from the trial records.

Methods

We developed a knowledge-based classifier allowing for automatic mapping of the health conditions studied in registered clinical trials to a 28- and 171-class grouping of the taxonomy of diseases and injuries defined by the GBD 2010 study. Our approach did not rely on statistical classification techniques but instead relied on text analysis and exploited the Unified Medical Language System® (UMLS®) as a domain knowledge resource. Specifically, the classification is based on the recognition of medical concepts in the free text description of trials and the mapping of concepts between medical taxonomies. The classifier allows deriving pathways between the clinical trial record and the taxonomy of diseases and injuries from the GBD study based on a succession of mathematical projections (also called normalization or entity linking). Finally, the classifier selects the relevant GBD classification based on rules of prioritization across the pathways found. We measured the classifier performance by comparing the automatic classifications to manual classifications with a large test set of registered clinical trials. Finally, we used the classifier to map the conditions studied by all trials registered at the WHO ICTRP.

From clinical trial records to the GBD cause list

GBD cause list

The GBD cause list is a set of 291 mutually exclusive and collectively exhaustive categories of diseases and injuries [28]. Each category is defined in terms of the codes of the International Classification of Diseases 9th and 10th versions (ICD9 and ICD10) [31]. We used the mapping from the ICD10 to the GBD cause list (Web Table 3 in [27]). Several residual categories, such as “Other infectious diseases”, are made up of ill-defined or residual causes from major disease groups. We excluded these because they are not informative from the perspective of a global analysis of the burden of diseases.

We developed a smaller list of categories by using a formal consensus method. Six experts independently defined a higher-level grouping of diseases and injuries that are sufficiently informative for developing a global mapping of clinical trials across health needs. The resulting list contained 28 categories that accounted for 98.8 % of the global burden in 2010 (Table 1). Moreover, we considered the list of aggregated categories defined by the GBD 2010 study to inform policy makers on the main health problems per country (Web Table 1 in [28]). This grouping contained 171 GBD categories that accounted for 90.6 % of the global burden of disease in 2010 (Additional file 1: Table S1). We report results of the mapping to the 28 categories; results of the mapping to the 171 categories are presented in the Additional file 1.

Table 1 Grouping of the Global Burden of Diseases (GBD) cause list in 28 GBD categories

GBD categories	Partition of the GBD cause list
Tuberculosis	Tuberculosis
HIV/AIDS	HIV/AIDS
Diarrhea, lower respiratory infections, meningitis, and other common infectious diseases	Diarrheal diseases; Typhoid and paratyphoid fevers; Lower respiratory infections; Upper respiratory infections; Otitis media: Meningitis; Encephalitis; Diphtheria; Whooping cough; Tetanus; Measles; Varicella
Malaria	Malaria
Neglected tropical diseases excluding malaria	Chagas disease; Leishmaniasis: African trypanosomiasis; Schistosomiasis; Cysticercosis; Echinococcosis; Lymphatic filariasis; Onchocerciasis; Trachoma; Dengue; Yellow fever; Rabies; Food-borne trematodiases; Intestinal nematode infections; Other neglected tropical diseases
Maternal disorders	Maternal hemorrhage; Maternal sepsis; Hypertensive disorders of pregnancy; Obstructed labor; Abortion; Other maternal disorders
Neonatal disorders	Preterm birth complications; Neonatal encephalopathy (birth asphyxia and birth trauma); Sepsis and other infectious disorders of the newborn baby; Other neonatal disorders
Nutritional deficiencies	Protein-energy malnutrition; Iodine deficiency; Vitamin A deficiency; Iron-deficiency anemia; Other nutritional deficiencies
Sexually transmitted diseases excluding HIV	Syphilis; Sexually transmitted chlamydial diseases; Gonococcal infection; Trichomoniasis; Other sexually transmitted diseases
Hepatitis	Acute hepatitis A; Acute hepatitis B; Acute hepatitis C; Acute hepatitis E
Leprosy	Leprosy
Neoplasms	Esophageal cancer; Stomach cancer; Liver cancer; Larynx cancer; Trachea, bronchus, and lung cancers; Breast cancer; Cervical cancer; Uterine cancer; Prostate cancer; Colon and rectum cancers; Mouth cancer; Nasopharynx cancer; Cancer of other part of pharynx and oropharynx; Gallbladder and biliary tract cancer; Pancreatic cancer; Malignant melanoma of skin; Non-melanoma skin cancer; Ovarian cancer; Testicular cancer; Kidney and other urinary organ cancers; Bladder cancer; Brain and nervous system cancers; Thyroid cancer; Hodgkin's disease; Non-Hodgkin lymphoma; Multiple myeloma; Leukemia; Other neoplasms
Cardiovascular and circulatory diseases	Rheumatic heart disease; Ischemic heart disease; Cerebrovascular disease; Hypertensive heart disease; Cardiomyopathy and myocarditis; Atrial fibrillation and flutter; Aortic aneurysm; Peripheral vascular disease; Endocarditis; Other cardiovascular and circulatory diseases
Chronic respiratory diseases	Chronic obstructive pulmonary disease; Pneumoconiosis; Asthma; Interstitial lung disease and pulmonary sarcoidosis; Other chronic respiratory diseases
Cirrhosis of the liver	Cirrhosis of the liver
Digestive diseases (except cirrhosis)	Peptic ulcer disease; Gastritis and duodenitis; Appendicitis; Paralytic ileus and intestinal obstruction without hernia; Inguinal or femoral hernia; Non-infective inflammatory bowel disease; Vascular disorders of intestine; Gall bladder and bile duct disease; Pancreatitis; Other digestive diseases
Neurological disorders	Alzheimer's disease and other dementias; Parkinson's disease; Epilepsy; Multiple sclerosis; Migraine; Tension-type headache; Other neurological disorders
Mental and behavioral disorders	Schizophrenia; Alcohol use disorders; Drug use disorders; Unipolar depressive disorders; Bipolar affective disorder; Anxiety disorders; Eating disorders; Pervasive development disorders; Childhood behavioral disorders; Idiopathic intellectual disability; Other mental and behavioral disorders
Diabetes, urinary diseases and male infertility	Diabetes mellitus; Acute glomerulonephritis; Chronic kidney diseases; Urinary diseases and male infertility
Gynecological diseases	Uterine fibroids; Polycystic ovarian syndrome; Female infertility; Endometriosis; Genital prolapse; Premenstrual syndrome; Other gynecological diseases
Hemoglobinopathies and hemolytic anemias	Hemoglobinopathies and hemolytic anemias; Thalassemias; Sickle cell disorders; G6PD deficiency; Other hemoglobinopathies and hemolytic anemias
Musculoskeletal disorders	Rheumatoid arthritis; Osteoarthritis; Low back and neck pain; Gout; Other muskuloskeletal disorders
Congenital anomalies	Congenital anomalies; Neural tube defects; Congenital heart anomalies; Cleft lip and cleft palate; Down's syndrome; Other chromosomal abnormalities; Other congenital anomalies
Skin and subcutaneous diseases	Eczema; Psoriasis; Cellulitis; Abscess, impetigo, and other bacterial skin diseases; Scabies; Fungal skin diseases; Viral skin diseases; Acne vulgaris; Alopecia areata; Pruritus; Urticaria; Decubitus ulcer; Other skin and subcutaneous diseases

Table 1 Grouping of the Global Burden of Diseases (GBD) cause list in 28 GBD categories *(Continued)*

Sense organ diseases	Glaucoma; Cataracts; Macular degeneration; Refraction and accommodation disorders; Other hearing loss; Other vision loss; Other sense organ diseases
Oral disorders	Dental caries; Periodontal disease; Edentulism
Sudden infant death syndrome	Sudden infant death syndrome
Injuries	Transport injuries; Unintentional injuries other than transport injuries; Self-harm and interpersonal violence; Forces of nature, war, and legal intervention
Excluded residual categories	Other infectious diseases; Other endocrine, nutritional, blood, and immune disorders

Grouping of the cause list of diseases and injuries from the Global Burden of Diseases 2010 study in 28 GBD categories, plus the excluded residual categories. This grouping was considered sufficiently informative for a global mapping of health research to a global mapping of health needs

Clinical trial records

In the WHO Trial Registration Dataset, the "Health Condition(s) or Problem(s) studied" field contains a natural language description of the primary condition or problem studied in any clinical trial. Figure 1 shows an example for which the health condition field is "Knee Osteoarthritis" and "Hip Osteoarthritis". This description is not captured by a coded field, with a standardized taxonomy of diseases, but is rather described in a free-text field. Moreover, the analysis of this free-text field alone may not be sufficient to identify the GBD categories of interest. Numerous health condition fields are empty, have entry errors, correspond to "Healthy volunteers", or the relevant GBD category may be difficult to identify because of synonymy. Thus, we also considered the "Public Title" and "Scientific Title" fields, which are most likely to bring additional information about the condition studied in the clinical trial and to enrich the mapping.

Classifier development

Because GBD categories are defined by ICD10 codes, we aimed to classify the text fields according to ICD10 codes.

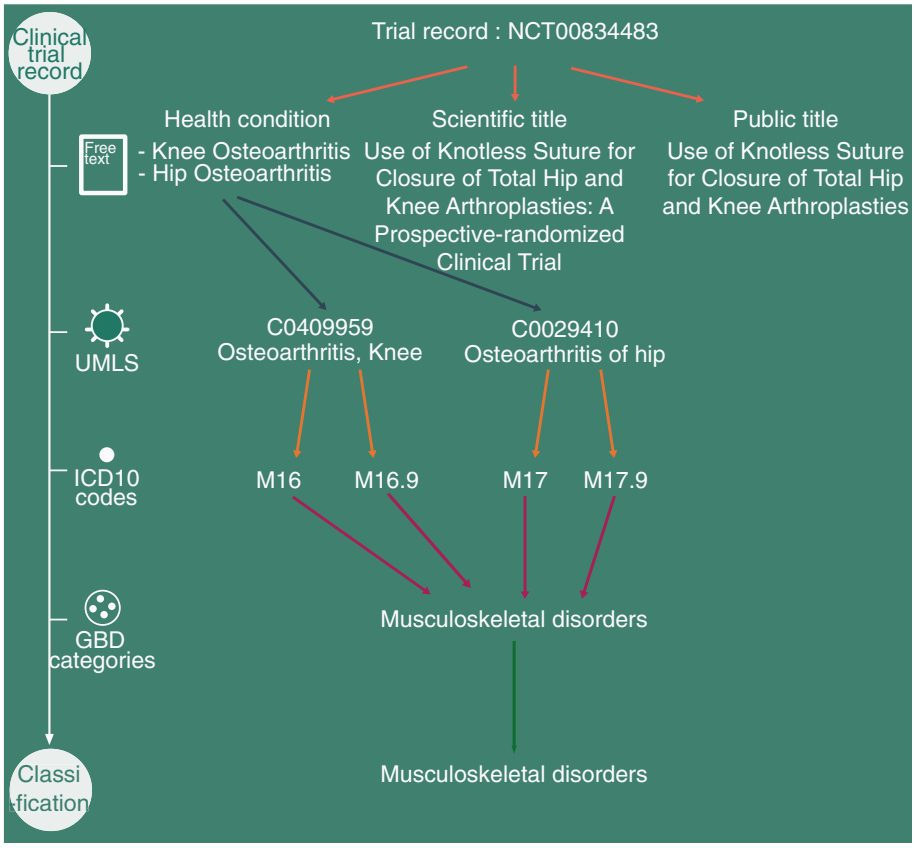


Fig. 1 Example of classification of a clinical trial record towards the GBD categories. The classification process is based on text extraction from the trial record, text annotation using UMLS concepts, projection of UMLS concepts to ICD10 codes, projection of ICD10 codes to candidate GBD categories among the 28 GBD categories, and GBD classification based on the candidate GBD categories. In this example, the text annotation involved use of the WSD server for MetaMap, and no expert-based enrichment was needed

The Unified Medical Language System® (UMLS®), developed at the US National Library of Medicine (NLM), is the most comprehensive metathesaurus to analyze biomedical text in English to date [32]. We based our classifier on established methods using the UMLS knowledge source to automatically annotate trial records with ICD10 codes.

Figure 2 illustrates the 5 methodological stages we defined for the classifier (interactive version at http://clinicalepidemio.fr/gbd_graph). The 4 initial stages allow for deriving pathways from the clinical trial record to candidate GBD categories. The 5th stage allows for deriving the GBD classification based on prioritization rules over the pathways found.

Free text annotation with concepts from the unified medical language system

We first annotated the text fields (health condition, public title and scientific title) with concepts from the UMLS metathesaurus [32]. The annotation involved use of MetaMap, a tool from the NLM for recognizing UMLS concepts in text [33]. We considered only UMLS concepts corresponding to diseases or injuries (MetaMap implementation in Additional file 1). A Word Sense Disambiguation (WSD) server can be used to select a single UMLS concept when a text is annotated with several UMLS concepts. We developed the classifier with and without using the WSD server. In Fig. 1, the health condition field was annotated with the concepts "Osteoarthritis, Knee" (C0409959) and "Osteoarthritis of hip" (C0029410).

Mapping of UMLS concepts to ICD10 codes

Each UMLS concept was then projected to one or several ICD10 codes. The projection involved a semantic-based approach to connect different terminologies present in the UMLS database, namely the Restrict-to-ICD10 algorithm, as implemented in the IntraMap program (IntraMap implementation in Additional file 1) [34]. In the example from Fig. 1, the concept "Osteoarthritis, Knee" was projected to the ICD10 codes "Coxarthrosis [arthrosis of hip]" and "Coxarthrosis, unspecified".

Mapping of ICD10 codes to candidate GBD categories

The resulting ICD10 codes were then projected to one or several candidate GBD categories. ICD10 codes could correspond to three- and four-character ICD10 codes (e.g. M16 and M16.9 in the example from Fig. 1), or to blocks of three- and four-character ICD10 codes (e.g. F30–F39.9). Three- and four-character ICD10 codes were projected to a GBD category only if it was totally included in an unique GBD category. For instance, the ICD10 code P37 could not be projected to a GBD category as P37.0 was included in the GBD category "Tuberculosis", and P37.3 was included in the GBD category "Neglected tropical diseases excluding malaria". Blocks of ICD10 codes were split into a list of three- and four-character ICD10 codes (e.g. F30–F39.9 was split into F30, F31, …, F39.9). The block of ICD10 codes was projected to the GBD category(ies) corresponding to the individual projections of the three- and four-character ICD10 codes. In the example from Fig. 1, the ICD10 codes were projected to the GBD category "Musculoskeletal disorders".

Expert-based enrichment

Some UMLS concepts were not mapped to any candidate GBD category. We manually reviewed those UMLS concepts appearing in more than 10 clinical trials registered at the WHO ICTRP database by February 2014 and projected them to candidate GBD categories when relevant. We manually reviewed 503 UMLS concepts, among which 62 could be projected to candidate GBD categories (Additional file 1: Datasets S1 and S2). We developed the classifier with and without the expert-based enrichment.

Prioritization rules for GBD classification

For each trial, the previous stages resulted in several pathways from the health condition, the public title and the scientific title fields to multiple candidate GBD categories, respectively. These pathways may pass through several UMLS concepts and ICD10 codes. We developed rules of prioritization to define the GBD classification.

We gave priority to pathways issued from the health condition field because, by definition, it contains the information about the health condition(s) studied in the clinical trial. We also gave priority to candidate GBD categories for which the trial record was consistently projected by several pathways versus candidate GBD categories reached by isolated pathways. This rule aims at discarding candidate GBD categories that may appear by noise (Prioritization rules in Additional file 1). We developed the classifier with and without the rule of giving priority to the health condition field. In the example from Fig. 1, all the pathways from the trial record arrived at the same GBD category, "Musculoskeletal disorders".

Note that for some trials, the classifier may not find any GBD category. These trials may study health conditions corresponding to residual categories or health conditions not relevant for the GBD 2010 study (eg, pain management). These trials were classified as "No GBD" category trials.

External validation

We compared the automatic classification to a manual classification (considered the gold standard) for a large test set of registered clinical trials. We measured the performance of 8 versions of the classifier, corresponding to the combinations of using or not the WSD server, using or not the expert-based enrichment, and giving or not priority to the health condition field.

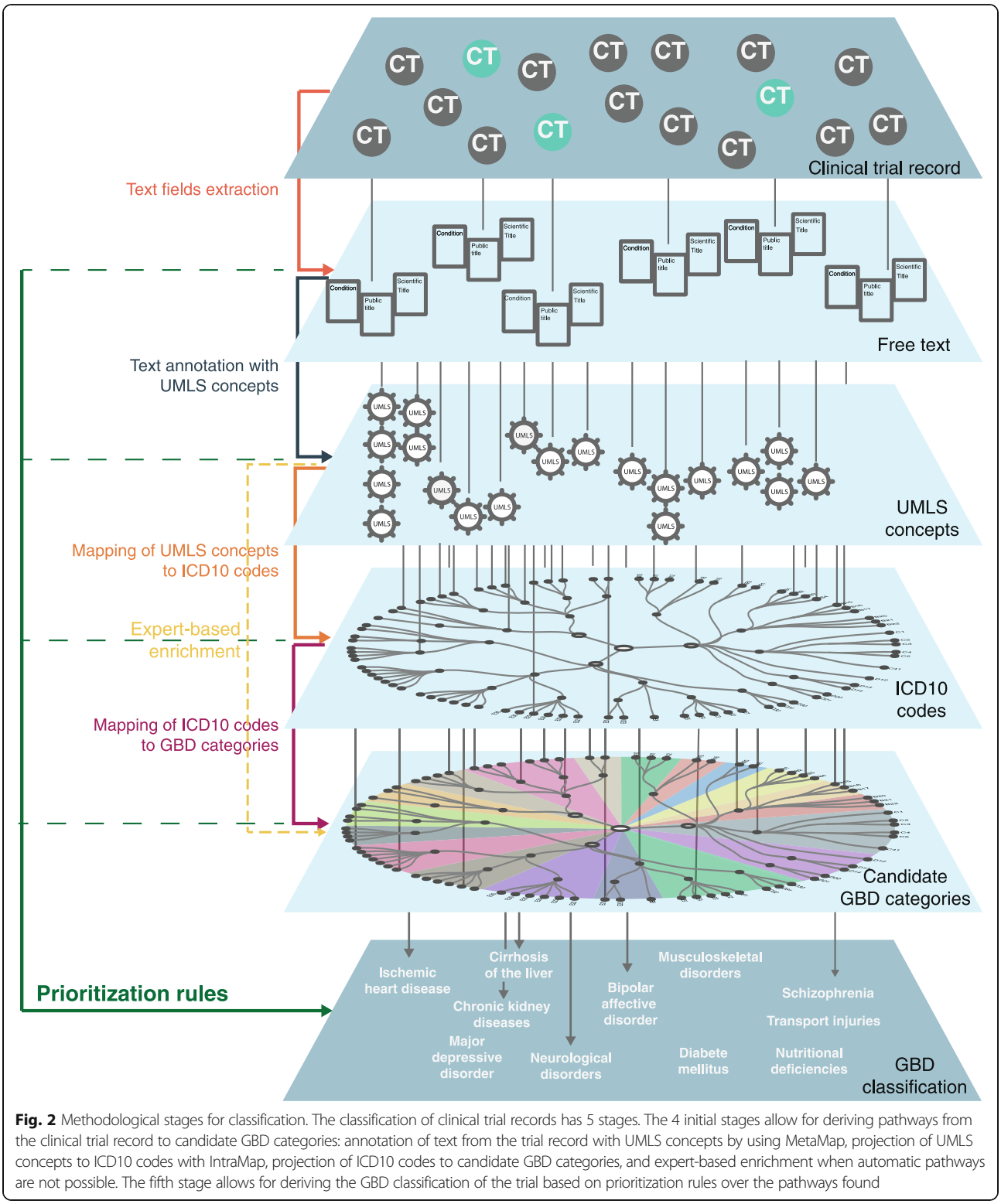


Fig. 2 Methodological stages for classification. The classification of clinical trial records has 5 stages. The 4 initial stages allow for deriving pathways from the clinical trial record to candidate GBD categories: annotation of text from the trial record with UMLS concepts by using MetaMap, projection of UMLS concepts to ICD10 codes with IntraMap, projection of ICD10 codes to candidate GBD categories, and expert-based enrichment when automatic pathways are not possible. The fifth stage allows for deriving the GBD classification of the trial based on prioritization rules over the pathways found

Clinical trial data used in our study

The test set included data from 3 different sources. First, we used data from the Epidemiological Study of Randomized Trials, which selected all primary publications of clinical trials published in December 2012 and indexed in PubMed by November 2013 [35]. Among the 1,351 publications, we identified 519 trials registered at the WHO ICTRP. Two independent physicians manually classified

each publication according to GBD categories. Second, we used data from a WHO study that extracted a random 5 % sample of clinical trials of interventions registered in the ICTRP by August 2012 [36]. One physician classified 2,381 trial records with GBD categories according to Table C3 in [37], with consensus with a second physician in case of ambiguity. We identified 1,271 trial records for which the classification could be unambiguously mapped to our grouping of GBD categories. Finally, we used data from an ongoing study from our team that involves 973 clinical trials of cancer registered at ICTRP before June 2015. One physician classified each record according to GBD categories, with consensus with a second physician in case of doubt. In total we included 2,763 trials in the external test set (Test set of clinical trials in Additional file 1).

Evaluation metrics

We assessed the performance of the classifier by measuring the proportion of trials for which the automatic classification corresponded exactly to the gold standard (exact-matching). We evaluated the exact-matching over trials concerning a unique GBD category, two or more GBD categories and no GBD categories. We computed the overall exact-matching separately for each source of data. We chose the best version of the classifier according to the overall exact-matching proportion. For the best version of the classifier, we evaluated the sensitivities, specificities and positive predictive values for each GBD category. The positive predictive value gives the probability that the trial truly concerned the GBD category identified. If the sensitivity is high for a GBD category, a negative result rules out the category; if the specificity is high, a positive result rules in the category. We derived the positive and negative likelihood ratios (LR+ and LR-); we considered that the classifier reliably identified GBD categories when $LR+ > 10$ (ruling in the disease), and $LR- < 0.1$ (ruling out the disease). We computed the weighted average of the sensitivities and specificities across categories.

Lastly, to put the performance measures of the knowledge-based classifier into context, we compared them to a baseline using a simple method of classification. The baseline did not used the UMLS knowledge source, but a clinical trial record was classified to a GBD category if at least one of the disease names defining that GBD category appeared verbatim in the condition field, the public or scientific titles, separately, or in at least one of these three text fields (for disease names used see Table 1 and Web Table 1 in [28]).

Classification of all clinical trials registered in the WHO ICTRP database

We downloaded all trial records available at the WHO ICTRP by February 1, 2014. We classified all interventional trials initiated between 2006 and 2012 by applying the best-performing version of the classifier. We evaluated the total number of trials mapped to each GBD category.

Research reproducibility

The classifier was coded by using R 3.2.2 (R Development Core Team, Vienna, Austria). The programs of the classifier is publicly available for the research community to use at the open source platform github (github.com/iatal/trial_gbd). It includes all the codes underlying the classification of clinical trial records downloaded at the WHO ICTRP or at clinicaltrials.gov websites towards the 28- or 171-class grouping of GBD categories. In addition, an online interface to optimize manual classification of clinical trials records registered at the WHO ICTRP is available at (http://www.clinicalepidemio.fr/gbd_study_who/). Finally, the classification using the best-performing version of the classifier is provided for all interventional trials registered at WHO ICTRP ($N = 109{,}603$ trials by February 2014, Additional file 2).

Results

Among 2,763 trials in the external test set, 2,328 (84.3 %) concerned a single GBD category, 28 (1.0 %) 2 or more GBD categories, and 407 (14.7 %) residual categories or health conditions not relevant in the GBD 2010 study. Many clinical trials studied "Neoplasms" (958 trials), followed by "Diabetes, urinary diseases and male infertility" (242 trials) and "Cardiovascular and circulatory diseases" (235 trials) (Table 2 and Additional file 1: Table S2).

Process of classification of trials

We describe how the classifier performed on the external test set (see Additional file 1 for the process of classification according to the 171 GBD categories).

Pathways from trial records to candidate GBD categories

MetaMap annotated 2,600/2,763 (94.1 %) of the trials with at least one UMLS concept. The median (Q1, Q3) number of UMLS concepts per trial was 3 (3, 5) when using the WSD server and 4 (3, 6) without the WSD server. The annotation of all trials involved 2,180 different UMLS concepts. IntraMap projected 1,995/2,180 (91.5 %) UMLS concepts. The median (Q1, Q3) number of ICD10 codes per UMLS concept was 2 (1, 2). The UMLS concepts were projected to 1,361 different ICD10 codes and 1,034/1,361 (76.0 %) ICD10 codes were projected to at least one GBD category.

At this stage, 573/2,180 (26.3 %) UMLS concepts could not be projected to a GBD category. The expert-based enrichment allowed for projecting an additional 41/573 (7.2 %) UMLS concepts.

Table 2 Distribution of the external test set ($n = 2,763$ trials) across the 28-class grouping of the GBD cause list, performance of the best performing version of the classifier in the external test set, and projection of all trials in the WHO ICTRP database ($n = 109,603$)

GBD categories	External test set						WHO ICTRP
	No. trials	Sen (%)	Spe (%)	PV+ (%)	LR+	LR-	No. trials (%)
Neoplasms	958	97.4 [96.7-97.7]	97.5 [97.0-97.7]	95.3 [94.4-95.8]	38.2 [28.7-50.8]	0.03 [0.02-0.04]	25,004 (22.8)
Diabetes, urinary diseases and male infertility	242	81.0 [78.0-83.0]	97.4 [97.0-97.7]	75.1 [72.1-77.4]	31.4 [24.5-40.2]	0.20 [0.15-0.25]	9,749 (8.9)
Cardiovascular and circulatory diseases	235	75.7 [72.5-78.1]	97.6 [97.2-97.9]	74.8 [71.6-77.2]	31.9 [24.6-41.4]	0.25 [0.20-0.31]	8,906 (8.1)
Mental and behavioral disorders	143	93.7 [90.5-94.7]	98.7 [98.4-98.9]	80.2 [76.5-82.6]	74.4 [52.9-104.7]	0.06 [0.03-0.12]	7,609 (6.9)
Musculoskeletal disorders	113	88.5 [84.2-90.3]	98.5 [98.2-98.7]	71.4 [67.1-74.6]	58.6 [42.8-80.3]	0.12 [0.07-0.19]	6,112 (5.6)
HIV/AIDS	97	88.7 [83.9-90.4]	99.7 [99.6-99.8]	92.5 [88.0-93.6]	337.7 [160.6-710.0]	0.11 [0.07-0.20]	2,295 (2.1)
Neurological disorders	93	84.9 [79.9-87.3]	98.5 [98.2-98.7]	66.4 [61.6-70.1]	56.7 [41.2-78.0]	0.15 [0.09-0.25]	6,355 (5.8)
Chronic respiratory diseases	81	93.8 [89.0-94.6]	99.4 [99.1-99.5]	81.7 [76.5-84.4]	148.0 [91.9-238.5]	0.06 [0.03-0.15]	4,104 (3.7)
Sense organ diseases	56	92.9 [86.5-93.7]	98.5 [98.2-98.7]	56.5 [51.2-61.3]	62.8 [45.8-86.2]	0.07 [0.03-0.19]	3,461 (3.2)
Injuries	56	16.1 [13.4-23.1]	99.5 [99.3-99.6]	39.1 [31.2-50.1]	31.1 [14.0-68.8]	0.07 [0.03-0.19]	655 (0.6)
Diarrhea, lower respiratory infections, meningitis, and other common infectious diseases	49	81.6 [73.9-84.8]	99.2 [99.0-99.3]	65.6 [58.7-70.6]	105.5 [67.5-164.8]	0.19 [0.10-0.33]	3,200 (2.9)
Maternal disorders	43	39.5 [33.2-47.6]	99.8 [99.7-99.8]	77.3 [64.7-81.7]	215.1 [83.1-556.4]	0.61 [0.48-0.77]	602 (0.5)
Digestive diseases (except cirrhosis)	32	75.0 [65.0-79.7]	99.0 [98.7-99.1]	46.2 [39.7-53.1]	73.2 [48.1-111.3]	0.25 [0.14-0.46]	4,454 (4.1)
Cirrhosis of the liver	23	82.6 [70.2-85.6]	99.4 [99.2-99.5]	52.8 [44.6-60.4]	133.1 [80.0-221.6]	0.17 [0.07-0.43]	1,412 (1.3)
Congenital anomalies	23	95.7 [78.1-99.9]	98.8 [98.5-98.9]	39.3 [33.7-46.3]	77.1 [54.6-108.9]	0.04 [0.01-0.30]	1,947 (1.8)
Skin and subcutaneous diseases	22	81.8 [69.1-85.1]	99.1 [98.9-99.2]	42.9 [36.1-50.8]	93.4 [59.9-145.7]	0.18 [0.08-0.45]	3,652 (3.3)
Hepatitis	17	82.4 [67.5-85.3]	99.9 [99.7-99.9]	77.8 [63.7-82.1]	565.4 [207.2-1542.5]	0.18 [0.06-0.49]	1,082 (1.0)
Tuberculosis	16	87.5 [71.9-88.5]	99.9 [99.8-99.9]	87.5 [71.9-88.5]	1201.8 [297.0-4862.5]	0.13 [0.03-0.46]	306 (0.3)
Nutritional deficiencies	16	68.8 [54.6-75.7]	99.5 [99.2-99.5]	42.3 [34.2-52.4]	125.9 [68.9-230.1]	0.31 [0.15-0.65]	1,226 (1.1)
Hemoglobinopathies and hemolytic anemias	16	62.5 [49.1-71.0]	99.9 [99.7-99.9]	71.4 [55.9-77.8]	429.2 [150.2-1226.9]	0.38 [0.20-0.71]	360 (0.3)
Malaria	14	100.0 [78.5-100.0]	100.0 [99.9-100.0]	93.3 [68.1-99.8]	2749.0 [387.4-19508.4]	-	442 (0.4)
Gynecological diseases	11	81.8 [62.7-84.4]	99.6 [99.4-99.7]	47.4 [37.4-58.3]	225.2 [114.2-443.8]	0.18 [0.05-0.64]	1,536 (1.4)
Neonatal disorders	10	40.0 [29.5-56.0]	99.7 [99.6-99.8]	36.4 [27.3-52.5]	157.3 [54.5-454.1]	0.60 [0.36-1.00]	718 (0.7)
Oral disorders	8	37.5 [27.3-55.8]	99.9 [99.7-99.9]	42.9 [30.3-60.5]	258.3 [68.6-973.0]	0.63 [0.37-1.07]	576 (0.5)
Neglected tropical diseases excluding malaria	7	85.7 [42.1-99.6]	100.0 [99.9-100.0]	100.0 [61.0-100.0]	-	0.14 [0.02-0.88]	361 (0.3)
Leprosy	2	100.0 [15.8-100.0]	100.0 [99.9-100.0]	66.7 [38.7-76.0]	2761.0 [389.1-19593.6]	-	74 (0.1)
Sexually transmitted diseases excluding HIV	1	0.0 [0.0-97.5]	99.8 [99.7-99.8]	0.0 [0.0-43.4]	-	-	187 (0.2)
Sudden infant death syndrome	0	-	100.0 [99.9-100.0]	-	-	-	5 (0.0)
No GBD category	407	53.1 [50.6-55.5]	92.9 [92.3-93.4]	56.4 [53.8-58.9]	7.5 [6.3-8.9]	0.51 [0.46-0.56]	22,450 (20.5)

Sen Sensitivity, *Spe* specificity, *PV+* positive predictive value, *LR+* positive likelihood ratio, *LR-* negative likelihood ratio. The version of the classifier used was: using the Word Server Disambiguation server, the expert-based enrichment, and giving priority to the health condition field

GBD classification

Depending on the version of the classifier, between 594 (21.5 %) and 648 trials (23.5 %) had several candidate GBD categories. With the rule giving priority to the health condition field, the number of trials actually classified with several GBD categories ranged from 177

(6.4 %) to 184 (6.7 %). Without the rule of giving priority to the health condition field, this number ranged from 244 (8.8 %) to 253 (9.2 %). Across all versions of the classifier, the number of trials without GBD classification ranged from 377 (13.6 %) to 414 (15.0 %).

Evaluation of the classifier

Overall performance

The performance of the 8 versions of the classifier is shown in Table 3. The exact-matching proportion was similar for all versions of the classifier. However, the best performance was achieved by using the WSD server, expert-based enrichment, and giving priority to the health condition field (77.8 % of exact-matching). The exact-matching proportion was larger for trials concerning a unique GBD category (82.7 %) and lowest for trials concerning two or more GBD categories (28.6 %). The best version of the classifier was the same for the 171 GBD categories (Additional file 1: Table S3). The performance varied across data sources; overall exact-matching ranged from 66.7 % to 82.2 % (Table 4). When classifying trial records without using the UMLS knowledge source but only using disease names defining the GBD categories, the proportion of clinical trial records from the test set correctly classified to GBD categories was of 51.8 % (Table 3). The knowledge-based classifier had sensitivity and specificity 29.6 % and 5.4 % higher as compared to the baseline not using the UMLS knowledge source.

Performance for each GBD category

The performance of the best-performing classifier to identify the "Neoplasms" category was excellent (Table 2). The positive likelihood ratio was 38.2 [28.7–50.8] and negative likelihood ratio 0.03 [0.02–0.04]; we can be confident that trials classified as studying "Neoplasms" actually concerned that GBD category, and conversely those not classified as studying "Neoplasms" did not concern the category.

The performance of the classifier in identifying the "Diabetes, urinary diseases and male infertility" and "Cardiovascular and circulatory diseases" categories was good. The specificity of these categories was very high, so a mapping of these categories based on the classifier will not overestimate the effort of research in these fields. However, the sensitivity for these categories was 81.0 % [78.0–83.0] and 75.7 % [72.5–78.1], respectively, so a mapping of these categories may underestimate the effort of research in these fields.

The performance of the classifier in identifying the "Mental and behavioral disorders", "Musculoskeletal disorders", "HIV/AIDS" and "Neurological disorders" categories was high. These categories also had high positive likelihood ratios and low negative likelihood ratios. However, the numbers of trials concerning these categories were lower. We cannot conclude on the performance in identifying the remaining GBD categories because of the very low numbers of trials in the external test set (<90 trials per category).

Table 3 Performance of the 8 versions of the classifier, compared to the baseline

	Word Sense Disambiguation	Expert-based enrichment	Priority to health condition field	Exact matching proportion				Weighted average across 28 GBD categories	
				All trials $N = 2,763$	One GBD category $N = 2,328$	Two or more GBD categories $N = 28$	No GBD category $N = 407$	Sensitivity	Specificity
1	Yes	Yes	Yes	77.8	82.7	28.6	53.1	81.9	97.4
2	Yes	Yes	No	77.5	82.5	28.6	52.1	81.8	97.4
3	Yes	No	Yes	76.9	81.4	28.6	54.8	81.0	97.2
4	Yes	No	No	76.9	81.5	28.6	53.8	81.1	97.2
5	No	Yes	Yes	75.6	80.1	28.6	53.1	81.9	97.0
6	No	Yes	No	75.3	79.9	28.6	52.1	81.8	97.0
7	No	No	Yes	74.8	79.0	25.0	54.8	81.0	96.9
8	No	No	No	74.8	79.1	25.0	53.8	81.2	96.9
Baselines		Condition field		48.7	40.5	10.7	98.5	49.3	91.4
		Public title		38.1	27.6	7.1	100.0	38.2	89.6
		Official title		38.0	27.6	7.1	99.3	38.2	89.6
		Three text fields		51.4	43.7	17.9	97.8	52.3	92.0

Exact-matching and weighted averaged sensitivities and specificities for 8 versions of the classifier for the 28 GBD categories, compared to the baseline. Exact-matching corresponds to the proportion (in %) of trials for which the automatic GBD classification is correct. Exact-matching was estimated over all trials ($N = 2,763$), trials concerning a unique GBD category ($N = 2,328$), trials concerning 2 or more GBD categories ($N = 28$), and trials not relevant for the GBD ($N = 407$). The weighted averaged sensitivity and specificity corresponds to the weighted average across GBD categories of the sensitivities and specificities for each GBD category plus the "No GBD" category (in %). The 8 versions correspond to the combinations of the use or not of the Word Sense Disambiguation server during the text annotation, the expert-based enrichment database, and the priority to the health condition field as a prioritization rule. The baseline did not used the UMLS knowledge source, but a clinical trial record was classified to a GBD category if at least one of the disease names defining that GBD category appeared verbatim in the condition field, the public or scientific titles, separately, or in at least one of these three text fields

Table 4 Performance of the classifier per source of data for the 28 GBD categories

Source	Exact-matching (% n/N)				Weighted average across 28 GBD categories	
	All trials	One GBD category	No GBD category	Two or more GBD categories	Sensitivity	Specificity
Emdin 2015	66.7 (346/519)	66.4 (300/452)	68.2 (45/66)	100.0 (1/1)	71.5	96.4
Viergever 2013	82.2 (1045/1271)	85.3 (925/1085)	64.5 (120/186)	0.0 (0/0)	86.6	97.8
On going work	77.9 (758/973)	88.5 (700/791)	32.9 (51/155)	25.9 (7/27)	81.3	97.2

Exact-matching and weighted averaged sensitivities and specificities for the classifier to the 28 GBD categories for each source of data. The version of the classifier used was: using the Word Sense Disambiguation server, the expert-based enrichment database and the priority to the health condition field. Exact-matching corresponds to the proportion (in %) of trials for which the automatic GBD classification is correct. Exact-matching was estimated over all trials, trials concerning a unique GBD category, trials concerning 2 or more GBD categories, and trials not relevant for the GBD. The weighted averaged sensitivity and specificity corresponds to the weighted average across GBD categories of the sensitivities and specificities for each GBD category plus the "No GBD" category (in %)

The lowest performance was for the "Injuries" and "Maternal disorders" categories. The "Injuries" category was studied by 56 clinical trials and the sensitivity was low (16.1 % [13.4–23.1]), so a high proportion of trials concerning injuries may not be detected by the classifier. Similarly, the sensitivity for "Maternal disorders" was 39.5 % [33.2–47.6], so the classifier may not detect correctly these trials.

Overall, our classifier identified 407 trials not concerning any GBD category. The sensitivity was low (53.1 % [50.6–55.5]), so half of the trials not concerning any relevant GBD category were actually classified by using GBD categories. The positive predictive value was also low (56.4 % [53.8–58.9]), so half of trials classified as "No GBD" category actually concerned a relevant GBD category.

When classifying trial records without using the UMLS knowledge source but only using disease names defining the GBD categories, the sensitivities were extremely low as compared to those of the knowledge-based classifier for all GBD categories but for semantically simple GBD categories: "HIV/AIDS", "Hepatitis", "Tuberculosis", "Malaria" and "Leprosy" (Additional file 1: Table S4).

Across the 171 GBD categories, the performance was appropriate for the GBD categories most represented in the test set. However, for a high proportion of GBD categories, the number of trials in the test set was not sufficient to conclude on the performance of the classifier in identifying them (Additional file 1: Table S2).

Classification of all trials registered at the WHO ICTRP

In total, 109,603 interventional trials were classified by using the best-performing version of the classifier (Additional file 2). The number of trials per GBD category is shown in Table 2. The "Neoplasms" category was the most used for classifying clinical trials (22.8 %), followed by "Diabetes, urinary diseases and male infertility" (8.9 %) and "Cardiovascular and circulatory diseases" (8.1 %). In total, 20.5 % of trials could not be classified by a relevant GBD category.

Discussion

We developed a knowledge-based classifier to automatically map clinical trial records to a 28- and 171-class grouping of the taxonomy of diseases and injuries from the GBD 2010 study. In a validation study, the performance of the classifier was very good for trials of major groups of diseases, including cancer, diabetes and cardiovascular diseases. Our classifier allowed for classifying all trials registered at the WHO ICTRP.

Comparison to related work

Several studies have previously evaluated the gap between health research and health needs [35, 36, 38–43]. However, in these studies, the classification of health R&D activities was always conducted manually. Manual classification inherently restricted those studies to limited sample sizes, specific medical areas, regions or types of studies. In addition, these studies were not updated. Our automatic classifier can allow for large-scale mapping of all clinical trials registered at the WHO ICTRP (more than 300,000 trials) about all diseases and all regions and the evolution over time.

Previous work used NLP methods to conduct curation of the eligibility criteria field from clinical trial records to improve the retrieval of relevant clinical trials for patients [14–26] In contrast to previous work, we conducted NLP analyses of the condition field and the public and scientific titles from clinical trial records to achieve a different objective, the classification of the condition studied in clinical trials according to a standardized taxonomy of diseases and injuries. Previous studies of automatic indexing used health topics in medical research. The Medical Text Indexer (MTI), developed at the NLM, is used for providing indexing recommendations for data sources such as MEDLINE, PubMed and ClinicalTrials.gov. [29, 44] MTI produces Medical Subject Headings (MeSH) recommendations by combining a statistical method and a natural language processing method based on MetaMap and the Restrict-to-MeSH implemented in IntraMap. This algorithm was shown to be successful for automatically assigning ICD9

codes to radiology reports [45]. To our knowledge no previous work has used the knowledge-based sequence MetaMap - IntraMap to assign GBD categories to clinical trials. The Aggregate Analysis of ClinicalTrials.gov project used indexing with MeSH terms to group trials by medical specialty [30]. However, the medical specialties cannot be connected to the burden of disease. Evans et al. projected all articles indexed in MEDLINE to GBD categories based on indexing publications with MeSH terms from the MTI [46]. The authors linked MeSH terms to ICD9 codes by using the UMLS database. In our work, we directly targeted a classification of texts from trial records by using ICD10 codes because GBD categories are defined with that terminology. Instead of using MeSH terms as an intermediate for projection, which may increase the error rate, we chose to develop our method for classifying automatically health topics according to GBD categories based on ICD10. In addition, we mapped ICD10 codes to GBD categories because the GBD 2010 study provides a burden estimate for each GBD category, and not for each ICD10 code. Moreover, these previous studies focused on the curation of health topics of clinical trials records registered at ClinicalTrials.gov, thereby excluding 31.2 % of trials in the WHO ICTRP [9]. Our method of classification was based on the processing of the condition field and public and scientific titles only, which are required by the WHO ICTRP [47]. Thus, our method can be transposed to any of the 16 clinical trial repositories included in the WHO ICTRP up to date, including clinicaltrials.gov. All these sources of registries are fundamental to conduct a worldwide mapping of registered clinical trials to be compared to global health needs. In addition, in our github repository we include codes to analyze clinical trial records downloaded from WHO ICTRP and clinicaltrials.gov websites.

Strength of the knowledge-based classifier

Our classifier has several strengths. First, it allows for developing a reliable region-specific mapping of trials, especially in fields such as cancer. Such a mapping can be compared to the region-specific burden of the corresponding diseases. Considering that the classification is imperfect, a region-specific mapping of research topics other than cancer with the classifier should take into account the possible misclassification. Second, the classifier of clinical trials we developed may be used for conducting semi- and fully-automatic classification recommendations. Machine learning methods based on the characteristics of trial records and on the pathways drawn between trials and GBD categories may allow for identifying trials for which the classifier does not show a confident classification. These trials may be considered for manual revision. Because the WHO ICTRP database is large and constantly growing, manual revisions may be expensive. Crowd-sourcing based on the interface for the manual classification we developed could be scaled up to divide the effort needed for revision. In addition, trial registries such as ClinicalTrials.gov could include the GBD classification as a mandatory field in trial records. The classifier we developed could provide an automatic recommendation for classification of newly registered trials by the GBD categories, thus reducing the burden of registration. Another strength of the classifier is that it is based on the UMLS Knowledge Source, a metathesaurus widely used for analyzing biomedical text, which increases the portability and reproducibility of the classification. The classification method development did not rely on data in the test set. Other approaches such as statistical methods of classification (e.g. support vector machines) may be used to address our objective. However, our knowledge-based classifier may be more resilient to the evolution of clinical trial records. Every year, about 20,000 new clinical trials are registered at WHO ICTRP [9]. Statistical methods of classification would need new training data to perform classification out of the rule space of a training dataset. Another strength is that our knowledge-based classifier allows understanding the process of classification of trial records (Fig. 1), as compared to statistical classifiers. For a public health project, it is of great value understanding the process of data curation [48, 49]. In addition, the approach is generalizable to other sources such as grants, articles, and systematic reviews.

Performance of the knowledge-based classifier

The evaluation of our classifier on a gold standard external test set yielded an overall performance of 81.9 % sensitivity and 97.6 % specificity. Overall, 77.8 % of trial records from the external test set were correctly classified towards a 28-class grouping of the GBD cause list. Pradhan et al. evaluated the performance of 17 systems to normalize disorder mentions in biomedical text using a standardized ontology, the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) [50, 51]. In that study, the best performing system correctly normalized 58.9 % of disorder mentions. It is hard to compare this performance to the performance of our classifier, as the input space (biomedical text vs clinical trial records) and the target spaces (SNOMED CT vs GBD categories) differ. However, we consider that the performance of the classifier was satisfactory for trials concerning majors groups of diseases as cancer, diabetes and cardiovascular diseases. In particular, we can be confident on the mapping provided by the classifier of clinical trials concerning cancer. In addition, the classifier may not overestimate the effort of research in diabetes and cardiovascular diseases. Our classifier performed differently

across data sources. This may be explained because the different these data sources can not be considered as random samples of clinical trials. However, we could identify some GBD categories for which the overall performance of the classifier was excellent.

Limitations

Our work has several limitations. First, the quality of the mapping of health research depends on the quality of the registration of clinical trials. Trial registration remains of low quality, but endorsements from WHO are attempting to improve the registration system [7, 47]. In addition, the misclassification of diseases may be correlated to trial location. For instance, our classifier only supports English language, as MetaMap identifies UMLS concepts in biomedical text written in English. This may increase the misclassification in non-English speaking countries. However, according to the International Standards for Clinical Trial Registries from the WHO, all items of trial records included in the WHO ICTRP (including the condition field and the public and scientific titles) must be available in English language [47]. Similarly, compliance to registration of clinical trials may vary across regions. However, it is unlikely that compliance on registration vary across diseases. Therefore, in regions with low compliance of registration, a lower number of clinical trials concerning a disease as compared to other diseases may effectively correspond to a gap of health research. Second, our classifier may poorly identify some categories. For instance, the sensitivity for the "Injuries" category, accounting for 10.7 % of the global burden in 2010, was low [27]. In our test set, clinical trials concerning injuries mainly studied the adverse effects of medical treatments (35/56). In these trials, the classifier is more likely to identify the health condition targeted by those medical treatments rather than considering that the clinical trials studied the adverse effects of the treatments. Thus, this misclassification may not be considered an error in the mapping because trials studying the adverse effects of the treatment used for a certain condition will be conducted in countries where that particular condition is a burden. Third, the classifier may poorly identify trials not concerning any relevant GBD category. For the classifier to identify a "No GBD" category trial, it needs to be unable to project the trial to any GBD category. However, any UMLS concept recognized in the trial record projected to a GBD category will lead to a classification of the trial. The suppression of noise candidate GBD categories by using the prioritization rules do not allow for suppressing all the candidate GBD categories but rather only choosing the most accurate classification among the candidates. However, the specificities of each of the 28 GBD categories were generally high, so the number of "No GBD" category trials wrongly classified remained low per GBD category.

In our 28-class grouping of diseases and injuries we excluded two residual categories from the GBD cause list, "Other infectious diseases" and "Other endocrine, nutritional, blood, and immune disorders", accounting for 1.2 % of the global burden in 2010. These residual categories are difficult to cover as they are defined using sets of ICD10 to complement the major diseases groups, and are thus particularly large and complex. We decided no to take into account these categories because these coverings may add much complexity to the classification tasks with very small benefits in terms of global mapping of clinical research. Actually, we considered that these categories would not be informative for the purposes of developing a global mapping of registered clinical trials across diseases to be compared to health needs. Finally, in our study, we considered the particular taxonomy of the US Institute for Health Metrics and Evaluation for the GBD 2010 Study. This taxonomy may not be perfectly suitable for conducting a mapping of health R&D. For instance, health conditions that may be considered public health priorities in some regions, such as obesity, venous thromboembolism or heart failure, are part of the residual categories. However, the GBD study is a worldwide effort to estimate the evolution of the burden of all diseases in all countries in the world. It provides a consensual taxonomy of diseases for use in comparing the research effort to the burden of diseases.

Conclusion

Herein, we presented a knowledge-based classifier to map the health conditions studied in registered clinical trials according to the taxonomy of diseases and injuries from the Global Burden of Diseases 2010 study. The overall performance of the classifier was 81.9 % sensitivity and 97.6 % specificity. We applied it to the entire WHO ICTRP database, which characterizes the global burden of disease addressed by the 109,603 clinical trials in the database. This classifier allows for comparing the research effort to the disease burden on a large scale for all diseases and all regions and studying the evolution over time.

Endnotes

[1]Body Mass Index

Additional files

Additional file 1: Includes details on the implementation of MetaMap and IntraMap, prioritization rules, the test set of clinical trials and the classification of the external test set according to the 171 GBD categories. **Dataset S1:** Expert-based enrichment database for the classification according to the 28 GBD categories. Manual classification of 503 UMLS

concepts that could not be mapped to any of the 28 GBD categories. **Dataset S2:** Expert-based enrichment database for the classification according to the 171 GBD categories. Manual classification of 655 UMLS concepts that could not be mapped to any of the 171 GBD categories, among which 108 could be projected to candidate GBD categories. **Table S1:** Excluded residual GBD categories for the grouping of the GBD cause list in 171 GBD categories. A grouping of 193 GBD categories was defined during the GBD 2010 study to inform policy makers about the main health problems per country. From these 193 GBD categories, we excluded the 22 residual categories listed in the Table. We developed a classifier for the remaining 171 GBD categories. Among these residual categories, the unique excluded categories in the grouping of 28 GBD categories were "Other infectious diseases" and "Other endocrine, nutritional, blood, and immune disorders". **Table S2:** Per-category evaluation of performance of the classifier for the 171 GBD categories plus the "No GBD" category. Number of trials per GBD category from the test set of 2,763 clinical trials. Sensitivities, specificities (in %) and likelihood ratios for each of the 171 GBD categories plus the "No GBD" category for the classifier using the Word Sense Disambiguation server, the expert-based enrichment database and the priority to the health condition field. **Table S3:** Performance of the 8 versions of the classifier for the 171 GBD categories. Exact-matching and weighted averaged sensitivities and specificities for 8 versions of the classifier for the 171 GBD categories. Exact-matching corresponds to the proportion (in %) of trials for which the automatic GBD classification is correct. Exact-matching was estimated over all trials ($N = 2,763$), trials concerning a unique GBD category ($N = 2,092$), trials concerning 2 or more GBD categories ($N = 187$), and trials not relevant for the GBD ($N = 484$). The weighted averaged sensitivity and specificity corresponds to the weighted average across GBD categories of the sensitivities and specificities for each GBD category plus the "No GBD" category (in %). The 8 versions correspond to the combinations of the use or not of the Word Sense Disambiguation server during the text annotation, the expert-based enrichment database, and the priority to the health condition field as a prioritization rule. **Table S4:** Per-category evaluation of the performance of the baseline for the 28 GBD categories plus the "No GBD" category. Number of trials per GBD category from the test set of 2,763 clinical trials. Sensitivities and specificities (in %) of the 28 GBD categories plus the "No GBD" category for the classification of clinical trial records towards GBD categories without using the UMLS knowledge source but based on the recognition in free text of the names of diseases defining in each GBD category only. For the baseline a clinical trial records was classified with a GBD category if at least one of the 291 disease names from the GBD cause list defining that GBD category appeared verbatim in the condition field, the public or scientific titles, separately, or in at least one of these three text fields. (DOCX 84 kb)

Additional file 2: Classification towards the 28-class grouping of GBD categories of all interventional trials registered at WHO ICTRP before February 2014. Classification of $N = 109,603$ clinical trials using the best-performing version of the classifier (using the Word Sense Disambiguation server, the expert-based enrichment database and the priority to the health condition field. (XLSX 12447 kb)

Abbreviations

GBD: Global Burden of Diseases; ICD10: International classification of diseases 10th version; ICTRP: International clinical trials registry platform; LR+/-: Positive and negative likelihood ratios; MeSH: Medical subject headings; MTI: Medical text indexer; NLM: National library of medicine; NLP: Natural language processing; R&D: Research and development; SNOMED CT: Systematized nomenclature of medicine—clinical terms; UMLS: Unified medical language system; WHO: World Health Organization; WSD: Word sense disambiguation

Acknowledgements

We thank Elise Diard for help with graphics and the interface for manual classification of trials. We thank Dr Bodenreider for running IntraMap at the NLM servers for a large amount of UMLS concepts. We thank Pr Altman, Dr Emdin, and Dr Odutayo and Dr Viergever for sharing their data. Finally, we thank Laura Smales for language revision of the manuscript.

Funding

This work did not receive any specific grant.

Availability of data and materials

The code of the classifier is available at github.com/iatal/trial_gbd. The expert-based enrichment databases and the classification of all interventional trials registered at the WHO ICTRP before February 2014 are available in the additional files. The test set is available upon request.

Authors contribution

IA, LT, RP, AN and PR conceived and designed the study. IA acquired and analyzed the data. JDZ conducted manual revisions and classification of data. IA, RP and LT interpreted data. The initial manuscript was drafted by IA. IA, RP, AN, JDZ and LT contributed to subsequent revisions. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable. The study only used data concerning the design and settings of clinical trials retrieved from publicly accessible clinical trial registries.

Author details

[1]Centre d'Épidémiologie Clinique, Hôpital Hôtel-Dieu, Paris, France. [2]INSERM U1153, Paris, France. [3]Université Paris Descartes, Paris, France. [4]LIMSI, CNRS UPR 3251, Université Paris-Saclay, Orsay, France. [5]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA.

Received: 24 February 2016 Accepted: 8 September 2016
Published online: 22 September 2016

References

1. Adam T, Røttingen J-A, Kieny M-P. Informing the establishment of the WHO Global Observatory on Health Research and Development: a call for papers. Heal Res Policy Syst. 2015;13:9.
2. Røttingen JA, Regmi S, Eide M, Young AJ, Viergever RF, Ardal C, Guzman J, Edwards D, Matlin SA, Terry RF. Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? Lancet. 2013;382:1286–307.
3. Terry RF, Salm JF, Nannei C, Dye C. Creating a global observatory for health R&D. Science. 2014;345:1302–4.
4. Ahmad N, Boutron I, Dechartres A, Durieux P, Ravaud P. Geographical representativeness of published and ongoing randomized controlled trials. the example of: Tobacco consumption and HIV infection. PLoS ONE. 2011;6:e16878.
5. Global Observatory on Health R&D [http://www.who.int/research-observatory/en/]
6. Atal I, Trinquart L, Porcher R, Ravaud P. Differential globalization of industry- and non-industry–sponsored clinical trials. PLoS ONE. 2015;10:e0145122.
7. Viergever RF, Karam G, Reis A, Ghersi D. The quality of registration of clinical trials: still a problem. PLoS ONE. 2014;9:e84727.
8. International Clinical Trials Registry Platform [http://www.who.int/ictrp/glossary/en/]. Accessed 1 Feb2014.
9. Viergever RF, Li K. Trends in global clinical trial registration: an analysis of numbers of registered clinical trials in different parts of the world from 2004 to 2013. BMJ Open. 2015;5:e008932.
10. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42(5):760–72.
11. Névéol A, Zweigenbaum P. Clinical natural language processing in 2014: foundational methods supporting efficient healthcare. Yearb Med Inf. 2014;2015(10):194–8.
12. McCray AT, Tse T. Understanding search failures in consumer health information systems. AMIA Annu Symp Proc 2003:430–4.
13. ClinicalTrials.gov [http://clinicaltrials.gov/]
14. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: A literature review. J Biomed Inform. 2010;43(3):451–67.
15. Besana P, Cuggia M, Zekri O, Bourde A, Burgun A. Using Semantic Web technologies for Clinical Trial Recruitment. In: The Semantic Web – ISWC 2010. Berlin Heidelberg: Springer; 2010. p. 34–49.

16. Milian K, Bucur A, Van Harmelen F. Building a library of eligibility criteria to support design of clinical trials. Knowl Eng Knowl Manag Lect Notes Comput Sci. 2012;7603:327–36.
17. Huang Z, ten Teije A, van Harmelen F: SemanticCT. A Semantically-Enabled System for Clinical Trials. Process Support and Knowledge Representation in Health Care. Murcia: Springer International Publishing; 2013.
18. Luo Z, Miotto R, Weng C. A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. J Biomed Inform. 2013;46:33–9.
19. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. J Biomed Inform. 2014;52:112–20.
20. He Z, Carini S, Hao T, Sim I, Weng C. A method for analyzing commonalities in clinical trial target populations. AMIA Annu Symp Proc. 2014;2014:1777–86.
21. He Z, Carini S, Sim I, Weng C. Visual aggregate analysis of eligibility features of clinical trials. J Biomed Inf. 2015;54:241–55.
22. Korkontzelos I, Mu T, Ananiadou S. ASCOT: a text mining-based web-service for efficient search and assisted creation of clinical trials. BMC Med Inform Decis Mak. 2012;12 Suppl 1:S3.
23. Boland MR, Miotto R, Gao J, Weng C. Feasibility of feature-based indexing, clustering, and search of clinical trials. Methods Inf Med. 2013;52:382–94.
24. Boland MR, Weng C. A method for probing disease relatedness using common clinical eligibility criteria. Stud Health Technol Inform. 2013;192:481–5.
25. Miotto R, Jiang S, Weng C. ETACTS: A method for dynamically filtering clinical trial search results. J Biomed Inform. 2013;46:1060–7.
26. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. J Am Med Inform Assoc. 2011;18:i116–24.
27. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY, Alvarado M, Anderson HR, Anderson LM, Andrews KG, Atkinson C, Baddour LM, Barker-Collo S, Bartels DH, Bell ML, Benjamin EJ, Bennett D, Bhalla K, Bikbov B, Bin AA, Birbeck G, Blyth F, Bolliger I, Boufous S, Bucello C, Burch M, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380:2095–128.
28. Murray CJL, Ezzati M, Flaxman AD, Lim S, Lozano R, Michaud C, Naghavi M, Salomon JA, Shibuya K, Vos T, Wikler D, Lopez AD. GBD 2010: Design, definitions, and metrics. Lancet. 2012;380:2063–6.
29. Ide NC, Loane RF, Demner-Fushman D. Essie: A concept-based search engine for structured biomedical text. J Am Med Informatics Assoc. 2007;14:253–63.
30. Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, McCourt BJ, Pietrobon R. The database for aggregate analysis of clinicaltrials.gov (AACT) and subsequent regrouping by clinical specialty. PLoS ONE. 2012;7:e33677.
31. International statistical classification of diseases and related health problems. -10th revision [http://www.who.int/classifications/icd/en]
32. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267–70.
33. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Informatics Assoc. 2010;17:229–36.
34. Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc 2005:266–270.
35. Emdin CA, Odutayo A, Hsiao AJ, Shakir M, Hopewell S, Rahimi K, Altman DG. Association between randomised trial evidence and global burden of disease : cross sectional study (Epidemiological Study of Randomized Trials — ESORT). BMJ. 2015;350:h117.
36. Viergever RF, Terry RF, Karam G. Use of data from registered clinical trials to identify gaps in health research and development. Bull World Heal Organ. 2013;91(6):416–425C.
37. World Health Organization. The Global Burden of Disease: 2004 Update. 2008
38. Bourgeois FT, Olson KL, Ioannidis JP A, Mandl KD. Association between pediatric clinical trials and global burden of disease. Pediatrics. 2014;133:78–87.
39. Isaakidis P, Swingler GH, Pienaar E, Volmink J, Ioannidis JP. Relation between burden of disease and randomised evidence in sub-Saharan Africa: survey of research. BMJ. 2002;324:702.
40. Swingler GH, Volmink J, Ioannidis JP. Number of published systematic reviews and global burden of disease: database analysis. BMJ. 2003;327:1083–4.
41. Karimkhani C, Boyers LN, Prescott L, Welch V, Delamere FM, Nasser M, Zaveri A, Hay RJ, Vos T, Murray CJL, Margolis DJ, Hilton J, Maclehose H, Williams HC, Dellavalle RP. Global burden of skin disease as reflected in cochrane database of systematic reviews. JAMA Dermatol. 2014;150:945–51.
42. Perel P, Miranda JJ, Ortiz Z, Casas JP. Relation between the global burden of disease and randomized clinical trials conducted in latin America published in the five leading medical journals. PLoS ONE. 2008;3:e1696.
43. Cottingham MD, Kalbaugh CA, Fisher JA. Tracking the pharmaceutical pipeline: clinical trials and global disease burden. Clin Transl Sci. 2014;7:297–9.
44. Mork JG, Yepes AJJ, Aronson AR. The NLM Medical Text Indexer System for Indexing Biomedical Literature. In BioASQ@ CLEF; 2013
45. Aronson AR, Bodenreider O, Demner-Fushman D, Fung KW, Lee VK, Mork JG, Névéol A, Peters L, Rogers WJ. From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches. Proc Work BioNLP 2007 Biol Transl Clin Lang Process 2007;105–12.
46. Evans JA, Shim J-M, Ioannidis JP. Attention to local health burden and the global disparity of health research. PLoS ONE. 2014;9:e90147.
47. World Health Organization. International Standards for Clinical Trial Registries. 2012.
48. Ruiz ME, Aronson A. User-Centered Evaluation of the Medical Text Indexing (MTI) System - Technical Report - US National Library of Medicine. 2007
49. Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Informatics Assoc. 2015;22:1009–19.
50. Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, Suominen H, Chapman WW, Savova G. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. J Am Med Inform Assoc. 2015;22:143–54.
51. NIH-NLM. SNOMED Clinical Terms® (SNOMED CT®). NIH-US National LIbrary of Medicine 2015.

Chapitre 4

Discussion

4.1 Résumé des résultats

Nous avons mis en place une analyse systémique pour cartographier la recherche clinique, c'est-à-dire pour savoir où est faite la recherche dans le monde, quelles maladies sont étudiées, et quels sont les principaux acteurs qui y participent. L'objectif des cartographies est d'élucider les forces mouvant les programmes de recherche et d'identifier des lacunes dans la production de connaissances sur l'effet des traitements en médecine.

De nombreuses bases de données permettent de tracer les investissements, les processus et les résultats de la recherche clinique dans le monde. Or, la plupart de ces bases ne permettent pas des analyses à échelle globale. Une base de données publiquement accessible et prête à l'utilisation pour cartographier les processus de la recherche clinique dans le monde est l'International Clinical Trials Registry Platform (ICTRP), créé par l'OMS pour donner un point d'accès unique aux registres d'essais cliniques de plusieurs registres nationaux, régionaux et internationaux. À partir des registres d'essais cliniques nous avons créé deux cartographies globales de la recherche clinique.

Notre première cartographie, présentée en Section 2.2 a analysé, pour 7 régions

dans le monde, l'alignement local entre la conduite d'essais cliniques randomisés et le fardeau pour 27 groupes de maladies. Pour chaque région, nous avons cherché à identifier des lacunes dans l'effort de recherche local relativement au fardeau des maladies. Pour cela, nous avons pre-spécifié un critère pour juger si une maladie était négligée par la recherche relativement au fardeau qu'elle impose : si sa part dans l'effort de recherche local est moins de la moitié de sa part dans le fardeau de la région. Nous avons cartographié 117,180 essais cliniques randomisés initiés entre 2006 et 2015, et 2.2 milliards d'années de vie corrigées à l'incapacité (DALY) en 2005 à partir de l'étude Global Burden of Diseases (GBD) 2010 (Lozano et al., 2012). Dans les pays à revenu élevé,130.9 essais randomisés par million de DALY ont été conduits contre 6.9 dans les autres régions. Les essais faits dans les pays à revenu élevé étaient bien alignés avec le fardeau des maladies. Pour toutes les autres régions nous avons identifié des lacunes dans l'effort de recherche relativement au fardeau. En particulier, en Afrique Subsaharienne, même si des causes majeures de fardeau comme le VIH et le paludisme ont reçu un effort de recherche à la hauteur du fardeau qu'elles imposent, d'autres causes de morbi-mortalité majeures ont été négligées par l'effort de recherche, notamment les maladies infectieuses communes et les pathologies du nouveau né. Ce travail nous a aussi permis d'identifier des lacunes de recherche non documentées ailleurs, en particulier en Asie du Sud où les maladies infectieuses et les pathologies du nouveau né sont négligées par la recherche locale, et en Europe de l'Est et l'Asie Centrale où les maladies cardiovasculaires sont sous-étudiées.

Notre deuxième cartographie, présentée en Section 2.3, a montré l'influence des promoteurs industriels et non-industriels (i.e. académiques, organismes, fondations) dans la globalisation des essais cliniques. Nous avons montré que 30% des essais cliniques à promoteurs industriels sont internationaux, i.e. recrutant des patients simultanément dans plusieurs pays, alors que seulement 3% des essais à promoteur non-industriel le sont. Ceci met en évidence les différentes capacités de ces deux acteurs à s'appuyer sur des réseaux de collaboration entre pays pour

globaliser leur recherche. Par ailleurs, nous avons étudié les réseaux de collaboration entre pays dérivés de la recherche internationale : dans ces réseaux, chaque nœud est un pays, et une arête entre deux pays correspond au nombre d'essais internationaux dans lesquels ces deux pays participent simultanément. À partir de ces réseaux, nous avons cherché des groupements de pays participant ensemble de façon surreprésentée aux mêmes essais internationaux. Ces analyses ont été faites séparément pour la recherche à promoteur industriel et non-industriel. Nous avons montré que les deux types de promoteurs s'appuient sur des réseaux de pays différents. En particulier, les pays d'Europe de l'Est participent dans leur ensemble de façon surreprésentée dans les essais cliniques internationaux à promoteur industriel, alors qu'ils ne sont pas significativement séparés du reste de l'Europe dans la recherche non-industrielle.

Pour mettre en place notre première cartographie, il a été nécessaire de rendre interopérables les registres d'essais cliniques présents dans l'ICTRP avec les estimations du fardeau données par l'étude GBD 2010. En fait, dans les registres d'essais cliniques la maladie étudiée n'est pas renseignée avec une nomenclature standardisée, mais en texte libre. Le troisième travail, présenté en Section 3.2, a consisté dans le développement et la validation d'un algorithme de classification automatique des essais cliniques enregistrés vers les catégories de morbi-mortalité utilisées dans l'étude GBD 2010 pour estimer le fardeau. L'algorithme, reposant sur le système de connaissances UMLS® et des méthodes de traitement automatique du langage, a classifié correctement 78% d'un ensemble test de 2,763 essais cliniques manuellement classifiés par des experts. Les performances ont été variables d'un groupe de maladie à l'autre, mais particulièrement bonnes pour le cancer (sensibilité 97.4%, spécificité 97.5%).

4.2 Limites

Plusieurs limites de ce travail de recherche méritent d'être traitées en détail. Tout d'abord, nous avons énoncé que les principaux objectifs des cartographies de la recherche clinique sont d'une part identifier des lacunes dans la production des connaissances sur l'effet des traitements, et d'autre part identifier des priorités pour diriger le programme de recherche. Or, il n'existe pas de méthode communément admise pour atteindre ces objectifs. En particulier, dans le travail présenté en Section 2.2, nous avons choisi de définir une lacune dans l'effort de recherche sur une maladie donnée lorsque la part de cette maladie dans le nombre d'essais cliniques représente moins de la moitié de la part de cette maladie dans le fardeau. Or, ce n'est pas forcément la recherche faite sur les maladies avec le plus grand fardeau celle qui portera le plus de gains en termes d'amélioration de santé (Prasad, 2016). En particulier, le gain en nombre de DALY diminués par investissement en recherche peut varier considérablement entre les maladies. Par exemple, le gain de fardeau diminué par investissement en recherche peut être très élevé pour une maladie rare si la découverte d'un remède pour celle-ci nécessite de moindres efforts. Nonobstant, dans notre travail nous avons cherché à identifier des différences conséquentes entre l'effort de recherche et le fardeau des maladies relatif au fardeau local de chaque maladie. C'est-à-dire, nous n'avons ni cherché un alignement parfait entre recherche et fardeau, ni identifié des lacunes uniquement pour les maladies à grand fardeau.

Nos travaux sont aussi limités par l'utilisation des registres d'essais cliniques pour cartographier la recherche. En effet, ces registres de ne reflètent pas l'ensemble de la recherche clinique, puisque pas toute la recherche clinique est enregistrée de façon prospective (Section 1.2.4 page 24). D'autre part, il peut exister des différences entre ce qui est annoncé dans le registre et ce qui est finalement fait, dues par exemple à des possibles changements dans le protocole qui n'auraient pas donné lieu à des mises à jour du registre (Zarin and Tse, 2013). Pour faire des cartographies, il est particulièrement préoccupant lorsque les localisations des

essais ne sont pas correctement rapportées (Patrone, 2010). Par ailleurs, la complétude et la qualité des données entrées dans chaque registre reste à améliorer (Zarin et al., 2007; Viergever et al., 2014). Des efforts de l'OMS et le NIH sont d'ailleurs en cours pour améliorer la qualité de l'enregistrement (Zarin et al., 2015; World Health Organization, 2012).

Finalement, nos travaux sont limités par l'échelle, par moments trop grande, de ceux-ci, alors que des analyses plus fines pourraient être plus informatives pour les décideurs des programmes de recherche. En effet, dans les cartographies présentées dans les Sections 2.2 et 2.3 nous avons respectivement utilisé tous les essais randomisés enregistrés dans l'ICTRP initiés entre 2006 et 2015, et tous les essais enregistrés dans l'ICTRP et initiés entre 2006 et 2013. L'hétérogénéité des formes de recherche qui ont été incluses dans ces cartographies, et les différents apports de ces différentes formes dans la production de connaissances sur l'effet des traitements justifie le besoin de faire des cartographies plus spécifiques. Par exemple, l'effet des méthodes de prévention et des traitements non-pharmacologiques sont plus susceptible à être dépendant de la localisation et du contexte dans lequel il est évalué (Rothwell, 2005). Aussi, des essais cliniques de phase IV peuvent viser, non pas à l'amélioration des connaissances sur l'effet d'un traitement, mais à changer les habitudes de prescription des médecins (Sox and Rennie, 2008). Des contextes médicaux tels que la pédiatrie sont particulièrement touchés par le besoin de cartographies spécifiques en vue de la sous-représentation de ces populations dans la recherche clinique (Bourgeois et al., 2014; Joseph et al., 2016). Or, la construction à grande échelle de cartographies plus spécifiques est limitée par le manque d'utilisation de nomenclatures standardisées dans les registres d'essais cliniques. Le travail supplémentaire pour classifier plus précisément la recherche s'avère être conséquent. Par exemple, alors que la première cartographie qu'on a faite (chronologiquement) (Section 2.3) amalgamait les essais de toutes les maladies, nous avons eu besoin de développer un algorithme de classification (Section 3.2) pour créer des cartographies par maladie (Section 2.2).

4.3 Implications de nos résultats

Les résultats de notre première cartographie mettent en avant les grandes problématiques médicales négligées par la recherche dans les régions à revenu faible, notamment les maladies infectieuses communes et les pathologies du nouveau né (Section 2.2). Dans ces régions, les ressources locales en recherche sont très limitées (Røttingen et al., 2013). En Afrique Subsaharienne par exemple, les essais cliniques sont majoritairement financés et conduits par des organismes venant de pays à revenu élevé (Ndounga Diakou et al., 2017). Notre travail montre que des maladies recevant une grande attention internationale comme le VIH et le paludisme se voient adjugées localement un effort de recherche à la hauteur du fardeau qu'elles imposent. Or, d'autres grandes causes de fardeau dans la région comme les maladies diarrhéiques, les infections respiratoires aiguës ou les complications liées aux accouchements prématurés sont négligées par la recherche. Ces maladies ne constituent pas un fardeau important dans les pays à revenu élevé, où des protocoles de prévention et de soins efficaces sont appliqués (comme par exemple se laver les mains pour prévenir des maladies diarrhéiques, ou la mise en place de protocoles de soins intensifs pour les nourrissons prématurés). L'étendue du fardeau de ces maladies en Afrique Subsaharienne montre le manque d'applicabilité de ces solutions dans cette région, et souligne le besoin de faire des efforts pour trouver des solutions locales.

Les résultats de notre deuxième cartographie montrent le manque de capacités opérationnelles des promoteurs non-industriels à mettre en place des essais multi-pays comparés aux promoteurs industriels (Section 2.3). Les bénéfices des essais multi-pays dans la recherche sont variés. Ils permettent en particulier d'améliorer la validité externe des résultats de recherche. Ils permettent aussi d'accélérer le recrutement de patients et ainsi diminuer la durée des essais, mais aussi d'augmenter les capacités de recrutement pour des maladies rares. Par ailleurs, des essais cliniques dans lesquels des équipes de recherche de plusieurs pays sont impliqués peuvent être de meilleur qualité, et peuvent avoir des bénéfices secondaires sur

la qualité des systèmes de soins des équipes impliquées (Trimble et al., 2009; Søreide et al., 2013). Nos résultats peuvent ainsi servir d'argument pour pousser à la simplification des procédures et régulations entre acteurs académiques ou institutionnels qui limiteraient leurs capacités à mettre en place des collaborations internationales en recherche clinique.

Nos recherches apportent aussi une réflexion générale sur le chemin à faire pour la conception d'un observatoire global de la recherche médicale (Terry et al., 2014). Les difficultés techniques nécessaires à surpasser pour réunir des bases de données hétérogènes dans un système unique d'information sur l'activité de recherche sont énormes. Nos travaux justifient l'utilisation de méthodes avancées d'extraction d'information sur des données non structurées comme le traitement automatique du langage pour créer des cartographies, mais justifient aussi le besoin de créer une communauté de chercheurs autour d'outils en accès libre pour s'attaquer à ces thématiques. En effet, pendant nos recherches, nous avons donné une importance particulière à donner accès libre aux outils d'analyse développés pour augmenter la transparence et la reproductibilité de nos travaux. Ceci peut servir d'exemple aux autres équipes apportant au développement de l'observatoire global de la recherche pour mener ce projet en tant que communauté.

4.4 Perspectives

Des cartographies utiles pour l'aide à la prise de décision sur l'avenir du programme de recherche nécessitent d'être modulables à l'échelle des besoins et intérêts des acteurs prenant les décisions (Terry et al., 2014). Nous avons ici fourni des cartographies à grande échelle, comparant d'une part la recherche faite par deux grands types d'acteurs (industrie versus non-industrie), d'autre part la recherche faite sur 27 grands groupes de maladies. Ces choix ont permis de faire des cartographies à l'échelle de tous les essais enregistrés avec des méthodes systématiques. Même dans ces cas "simples", il a été impératif de tenir en compte l'incertitude liée

à la mauvaise classification de la recherche pour assurer la qualité de l'information fournie par ces cartographies. Un travail supplémentaire est nécessaire pour construire des cartographies plus détaillées de façon systématique et reproductible, en particulier au niveau des pays, des maladies, des acteurs, des populations ciblées et des formes de recherche.

Dans le cas de cartographies plus fines, l'exhaustivité de l'information recueillie doit être garantie. Un exemple de procédure systématique visant à donner un unique point d'accès à toutes les données de recherche existantes sur une problématique médicale précise est la *meta-analyse en réseau cumulative et dynamique* (LCNMA de l'anglais *live-cumulative network meta-analysis*) (Créquit et al., 2016a). En faisant des requêtes périodiques sur un grand ensemble de bases de données, les LCNMA permettent de connaître l'état de l'art sur l'effet des interventions pour un état de santé précis. Elles pourraient aussi être utilisées pour identifier des lacunes dans la production de connaissances sur des question précises, pour identifier par exemple des populations sous-représentées dans les essais, ou des types d'interventions sous-évaluées (Créquit et al., 2016b).

Nous pouvons ainsi imaginer un observatoire global de la recherche médicale dans lequel des cartographies de la recherche seraient faites depuis le niveau des grandes problématiques de santé publique, jusqu'au niveau des problématiques médicales précises sous forme de LCNMA. Toutes ces cartographies seraient regroupées dans un unique système de connaissances tenant en compte la complexité du système de recherche clinique. Dans ce système multi-échelle, les cartographies à grande échelle nourriraient la totalité de l'observatoire par gravité, et les cartographies à petite échelle le nourriraient par capillarité. Or, pour mettre en place un tel observatoire, les méthodes automatiques d'extraction des connaissances sont à ce jour trop limitées pour assurer une bonne qualité de l'information à toutes les échelles, en raison de la taille, de l'hétérogénéité et du manque de structure des bases de données renseignant l'activité globale de la recherche.

Pour s'attaquer à un objectif aussi colossal, il faut tout d'abord profiter de

ce qui a déjà été fait. Toutes les cartographies à petite échelle présentées dans la Section 1.3 peuvent être recyclées en les unifiant sur une nomenclature commune. Dans 28 publications présentées dans notre travail bibliographique, un total de 32,201 essais cliniques, 6,780 publications et 2,419 revues systématiques ont été manuellement analysés et classifiés pour faire des cartographies. L'incorporation de toutes ces données à un même observatoire aiderait à la diminution du gâchis de la méta-recherche. Suivant le même état d'esprit, des communautés de chercheurs, patients ou acteurs de la recherche médicale pourraient aider au développement de cet observatoire grâce à des interfaces de *crowdsourcing*. De plus, le traitement automatique du langage et l'apprentissage statistique (*machine learning*) pourraient être utilisés pour mettre en place des systèmes de classification semi-automatique nécessitant des validations simples et rapides par des humains, comme c'est le cas par exemple de l'indexation MeSH des articles accessibles par PubMed (Aronson et al., 2007). Par ailleurs, les mêmes technologies pourraient être utilisées pour améliorer la qualité des nouveaux rapports (registres, publications, ...) tout en diminuant le fardeau de l'enregistrement. Par exemple, lors de l'enregistrement d'un nouvel essai clinique, l'algorithme de classification décrit en Section 3.2 pourrait être utilisé pour proposer automatiquement à la personne qui enregistre l'essai une classification de la maladie étudiée dans l'essai selon les catégories de l'étude GBD.

Ces nouveaux outils de classification automatique et semi-automatique deviennent nécessaires pour extraire des connaissances sur ces grandes bases de données, dont un grand nombre est accessible publiquement. Une communauté de chercheurs en sciences de données pour le développement d'un observatoire global de la recherche doit ainsi se former autour de ces problématiques pour le développement d'outils communs en accès libre.

Des cartographies spécifiques sont particulièrement nécessaires dans les pays avec des ressources limitées, les populations vulnérables et les maladies rares. Dans ces cas, un accès compréhensible à toute l'information existante sur l'état de la recherche est indispensable pour allouer les ressources vers les priorités de recherche.

Pour les maladies rares par exemple, l'INSERM a développé en 1997 Orphanet, portail d'accès unique à une grande quantité de données concernant la recherche sur les maladies rares et les médicaments orphelins en Europe, dans lequel sont aujourd'hui accessibles des données d'intérêt pour les chercheurs, les investisseurs et les patients (INSERM, 1997; Rath et al., 2012). Par ailleurs, des cartographies doivent servir de système de veille aux innovations frugales (Burke et al., 2017; Tran and Ravaud, 2016).

Finalement, avec des données renseignant l'activité de recherche clinique réunies dans un observatoire global de la recherche muni de nomenclatures standardisées, il serait possible d'effectuer des analyses plus fines sur les forces influençant la production des connaissances médicales. Des méthodes d'analyses issues de la science des systèmes complexes pourraient être utilisées par exemple pour catégoriser les maladies selon la complexité de la recherche l'étudiant, et comparer l'implication des acteur dans la recherche sur des maladies en fonction de leur complexité (Hausmann et al., 2011; Coscia et al., 2013).

Conclusion

Dans ce travail nous avons construit des cartographies de la recherche clinique, c'est-à-dire des analyses agrégées des bases de données traçant l'activité de recherche dans le monde, pour apporter de l'information sur quelle recherche est faite et où, quelles maladies sont étudiées, et quels acteurs la mettent en place.

À partir de la totalité des essais cliniques enregistrés dans l'International Clinical Trials Registry Platform de l'Organisation Mondiale de la Santé, nous avons évalué pour 7 régions dans le monde l'alignement local entre l'effort de recherche sur 10 ans et le fardeau des maladies. Nous avons montré que la recherche faite dans les pays à revenu élevé est bien alignée avec leurs besoins, mais aussi que dans les régions à revenu moyen ou faible il existe des maladies sous-étudiées relativement à l'ampleur de leur fardeau. Par exemple, alors que l'effort local de recherche sur le VIH et le paludisme en Afrique Subsaharienne est à la hauteur du fardeau qu'elles imposent, d'autres causes majeures de fardeau dans cette région sont négligées par la recherche, notamment les maladies infectieuses communes et les pathologies du nouveau né.

Ce travail a nécessité le développement d'un algorithme de classification automatique des registres d'essais cliniques pour identifier les maladies étudiées dans les essais selon des catégories de morbi-mortalité pour lesquelles des estimations du fardeau dans le monde existent. Cet algorithme est basé sur des méthodes de traitement automatique du langage et des systèmes ontologiques de l'information biomédicale pour classifier les essais cliniques à partir du texte libre présent dans les registres.

Finalement, nous avons évalué l'influence du type de promoteur (industriel ou non-industriel) dans la globalisation des essais cliniques. Nous avons en particulier comparé les différentes capacités des deux types de promoteurs à mettre en place des essais cliniques multi-pays, et comparé les réseaux de collaboration entre pays participant aux essais multi-pays industriels et non-industriels. Nous avons montré la capacité de l'industrie à globaliser leur recherche par rapport aux acteurs non-industriels : 30% de leur recherche est multi-pays contre 3% seulement pour les autres promoteurs. Nous avons aussi montré que des régions, comme les pays d'Europe de l'Est, sont surreprésentés dans la recherche internationale industrielle.

Nos travaux ont ainsi mis en évidence, à partir d'analyses à grande échelle, des lacunes majeures dans l'effort de recherche dans le monde, et ont permis d'identifier l'influence des différents types de promoteurs dans la globalisation de la recherche médicale. Ces travaux ont donné lieu au développement d'outils en accès libre pour permettre l'opérabilité des bases de données traçant l'activité de recherche, aidant ainsi au développement d'un observatoire global de la recherche médicale.

Table des figures

Liste des tableaux

Bibliographie

Adams, J. and Light, R. (2014). Mapping interdisciplinary fields : efficiencies, gaps and redundancies in HIV/AIDS research. *PloS One*, 9(12) :e115092.

Ahmad, N., Boutron, I., Dechartres, A., Durieux, P., and Ravaud, P. (2011). Geographical representativeness of published and ongoing randomized controlled trials. the example of : Tobacco consumption and HIV infection. *PLoS One*, 6(2) :e16878.

Albert, R. and Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1) :47–97.

Alexander, K. P., Kong, D. F., Starr, A. Z., et al. (2013). Portfolio of clinical research in adult cardiovascular disease as reflected in ClinicalTrials.gov. *Journal of the American Heart Association*, 2(5) :e000009.

Anand, S. and Hanson, K. (1997). Disability-adjusted life years : a critical review. *Journal of Health Economics*, 16(6) :685–702.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus : the MetaMap program. *AMIA Annual Symposium Proceedings*, pages 17–21.

Aronson, A. R., Bodenreider, O., Demner-Fushman, D., et al. (2007). From indexing the biomedical literature to coding clinical text : experience with MTI and machine learning approaches. In *BioNLP 2007 : Biological, translational, and clinical language processing*, pages 105–12.

Atal, I., Trinquart, L., Porcher, R., and Ravaud, P. (2015). Differential globalization of industry- and non-industry–sponsored clinical trials. *PLoS One*, 10(12) :e0145122.

Atal, I., Zeitoun, J.-D., Névéol, A., et al. (2016). Automatic classification of registered clinical trials towards the Global Burden of Diseases taxonomy of diseases and injuries. *BMC Bioinformatics*, 17(1) :392.

Babu, A. S., Veluswamy, S. K., Rao, P. T., and Maiya, A. G. (2014). Clinical Trial Registration in physical therapy journals : a cross-sectional study. *Physical Therapy*, 94(1) :83–90.

Bell, S. A. and Tudur Smith, C. (2014). A comparison of interventional clinical trials in rare versus non-rare diseases : an analysis of ClinicalTrials.gov. *Orphanet Journal of Rare Diseases*, 9 :170.

Benchoufi, M. and Atal, I. (2016). Open tool development for clustering biomedical text indexing tools at large scale (unpublished).

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2008(10) :P10008. arXiv : 0803.0476.

Boccia, S., Rothman, K. J., Panic, N., et al. (2015). Registration practices for observational studies on clinicaltrials.gov indicated low adherence. *Journal of Clinical Epidemiology*, 70 :176–82.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS) : integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue) :D267–D270.

Bourgeois, F. T., Olson, K. L., Ioannidis, J. P. A., and Mandl, K. D. (2014). Association between pediatric clinical trials and global burden of disease. *Pediatrics*, 133(1) :78–87.

Boyers, L. N., Karimkhani, C., Hilton, J., Richheimer, W., and Dellavalle, R. P. (2015). Global Burden of Eye and Vision Disease as reflected in the Cochrane Database of Systematic Reviews. *JAMA Ophthalmology*, 133(1) :25.

Brower, V. (2005). The squeaky wheel gets the grease. Research funding is not necessarily allocated to those who need it most. *EMBO reports*, 6(11) :1014–1017.

Burke, T. F., Danso-Bamfo, S., Guha, M., et al. (2017). Shock progression and survival after use of a condom uterine balloon tamponade package in women with

uncontrolled postpartum hemorrhage. *International Journal of Gynaecology and Obstetrics : The Official Organ of the International Federation of Gynaecology and Obstetrics*.

Caldron, P. H., Gavrilova, S. I., and Kropf, S. (2012). Why (not) go east ? Comparison of findings from FDA Investigational New Drug study site inspections performed in Central and Eastern Europe with results from the USA, Western Europe, and other parts of the world. *Drug Design, Development and Therapy*, 6 :53–60.

Califf, R. M., Zarin, D. A., Kramer, J. M., et al. (2012). Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. *JAMA : The Journal of the American Medical Association*, 307(17) :1838–1847.

Canario, J. A., Lizardo, J., Espinal, R., and Colomé, M. (2016). Gaps in health research in the Dominican Republic. *Revista Panamericana De Salud Publica = Pan American Journal of Public Health*, 39(4) :179–185.

Carter, A. J. and Nguyen, C. N. (2012). A comparison of cancer burden and research spending reveals discrepancies in the distribution of research funding. *BMC Public Health*, 12 :526.

Carter, P. H., Berndt, E. R., DiMasi, J. A., and Trusheim, M. (2016). Investigating investment in biopharmaceutical R&D. *Nature Reviews. Drug Discovery*, 15(10) :673–674.

Chakma, J., Sun, G. H., Steinberg, J. D., Sammut, S. M., and Jagsi, R. (2014). Asia's Ascent - Global Trends in Biomedical R&D Expenditures. *New England Journal of Medicine*, 370(1) :3–6.

Chalmers, I., Bracken, M. B., Djulbegovic, B., et al. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet*, 383(9912) :156–165.

Chan, A.-W., Song, F., Vickers, A., et al. (2014). Increasing value and reducing waste : addressing inaccessible research. *The Lancet*, 383(9913) :257–266.

Chang, M., Chang, M., Reed, J. Z., et al. (2016). Developing timely insights into comparative effectiveness research with a text-mining pipeline. *Drug Discovery Today*, 21(3) :473–480.

Cheng, S. K., Hirsch, B. R., Califf, R. M., et al. (2012). Geographic and network analysis of oncology trials : Portfolio assessment of ClinicalTrials.gov. *Journal of Clinical Oncology*, 30(15_suppl) :6047–6047.

Chersich, M. F., Blaauw, D., Dumbaugh, M., et al. (2016). Local and foreign authorship of maternal health interventional research in low- and middle-income countries : systematic mapping of publications 2000-2012. *Globalization and Health*, 12(1) :35.

Chersich, M. F. and Martin, G. (2017). Priority gaps and promising areas in maternal health research in low- and middle-income countries : summary findings of a mapping of 2292 publications between 2000 and 2012. *Globalization and Health*, 13(1) :6.

Connor, E. F. and Simberloff, D. (1979). The assembly of species communities - chance or competition. *Ecology*, 60(6) :1132–1140.

Coscia, M., Hausmann, R., and Hidalgo, C. a. (2013). The Structure and Dynamics of International Development Assistance. *Journal of Globalization and Development*.

Cottingham, M. D., Kalbaugh, C. A., and Fisher, J. A. (2014). Tracking the pharmaceutical pipeline : clinical trials and global disease burden. *Clinical and Translational Science*, 7(4) :297–9.

Créquit, P., Trinquart, L., and Ravaud, P. (2016a). Live cumulative network meta-analysis : protocol for second-line treatments in advanced non-small-cell lung cancer with wild-type or unknown status for epidermal growth factor receptor. *BMJ Open*, 6(8) :e011841.

Créquit, P., Trinquart, L., Yavchitz, A., and Ravaud, P. (2016b). Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis : the example of lung cancer. *BMC Medicine*, 14 :8.

Cummings, J. L., Morstorf, T., and Zhong, K. (2014). Alzheimer's disease drug-development pipeline : few candidates, frequent failures. *Alzheimer's Research & Therapy*, 6(4) :37.

Curley, M. A. Q., Hibberd, P. L., Fineman, L. D., et al. (2005). Effect of prone positioning on clinical outcomes in children with acute lung injury : a randomized controlled trial. *JAMA : The Journal of the American Medical Association*, 294(2) :229–237.

Dal-Ré, R. (2011). Worldwide clinical interventional studies on leading causes of death : a descriptive analysis. *Annals of Epidemiology*, 21(10) :727–731.

Dal-Ré, R., Ioannidis, J. P., Bracken, M. B., et al. (2014). Making prospective registration of observational research a reality. *Science Translational Medicine*, 6(224) :224cm1.

De Angelis, C., Drazen, J. M., Frizelle, F. A., et al. (2004). Clinical trial registration : a statement from the International Committee of Medical Journal Editors. *Annals of Internal Medicine*, 141(6) :477–478.

Dear, R. F., Barratt, A. L., Evans, A., et al. (2012). Identifying and prioritising gaps in colorectal cancer trials research in Australia. *The Medical Journal of Australia*, 197(9) :507–511.

Dear, R. F., Barratt, A. L., McGeechan, K., et al. (2011). Landscape of cancer clinical trials in Australia : using trial registries to guide future research. *The Medical Journal of Australia*, 194(8) :387–391.

Dechartres, A., Ravaud, P., Atal, I., Riveros, C., and Boutron, I. (2016). Association between trial registration and treatment effect estimates : a meta-epidemiological study. *BMC Medicine*, 14(1) :100.

Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support ? *Journal of Biomedical Informatics*, 42(5) :760–772.

Demotes-Mainard, J. and Ohmann, C. (2005). European Clinical Research Infrastructures Network : promoting harmonisation and quality in European clinical research. *The Lancet*, 365 :107–108.

Devleesschauwer, B., Havelaar, A. H., Maertens de Noordhout, C., et al. (2014). DALY calculation in practice : a stepwise approach. *International Journal of Public Health*, 59(3) :571–574.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA : The Journal of the American Medical Association*, 263(10) :1385–1389.

Dickersin, K. and Rennie, D. (2012). The evolution of trial registries and their use to assess the clinical trial enterprise. *JAMA : The Journal of the American Medical Association*, 307(17) :1861–4.

Drain, P. K., Robine, M., Holmes, K. K., and Bassett, I. V. (2014). Trail watch : global migration of clinical trials. *Nature reviews. Drug discovery*, 13(3) :166–7.

Dunn, A. G., Gallego, B., and Coiera, E. (2012). Industry influenced evidence production in collaborative research communities : a network analysis. *Journal of clinical epidemiology*, 65(5) :535–43.

Emdin, C. A., Odutayo, A., Hsiao, A. J., et al. (2015). Association between randomised trial evidence and global burden of disease : cross sectional study (Epidemiological Study of Randomized Trials — ESORT). *BMJ (Clinical research ed.)*, 350 :h117.

Emmons, S., Kobourov, S., Gallant, M., and Börner, K. (2016). Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. *PLoS One*, 11(7) :e0159161.

Evans, J. A., Shim, J.-M., and Ioannidis, J. P. A. (2014). Attention to local health burden and the global disparity of health research. *PLoS One*, 9(4) :e90147.

Fisk, N. M. and Atun, R. (2008). Market Failure and the Poverty of New Drugs in Maternal Health. *PLoS Medicine*, 5(1) :e22.

Fung, K. W. and Bodenreider, O. (2005). Utilizing the UMLS for semantic mapping between terminologies. *AMIA Annual Symposium Proceedings*, pages 266–270.

GBD 2015 DALYs and HALE Collaborators (2016). Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015 : a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053) :1603–1658.

GESICA Investigators (2005). Randomised trial of telephone intervention in chronic heart failure : DIAL trial. *BMJ (Clinical research ed.)*, 331(7514) :425.

Gillum, L. A., Gouveia, C., Dorsey, E. R., et al. (2011). NIH disease funding levels and burden of disease. *PLoS One*, 6(2) :e16837.

Global Burden of Disease Health Financing Collaborator Network (2017). Evolution and patterns of global health financing 1995-2014 : development assistance for health, and government, prepaid private, and out-of-pocket health spending in 184 countries. *The Lancet*, 389(10083) :1981–2004.

Goldacre, B. and Gray, J. (2016). OpenTrials : towards a collaborative open database of all available information on all clinical trials. *Trials*, 17 :164.

Gotelli, N. J. (2000). Null model analysis of species co-occurrence patterns. *Ecology*, 81(9) :2606–2621.

Gross, C. P., Anderson, G. F., and Powe, N. R. (1999). The relation between funding by the National Institutes of Health and the burden of disease. *The New England Journal of Medicine*, 340(24) :1881–1887.

Guegan, E. W., Dorling, H., Ollerhead, L., and Westmore, M. (2016). Mapping public health research across the National Institute for Health Research 2006-2013. *BMC Public Health*, 16 :911.

Hausmann, R., Hidalgo, C. A., Bustos, S., et al. (2011). *The Atlas of economic complexity : mapping paths to prosperity*. Center for International Development, Harvard University : Harvard Kennedy School : Macro Connections, MIT : Massachusetts Institute of Technology, Cambridge, Mass. OCLC : 775005639.

Hazo, J.-B., Gervaix, J., Gandré, C., et al. (2016). European Union investment and countries' involvement in mental health research between 2007 and 2013. *Acta Psychiatrica Scandinavica*, 134(2) :138–149.

Herrera, A. P., Snipes, S. A., King, D. W., et al. (2010). Disparate inclusion of older adults in clinical trials : priorities and opportunities for policy and practice change. *American Journal of Public Health*, 100 Suppl 1 :S105–112.

Hill, K. D., Chiswell, K., Califf, R. M., Pearson, G., and Li, J. S. (2014). Characteristics of pediatric cardiovascular clinical trials registered on ClinicalTrials.gov. *American Heart Journal*, 167(6) :921–929.e2.

Hirsch, B. R., Califf, R. M., Cheng, S. K., et al. (2013). Characteristics of oncology clinical trials : insights from a systematic analysis of ClinicalTrials.gov. *JAMA Internal Medicine*, 173(11) :972–979.

Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., and Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *The Cochrane Database of Systematic Reviews*, (1) :MR000006.

Inrig, J. K., Califf, R. M., Tasneem, A., et al. (2014). The landscape of clinical trials in nephrology : a systematic review of Clinicaltrials.gov. *American Journal*

of Kidney Diseases : The Official Journal of the National Kidney Foundation, 63(5) :771–780.

INSERM (1997). Orphanet (http ://www.orpha.net).

Isaakidis, P., Swingler, G. H., Pienaar, E., Volmink, J., and Ioannidis, J. P. A. (2002). Relation between burden of disease and randomised evidence in sub-Saharan Africa : survey of research. *BMJ (Clinical research ed.)*, 324(7339) :702.

Jones, B. F., Wuchty, S., and Uzzi, B. (2008). Multi-university research teams : shifting impact, geography, and stratification in science. *Science*, 322(November) :1259–1262.

Jones, C. W., Handler, L., Crowell, K. E., et al. (2013). Non-publication of large randomized clinical trials : cross sectional analysis. *BMJ (Clinical research ed.)*, 347(oct28 9) :f6104–f6104.

Joseph, P. D., Caldwell, P. H., Barnes, E. H., and Craig, J. C. (2017). Disease burden-research match ? Registered trials in child health from low- and middle-income and high-income countries. *Journal of Paediatrics and Child Health*, 53(7) :667–674.

Joseph, P. D., Caldwell, P. H. Y., Tong, A., Hanson, C. S., and Craig, J. C. (2016). Stakeholder Views of Clinical Trials in Low- and Middle-Income Countries : A Systematic Review. *Pediatrics*, 137(2) :e20152800.

Joseph, P. D., Craig, J. C., and Caldwell, P. H. Y. (2015). Clinical trials in children. *British Journal of Clinical Pharmacology*, 79(3) :357–369.

Kannan, R., Tetali, P., and Vempala, S. (1999). Simple Markov-chain algorithms for generating bipartite graphs and tournaments. *Random Structures & Algorithms*, 14(4) :293–308.

Karimkhani, C., Boyers, L. N., Prescott, L., et al. (2014). Global Burden of Skin Disease as reflected in Cochrane Database of Systematic Reviews. *JAMA Dermatology*, 150(9) :945–51.

Karimkhani, C., Trikha, R., Aksut, B., et al. (2016). Identifying gaps for research prioritisation : Global burden of external causes of injury as reflected in the Cochrane Database of Systematic Reviews. *Injury*, 47(5) :1151–1157.

Korevaar, D. A., Bossuyt, P. M. M., and Hooft, L. (2014). Infrequent and incomplete registration of test accuracy studies : analysis of recent study reports. *BMJ Open*, 4(1) :e004596.

Kotseva, K., Wood, D., De Backer, G., et al. (2009). EUROASPIRE III : a survey on the lifestyle, risk factors and use of cardioprotective drug therapies in coronary patients from 22 European countries. *European Journal of Cardiovascular Prevention and Rehabilitation*, 16(2) :121–137.

Kunath, F., Grobe, H. R., Keck, B., et al. (2011). Do urology journals enforce trial registration ? A cross-sectional study of published trials. *BMJ Open*, 1(2) :e000430–e000430.

Lakey, W. C., Barnard, K., Batch, B. C., et al. (2013). Are current clinical trials in diabetes addressing important issues in diabetes care ? *Diabetologia*, 56(6) :1226–1235.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms : A comparative analysis. *Physical Review E*, 80(5) :056117.

Larsen, P. O. and Ins, M. v. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3) :575–603.

Leopold, S. S., Swiontkowski, M., and Haddad, F. (2016). JBJS, The Bone & Joint Journal, and Clinical Orthopaedics and Related Research Require Prospective Registration of Randomized Clinical Trials* : Why Is This Important ? *The Journal of Bone and Joint Surgery. American Volume*, 98(23) :1947–1948.

Lozano, R., Naghavi, M., Foreman, K., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010 : A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859) :2095–128.

Mahaffey, K. W., Wojdyla, D. M., Carroll, K., et al. (2011). Ticagrelor compared with clopidogrel by geographic region in the Platelet Inhibition and Patient Outcomes (PLATO) Trial. *Circulation*, 124(5) :544–554.

Maitland, K., Kiguli, S., Opoka, R. O., et al. (2011). Mortality after fluid bolus in African children with severe infection. *The New England Journal of Medicine*, 364(26) :2483–2495.

Matee, M. I., Manyando, C., Ndumbe, P. M., et al. (2009). European and Developing Countries Clinical Trials Partnership (EDCTP) : the path towards a true partnership. *BMC Public Health*, 9(Dccc) :249.

Miklós, I. and Podani, J. (2004). Randomization of presence-absence matrices : Comments and new algorithms. *Ecology*, 85(1) :86–92.

Miller, J. W. and Harrison, M. T. (2013). Exact sampling and counting for fixed-margin matrices. *The Annals of Statistics*, 41(3) :1569–1592.

Mok, T. S., Wu, Y.-L., Thongprasert, S., et al. (2009). Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *The New England Journal of Medicine*, 361(10) :947–957.

Moran, M. (2016). The Grand Convergence : Closing the Divide between Public Health Funding and Global Health Needs. *PLoS Biology*, 14(3) :e1002363.

Mork, J., Aronson, A., and Demner-Fushman, D. (2017). 12 years on - Is the NLM medical text indexer still useful and relevant ? *Journal of Biomedical Semantics*, 8(1) :8.

Mork, J. G., Yepes, A. J. J., and Aronson, A. R. (2013). The NLM Medical Text Indexer System for Indexing Biomedical Literature. In *BioASQ@ CLEF*.

Murray, C. J. L., Ezzati, M., Flaxman, A. D., et al. (2012). GBD 2010 : Design, definitions, and metrics. *The Lancet*, 380(9859) :2063–6.

Murthy, S., Mandl, K. D., and Bourgeois, F. T. (2015). Industry-sponsored clinical research outside high-income countries : an empirical analysis of registered clinical trials from 2006 to 2013. *Health research policy and systems / BioMed Central*, 13(1) :28.

Ndounga Diakou, L. A., Ntoumi, F., Ravaud, P., and Boutron, I. (2017). Published randomized trials performed in Sub-Saharan Africa focus on high-burden diseases but are frequently funded and led by high-income countries. *Journal of Clinical Epidemiology*, 82 :29–36.e6.

NEPAD (2005). African Science Technology and Innovation Indicators (ASTII) (http ://www.nepad.org/programme/african-science-technology-and-innovation-indicators-astii).

Neumann, P. J., Thorat, T., Zhong, Y., et al. (2016). A Systematic Review of Cost-Effectiveness Studies Reporting Cost-per-DALY Averted. *PLoS One*, 11(12) :e0168512.

Névéol, A. and Zweigenbaum, P. (2015). Clinical Natural Language Processing in 2014 : Foundational Methods Supporting Efficient Healthcare. *Yearb Med Inform*, 10(1) :194–8.

Okhovati, M., Zare, M., and Bazrafshan, A. (2015). Variations in Ischemic Heart Disease Research by Country, Income, Development and Burden of Disease : A Scientometric Approach. *Journal of Cardiovascular and Thoracic Research*, 7(4) :164–167.

Pan, R. K., Kaski, K., and Fortunato, S. (2012). World citation and collaboration networks : uncovering the role of geography in science. *Scientific Reports*, 2 :1–7.

Pasquali, S. K., Lam, W. K., Chiswell, K., Kemper, A. R., and Li, J. S. (2012). Status of the pediatric clinical trials enterprise : an analysis of the US ClinicalTrials.gov registry. *Pediatrics*, 130(5) :e1269–1277.

Patrone, D. (2010). Discrepancies between research advertisements and disclosure of study locations in trial registrations for USA-sponsored research in Russia. *Journal of Medical Ethics*, 36(7) :431–434.

Pederson, H., Okland, T., Boyers, L. N., et al. (2015). Identifying Otolaryngology systematic review research gaps. *JAMA Otolaryngology–Head & Neck Surgery*, 141(1) :67.

Pedrique, B., Strub-Wourgaft, N., Some, C., et al. (2013). The drug and vaccine landscape for neglected diseases (2000-11) : a systematic assessment. *The Lancet Global Health*, 1(6) :e371–379.

Perel, P., Miranda, J. J., Ortiz, Z., and Casas, J. P. (2008). Relation between the Global Burden of Disease and randomized clinical trials conducted in Latin America published in the five leading medical journals. *PLoS One*, 3(2) :e1696.

Policy Cures (2007). G-FINDER (http ://policycures.org/gfinder.html).

Powell-Smith, A. and Goldacre, B. (2016). The TrialsTracker : Automated ongoing monitoring of failure to share clinical trial results by all major companies and research institutions. *F1000Research*, 5 :2629.

Prasad, V. (2016). How should research be funded ? Difficulties with the argument for proportionality to causes of death or years of life lost. *Journal of the National Comprehensive Cancer Network*, 14(3) :365–366.

Rath, A., Olry, A., Dhombres, F., et al. (2012). Representation of rare diseases in health information systems : the Orphanet approach to serve a wide range of end users. *Human Mutation*, 33(5) :803–808.

Resource Tracking Working Group (2004). HIV Prevention Research and Development Investments (http ://www.hivresourcetracking.org/).

Reveiz, L., Bonfill, X., Glujovsky, D., et al. (2012). Trial registration in Latin America and the Caribbean's : study of randomized trials published in 2010. *Journal of Clinical Epidemiology*, 65(5) :482–487.

RICYT (1990). Ibero-American and Inter-American Network for Science and Technology Indicators (http ://www.ricyt.org/).

Rochon, P. A., Mashari, A., Cohen, A., et al. (2004). Relation between randomized controlled trials published in leading general medical journals and the global burden of disease. *CMAJ*, 170(11) :1673–1677.

Ross, J. S., Mulvey, G. K., Hines, E. M., Nissen, S. E., and Krumholz, H. M. (2009). Trial Publication after Registration in ClinicalTrials.Gov : A Cross-Sectional Analysis. *PLoS Medicine*, 6(9) :e1000144.

Rosvall, M. and Bergstrom, C. (2007). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4) :1118–23.

Rothwell, P. M. (2005). External validity of randomised controlled trials : "to whom do the results of this trial apply ?". *The Lancet*, 365(9453) :82–93.

Rothwell, P. M. (2007). *Treating Individuals : From Randomised Trials to Personalised Medicine*. Elsevier Health Sciences. Google-Books-ID : aHfBF8GoL8UC.

Roumiantseva, D., Carini, S., Sim, I., and Wagner, T. H. (2013). Sponsorship and design characteristics of trials registered in ClinicalTrials.gov. *Contemporary Clinical Trials*, 34(2) :348–355.

Røttingen, J. A., Regmi, S., Eide, M., et al. (2013). Mapping of available health research and development data : what's there, what's missing, and what role is there for a global observatory ? *The Lancet*, 382(9900) :1286–307.

Sampat, B. N., Buterbaugh, K., and Perl, M. (2013). New evidence on the allocation of NIH funds across diseases. *The Milbank Quarterly*, 91(1) :163–185.

Say, L., Chou, D., Gemmill, A., et al. (2014). Global causes of maternal death : a WHO systematic analysis. *The Lancet Global Health*, 2(6) :e323–333.

Sox, H. C. and Rennie, D. (2008). Seeding trials : just say "no". *Annals of Internal Medicine*, 149(4) :279–280.

Stamatakis, E., Weiler, R., and Ioannidis, J. P. A. (2013). Undue industry influences that distort healthcare research, strategy, expenditure and practice : a review. *European Journal of Clinical Investigation*, 43(5) :469–475.

Subherwal, S., Patel, M. R., Chiswell, K., et al. (2014). Clinical trials in peripheral vascular disease : pipeline and trial designs : an evaluation of the ClinicalTrials.gov database. *Circulation*, 130(20) :1812–1819.

Swingler, G. H., Pillay, V., Pienaar, E. D., and Ioannidis, J. P. A. (2005). International collaboration, funding and association with burden of disease in randomized controlled trials in Africa. *Bulletin of the World Health Organization*, 83(7) :511–517.

Swingler, G. H., Volmink, J., and Ioannidis, J. P. a. (2003). Number of published systematic reviews and global burden of disease : database analysis. *BMJ (Clinical research ed.)*, 327(7423) :1083–4.

Søreide, K., Alderson, D., Bergenfelz, A., et al. (2013). Strategies to improve clinical research in surgery through international collaboration. *The Lancet*, 382(9898) :1140–1151.

Tasneem, A., Aberle, L., Ananth, H., et al. (2012). The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty. *PLoS One*, 7(3) :e33677.

Terry, R. F., Allen, L., Gardner, C. A., et al. (2012). Mapping global health research investments, time for new thinking–a Babel Fish for research data. *Health Research Policy and Systems*, 10 :28.

Terry, R. F., Salm, J. F., Nannei, C., and Dye, C. (2014). Creating a global observatory for health R&D. *Science*, 345(6202) :1302–4.

The PLOS Medicine Editors (2014). Observational Studies : Getting Clear about Transparency. *PLoS Medicine*, 11(8) :e1001711.

Thiers, F. a., Sinskey, A. J., and Berndt, E. R. (2008). Trends in the globalization of clinical trials. *Nature Reviews Drug Discovery*, 7(1) :13–14.

Todd, J. L., White, K. R., Chiswell, K., Tasneem, A., and Palmer, S. M. (2013). Using ClinicalTrials.gov to understand the state of clinical research in pulmonary, critical care, and sleep medicine. *Annals of the American Thoracic Society*, 10(5) :411–417.

Torloni, M., Gomes Freitas, C., Kartoglu, U., Metin Gülmezoglu, A., and Widmer, M. (2016). Quality of oxytocin available in low- and middle-income countries : a systematic review of the literature. *BJOG : An International Journal of Obstetrics & Gynaecology*, 123(13) :2076–2086.

Tran, V.-T. and Ravaud, P. (2016). Frugal innovation in medicine for low resource settings. *BMC Medicine*, 14 :102.

Treatment Action Group (2005). TAG (http ://www.treatmentactiongroup.org/).

Trimble, E. L., Abrams, J. S., Meyer, R. M., et al. (2009). Improving Cancer Outcomes Through International Collaboration in Academic Cancer Treatment Trials. *Journal of Clinical Oncology*, 27(30) :5109–5114.

Trinquart, L., Johns, D. M., and Galea, S. (2016). Why do we think we know what we know ? A metaknowledge analysis of the salt controversy. *International Journal of Epidemiology*, 45(1) :251–260.

Trouiller, P., Olliaro, P., Torreele, E., et al. (2002). Drug development for neglected diseases : a deficient market and a public-health policy failure. *Lancet (London, England)*, 359(9324) :2188–2194.

Tsalatsanis, A., Barnes, L., Hozo, I., Skvoretz, J., and Djulbegovic, B. (2011). A social network analysis of treatment discoveries in cancer. *PLoS One*, 6(3) :e18060.

U.S. National Institutes of Health (2000). ClinicalTrials.gov (https ://clinicaltrials.gov/).

van de Wetering, F. T., Scholten, R. J. P. M., Haring, T., Clarke, M., and Hooft, L. (2012). Trial registration numbers are underreported in biomedical publications. *PLoS One*, 7(11) :e49599.

Vanderelst, D. and Speybroeck, N. (2013). Scientometrics reveals funding priorities in medical research policy. *Journal of Informetrics*, 7(1) :240–247.

Viergever, R. F. and Hendriks, T. C. C. (2016). The 10 largest public and philanthropic funders of health research in the world : what they fund and how they distribute their funds. *Health Research Policy and Systems*, 14 :12.

Viergever, R. F., Karam, G., Reis, A., and Ghersi, D. (2014). The Quality of Registration of Clinical Trials : Still a Problem. *PLoS One*, 9(1) :e84727.

Viergever, R. F. and Li, K. (2015). Trends in global clinical trial registration : an analysis of numbers of registered clinical trials in different parts of the world from 2004 to 2013. *BMJ Open*, 5 :e008932.

Viergever, R. F., Terry, R. F., and Karam, G. (2013). Use of data from registered clinical trials to identify gaps in health research and development. *Bulletin of the World Health Organization*, 91(6) :416–425C.

Weber, W. E. J., Merino, J. G., and Loder, E. (2015). Trial registration 10 years on. *BMJ (Clinical research ed.)*, 3572(July) :h3572.

Witsell, D. L., Schulz, K. A., Lee, W. T., and Chiswell, K. (2013). An analysis of registered clinical trials in otolaryngology from 2007 to 2010 : ClinicalTrials.gov. *Otolaryngology–Head and Neck Surgery : Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, 149(5) :692–699.

World Health Organization (2005). International Clinical Trials Registry Platform (ICTRP) (http ://www.who.int/ictrp/en/).

World Health Organization (2010). *World Health Report, 2010 : health systems financing the path to universal coverage.* WHO Library Cataloguing - in - Publication Data.

World Health Organization (2011). *International statistical classification of diseases and related health problems. -10th revision.* WHO Library Cataloguing-in-Publication Data, 2010 edition.

World Health Organization (2012). *International standards for clinical trial registries.* WHO Library Cataloguing - in - Publication Data.

Yoong, S. L., Hall, A., Williams, C. M., et al. (2015). Alignment of systematic reviews published in the Cochrane Database of Systematic Reviews and the Database of Abstracts and Reviews of Effectiveness with global burden-of-disease data : a bibliographic analysis. *Journal of Epidemiology & Community Health*, 69(7) :708–714.

Zarin, D. A., Ide, N. C., Tse, T., et al. (2007). Issues in the registration of clinical trials. *JAMA : The Journal of the American Medical Association*, 297(19) :2112–20.

Zarin, D. A. and Tse, T. (2013). Trust but Verify : Trial Registration and Determining Fidelity to the Protocol. *Annals of Internal Medicine*, 159(1) :63.

Zarin, D. A., Tse, T., and Sheehan, J. (2015). The proposed rule for U.S. clinical trial registration and results submission. *The New England Journal of Medicine*, 372(2) :174–180.

Zeitoun, J.-D., Baron, G., Vivot, A., et al. (2017). Post- marketing research and its outcome for novel anticancer agents approved by both the FDA and EMA between 2005 and 2010 : a cross-sectional study. *International Journal of Cancer*, (in press).

Does clinical research effort match public health needs? A large-scale mapping of 115,000 randomized trials and 2.2 billion disability-adjusted life years

Ignacio ATAL, Ludovic TRINQUART, Philippe RAVAUD, Raphaël PORCHER

Supplementary Information

TABLE OF CONTENTS:

APPENDIX S1: Supplementary Information for methods

Identification of randomized controlled trials (RCTs)

We identified clinical trial records corresponding to RCTs based on the study type and study design fields from the trial records. Both fields correspond to free text. First, we identified interventional trials (as opposed to observational studies) based on the study type field. We included records for which the study type field included the string “interv”. Second, we identified non-randomized and non-controlled clinical trials based on the study design field. We identified records for which the study design field included at least one of the following strings: "not random", "non random", "not-random", "non-random", "without random", "nonrandom", "no random", "randomised: no", "randomized: no", "not control", "non control", "not-control", "non-control", "without control", "noncontrol", "no control", "control: no", "controlled: no" and "pharmacokinetics". Finally, we considered RCTs based on trial records identified as interventional, not identified as non-randomized or non-controlled, and for which the study design field included the string “random”.

Sample size extraction from clinical trial records and count of patients enrolled in RCTs within regions across groups of diseases

In clinical trial records, the planned sample size is reported in a sentence. We extracted numbers from that text. We defined several rules to automatically assigning a sample size to each trial record depending on the nature of the numbers extracted. Rules were defined to deal with cases such as “150 patients”, “Group A: 30; Group B: 30; Group C: 30”, “Group 1: 30; Group 2: 30; Group 3: 30”, “150 patients (control: 75; experimental: 75)”.

The number of patients planned to be enrolled was divided equally across countries of

recruitment. RCTs concerning several groups of diseases were considered in the mapping of each group of diseases, and the total planned sample size was considered for each group of diseases.

Incorporation of uncertainty in measures of research effort

We incorporated classification uncertainty of RCTs across groups of diseases in measures concerning the health research effort in each region and for each group of diseases. The method used was inspired by *Fox et al. (2005), International Journal of Epidemiology* and was as follows:

We consider a region *R* and a group of diseases *d* included in the set *D* of all 27 groups of diseases. In our analyses we were interested in the following measures concerning research effort:

- $N_{R,d}$ the number of RCTs conducted in *R* concerning *d*

- $N_{R,D}$ the total number of RCTs conducted in *R* relevant to the burden of diseases

- $LRE_{R,d} = N_{R,d} / N_{R,D}$ the regional research effort in *R* concerning *d*

- $LRG_R = \sum_{d \text{ with Gap in } R} | LRE_{R,d} - ½ * LRN_{R,d} |$ the proportion of research effort that need to be reallocated in *R* to eliminate regional gaps (where $LRN_{R,d} = Burden_{R,d} / Burden_{R,D}$ is the regional health needs in *R* attributable to *d)*

- $RRE_{R,d} = N_{R,d} / N_{R \neq \text{High-Income}, d}$ the disease-specific research effort on *d* in *R (*for *R* among the 6 non-high-income regions)

- $RRG_d = \sum_{R \text{ with Gap on } d} | RRE_{R,d} - ½ * RRN_{R,d} |$ the proportion of research effort that need to be

reallocated across non-high-income regions to eliminate disease-specific research gaps on *d* (where $RRN_{R,d} = Burden_{R,d} / Burden_{R \neq High\text{-}Income,\ d}$ is the disease-specific health needs in *R)*

For each group of diseases *d*, we first simulated *N=10,000* pairs of sensitivities and specificities of the classifier to identify *d* using beta distributions on the true positive and negative rates:

$$Sens_d \sim \beta(1 + TP_d,\ 1 + TP_d + FN_d) \text{ and } Sp_d \sim \beta(1 + TN_d,\ 1 + FP_d + Tn_d).$$

Second, we derived *N* pairs of positive and negative predictive values PPV_d and NPV_d for the classification toward *d.* Third, for each pair of positive and negative predictive values, we corrected the classification for each RCT using as probability of reclassification Bernouilli trials with parameter PPV_d (resp. NPV_d) for RCTs initially classified as concerning *d* (resp. not concerning *d*). This gave us *N* simulated datasets, over which we counted the corrected number of RCTs concerning *d* in each region *R*. Therefore, we derived medians, 2.5th and 97.5th percentiles for $N_{R,d}$ for each *R* and *d* that we used as estimates and 95% uncertainty intervals (UIs).

To derive estimates and 95% UIs for $LRE_{R,d} = N_{R,d} / N_{R,D}$, we simultaneously accounted for the uncertainty on $N_{R,d}$ and on $N_{R,D}$. These two numbers are dependent, because RCTs concerning *d* are therefore relevant to the burden of diseases. In each simulated dataset mentioned previously, we used a similar method to incorporate a corrected classification of the status “does the RCT concern another group of diseases than *d*” for each RCT. For each simulated dataset we then counted the number of RCTs relevant to the burden of diseases in each region by accounting for RCTs concerning *d* or concerning a group of diseases other than *d*. We then evaluated over each simulated dataset the proportions $LRE_{R,d} = N_{R,d} / N_{R,D}$ for

each *R* to derive estimates and 95% UIs for these values. Over each simulated dataset, we also evaluated $RRE_{R,d}$ for each *R* and *d,* and RRG_d for each *d,* to derive estimates and 95% UIs for these values.

Because we derived *N* simulated datasets separately for each *d*, we could not directly derive estimates and 95% UIs for LRG_R because this measure needs to incorporate the uncertainty of classification simultaneously for all diseases. Simultaneously for all groups of diseases we sampled a simulated dataset. Over each sampled dataset, we evaluated for each region the proportion of research effort that need to be reallocated to eliminate a regional research gap on *d* (if it exists in the sampled dataset). We then evaluated LRG_R for each *R* by summing the regional proportions of reallocation across diseases. We repeated this process 10,000 times and evaluated the median and 2.5th and 97.5th percentiles for $N_{R,d}$ for each *R* that we used as estimates and 95% UIs.

We used the same method to incorporate classification uncertainty in measures of research effort based on the number of patients planned to be enrolled in RCTs.

As described in *Fox et al.* (2005) some sampled values of sensitivity and specificity may lead to impossible values for the positive and negative predictive values (< 0 or > 1). When this was the case for more than 10% of the iterations, it is recommended to modify the distributions for sampling sensitivities and specificities. For sexually transmitted diseases excluding HIV, and congenital anomalies, sampled specificities were too low. We considered a new beta distribution using the lower-bound of a 50% exact confidence interval of the number of false positives. For leprosy, hemoglobinopathies and hemolytic anemias, and sudden infant death syndrome, this method still led to more than 10% of impossible iterations,

so we did not conduct simulations for these diseases.

This method is based on the assumption that misclassification of RCTs is not correlated between groups of diseases and is independent of the countries of location and the targeted sample size. In addition, it relies on the assumption that for a given group of diseases, the sensitivities and specificities of classification are independent.

SUPPLEMENTARY TABLES

Table S1: Confusion matrices with the performances on a test set of the knowledge-based classifier to identify clinical trials for each of the 27 groups of diseases

Group of diseases	Classification towards the group of diseases				Classification towards another group of diseases			
	TP_d	FP_d	TN_d	FN_d	$Tp_{D\setminus d}$	$FP_{D\setminus d}$	$TN_{D\setminus d}$	$FN_{D\setminus d}$
Tuberculosis	14	2	2745	2	2142	204	267	150
HIV/AIDS	86	7	2659	11	2072	214	333	144
Common infectious diseases	40	21	2693	9	2113	207	299	144
Malaria	14	1	2748	0	2142	204	267	150
Neglected tropical diseases excluding malaria	6	0	2756	1	2150	203	261	149
Maternal disorders	17	5	2715	26	2130	210	289	134
Neonatal disorders	4	7	2746	6	2148	205	262	148
Nutritional deficiencies	11	15	2732	5	2140	201	272	150
Sexually transmitted diseases excluding HIV	0	3	2759	1	2155	203	255	150
Hepatitis	14	4	2742	3	2141	208	262	152
Leprosy	2	1	2760	0	2154	203	256	150
Neoplasms	933	42	1763	25	1213	214	1198	138
Cardiovascular and circulatory diseases	178	60	2468	57	1951	217	466	129
Chronic respiratory diseases	76	17	2665	5	2074	209	328	152
Cirrhosis of the liver	19	17	2723	4	2133	211	267	152
Digestive diseases (except cirrhosis)	24	28	2703	8	2129	199	289	146
Neurological disorders	79	40	2630	14	2060	211	339	153
Mental and behavioral disorders	134	33	2587	9	2014	198	402	149
Diabetes, urinary diseases and male infertility	196	63	2458	46	1930	213	473	147
Gynecological diseases	9	8	2744	2	2146	206	262	149

Hemoglobinopathies and hemolytic anemias	10	4	2743	6	2143	203	270	147
Musculoskeletal disorders	100	40	2610	13	2046	188	382	147
Congenital anomalies	22	34	2706	1	2121	205	275	162
Skin and subcutaneous diseases	18	24	2717	4	2134	198	281	150
Sense organ diseases	52	40	2667	4	2085	190	322	166
Oral disorders	3	4	2751	5	2150	207	258	148
Sudden infant death syndrome	0	0	2763	0	2156	203	254	150

TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives.

The confusion matrices were derived from a test set of 2,763 registered clinical trials described in Atal et al. (2016 BMC Bioinformatics).

Figure S1: Flow of trials in the study

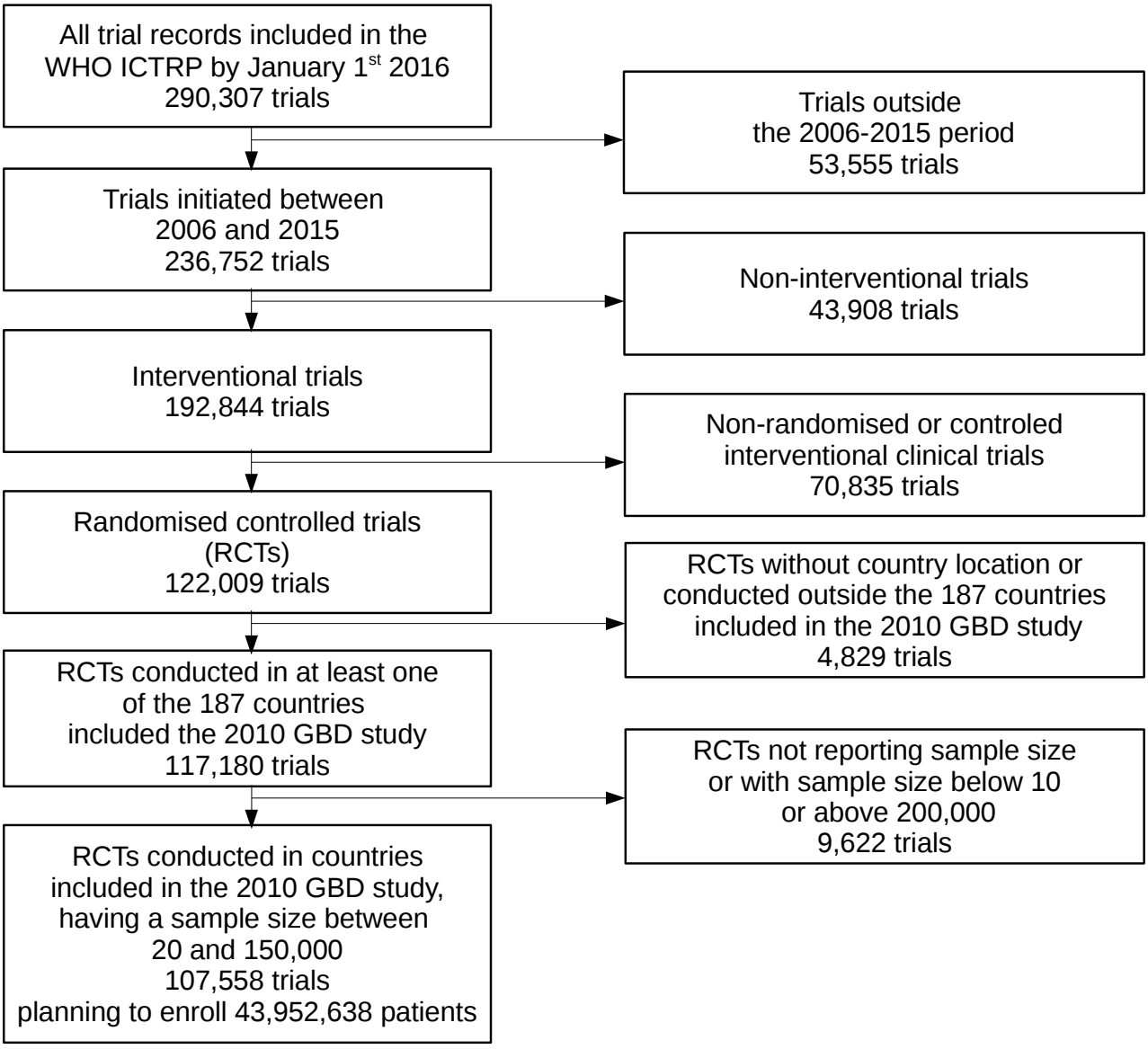
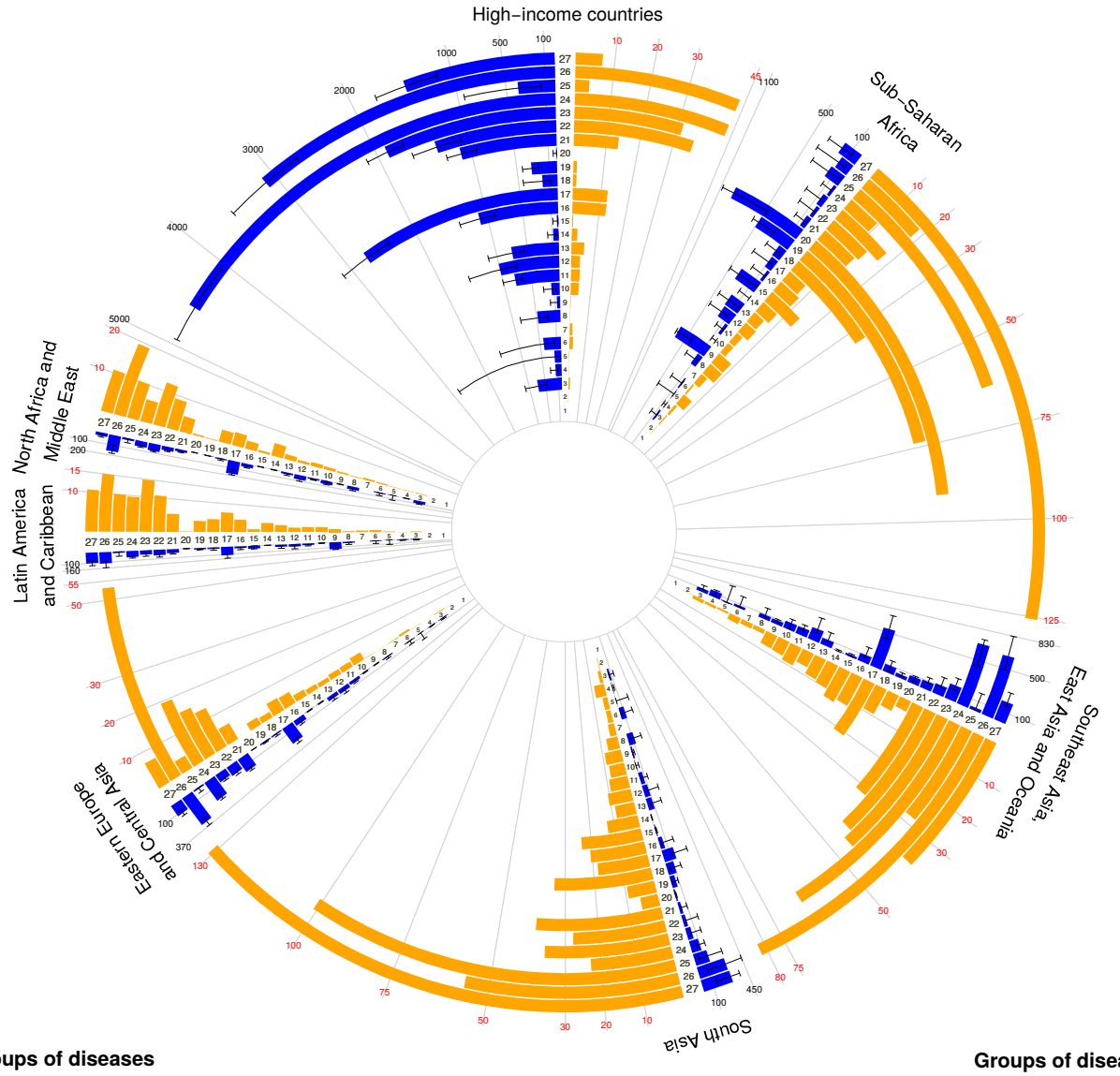


Figure S2: Number of patients planned to be enrolled in randomized controlled trials (RCTs) versus number of disability-adjusted life years (DALYs) for the 7 regions and the 27 groups of diseases



Groups of diseases

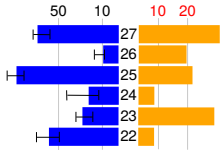
27 Common infectious diseases
26 Cardiovascular and circulatory diseases
25 Neonatal disorders
24 Neoplasms
23 Mental and behavioral disorders
22 Musculoskeletal disorders
21 Chronic respiratory diseases
20 Malaria
19 HIV/AIDS
18 Nutritional de ciencies
17 Diabetes, urinary diseases and male infertility
16 Neurological disorders
15 Tuberculosis
14 Congenital anomalies

Regional research e ort

(Thousands of patients planned to be enrolled in RCTs [95% UI])

Regional health needs

(Million DALYs)



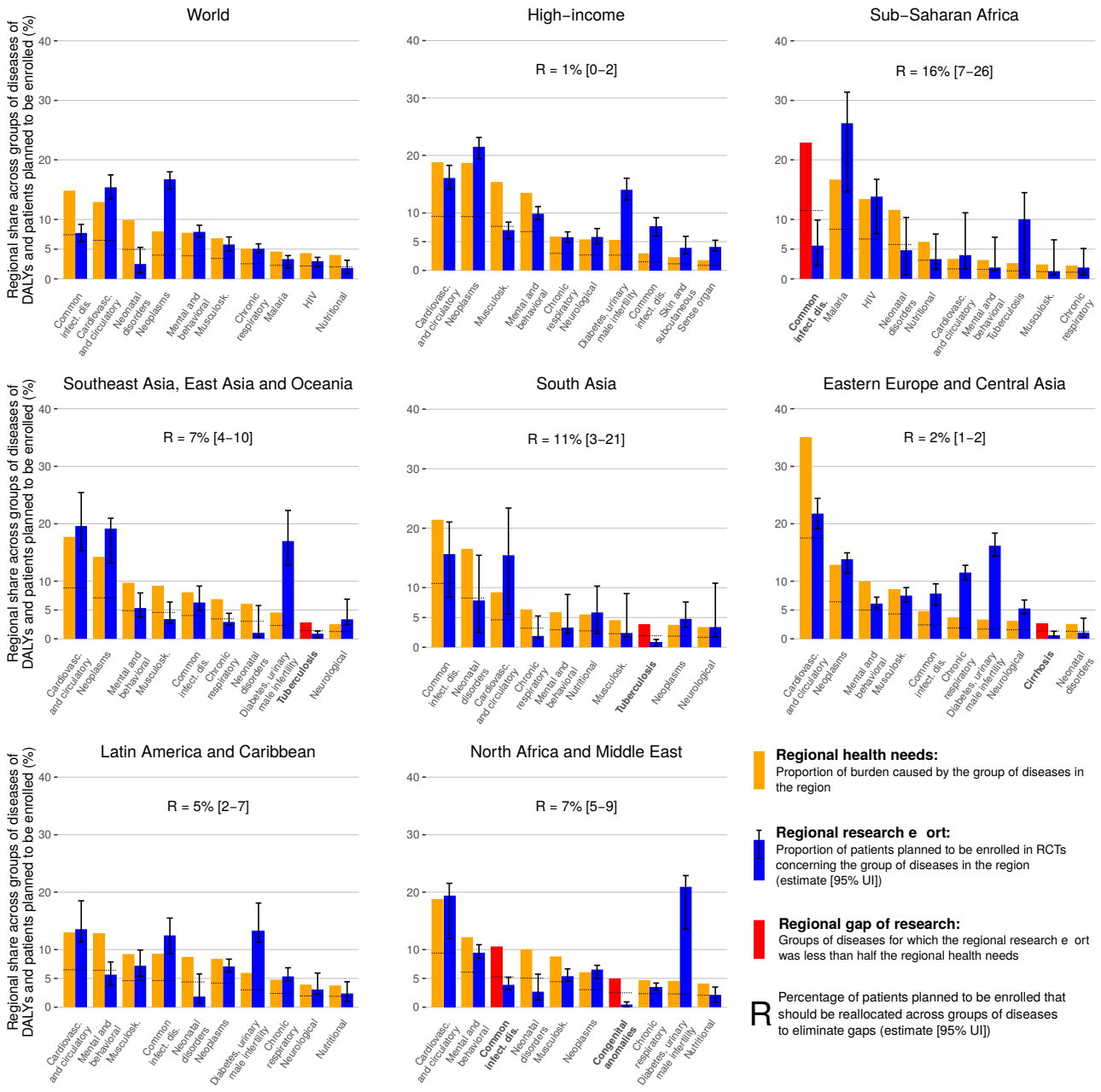
Groups of diseases

Skin and subcutaneous diseases 13
Digestive diseases (except cirrhosis) 12
Sense organ diseases 11
Cirrhosis of the liver 10
Neglected tropical diseases excluding malaria 9
Maternal disorders 8
Hemoglobinopathies and hemolytic anemias 7
Oral disorders 6
Sexually transmitted diseases excluding HIV 5
Hepatitis 4
Gynecological diseases 3
Sudden infant death syndrome 2
Leprosy 1

Number of patients planned to be enrolled in registered RCTs (per thousands) initiated in 2006-2015, and number of DALYs in 2005 within the 7 regions across the 27 groups of

diseases. The estimates and 95% uncertainty intervals (Uis) for the number of patients planned to be enrolled in RCTs are shown after incorporating classification uncertainty of RCTs across groups of diseases. Regions are ordered clockwise by the total number of patients planned to be enrolled in RCTs relevant to the burden of diseases. Groups of diseases are ordered by the global burden.

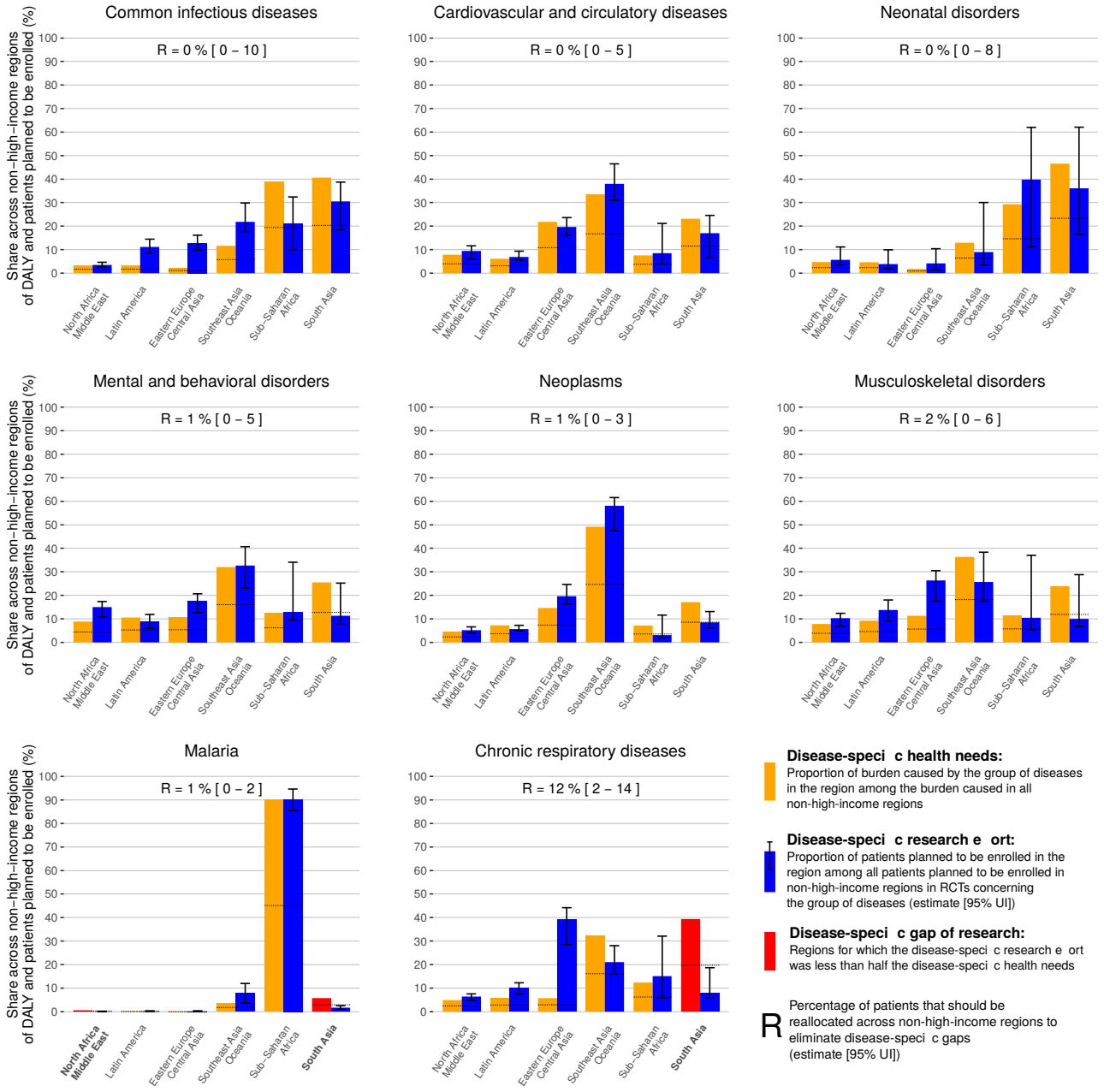
Figure S3: Regional proportion of patients planned to be enrolled in randomized controlled trials (RCTs) and disability-adjusted life years (DALYs) within regions across groups of diseases and regional research gaps



Worldwide, and for each region, compares the regional research effort (measured as the proportion of patients planned to be enrolled in registered RCTs initiated in 2006-2015), to the

regional health needs (measured as the proportion of DALYs in 2005) for the 10 groups of diseases causing the highest burden in the region. For regional research effort, shows the estimates and 95% uncertainty intervals (UIs) after incorporating classification uncertainty of RCTs across groups of diseases. Regions are ordered by the total number of patients planned to be enrolled in RCTs relevant to the burden of diseases. Groups of diseases are ordered by the regional burden. For each region, highlights groups of diseases showing regional research gaps (i.e., for which the regional research effort was less than half the regional health needs). For each region, annotates the proportion of patients planned to be enrolled in RCTs that should be reallocated (R) across groups of diseases to eliminate regional research gaps. Visualization for other groups of diseases and other measures of burden are available at https://clinicalepidemio.fr/RCTvsBurden/.

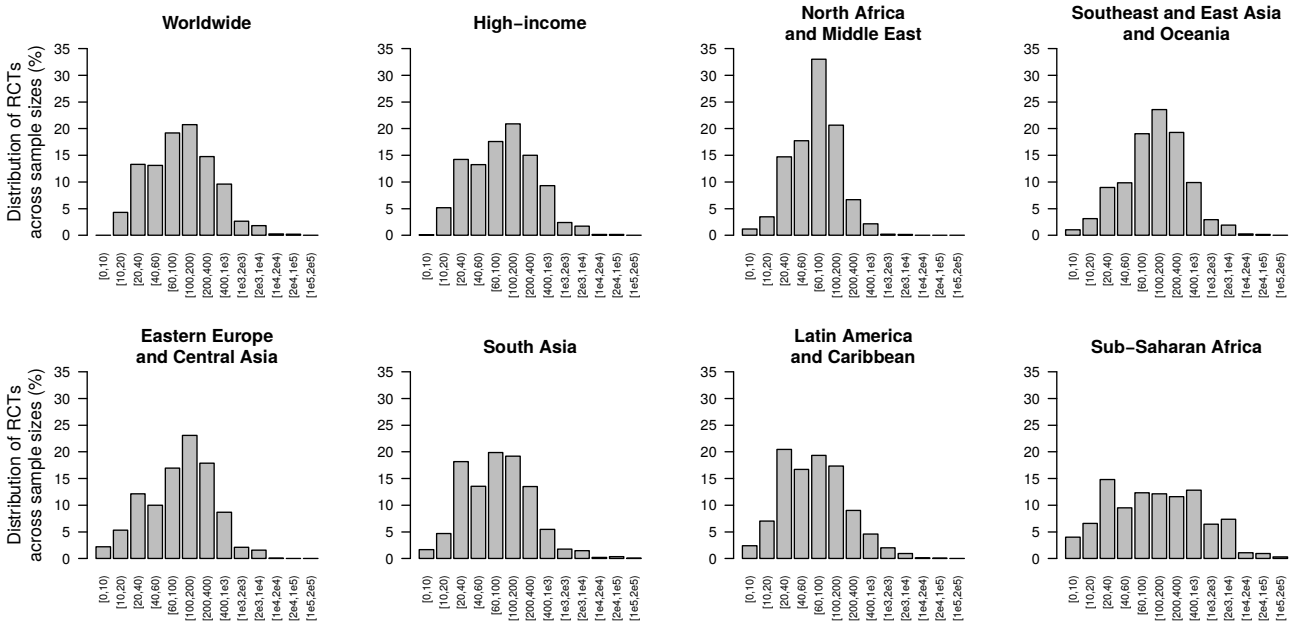
Figure S4: Disease-specific proportion of patients planned to be enrolled in randomized controlled trials (RCTs) and disability-adjusted life years (DALYs) across non–high-income regions and disease-specific research gaps



For the 8 groups of diseases causing the highest number of DALYs in 2005, compares, for each region, excluding high-income countries, the disease-specific research efforts

(measured as the proportion of patients planned to be enrolled in RCTs initiated in 2006-2015) to the disease-specific health needs (measured as the proportion of DALYs in 2005) among all non–high-income regions. For disease-specific research effort in each region, shows the estimates and 95% uncertainty intervals (UIs) after incorporating classification uncertainty of RCTs across groups of diseases. Groups of diseases are ordered by the total burden in non–high-income regions. Regions are ordered by the total burden. For each group of diseases, highlights regions showing disease-specific research gaps (i.e., for which the disease-specific research effort in the region was less than half the disease-specific health needs of the region). For each group of diseases, annotates the proportion of patients planned to be enrolled in RCTs that should be reallocated (R) across non–high-income regions to eliminate disease-specific research gaps. Visualization for other groups of diseases and other measures of burden are available at https://clinicalepidemio.fr/RCTvsBurden/.

Figure S5: Distribution of sample sizes of randomized controlled trials (RCTs) worldwide and within the 7 regions



Differential globalization of industry- and non-industry–sponsored clinical trials

Ignacio ATAL, Ludovic TRINQUART, Raphaël PORCHER, Philippe RAVAUD

SUPPLEMENTARY INFORMATION

TABLE OF CONTENTS:

Appendix S1: Supplementary information for methods

Trial data extraction and management

All trial registries were extracted in XML format after empty searches from ClinicalTrials.gov and from WHO International Clinical Trials Registry Platform on February 2, 2014. Data was managed using R software.

For each trial we extracted the following fields present in the XML documents:

- Main ID
- Secondary ID
- Primary sponsor
- Sponsor type (only in ClinicalTrials.gov registries)
- Country location(s)
- Start date

Duplicates were identified using the Main ID and Secondary ID fields.

Sponsor classification

Trial sponsors that did not appear in the sponsor list of ClinicalTrials.gov were classified as industry or non-industry if one of the following keywords was included in the sponsor name (case sensitive search).

Keywords for industry sponsors: "Inc", "INC", "LTD", "Limited", "LIMITED", "Ltd", "Co.", "Corporation", "Company", "LLC", "S.A", "A/S", "a.s", "S A", "S. A", "S.p.A", "S.L", "S. L", "SAS", "GmbH", "GMBH", "Pvt", "Private", "Pharma", "pharma", "PHARMA", "ROCHE", "Praxis", "Laborato", "LABORATO", "Plc", "plc", "FARMEX".

Keywords for non-industry sponsors: "institut", "Institut", "INSTITUT", "Univ", "univ", "UNIV", "UMC", "UH", "Uniklinik", "College", "COLLEGE", "college", "World Health Organization", "World health Organization", "NHS", "Public", "UCL", "Hospital", "HOSPITAL", "Hospices", "hospices", "Hôpital", "hospital", "Hopital", "World", "WORLD", "national", "National", "Social", "Zon", "Minist", "Foundation", "Fundación", "FONDAZIONE", "Fondation", "Medical Research Council", "MRC", "School", "Klinik", "ISTITUT", "Istitut", "UZ", "Medical Center", "Medisch Centrum", "Council", "council", "COUNCIL", "Grupo Español", "GRUPO ESPAÑOL", "Grupo de ", "GRUPO DE", "GRUPPO ITALIANO",

"Pública", "Fédération", "Facul", "Research Center", "Research Centre", "research center", "research

centre", "Zentrum", "Wellcome Trust", "medical center", "Medical center".

When the algorithm classified a trial as both industry- and non-industry–sponsored (209/15273 trials), the sponsor was screened and manually classified.

SUPPLEMENTARY TABLES

Table S1: Summary of the number of registered trials initiated in 2006-2013 per million inhabitants per geographical region.

Region	Median	Minimum	Maximum
Africa	2·08	0·05	25·92
South America	6·84	1·26	41·51
Oceania	3·65	1·40	180·20
North America	9·99	1·00	253·60
Western Europe	166·59	33·66	645·70
Eastern Europe	76·24	0·55	415·10
Asia	27·54	3·06	475·20

Table S2: Proportion of industry-sponsored trials per year for each geographical region.

Region	2006	2007	2008	2009	2010	2011	2012
Africa	60·1	54·2	55·9	50·9	42·7	47·6	39·4
South America	59·4	59·3	58·4	52·9	48·9	43·5	51·2
Oceania	73·6	75·8	77·3	76·3	74·2	79·0	76·5
North America	32·0	33·0	34·3	32·5	31·6	34·1	32·8
Western Europe	43·9	39·9	38·6	35·9	33·8	31·0	31·2
Eastern Europe	89·2	85·0	86·0	85·7	85·9	83·2	82·7
Asia	42·5	42·5	40·5	40·5	38·5	38·1	36·4

Table S3: Proportion of industry-sponsored trials per year for each income group.

Income	2006	2007	2008	2009	2010	2011	2012
High income	33·5	33·4	33·3	31·2	30·0	30·0	29·5
Upper middle income	59·3	54·8	51·7	49·4	47·1	44·3	40·3
Low middle income	69·2	69·6	71·0	66·7	64·3	53·3	52·2
Low income	7·3	7·3	5·5	9·9	7·3	11·5	3·3

Table S4: Proportion of industry-sponsored trials for each country.

Country	Ratio
Argentina	0.899
Australia	0.765
Austria	0.626
Bangladesh	0.133
Belarus	0.951
Belgium	0.706
Bosnia and H.	0.913
Brazil	0.438
Bulgaria	0.970
Burkina Faso	0.219
Canada	0.447
Chile	0.811
China	0.319
Colombia	0.835
Costa Rica	0.934
Croatia	0.838
Czech Rep.	0.920
Denmark	0.387
Dominican R.	0.798
Ecuador	0.712
Egypt	0.413
Estonia	0.926
Finland	0.638
France	0.501
Georgia	0.870
Germany	0.626
Ghana	0.200
Greece	0.694
Guatemala	0.865
Hong Kong	0.715
Hungary	0.939
Iceland	0.670
India	0.597
Indonesia	0.649
Iran	0.021
Ireland	0.626
Israel	0.400
Italy	0.584
Japan	0.796
Jordan	0.733
Kenya	0.116
South Korea	0.472
Latvia	0.972
Lebanon	0.802

Country	Ratio
Lithuania	0.933
Macedonia	0.938
Malawi	0.067
Malaysia	0.797
Mali	0.147
Mexico	0.812
Moldova	0.900
Morocco	0.810
Netherlands	0.451
New Zealand	0.820
Nigeria	0.161
Norway	0.382
Pakistan	0.330
Panama	0.819
Peru	0.841
Philippines	0.920
Poland	0.899
Portugal	0.851
Puerto Rico	0.816
Romania	0.954
Russia	0.923
Saudi Arabia	0.455
Serbia	0.918
Singapore	0.647
Slovakia	0.968
Slovenia	0.641
South Africa	0.808
Spain	0.672
Sweden	0.607
Switzerland	0.450
Taiwan	0.436
Tanzania	0.106
Thailand	0.485
Tunisia	0.705
Turkey	0.635
Uganda	0.033
Ukraine	0.972
United Arab Emirates	0.831
U.K.	0.547
United States	0.336
Venezuela	0.875
Vietnam	0.512
Zambia	0.012

Table S5: Proportion of international clinical trials among industry-sponsored trials for each country.

Country	Ratio
Argentina	0.953
Australia	0.867
Austria	0.912
Belarus	0.990
Belgium	0.871
Bosnia and H.	0.947
Brazil	0.777
Bulgaria	0.975
Canada	0.856
Chile	0.966
China	0.406
Colombia	0.965
Costa Rica	0.958
Croatia	0.978
Czech Rep.	0.957
Denmark	0.893
Dominican Rep.	0.866
Egypt	0.848
Estonia	0.990
Finland	0.887
France	0.827
Georgia	0.931
Germany	0.752
Greece	0.920
Guatemala	0.987
Hong Kong	0.967
Hungary	0.951
Iceland	0.898
India	0.621
Indonesia	0.643
Ireland	0.953
Israel	0.728
Italy	0.905
Japan	0.310
South Korea	0.603
Latvia	0.979

Country	Ratio
Lebanon	0.946
Lithuania	0.994
Macedonia	0.921
Malaysia	0.948
Mexico	0.906
Netherlands	0.867
New Zealand	0.927
Norway	0.953
Pakistan	0.825
Panama	0.890
Peru	0.978
Philippines	0.853
Poland	0.965
Portugal	0.972
Puerto Rico	0.993
Romania	0.958
Russia	0.935
Saudi Arabia	0.911
Serbia	0.954
Singapore	0.785
Slovakia	0.960
Slovenia	0.954
South Africa	0.951
Spain	0.876
Sweden	0.854
Switzerland	0.863
Taiwan	0.834
Thailand	0.904
Tunisia	0.899
Turkey	0.928
Ukraine	0.991
United Arab Emirates	0.959
United Kingdom	0.749
United States	0.333
Venezuela	0.934
Vietnam	0.750

Table S6: Proportion of international clinical trials among non-industry-sponsored trials for each country.

Country	Ratio
Argentina	0.482
Australia	0.399
Austria	0.213
Bangladesh	0.103
Belgium	0.259
Brazil	0.070
Burkina Faso	0.316
Canada	0.158
Chile	0.314
China	0.035
Colombia	0.250
Croatia	0.342
Bosnia and H.	0.516
Denmark	0.097
Egypt	0.080
Finland	0.170
France	0.086
Germany	0.143
Ghana	0.339
Greece	0.210
Hong Kong	0.125
Hungary	0.609
India	0.125
Indonesia	0.397
Iran	0.006
Ireland	0.375
Israel	0.054
Italy	0.157
Japan	0.103
Kenya	0.337
Czech Rep.	0.029

Country	Ratio
Malawi	0.429
Malaysia	0.321
Mali	0.241
Mexico	0.157
Netherlands	0.112
New Zealand	0.778
Norway	0.115
Pakistan	0.172
Peru	0.563
Poland	0.388
Portugal	0.558
Puerto Rico	0.863
Romania	0.571
Dominican Rep.	0.333
Saudi Arabia	0.216
Singapore	0.152
Slovenia	0.373
South Africa	0.486
Spain	0.138
Sweden	0.191
Switzerland	0.188
Taiwan	0.029
Tanzania	0.339
Thailand	0.124
Turkey	0.071
Uganda	0.327
United Kingdom	0.109
United States	0.035
Vietnam	0.300
Zambia	0.494

Table S7: Distribution of country trial location of industry-sponsored trial over geographical regions per year.

Region	2006	2007	2008	2009	2010	2011	2012
Africa	0.019	0.015	0.015	0.015	0.013	0.017	0.015
South America	0.045	0.045	0.043	0.043	0.039	0.044	0.041
Oceania	0.030	0.027	0.027	0.027	0.028	0.028	0.031
North America	0.227	0.237	0.232	0.227	0.222	0.230	0.227
Western Europe	0.423	0.401	0.401	0.394	0.390	0.362	0.372
Eastern Europe	0.148	0.151	0.156	0.152	0.163	0.172	0.163
Asia	0.108	0.124	0.127	0.142	0.145	0.146	0.152

Table S8: Distribution of country trial location of non-industry-sponsored trial over geographical regions per year.

Region	2006	2007	2008	2009	2010	2011	2012
Africa	0.021	0.019	0.015	0.020	0.021	0.021	0.024
South America	0.024	0.023	0.025	0.027	0.027	0.035	0.023
Oceania	0.014	0.011	0.010	0.009	0.010	0.007	0.009
North America	0.531	0.494	0.477	0.450	0.433	0.389	0.397
Western Europe	0.284	0.317	0.313	0.333	0.335	0.362	0.353
Eastern Europe	0.011	0.014	0.017	0.014	0.013	0.017	0.015
Asia	0.115	0.122	0.144	0.146	0.162	0.168	0.178

Table S9: Distribution of country trial location of industry-sponsored trial over income groups per year.

Income	2006	2007	2008	2009	2010	2011	2012
United States	0.160	0.174	0.169	0.164	0.162	0.165	0.162
High income	0.662	0.645	0.646	0.646	0.653	0.635	0.647
Upper-middle income	0.141	0.140	0.144	0.150	0.142	0.157	0.150
Lower-middle income	0.036	0.040	0.040	0.039	0.041	0.041	0.042
Low income	0.002	0.001	0.001	0.001	0.001	0.002	0.001

Table S10: Distribution of country trial location of non-industry-sponsored trial over income groups per year.

Income	2006	2007	2008	2009	2010	2011	2012
United States	0.464	0.429	0.414	0.395	0.374	0.333	0.334
High income	0.442	0.474	0.476	0.489	0.503	0.521	0.516
Upper-middle income	0.061	0.067	0.083	0.086	0.088	0.101	0.102
Lower-middle income	0.018	0.019	0.017	0.019	0.021	0.031	0.033
Low income	0.016	0.011	0.009	0.011	0.013	0.013	0.015

Figure S1: Flowchart.

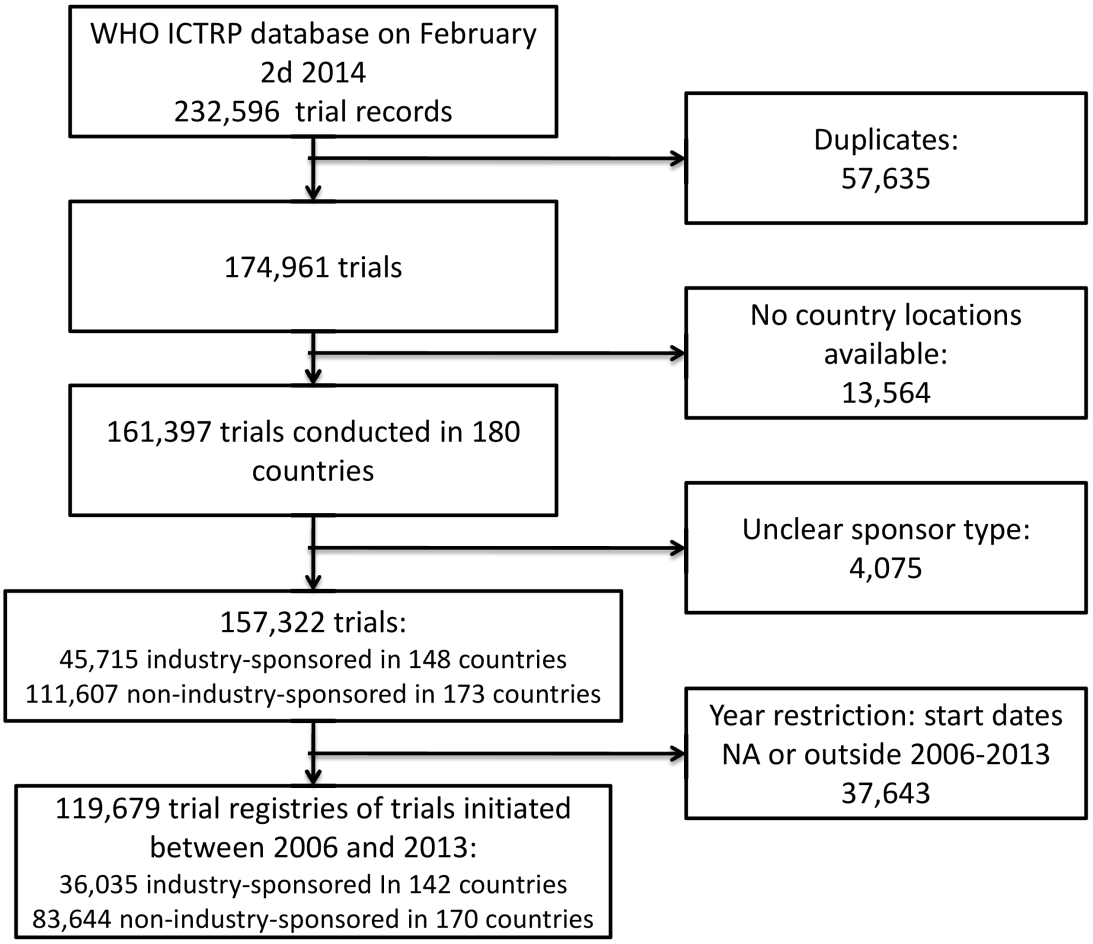
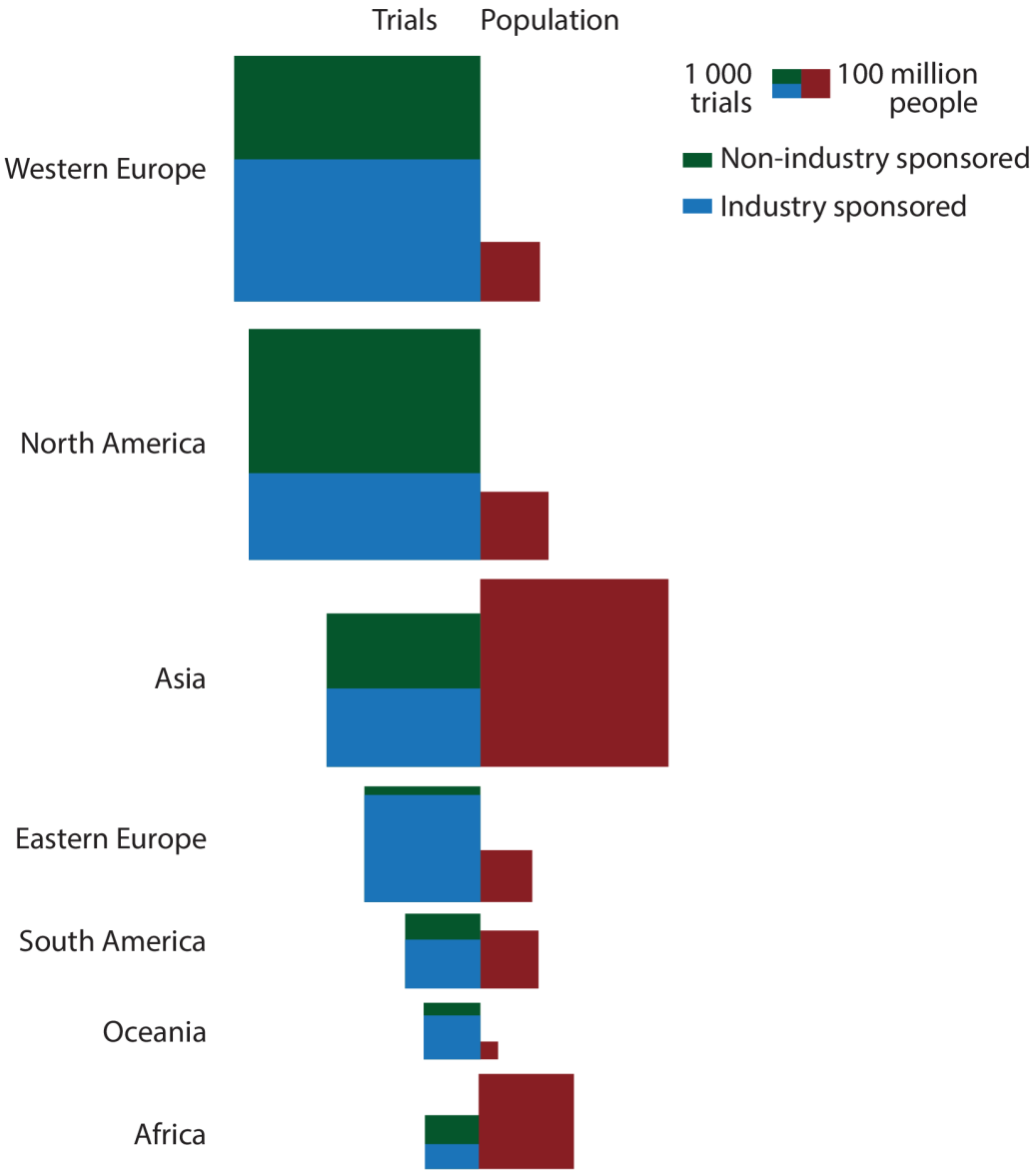
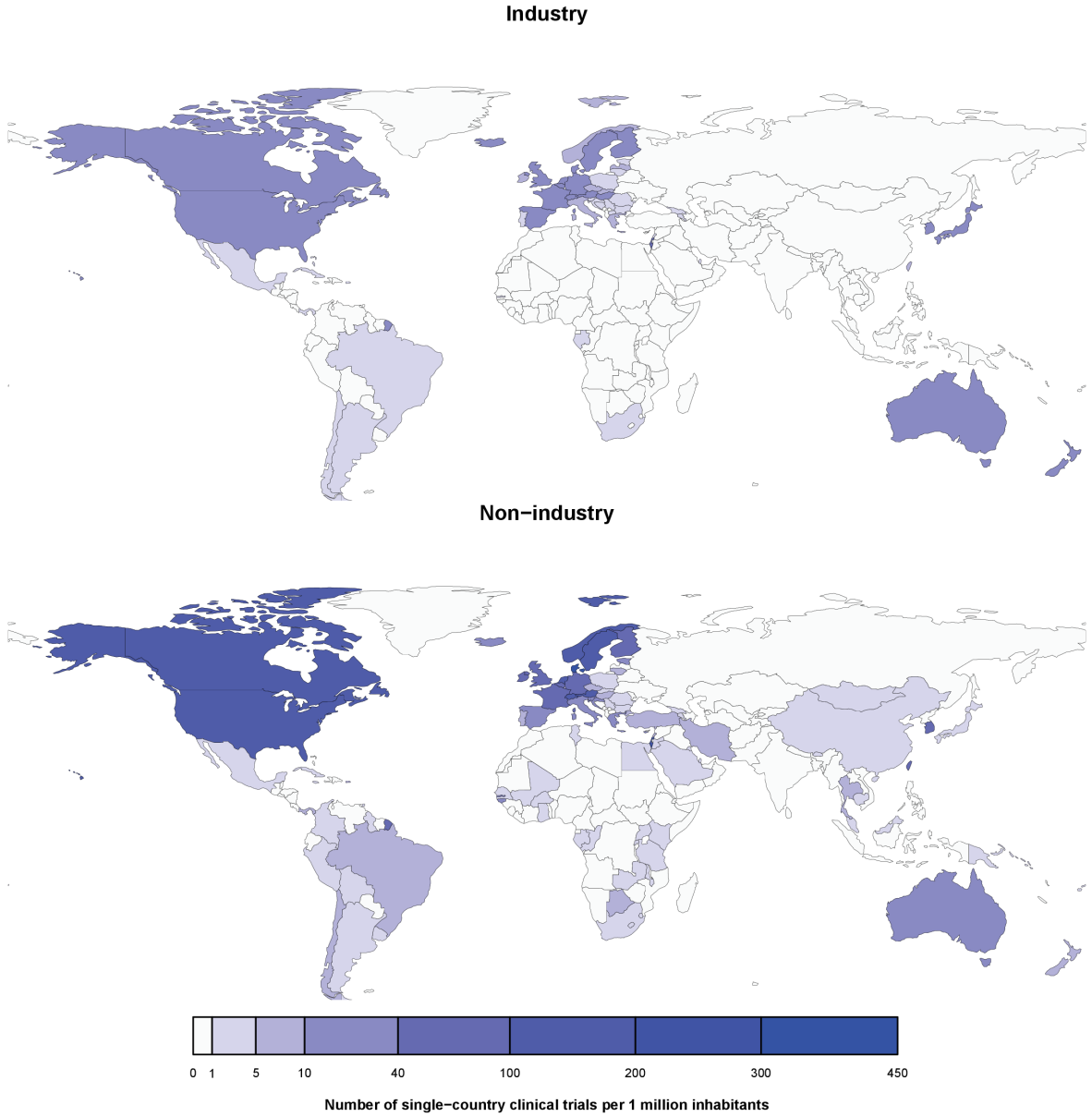


Figure S2: Distribution of clinical trials and population per geographical regions.



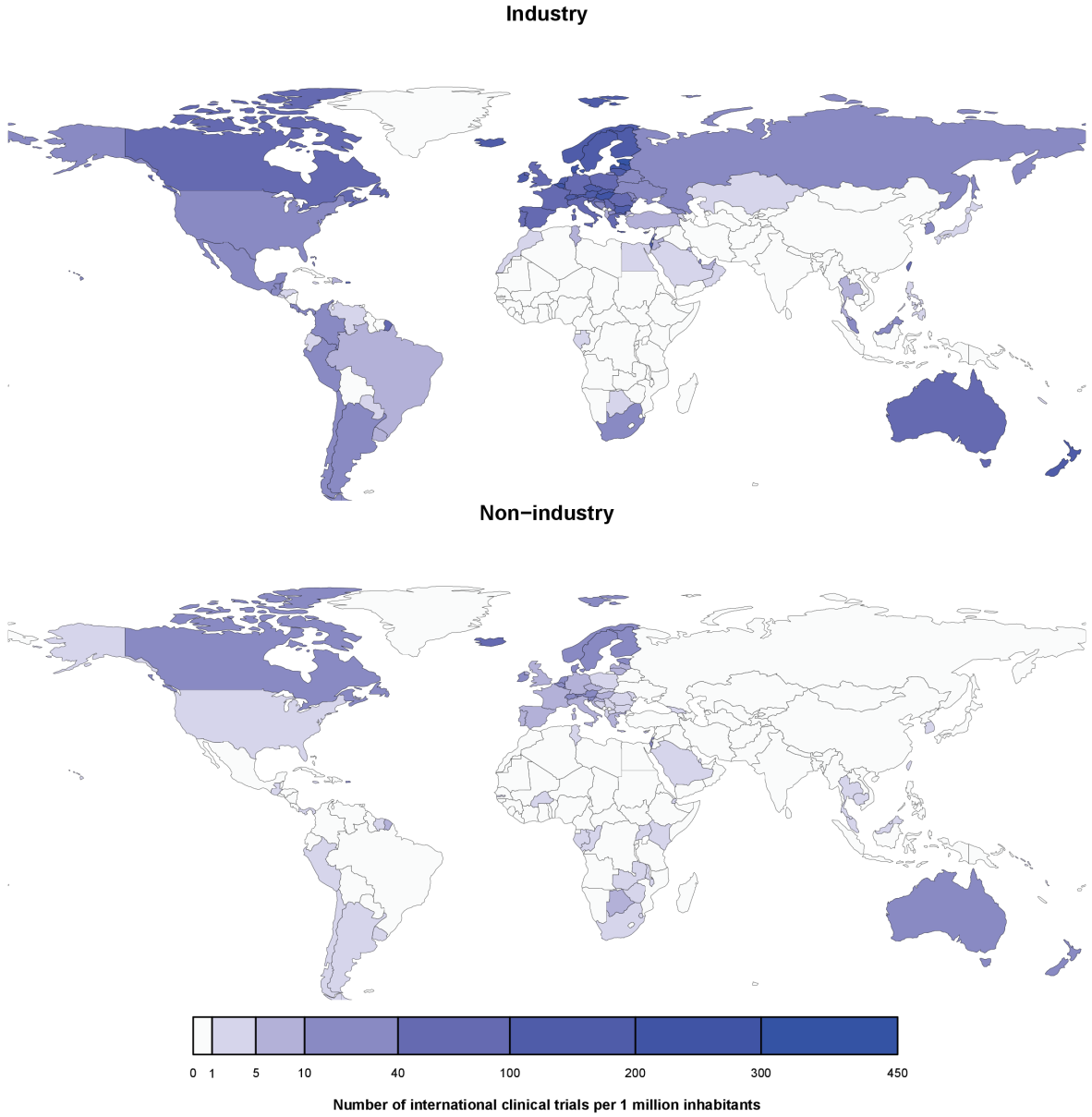
For each geographic region, the size of the green (blue, respectively) area is proportional to the number of industry- (non-industry–, respectively) sponsored trials initiated during the 2006-2013 period, and the size of the red area is proportional to the population as of 2012. Equal sized trial and population squares correspond to an overall density of 10 trials per million inhabitants. The proportion of industry-sponsored clinical trials was 57.0%, 37.5%, 51.0%, 92.5%, 65.4%, 77.2% and 45.8% in Western Europe, North America, Asia, Eastern Europe, South America, Oceania and Africa, respectively.

Figure S3: Mapping of single-country for industry- and non-industry–sponsored clinical trials.



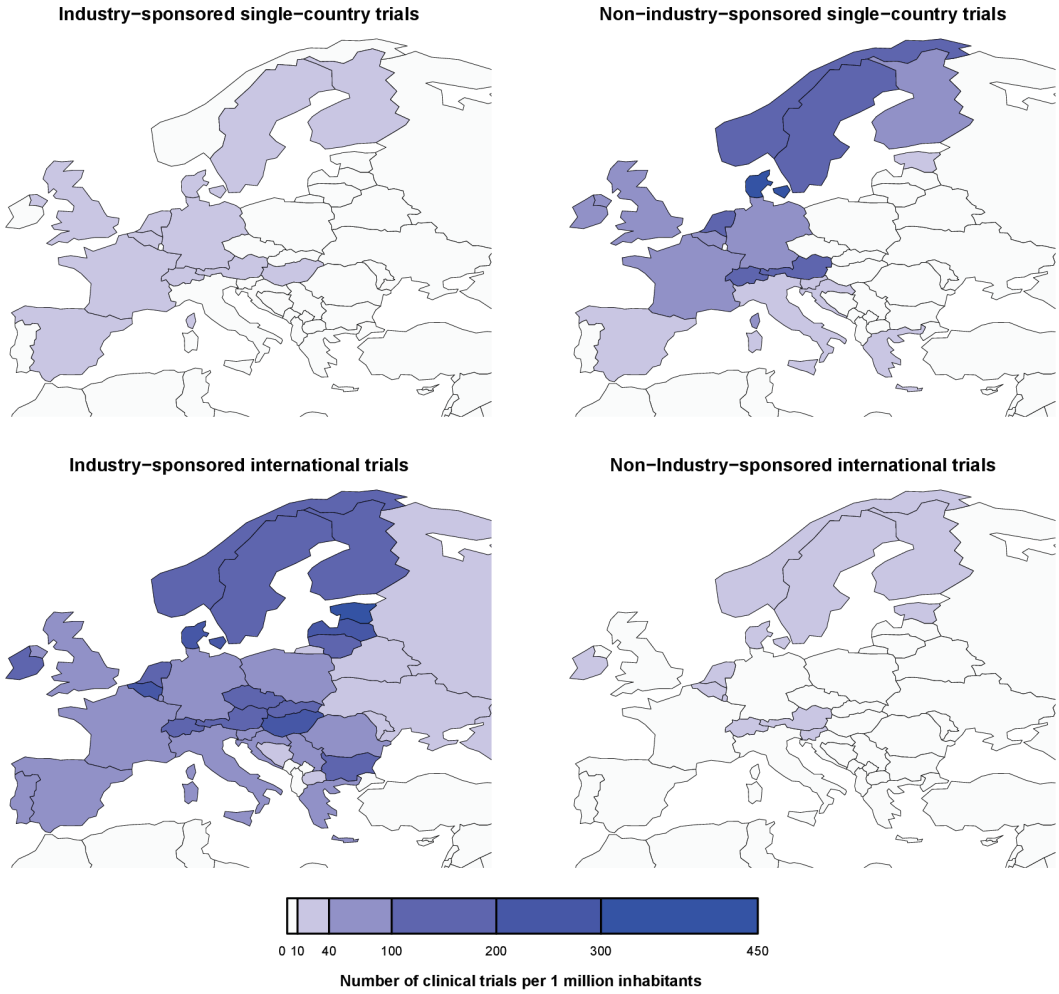
The number of single-country clinical trials per million inhabitants for industry-sponsored (top) and non-industry–sponsored (bottom) research for registered trials initiated between 2006 and 2013.

Figure S4: Mapping of international trials for industry- and non-industry–sponsored clinical trials.



The number of international clinical trials per million inhabitants for industry-sponsored (top) and non-industry–sponsored (bottom) research for registered trials initiated between 2006 and 2013.

Figure S5: Mapping of single-country and international clinical trials for industry- and non-industry–sponsored clinical trials in Europe.



The number of single-country (top) and international (bottom) clinical trials per million inhabitants for industry-sponsored (left) and non-industry–sponsored (rigth) research for registered trials initi- ated between 2006 and 2013 in Europe.

Automatic classification of registered clinical trials to the Global Burden of Diseases taxonomy of diseases and injuries

Ignacio ATAL, Jean-David Zeitoun, Aurélie NÉVÉOL, Philippe RAVAUD,

Raphaël PORCHER, Ludovic TRINQUART

Supplementary Information

TABLE OF CONTENTS

APPENDIX S1: Supplementary information for methods

MetaMap implementation

We used the MetaMap version metamap14, the lexicon version used was 2014 and the database used was USAbase 2014AB.

We restricted the output of MetaMap to concepts included in the semantic group DISORDERS, with the exception of concepts of semantic type “Findings”. We considered the following semantic types:

- Acquired Abnormality
- Anatomical Abnormality
- Cell or Molecular Dysfunction
- Congenital Abnormality
- Disease or Syndrome
- Experimental Model of Disease
- Injury or Poisoning
- Mental or Behavioral Dysfunction
- Neoplastic Process
- Pathologic Function
- Sign or Symptom

We used the strict-model parametrization of MetaMap.

We developed the classifier using and not the Word Sense Disambiguation server.

The metamap options used were: -J acab,anab,comd,cgab,dsyn,emod,inpo,mobd,neop,patf,sosy -AvN -V USAbase (-y)

For a high proportion of clinical trial records, the health condition field corresponds to a list of diseases. We indexed separately each item of the list. For the public title and scientific title, the entire text was indexed.

IntraMap implementation

We used the IntraMap version 2014 based on the USAbase 2014AB.

For each UMLS concept, the IntraMap output is a list of ICD10 codes. This list can be found by four different means: by synonymy, by built-in relations, through the graph of ancestors and based on other mappings. Mappings through synonymy are expected to have the highest quality, followed by mappings through built-in relations. Follows mappings through the graph of ancestors, which generally mapped to a more general concept. We excluded mappings using other means than the three aforementioned.

In some cases, IntraMap's output is empty. For instance, concepts C0012634 “Disease”, or C0009566 “Complication”, could not be projected by IntraMap to ICD10 codes.

Punctual reparations to MetaMap and IntraMap

During the development of the classifier, we identified several errors of MetaMap or IntraMap leading to misclassification. These reparations are explained bellow:

- IntraMap projected the UMLS concept C1306459, “Primary malignant neoplasms”, to the ICD10 code C72.9, “Malignant neoplasm of central nervous system, unspecified”. We changed this projection to C00-C97, “Malignant neoplasms”.

- IntraMap projected the UMLS concept C0876994, “Cardiotoxicity” to the ICD10 code M10.2, “Drug-induced gout”. We suppressed this projection and leaved the concept without projection.

- IntraMap projected the UMLS concept C2825032, “Withdrawal (dysfunction)” to the ICD10 code F91.2 “Socialized conduct disorder”, leading to the GBD category “Mental and behavioral disorders”. However, the annotation from MetaMap using the “Withdrawal (dysfunction)” UMLS concept arrived when recognizing in text the term “withdrawal”, generally corresponding to a specification of the design of the study rather than the health condition studied. We suppressed this projection and leaved the concept without projection.

- The UMLS concept C0949179, “Edentulism” could not be projected to any ICD10 code. However, the GBD study reserves a special GBD category to that health condition. We projected that ICD10 code directly to the “Oral disorders” category.

- IntraMap projected the UMLS concept C0878773, “Overactive Bladder” to several ICD10 codes, among which G00-G99.9, “Diseases of the nervous system”. We suppressed the projection to that ICD10 code.
- IntraMap projected the UMLS concept C0242656, “Disease Progression” to the ICD10 code C00-D48.9, “Neoplasms”. We suppressed this projection and leaved the concept without projection.
- IntraMap projected the UMLS concept C0687702, “Cancer Remission” to the ICD10 code C00-D48.9, “Neoplasms”. However, the annotation from MetaMap using the “Cancer Remission” UMLS concept arrived when recognizing in text the term “remission”, which could correspond to the remission of any disease other than cancer. We suppressed this projection and leaved the concept without projection.

ICD10 codes manipulation

Each ICD10 code output from IntraMap was projected to one or several GBD categories. These ICD10 codes could correspond to different levels of the ICD10 taxonomy. We could have ICD10 codes corresponding to groups of codes (e.g. F00-F99.9: “Mental and behavioral disorders”), or single codes (e.g. A00 or A00.1). We developed an algorithm for projecting every ICD10 code to one or several groups of GBD categories.

First, ICD10 codes corresponding to groups of codes were transformed into lists of single codes. For instance, the ICD10 code C00-D48.9: “Neoplasms” was transformed to the list C00, C01, … C99, D00, D01, … D47, D47.1, … D47.9.

Second, each single code was projected to a *unique* GBD category if possible. The projection was conducted only if we had 100% certainty that that single ICD10 code was included in a unique GBD category. For instance, the code E11 (“Non-insulin dependent diabetes mellitus”) could not be assigned to a unique GBD category among the 171 GBD categories because the GBD category “Diabetes Mellitus” included the ICD10 code E11 except E11.2 (“Non-insulin dependent diabetes mellitus with renal complications”), which was included in the GBD category “Chronic kidney diseases”.

Finally, the total ICD10 code was projected to the group of GBD categories corresponding to the projection of its corresponding single ICD10 codes.

Prioritization rules

Prioritization rules corresponded to a set of algorithms to derive the GBD classification of a trial based on the pathways issued from the trial record to candidate GBD categories. Pathways from the trial record to candidate GBD categories were derived from successive projections: from the trial record to free text, from free text to UMLS concepts, from UMLS concepts to ICD10 codes, and from ICD10 codes to GBD categories. The rules of prioritization gave priority to candidate GBD categories consistently achieved by the pathways, as compared to candidate GBD categories achieved by isolated pathways. At each projection, noise GBD categories may appear. For instance, noise candidate GBD categories may appear when annotating the scientific title of the trial using a UMLS concept revealing of the design of the study rather than the health condition studied. Similarly, the UMLS concept “Breast neoplasms” may be projected to both ICD10 codes “Malignant neoplasms” and “Malignant breast neoplasms”, leading to an UMLS concept having as candidate GBD categories all the 27/171 GBD categories corresponding to cancers.

Prioritization rules were then used at different stages of the projections. First, to deriving a projection of each UMLS concept to GBD categories without noise, based on the projection to GBD categories of the ICD10 codes corresponding to that UMLS concept. Second, to deriving a projection of each text field (health condition, public title and scientific title) without noise based on the projection to GBD categories of the UMLS concepts used for annotating each text. Third and finally, to deriving a projection of the trial record based on the projection to GBD categories of each of its text fields.

The first two utilizations of the prioritization rules used the same algorithm described bellow. The final prioritization rule used an algorithm giving particular priority to the projection to GBD categories of the health condition field as compared to the public and scientific titles.

The algorithm used for the first two stages of projection was based on gaving priority to smaller lists and to intersections of GBD categories between lists. Given a set E of lists of GBD categories (e.g. the list of candidate GBD categories per UMLS concept annotating a given text field), the algorithm derived a final list L of GBD categories (e.g. corresponding to the GBD classification of the text field) by adding progressively to L eligible GBD categories. We initialized the list of GBD categories L with all the elements of E of size k=1. Then, for k>1 in increasing order:

1. We considered E_k as the set of lists of E of size k.

2. We suppressed from *E_k* all the lists having a non-empty intersection with *L*.
3. We considered *I_k* as the global intersection of all the lists of *E_k*. If *I_k* was non empty, we added *I_k* to *L*. If *I_k* was empty, we added to *L* all the elements included in the lists of *E_k*.

The algorithm used for the final prioritization rule to derive the GBD classification of the trial record based on the projection to GBD categories of the corresponding text fields was as follows:

1. If the condition field had a unique candidate GBD category, we considered it as the GBD classification of the trial.
2. If the condition field had multiple candidate GBD categories, we tested if the intersection of these categories with those appearing in the other text fields was non-empty. If it was non-empty, the GBD classification of the trail was the aforementioned intersection. If it was empty, the GBD classification of the trial was the set of candidate GBD categories of the condition field.
3. If the condition field did not have any candidate GBD category, we derived the GBD classification by considering the candidate GBD categories from the public title and the scientific title fields. We gave priority to the GBD categories appearing in both text fields, and if the intersection was empty, to the union.

When we suppressed the priority to the condition field from the classifier, the final GBD classification of the trial was derived from GBD classifications of each text field using the same algorithm as in the first two prioritization rules.

Sample of clinical trials

For trials classified in more than one data source, we gave priority to the classification from the ESORT study, then from *Viergever et al*, and finally from the ongoing study from our team.

From the ESORT study, we had a manual classification of 519 unique trials having valid registration number and included in ICTRP. In the study from *Viergever et al.,* trial records were classified using the GBD categories as defined in the GBD study conducted by the World Health Organization (WHO) (Table C3 in *The Global Burden of Disease: 2004 update (2008))*. The GBD categories defined by the WHO may differ with the GBD cause list

from the GBD 2010 study conducted by the Institute for Health Metrics and Evaluation (IHME). In the GBD study from the WHO, GBD categories are also defined using ICD10 codes. We identified the GBD categories from the WHO for which all ICD10 codes were included in a unique GBD category among the 171 groups of GBD categories we defined from the GBD cause list of the IHME. We excluded all trials classified using a GBD category from the WHO for which the ICD10 codes could be projected to two or more GBD categories. For instance, the GBD category from the WHO “Melanoma and other skin cancers” (C43-C44) could correspond to the GBD categories from the IHME “Malignant melanoma of skin” (C43,D03,D48.5) or “Non-melanoma skin cancer” (C44,D04). From that sample, we did not exclude trials classified as “No GBD” category, as we considered they corresponded to trials without GBD category for both taxonomies, WHO and IHME. We excluded 1,105 trials. We also excluded 5 trials already classified during the ESORT study. From the manual classification of the researcher of our team, we had 1,001 trials. In total, 28 trials were classified in the other data sources and were excluded. In total, we disposed of 2,381 unique clinical trials.

During the ESORT study, trials were classified using the 193 categories from the GBD cause list. In the study from *Viergever et al.* trials were reclassified to the 171 GBD categories defined in our study using the aforementioned rule. The physician from our team classified trials directly using the 171 GBD categories defined in our study.

For each trial we identified the GBD categories corresponding to the classification among the 28 and 171 GBD categories. For instance, when a trial from the ESORT study was classified using a residual category excluded from the 28 or 171 GBD categories, we classified it as “No GBD” category.

APPENDIX S2: Results for the classification to the 171 GBD categories

Across 2,763 trial records, 2,092 (75.7%) concerned a unique GBD category, 187 (6.8%) concerned two or more GBD categories, and 484 (17.5%) concerned health conditions from the residual categories that we excluded or health conditions not relevant for the GBD 2010 study. A majority of clinical trials studied the “Breast cancer” category (232 trials), followed by “Leukemia” (225 trials) and “Diabetes mellitus” (202 trials) (Table S2). In our sample, trials concerned 128/171 GBD categories.

Process of classification of trials

The stages of text annotation and projection of UMLS concepts to ICD10 codes are identical than for the classification to 28 GBD categories.

In total, 623/1361 (45.8%) ICD10 codes were projected to at least one GBD category. The median (Q1, Q3) number of GBD categories per projected ICD10 code was 1 (1, 1).

At this stage, 965/2180 (44.2%) UMLS concepts could not be projected to a GBD category. In the expert-based enrichment database we found manual revision for 403/965 (41.8%) not projected UMLS concepts, over which 68 were manually projected to a GBD category.

Evaluation of the classifier

Overall performances

The performances of classification of the 2,763 trial records for the eight versions of the classifier are shown in Table S3. The unique option of the versions of the classifier that substantially improved the performances of classification was the use of the expert-based enrichment database (between 6 and 7% of improvement). The best performances (74.0% of exact-matching) were achieved using the WSD server, the expert-based enrichment database and giving priority to the health condition field. The exact-matching was higher for trials concerning a unique GBD category (77.4%) and was the lowest for trials concerning two or more GBD categories (43.3%).

Performances for each GBD category

The performances for each GBD category for the version of the classifier using the WSD, the expert-based enrichment and the priority to the condition are shown in Table S2.

For all GBD categories with a sufficient high number of trials, the specificity was consistently high (more than 50 trials). This means that the classifier generally does not underestimate the effort of research for these GBD categories. The 27 types of cancers among the 171 GBD categories have generally a low sensitivity, and are studied by a similar amount of trials. This is because trials in the sample we considered 95 trials were classified as “Neoplasms”, meaning that they were classified using all the 27 types of cancers among the 171 categories. The low sensitivity of these categories may be explained because some of the trials concerning all cancers may be classified using specific cancers.

SUPPLEMENTARY TABLES

Table S1: Excluded residual GBD categories for the grouping of the GBD cause list in 171 GBD categories

Excluded GBD categories
Other neglected tropical diseases
Other neonatal disorders
Other nutritional deficiencies
Other sexually transmitted diseases
Other infectious diseases
Other neoplasms
Other cardiovascular and circulatory diseases
Other chronic respiratory diseases
Other digestive diseases
Other neurological disorders
Other mental and behavioral disorders
Other urinary diseases
Other gynecological diseases
Other hemoglobinopathies and hemolytic anemias
Other endocrine, nutritional, blood, and immune disorders
Other muskuloskeletal disorders
Other skin and subcutaneous diseases
Other hearing loss
Other vision loss
Other sense organ diseases
Other transport injury
Unintentional injuries not classified elsewhere

A grouping of 193 GBD categories was defined during the GBD 2010 study to inform policy makers on the main health problems per country. From these 193 GBD categories, we excluded the 22 residual categories listed in the Table. We developed a classifier to the remaining 171 GBD categories. Among these residual categories, the unique excluded categories in the grouping of 28 GBD categories were “Other infectious diseases” and “Other endocrine, nutritional, blood, and immune disorders”.

Table S2: Per category evaluation performances of the classifier for the 171 GBD categories plus the "No GBD" category

GBD categories	Number of trials	Sensitivity	Specificity	Positive Likelihood Ratio	Negative Likelihood Ratio
Tuberculosis	16	87.5 [71.9-88.5]	99.9 [99.8-99.9]	1201.8 [297.0-4862.5]	0.13 [0.03-0.46]
HIV/AIDS	97	88.7 [83.9-90.4]	99.7 [99.5-99.7]	295.5 [147.4-592.3]	0.11 [0.07-0.20]
Diarrhea, lower respiratory infections, meningitis, and other common infectious diseases					
Diarrheal diseases	1	100.0 [20.7-100.0]	99.9 [99.8-99.9]	1381.0 [345.6-5519.1]	NaN
Typhoid and paratyphoid fevers	0	NaN	100.0 [99.9-100.0]	NaN	NaN [NaN-NaN]
Lower respiratory infections	25	84.0 [72.3-86.7]	99.5 [99.3-99.6]	176.9 [100.2-312.4]	0.16 [0.07-0.39]
Upper respiratory infections	13	53.8 [40.8-65.1]	99.9 [99.7-99.9]	493.6 [143.1-1702.1]	0.46 [0.26-0.83]
Otitis media	1	100.0 [20.7-100.0]	100.0 [99.9-100.0]	NaN	NaN
Meningitis	4	100.0 [51.0-100.0]	99.3 [99.0-99.4]	138.0 [89.1-213.5]	NaN
Encephalitis	0	NaN	99.3 [99.1-99.4]	NaN	NaN [NaN-NaN]
Diphtheria	1	100.0 [20.7-100.0]	99.9 [99.8-99.9]	1381.0 [345.6-5519.1]	NaN
Whooping cough	0	NaN	99.9 [99.8-99.9]	NaN	NaN [NaN-NaN]
Tetanus	2	100.0 [34.2-100.0]	99.9 [99.8-99.9]	1380.5 [345.4-5517.1]	NaN
Measles	2	50.0 [29.3-70.7]	100.0 [99.9-100.0]	NaN	0.50 [0.13-2.00]
Varicella	0	NaN	100.0 [99.9-100.0]	NaN	NaN [NaN-NaN]
Malaria	14	100.0 [78.5-100.0]	100.0 [99.8-99.9]	2749.0 [387.4-19508.4]	NaN
Neglected tropical diseases excluding malaria					
Chagas disease	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Leishmaniasis	0	NaN	100.0 [99.9-100.0]	NaN	NaN
African trypanosomiasis	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Schistosomiasis	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Cysticercosis	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Echinococcosis	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Lymphatic filariasis	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Onchocerciasis	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Trachoma	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Dengue	3	100.0 [43.9-100.0]	100.0 [99.9-100.0]	NaN	NaN
Yellow fever	1	100.0 [20.7-100.0]	100.0 [99.9-100.0]	NaN	NaN
Rabies	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Ascariasis	1	0.0 [0.0-79.3]	100.0 [99.8-99.9]	NaN	NaN
Trichuriasis	0	NaN	100.0 [99.8-99.9]	NaN	NaN

Hookworm disease	1	100.0 [20.7-100.0]	100.0 [99.9-100.0]	NaN	NaN
Food-borne trematodiases	1	100.0 [20.7-100.0]	100.0 [99.8-99.9]	2762.0 [389.2-19600.7]	NaN
Maternal disorders	43	39.5 [33.2-47.6]	99.8 [99.6-99.8]	179.2 [74.3-432.4]	0.61 [0.48-0.77]
Neonatal disorders					
Preterm birth complications	8	50.0 [35.5-64.5]	100.0 [99.8-99.9]	1377.5 [172.3-11009.9]	0.50 [0.25-1.00]
Neonatal encephalopathy (birth asphyxia and birth trauma)	1	0.0 [0.0-79.3]	100.0 [99.9-100.0]	NaN	NaN
Sepsis and other infectious disorders of the newborn baby	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Nutritional deficiencies					
Protein-energy malnutrition	3	33.3 [24.0-61.3]	99.8 [99.7-99.8]	184.0 [29.7-1140.5]	0.67 [0.30-1.49]
Iodine deficiency	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Vitamin A deficiency	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Iron-deficiency anemia	9	88.9 [66.7-87.8]	99.6 [99.4-99.7]	244.8 [126.5-473.8]	0.11 [0.02-0.71]
Sexually transmitted diseases excluding HIV					
Syphilis	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Sexually transmitted chlamydial diseases	1	0.0 [0.0-79.3]	100.0 [99.8-99.9]	NaN	NaN
Gonococcal infection	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Trichomoniasis	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Hepatitis					
Acute hepatitis A	17	47.1 [36.7-58.5]	100.0 [99.8-99.9]	1292.2 [170.8-9774.8]	0.53 [0.34-0.83]
Acute hepatitis B	17	70.6 [56.6-77.0]	99.9 [99.7-99.9]	484.6 [173.7-1352.2]	0.29 [0.14-0.62]
Acute hepatitis C	17	0.0 [0.0-18.4]	100.0 [99.9-100.0]	NaN	NaN
Acute hepatitis E	17	0.0 [0.0-18.4]	100.0 [99.9-100.0]	NaN	NaN
Leprosy	2	100.0 [34.2-100.0]	100.0 [99.8-99.9]	2761.0 [389.1-19593.6]	NaN
Neoplasms					
Esophageal cancer	111	73.0 [68.0-76.4]	98.2 [97.9-98.4]	41.2 [30.3-55.9]	0.28 [0.20-0.37]
Stomach cancer	135	77.8 [73.5-80.6]	98.6 [98.3-98.8]	55.2 [39.6-77.0]	0.23 [0.16-0.31]
Liver cancer	169	81.1 [77.4-83.4]	98.1 [97.8-98.3]	42.9 [32.2-57.2]	0.19 [0.14-0.26]
Larynx cancer	103	71.8 [66.7-75.4]	98.2 [97.8-98.4]	39.0 [28.8-52.8]	0.29 [0.21-0.39]
Trachea, bronchus, and lung cancers	190	83.7 [80.3-85.7]	98.3 [98.0-98.5]	50.1 [37.0-67.8]	0.17 [0.12-0.23]
Breast cancer	232	84.9 [82.0-86.7]	98.2 [97.9-98.4]	47.8 [35.6-64.1]	0.15 [0.11-0.21]
Cervical cancer	110	71.8 [66.8-75.3]	98.3 [98.0-98.5]	42.3 [31.0-57.9]	0.29 [0.21-0.39]
Uterine cancer	97	70.1 [64.8-73.9]	98.3 [98.0-98.5]	40.6 [29.7-55.6]	0.30 [0.22-0.41]
Prostate cancer	151	80.1 [76.2-82.6]	98.3 [98.0-98.5]	47.6 [35.1-64.4]	0.20 [0.15-0.28]
Colon and rectum cancers	165	80.0 [76.2-82.4]	98.1 [97.7-98.3]	41.6 [31.3-55.3]	0.20 [0.15-0.28]
Mouth cancer	104	72.1 [67.0-75.7]	98.2 [97.8-98.4]	39.1 [28.9-52.9]	0.28 [0.21-0.39]

Nasopharynx cancer	99	68.7 [63.4-72.6]	98.3 [97.9-98.5]	39.8 [29.0-54.6]	0.32 [0.24-0.43]
Cancer of other part of pharynx and oropharynx	103	72.8 [67.7-76.3]	98.2 [97.9-98.4]	40.4 [29.8-54.7]	0.28 [0.20-0.38]
Gallbladder and biliary tract cancer	100	70.0 [64.7-73.8]	98.3 [97.9-98.5]	40.5 [29.6-55.5]	0.31 [0.23-0.41]
Pancreatic cancer	127	75.6 [71.1-78.6]	98.2 [97.9-98.4]	42.4 [31.4-57.2]	0.25 [0.18-0.34]
Malignant melanoma of skin	104	76.0 [70.9-79.2]	98.1 [97.8-98.3]	40.4 [30.1-54.3]	0.24 [0.17-0.34]
Non-melanoma skin cancer	96	69.8 [64.4-73.6]	98.3 [98.0-98.5]	40.5 [29.5-55.5]	0.31 [0.23-0.42]
Ovarian cancer	122	74.6 [69.9-77.7]	98.3 [97.9-98.4]	42.8 [31.6-58.1]	0.26 [0.19-0.35]
Testicular cancer	96	68.8 [63.4-72.7]	98.3 [98.0-98.5]	40.7 [29.6-56.1]	0.32 [0.24-0.43]
Kidney and other urinary organ cancers	133	78.9 [74.6-81.6]	98.3 [98.0-98.5]	46.1 [34.1-62.4]	0.21 [0.15-0.30]
Bladder cancer	111	73.0 [68.0-76.4]	98.3 [97.9-98.5]	42.1 [30.9-57.2]	0.28 [0.20-0.37]
Brain and nervous system cancers	111	82.9 [78.2-85.3]	98.0 [97.6-98.2]	40.7 [30.9-53.7]	0.17 [0.12-0.26]
Thyroid cancer	102	69.6 [64.4-73.4]	98.3 [97.9-98.5]	40.3 [29.4-55.1]	0.31 [0.23-0.41]
Hodgkin's disease	98	69.4 [64.1-73.2]	98.0 [97.6-98.2]	34.2 [25.5-46.0]	0.31 [0.23-0.42]
Non-Hodgkin lymphoma	145	76.6 [72.4-79.4]	97.6 [97.3-97.9]	32.3 [24.9-42.0]	0.24 [0.18-0.32]
Multiple myeloma	138	77.5 [73.3-80.3]	97.9 [97.5-98.1]	36.3 [27.6-47.8]	0.23 [0.17-0.31]
Leukemia	225	83.6 [80.5-85.5]	98.1 [97.7-98.3]	43.3 [32.6-57.5]	0.17 [0.12-0.23]
Cardiovascular and circulatory diseases					
Rheumatic heart disease	1	0.0 [0.0-79.3]	99.6 [99.4-99.7]	NaN	NaN
Ischemic heart disease	191	77.0 [73.4-79.5]	99.5 [99.2-99.5]	141.4 [83.4-239.8]	0.23 [0.18-0.30]
Cerebrovascular disease	14	85.7 [68.9-87.2]	99.3 [99.1-99.4]	124.0 [75.5-203.8]	0.14 [0.04-0.52]
Hypertensive heart disease	2	0.0 [0.0-65.8]	98.8 [98.5-98.9]	NaN	NaN
Cardiomyopathy and myocarditis	10	20.0 [16.8-39.9]	99.5 [99.3-99.6]	42.4 [10.9-163.9]	0.80 [0.59-1.10]
Atrial fibrillation and flutter	10	70.0 [51.8-77.1]	99.6 [99.4-99.7]	175.2 [85.6-358.4]	0.30 [0.12-0.78]
Aortic aneurysm	2	50.0 [29.3-70.7]	99.5 [99.3-99.6]	106.2 [24.0-470.4]	0.50 [0.13-2.01]
Peripheral vascular disease	9	66.7 [48.3-75.1]	99.5 [99.3-99.6]	141.2 [69.3-288.0]	0.33 [0.13-0.84]
Endocarditis	1	0.0 [0.0-79.3]	99.6 [99.4-99.7]	NaN	NaN
Chronic respiratory diseases					
Chronic obstructive pulmonary disease	34	70.6 [61.0-76.0]	99.7 [99.6-99.8]	275.2 [127.3-595.0]	0.18 [0.07-0.43]
Pneumoconiosis	0	NaN	99.9 [99.7-99.9]	NaN	0.18 [0.07-0.43]
Asthma	40	92.5 [84.4-93.2]	99.9 [99.7-99.9]	839.6 [270.0-2610.5]	0.18 [0.07-0.43]
Interstitial lung disease and pulmonary sarcoidosis	0	NaN	99.8 [99.6-99.8]	NaN	0.18 [0.07-0.43]
Cirrhosis of the liver	23	82.6 [70.2-85.6]	98.9 [98.7-99.1]	78.1 [51.9-117.3]	0.18 [0.07-0.43]
Digestive diseases (except cirrhosis)					
Peptic ulcer disease	5	40.0 [27.7-61.0]	99.8 [99.7-99.8]	220.6 [55.2-881.8]	0.60 [0.29-1.23]
Gastritis and duodenitis	1	100.0 [20.7-100.0]	99.8 [99.7-99.8]	552.4 [230.1-1326.1]	NaN
Appendicitis	2	100.0 [34.2-100.0]	99.8 [99.6-99.8]	460.2 [206.9-1023.4]	NaN

Paralytic ileus and intestinal obstruction without hernia	0	NaN	99.8 [99.6-99.8]	NaN	NaN
Inguinal or femoral hernia	1	100.0 [20.7-100.0]	99.8 [99.6-99.8]	460.3 [207.0-1023.8]	NaN
Non-infective inflammatory bowel disease	9	100.0 [70.1-100.0]	99.8 [99.6-99.8]	459.0 [206.4-1020.8]	NaN
Vascular disorders of intestine	1	0.0 [0.0-79.3]	99.8 [99.7-99.8]	NaN	NaN
Gall bladder and bile duct disease	4	50.0 [32.1-67.9]	99.7 [99.6-99.8]	197.1 [57.7-672.8]	0.50 [0.19-1.34]
Pancreatitis	2	100.0 [34.2-100.0]	99.8 [99.6-99.8]	460.2 [206.9-1023.4]	NaN
Neurological disorders					
Alzheimer's disease and other dementias	22	81.8 [69.1-85.1]	99.3 [99.0-99.4]	112.1 [69.5-181.0]	0.18 [0.08-0.44]
Parkinson's disease	17	94.1 [79.4-92.6]	99.3 [99.1-99.4]	136.0 [85.6-216.2]	0.06 [0.01-0.40]
Epilepsy	15	86.7 [70.5-87.9]	99.4 [99.2-99.5]	140.1 [83.8-234.2]	0.13 [0.04-0.49]
Multiple sclerosis	24	95.8 [84.5-94.5]	99.3 [99.1-99.4]	145.8 [91.3-232.8]	0.04 [0.01-0.29]
Migraine	7	85.7 [60.6-85.5]	99.4 [99.2-99.5]	139.0 [79.2-243.8]	0.14 [0.02-0.88]
Tension-type headache	0	NaN	99.3 [99.1-99.4]	NaN	NaN
Mental and behavioral disorders					
Schizophrenia	49	100.0 [92.7-100.0]	99.7 [99.6-99.8]	387.7 [185.0-812.5]	NaN
Alcohol use disorders	13	76.9 [60.0-81.5]	99.8 [99.7-99.8]	423.1 [167.8-1066.9]	0.23 [0.09-0.62]
Drug use disorders	21	81.0 [67.9-84.4]	99.9 [99.7-99.9]	554.9 [203.9-1510.0]	0.19 [0.08-0.46]
Major depressive disorder	15	80.0 [64.2-83.6]	99.5 [99.3-99.6]	157.0 [87.9-280.6]	0.20 [0.07-0.55]
Dysthymia	0	NaN	99.2 [98.9-99.3]	NaN	NaN
Bipolar affective disorder	13	76.9 [60.0-81.5]	99.8 [99.6-99.8]	352.6 [150.2-827.3]	0.23 [0.09-0.62]
Anxiety disorders	17	94.1 [79.4-92.6]	99.6 [99.4-99.7]	235.0 [128.7-428.8]	0.06 [0.01-0.40]
Eating disorders	2	50.0 [29.3-70.7]	100.0 [99.9-100.0]	NaN	0.50 [0.13-2.00]
Autism	2	100.0 [34.2-100.0]	99.9 [99.8-99.9]	1380.5 [345.4-5517.1]	NaN
Asperger's syndrome	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Attention-deficit hyperactivity disorder	2	100.0 [34.2-100.0]	100.0 [99.9-100.0]	NaN	NaN
Conduct disorder	0	NaN	99.9 [99.8-99.9]	NaN	NaN
Idiopathic intellectual disability	1	0.0 [0.0-79.3]	100.0 [99.8-99.9]	NaN	NaN
Diabetes, urinary diseases and male infertility					
Diabetes mellitus	202	87.1 [84.1-88.8]	99.6 [99.4-99.6]	202.9 [112.2-366.7]	0.13 [0.09-0.19]
Acute glomerulonephritis	0	NaN	99.9 [99.7-99.9]	NaN	NaN
Chronic kidney diseases	29	48.3 [39.8-57.2]	98.0 [97.7-98.2]	24.0 [15.2-38.0]	0.53 [0.37-0.75]
Tubulointerstitial nephritis, pyelonephritis, and urinary tract infections	4	100.0 [51.0-100.0]	99.8 [99.7-99.8]	551.8 [229.9-1324.7]	NaN
Urolithiasis	1	100.0 [20.7-100.0]	99.8 [99.6-99.8]	460.3 [207.0-1023.8]	NaN
Benign prostatic hyperplasia	4	100.0 [51.0-100.0]	99.8 [99.6-99.8]	459.8 [206.8-1022.6]	NaN
Male infertility	1	100.0 [20.7-100.0]	99.8 [99.6-99.8]	460.3 [207.0-1023.8]	NaN

Gynecological diseases					
Uterine fibroids	1	100.0 [20.7-100.0]	99.9 [99.8-99.9]	1381.0 [345.6-5519.1]	NaN
Polycystic ovarian syndrome	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Female infertility	6	16.7 [16.1-43.3]	99.8 [99.7-99.8]	91.9 [12.5-673.7]	0.83 [0.58-1.19]
Endometriosis	2	100.0 [34.2-100.0]	99.9 [99.7-99.9]	690.3 [259.2-1837.8]	NaN
Genital prolapse	0	NaN	99.8 [99.7-99.8]	NaN	NaN
Premenstrual syndrome	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Hemoglobinopathies and hemolytic anemias					
Thalassemias	14	42.9 [32.7-56.1]	99.9 [99.8-99.9]	589.1 [129.9-2671.1]	0.57 [0.36-0.90]
Sickle cell disorders	14	14.3 [12.8-31.1]	100.0 [99.8-99.9]	392.7 [37.7-4086.3]	0.86 [0.69-1.06]
G6PD deficiency	14	0.0 [0.0-21.5]	100.0 [99.8-99.9]	NaN	NaN
Musculoskeletal disorders					
Rheumatoid arthritis	45	88.9 [81.1-90.6]	99.6 [99.4-99.7]	241.6 [129.0-452.4]	0.11 [0.05-0.25]
Osteoarthritis	40	70.0 [61.3-75.2]	100.0 [99.8-99.9]	1906.1 [265.8-13669.1]	0.30 [0.19-0.48]
Low back pain	13	92.3 [74.5-90.8]	99.7 [99.5-99.7]	317.3 [156.1-645.1]	0.08 [0.01-0.51]
Neck pain	3	33.3 [24.0-61.3]	100.0 [99.8-99.9]	920.0 [73.3-11549.5]	0.67 [0.30-1.48]
Gout	4	50.0 [32.1-67.9]	99.6 [99.4-99.7]	137.9 [43.3-439.6]	0.50 [0.19-1.34]
Congenital anomalies	23	95.7 [84.0-94.3]	98.3 [98.0-98.5]	55.8 [41.5-75.0]	0.04 [0.01-0.30]
Skin and subcutaneous diseases					
Eczema	3	33.3 [24.0-61.3]	99.9 [99.8-99.9]	460.0 [55.4-3819.6]	0.67 [0.30-1.48]
Psoriasis	5	80.0 [52.0-82.0]	100.0 [99.9-100.0]	NaN	0.20 [0.03-1.15]
Cellulitis	3	0.0 [0.0-56.1]	100.0 [99.9-100.0]	NaN	NaN
Abscess, impetigo, and other bacterial skin diseases	2	100.0 [34.2-100.0]	99.7 [99.6-99.8]	394.4 [188.2-826.6]	NaN
Scabies	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Fungal skin diseases	1	100.0 [20.7-100.0]	99.8 [99.6-99.8]	460.3 [207.0-1023.8]	NaN
Viral skin diseases	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Acne vulgaris	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Alopecia areata	1	100.0 [20.7-100.0]	100.0 [99.8-99.9]	2762.0 [389.2-19600.7]	NaN
Pruritus	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Urticaria	1	100.0 [20.7-100.0]	100.0 [99.9-100.0]	NaN	NaN
Decubitus ulcer	0	NaN	100.0 [99.8-99.9]	NaN	NaN
Sense organ diseases					
Glaucoma	16	100.0 [80.6-100.0]	99.9 [99.8-99.9]	1373.5 [343.7-5489.1]	NaN
Cataracts	12	91.7 [72.9-90.2]	100.0 [99.8-99.9]	2521.8 [352.7-18028.8]	0.08 [0.01-0.54]
Macular degeneration	18	100.0 [82.4-100.0]	100.0 [99.8-99.9]	2745.0 [386.8-19480.0]	NaN
Refraction and accommodation disorders	4	50.0 [32.1-67.9]	100.0 [99.8-99.9]	1379.5 [154.2-12338.3]	0.50 [0.19-1.33]

Oral disorders					
Dental caries	1	0.0 [0.0-79.3]	99.8 [99.6-99.8]	NaN	NaN
Periodontal disease	4	50.0 [32.1-67.9]	99.8 [99.7-99.8]	275.9 [74.1-1026.9]	0.50 [0.19-1.33]
Edentulism	3	0.0 [0.0-56.1]	100.0 [99.9-100.0]	NaN	NaN
Sudden infant death syndrome	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Injuries					
Road injury	3	0.0 [0.0-56.1]	99.9 [99.8-99.9]	NaN	NaN
Falls	7	14.3 [14.5-39.4]	100.0 [99.9-100.0]	NaN	0.86 [0.63-1.16]
Drowning	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Fire, heat and hot substances	3	66.7 [38.7-76.0]	100.0 [99.9-100.0]	NaN	0.33 [0.07-1.65]
Poisonings	0	NaN	99.9 [99.8-99.9]	NaN	NaN
Exposure to mechanical forces	2	0.0 [0.0-65.8]	99.9 [99.8-99.9]	NaN	NaN
Adverse effects of medical treatment	35	8.6 [7.7-17.6]	99.6 [99.4-99.6]	19.5 [5.8-66.0]	0.92 [0.83-1.02]
Animal contact	0	NaN	100.0 [99.9-100.0]	NaN	NaN
Self-harm	4	25.0 [20.6-53.9]	100.0 [99.9-100.0]	NaN	0.75 [0.43-1.32]
Interpersonal violence	0	NaN	99.9 [99.8-99.9]	NaN	NaN
Exposure to forces of nature	1	0.0 [0.0-79.3]	100.0 [99.9-100.0]	NaN	NaN
Collective violence and legal intervention	0	NaN	100.0 [99.9-100.0]	NaN	NaN
No GBD category	484	71.3 [69.1-73.2]	91.4 [90.7-91.9]	8.3 [7.2-9.6]	0.31 [0.27-0.36]

Number of trials per GBD category from the sample of 2,763 clinical trials. Sensitivities, specificities (in %) and likelihood ratios for each of the 171 GBD categories plus the “No GBD” category for the classifier using the Word Sense Disambiguation, the expert-based enrichment database and the priority to the health condition field.

Table S3: Performances of the eight versions of the classifier to the 171 GBD categories

Word Sense Disambiguation	Expert-based enrichment	Priority to health condition field	Proportion of trials with correct automatic classification				Weighted average across GBD categories	
			All trials N=2763	One GBD category N=2092	Two or more GBD categories N=187	No GBD category N=484	Sensitivity	Specificity
Yes	Yes	Yes	74.0	77.4	43.3	71.3	75.2	98.0
Yes	Yes	No	74.0	77.4	44.4	70.5	75.2	98.0
Yes	No	Yes	67.1	68.4	36.9	73.6	67.3	97.0
Yes	No	No	67.8	69.3	38.5	72.7	68.6	97.1
No	Yes	Yes	72.3	75.2	42.2	71.3	74.5	97.8
No	Yes	No	72.2	75.2	43.3	70.5	74.5	97.8
No	No	Yes	65.7	66.6	35.3	73.6	66.6	96.8
No	No	No	66.3	67.5	36.9	72.7	67.9	96.9

Exact-matching and weighted averaged sensitivities and specificities for eight versions of the classifier to the 171 GBD categories. Exact-matching corresponds to the proportion (in %) of trials for which the automatic GBD classification is correct. Exact-matching was estimated over all trials (N=2,763), over trials concerning a unique GBD category (N=2,092), over trials concerning two or more GBD categories (N=187), and over trials not relevant for the GBD (N=484). The weighted averaged sensitivity and specificity corresponds to the weighted average across GBD categories of the sensitivities and specificities of each GBD category plus the “No GBD” category (in %). The eight versions correspond to the combinations of: the use or not of the Word Sense Disambiguation during the text annotation, the use or not of the expert-based enrichment database, and the use or not of the priority to the health condition field as a prioritization rule.