

Émulation d'essais cibles à partir de données observationnelles : variables instrumentales, apprentissage fédéré et estimands causaux & applications en médecine cardio-métabolique et en réanimation

Contexte

L'évaluation rigoureuse de l'efficacité et de la sécurité des traitements en conditions réelles constitue un enjeu central de la médecine fondée sur les preuves. Les essais contrôlés randomisés (ECR) demeurent le *gold standard* pour établir la causalité, mais leur réalisation est limitée par des contraintes de faisabilité et de coût ; de plus, la question de la généralisabilité des résultats en vie réelle et/ou à d'autres populations demeure. Le cadre de l'émulation d'essais cibles s'est imposé comme une approche structurée permettant de reproduire, à partir de données observationnelles complexes et de très grande taille (*e.g.* dossiers médicaux électroniques, bases médico-administratives, cohortes *ad hoc*), le protocole de l'ECR qui aurait répondu à la question d'intérêt [1]. Cette démarche, malgré son apparente simplicité conceptuelle, requiert des méthodes statistiques avancées pour estimer des effets causaux valides. Si le cadre conceptuel de l'émulation d'essais cibles gagne en popularité, les méthodes statistiques permettant de l'implémenter rigoureusement restent un champ de recherche actif. En particulier, le choix de l'*estimand* (*i.e.* la quantité causale précise que l'on cherche à estimer) conditionne l'ensemble de la stratégie d'analyse. Par ailleurs, l'émulation d'un essai cible est le plus souvent conduite sur une source de données unique, ce qui limite la puissance statistique pour les pathologies rares, les expositions peu fréquentes ou les critères de jugement avec peu d'événements. L'apprentissage fédéré, qui permet d'agréger les analyses conduites sur des bases de données distinctes sans transfert des données individuelles, ouvre ici des perspectives majeures et encore peu explorées. En effet, le contexte de l'émulation d'essais cibles nécessite d'étendre et d'adapter les méthodes d'agrégation d'estimations locales existantes. Il introduit des défis méthodologiques propres, notamment la prise en compte de l'hétérogénéité des populations entre centres médicaux et/ou pays et de la variabilité des pratiques de soins, de diagnostic et de codage. Dans le même temps, la disponibilité de larges bases de données de santé, dont le Système National des Données de Santé (SNDS) en France (plus de 65 millions d'individus), UK Biobank au Royaume-Uni (environ 500 000 participants) et NIH All of Us aux Etats-Unis (plus de 620 000 participants), offre une opportunité inédite de conduire des émulations à grande échelle. Ainsi, les choix méthodologiques opérés à chaque étape de l'émulation, qu'il s'agisse de la méthode d'ajustement pour les facteurs de confusion, de la définition de l'estimand ou de la stratégie d'agrégation multi-sources, peuvent conduire à des estimations substantiellement différentes des effets thérapeutiques, avec des conséquences directes sur les recommandations cliniques en vie réelle.

Ce projet propose de contribuer au développement et à l'évaluation critique de méthodes d'émulation d'essais cibles, en les appliquant à deux domaines cliniques distincts (*i.e.* médecine cardio-métabolique et réanimation).

Objectifs scientifiques

L'objectif général de cette thèse est de comparer et d'appliquer des approches d'inférence causale dans le cadre de l'émulation d'essais cibles. Nous nous appuierons tout d'abord sur des simulations de données fictives afin de vérifier la performance des approches proposées dans un cadre maîtrisé où les effets réels sont connus, puis sur des applications pratiques à l'aide de données observationnelles issues de différentes sources. Trois axes méthodologiques structurent le projet :

Axe 1 : Comparaison des méthodes d'ajustement pour les facteurs de confusion. Les méthodes fondées sur le score de propension, les plus couramment utilisées dans l'émulation d'essais cibles, reposent sur l'hypothèse qu'il n'existe pas de facteurs de confusion non mesurés — une hypothèse rarement vérifiable en pratique. Les méthodes fondées sur l'utilisation des variables instrumentales constituent une approche alternative qui permet de s'affranchir de cette limite, en exploitant une source de variation exogène dans l'exposition au traitement, telle que les préférences de prescription médicamenteuse des médecins. Ces deux familles de méthodes reposant sur des hypothèses distinctes, elles peuvent aboutir à des estimations divergentes [2] : comprendre quand et pourquoi ces divergences surviennent constitue un enjeu méthodologique majeur. Ainsi, le premier axe de la thèse consistera à comparer ces deux familles de méthodes : d'une part, les approches fondées sur le score de propension, implémentées avec des modèles de complexité croissante dont des techniques de machine learning (*e.g.* forêts aléatoires, gradient boosting)

; d'autre part, les approches par variables instrumentales de nature diverse (e.g. préférences de prescription médicamenteuse, distance géographique aux centres de soins, changements de recommandations cliniques au cours du temps). Leur comparaison systématique, d'abord à l'aide de simulations statistiques reflétant différents scénarios rencontrés en vie réelle, puis sur des données observationnelles, permettra d'identifier les situations dans lesquelles chaque famille de méthodes est préférable.

Axe 2 : Fédération de l'émulation d'essais cibles entre bases de données de santé. Le partage de données individuelles de patients issues de plusieurs bases de données et/ou pays se heurte fréquemment à des obstacles réglementaires et techniques : le Règlement Général sur la Protection des Données (RGPD) en Europe et les restrictions propres à chaque base de données de santé limitent de facto la constitution d'une base poolée. Le développement et la comparaison de méthodes permettant de fédérer des émulations d'essais cibles sans transfert de données individuelles constituent donc un enjeu méthodologique à part entière, que cet axe entend explorer. La thèse consistera à étudier et développer différentes stratégies pour conduire des émulations d'essais cibles dans plusieurs bases de données nationales et internationales (e.g. SNDS, UK Biobank, NIH All of Us, MIMIC-IV, eICU) : analyses séparées, méta-analyses à effets fixes ou aléatoires sur données agrégées et apprentissage fédéré sans partage de données individuelles [3].

Axe 3 : Diversité des estimands causaux. L'effet en intention de traiter est l'estimand le plus utilisé dans un ECR, mais il est difficile à estimer fidèlement dans une émulation d'essai cible réalisée à partir de données observationnelles. De fait, le choix de l'estimand est une décision structurante qui conditionne l'ensemble de la stratégie d'analyse et l'interprétation des résultats d'une émulation d'essai cible. Dans ce projet, nous proposons d'étudier systématiquement, pour une même question clinique, comment différents estimands influencent les estimations obtenues et leur interprétabilité pour la décision clinique. Cet axe est d'autant plus pertinent que le nombre d'estimands présentés dans la littérature ne cesse de croître. Nous nous intéresserons notamment aux effets de stratégies thérapeutiques modifiées (*modified treatment policies*). Cette famille d'estimands, d'apparition récente dans la littérature, pose la question suivante : qu'advierait-il si, au lieu d'imposer un traitement fixe à tous les patients, on modulait le traitement observé de chaque patient (e.g. en réduisant la dose prescrite d'une quantité donnée, en raccourcissant la durée d'exposition à un médicament, en avançant l'initiation du traitement) ? Contrairement aux estimands classiques qui comparent des stratégies fixes et identiques pour tous les patients, ces estimands tiennent compte du traitement naturellement reçu par chaque patient, en vie réelle. Ils sont donc particulièrement pertinents lorsque les caractéristiques des individus, les stratégies thérapeutiques considérées et/ou l'adhérence à ces dernières présentent une hétérogénéité accrue. Cet axe est transversal aux deux autres : le choix de l'estimand interagit à la fois avec la méthode d'ajustement sur les facteurs de confusion retenue et avec la stratégie d'agrégation choisie dans un cadre fédéré, rendant leur étude conjointe innovante et impactante.

Applications cliniques

Les développements méthodologiques seront illustrés et validés sur deux cas d'usage cliniques complémentaires, choisis pour leur pertinence en santé publique et la disponibilité des données. Ils sont susceptibles d'évoluer au fil du doctorat, dès lors que des données adaptées seront disponibles et que la question causale posée se prêtera à une émulation d'essai cible.

La première application concerne l'efficacité cardiovasculaire des agonistes du récepteur GLP-1 dans le diabète de type 2, un contexte caractérisé par une observance hétérogène particulièrement adaptée à l'étude des estimands causaux brièvement décrits dans l'axe 3. À partir du SNDS, de UK Biobank et de NIH All of Us, la thèse caractérisera les profils temporels de dispensation des GLP-1, puis estimera les effets cardiovasculaires de différentes stratégies de traitement, déterminées en concertation avec des médecins.

La seconde application envisagée porte sur la comparaison de deux antibiothérapies, le céfépime et l'association pipéracilline-tazobactam, dans le choc septique. La thèse cherchera à émuler l'essai ACORN [4], étude monocentrique menée pendant la pandémie COVID-19, dont la généralisabilité reste incertaine et qui se prête à une émulation fédérée sur plusieurs bases de données (e.g. MIMIC-IV, eICU).

Justification de l'approche scientifique

L'association d'un développement méthodologique rigoureux et d'applications cliniques concrètes répond à un besoin identifié dans la littérature : de nombreuses études déclarant émuler un essai cible ne respectent pas entièrement les principes de ce cadre méthodologique, ce qui fragilise leurs conclusions [5]. La comparaison systématique à la fois des méthodes de prise en compte des facteurs de confusion et des estimands causaux considérés dans des scénarios cliniques plausibles constitue une contribution originale permettant de produire des recommandations pratiques à destination de la communauté épidémiologique. Le choix de deux domaines cliniques distincts, l'un caractérisé par une exposition médicamenteuse chronique et variable (*i.e.* traitement aux agonistes du récepteur GLP-1 pour le diabète de type 2), l'autre par une exposition aiguë en réanimation, renforcera la possibilité de généraliser nos conclusions méthodologiques. Enfin, l'axe 2 portant sur l'apprentissage fédéré à l'aide de multiples bases de données offrira une solution aux situations où une source unique ne fournit pas une puissance statistique suffisante (*e.g.* pathologies rares, expositions peu fréquentes, critères de jugement avec peu d'événements).

La faisabilité de ce projet repose sur un accès aux données nécessaires : le SNDS et UK Biobank, avec lesquels l'équipe d'accueil travaille régulièrement, ainsi que les bases de réanimation MIMIC-IV et eICU, disponibles en accès libre après approbation. Ces accès sont donc compatibles avec le calendrier doctoral.

Adéquation à l'IUIS

Ce projet s'inscrit pleinement dans la mission de l'Institut Universitaire pour l'Ingénierie en Santé (IUIS), en articulant ingénierie des données de santé, développement et évaluation de méthodes statistiques avancées, et mesure de l'impact clinique. Il repose sur la co-direction d'une ingénieure, enseignante-chercheuse en épidémiologie expérimentée en gestion et analyse de données de santé en grande dimension (Dr. Marie-Laure Charpignon), et d'un médecin de santé publique de l'AP-HP, également enseignant-chercheur en biostatistiques (Prof. David Hajage), en cohérence avec les besoins du projet. Ainsi, David Hajage apportera son expertise dans le domaine de l'inférence causale et guidera la spécification des essais cibles ainsi que l'interprétation médicale des résultats. Marie-Laure Charpignon contribuera au développement et à l'implémentation des algorithmes de fédération, ainsi qu'à la gestion des bases de données internationales. Ensemble, ils accompagneront le doctorant sélectionné dans l'élaboration de simulations et la construction d'estimands causaux pertinents pouvant informer la pratique clinique. Les résultats produits auront vocation à être diffusés sous forme de publications dans des revues de biostatistiques et d'épidémiologie de haut rang, et à alimenter des recommandations pour l'élaboration et l'analyse d'émulations d'essais cibles à partir de larges bases de données analysées rétrospectivement.

Bibliographie

- [1] Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol.* 2016 Apr 15;183(8):758-64. doi: 10.1093/aje/kwv254. Epub 2016 Mar 18. PMID: 26994063; PMCID: PMC4832051.
- [2] Laborde-Castérot H, Agrinier N, Thilly N. Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review. *J Clin Epidemiol.* 2015 Oct;68(10):1232-40. doi: 10.1016/j.jclinepi.2015.04.003. Epub 2015 Apr 8. PMID: 26026496.
- [3] Li H, Zang C, Xu Z, Pan W, Rajendran S, Chen Y, Wang F. Federated target trial emulation using distributed observational data for treatment effect estimation. *NPJ Digit Med.* 2025 Jul 1;8(1):387. doi: 10.1038/s41746-025-01803-y. PMID: 40593099; PMCID: PMC12214564.
- [4] Qian ET, Casey JD, Wright A, Wang L, Shotwell MS, Siemann JK, Dear ML, Stollings JL, Lloyd BD, Marvi TK, Seitz KP, Nelson GE, Wright PW, Siew ED, Dennis BM, Wrenn JO, Andereck JW, Han JH, Self WH, Semler MW, Rice TW; Vanderbilt Center for Learning Healthcare and the Pragmatic Critical Care Research Group. Cefepime vs Piperacillin-Tazobactam in Adults Hospitalized With Acute Infection: The ACORN Randomized Clinical Trial. *JAMA.* 2023 Oct 24;330(16):1557-1567. doi: 10.1001/jama.2023.20583. PMID: 37837651; PMCID: PMC10576861.
- [5] Simon-Tillaux N, Martin GL, Hajage D, Scheifer C, Beydon M, Dechartres A, Tubach F. Conducting observational analyses with the target trial emulation approach: a methodological systematic review. *BMJ Open.* 2024 Nov 12;14(11):e086595. doi: 10.1136/bmjopen-2024-086595. PMID: 39532374; PMCID: PMC11574403.