1.6 Résumé extensif en Français

Pourquoi étudier l'inférence causale à partir des dossiers patients informatisés

***Des données qui cherchent une question [...] et posent des énigmes* (Cox, 2001)**

L'apprentissage automatique a connu de grands succès en traitement automatique du langage (TAL) et en analyse d'image, grâce à l'exploitation de grandes quantités de données de piètre qualité, faiblement labellisées. D'autres champs d'applications peuvent-ils bénéficier d'une telle approche ? Parce que *la pratique médicale et la recherche biomédicale, [sont] intrinsèquement des tâches de gestion de l'information* (Patel et al., 2009), de nombreux chercheurs prévoient un grand potentiel d'amélioration des soins grâce à l'apprentissage automatique appliqué à de nouvelles collections de données (Topol, 2019; Rajkomar et al., 2019).

Il existe actuellement une tension entre la collecte massive de données de soins de routine, telles que les données de remboursement ou les dossiers patients informatiques (DPI), un nouveau cadre statistique (l'apprentissage automatique (Breiman, 2001b)), et des questions analytiques urgentes en matière de santé publique/ Pour comprendre l'importance et les défis de l'inférence causale à partir des DPI, examinons les trois exemples concrets ci-dessous.

Etudier les statistiques pendant la révolution du traitement automatique du langage

Dans les années 2010, Halevy et al., 2009 a proposé de tirer parti des régularités présentes dans de grandes masses de données pour concevoir automatiquement des variables pertinentes pour de multiples tâches applicatives. Considérons un modèle destiné à classifié des images à partir d'une grande collection de pairs d'images et d'étiquettes. Au cours de l'entraînement, le modèle apprend progressivement des représentations internes d'images naturelles telles que des pieds, des visages ou des mains. Ces représentations intermédiaires peuvent être réutilisées –transférées– à des tâches applicatives ayant des objectifs différents, telles que prédire la gravité d'un traumatisme à partir d'une photo. Ce paradigme, appelé préapprentissage a connu un grand succès dans des domaines appliqués tels que la vision par ordinateur (Krizhevsky et al., 2012), puis le traitement du langage naturel (TAL) (Devlin et al., 2018) et la biologie structurelle (Jumper et al., 2021). Il est pertinent de savoir si le pré-entraînement pourrait être appliqué à d'autres domaines avec des données complexes, tels que le soin.

Les hôpitaux de Paris, un vaste dépôt de notes cliniques

En 2017, le nouvel entrepôt de données de santé des Hôpitaux de Paris (AP-HP), ambitionnait d'exploiter son vaste référentiel de notes cliniques à des fins de recherche. Pourtant, les données de soins de routine ne sont pas un matériau familier pour la recherche médicale. Pour les ingénieurs et les chercheurs en TAL, ces données diffèrent de celles de la recherche traditionnelle car elles requièrent de nouveaux outils, par exemple provenant du TAL afin de gérer la complexité et l'échelle inédites des processus de collecte de données. Pour les épidémiologistes et les médecins, la principale différence réside dans l'opposition entre données expérimentales et données observationnelles. Cette difference de nature apparaît comme plus importante pour comprendre comment bien mobiliser les données de routine pour améliorer le soin.

Les données de facturation : de nouvelles données pour la santé publique?

En 2018, la direction des statistiques du ministère français de la santé a extrait des milliards de données de remboursement depuis la plateforme d'exploitation de l'assurance maladie au sein d'un serveur dédié. Pour améliorer et accélérer l'analyse sur cette nouvelle plateforme, elle s'est appuyée sur les travaux de Bacry et al., 2020 qui tirent parti du calcul parallèle et sur de la documentation collaborative (HDH, 2023a). Cependant, la description des séquences de soins est restée complexe en raison du grand nombre d'événements médicaux différents utilisés dans ces données. Ce problème, connu sous le nom de *malédication de la grande dimension* a été introduit dans la recherche opérationnelle (Bellman, 1957) et est bien connu en statistique (**breiman2001statistique**) : Le nombre d'échantillons requis pour développer des algorithmes prédictifs croît généralement de manière exponentielle avec le nombre de dimensions. Les séquences d'événements de facturation peuvent être considérées comme une séquences de signes, tout comme le texte. Ainsi, des travaux en informatique médicale, s'inspirant des techniques de TAL, consiste à créer des représentations de faible dimension –embeddings– de concepts médicaux (Beam et al., 2019). Les relations entre les événements capturés par ces sont étonnamment proches des associations connues, comme le montre la figure 1.1. Cependant, l'utilité avale de ces représentations reste incertaine. La santé publique est peu intéressée par les domaines de représentation de l'information ou même par des tâches de prédictions. Les services publics cherchent à comprendre l'hétérogénéités des consommations de soins : qu'est-ce qu'un soin approprié et comment le repérer à partir des données ? (Canadian Medical Association, 2015) ? Cette question est motivée par la nécessité croissante d'adapter le financement des soins à un système aux ressources limitées (McGinnis et al., 2013; Aubert et al., 2019).

Contributions

Initialement influencé par les succès croissants de l'apprentissage automatique pour la modélisation prédictive, ce travail cherche à comprendre quels modèles et quel cadre sont appropriés pour évaluer l'efficacité des recommendations de bonnes pratiques en santé à paritr des données de vie réelle. Quels sont les modèles prédictifs utiles ? Pourquoi la prédiction n'est-elle pas suffisante ? Comment des modèles flexibles peuvent-ils contribuer aux véritables objectifs du soin : fournir un traitement approprié à chaque patient pour améliorer sa santé (Canadian Medical Association, 2015) ?

Les contributions de chaque chapitre –résumées ci-dessous, ont donné lieu à trois articles en tant que premier auteur et à un travail en cours :

- Le chapitre 2 est publié dans *PLOS Digital Health*,
- Le chapitre 3 est un travail en cours,
- Le chapitre 4 est en cours de finalisation pour soumission,
- Le chapitre 5 est soumis à *Artificial Intelligence in Medicine*.

Chapitre 2 : Opportunités et obstacles rencontrés par les entrepôts de données cliniques cliniques, une étude de cas en France Ce chapitre présente la première vue d'ensemble des entrepôts de données de santé hospitaliers (EHDS) en France. Ces infrastructures techniques et organisationnelles émergent dans les hôpitaux afin de collecter et analyser les données produites en routine. Ce travail tente de mieux caractériser la réalité des réutilisations de données dans les Centres Hospitaliers Universitaires. Il documente les

aspects clés de la collecte et de l'organisation des données de soins de routine dans des bases de données homogènes: gouvernance, transparence, types de données, objectifs principaux de la réutilisation des données, outils techniques, types d'analyse, documentation et processus de contrôle de la qualité des données.

A partir d'entretiens semi-dirigés, nous montrons que l'écosystème naissant des EDSH en France est très hétérogène et principalement axé sur la recherche ou le pilotage. Nous soulignons la nécessité de créer ou de pérenniser des équipes d'entrepôts pluridisciplinaires capables d'opérer et d'exploiter l'EDSH afin de soutenir les différents projets de données. Les collaborations à plusieurs échelles permettent de mutualiser les ressources et les compétences au niveau régional ou national. Nous constatons une faible documentation des données et une adoption inégale des modèles de données communs pourtant reconnus au niveau international. Enfin, nous encourageons une extension du champ des données au-delà de l'hôpital pour mieux inclure les soins de ville. L'aspect qualitatif de ce chapitre contraste avec le contexte mathématique général de la thèse.

Chapitre 3 : Exploration d'un gradient de complexité pour les modèles prédictifs à partir de DPI Constatant l'intérêt croissant pour les algorithmes prédictifs à partir de données de DPI, ce chapitre introduit deux méthodes simples de construction de variables prenant les événements médicaux bruts en entrée avant d'alimenter un modèle prédictif. Il compare quatre pipelines prédictives de complexité croissante sur trois tâches médicales : classification de la durée du séjour, pronostic de la prochaine visite et prédiction d'événements cardiovasculaires indésirables. Ce travail se concentre sur des ensembles de données de taille moyenne où la population à risque (après les règles d'inclusion et d'exclusion) se situe entre 10 000 et 20 000 échantillons. Dans ces configurations, ce travail explore le compromis complexité-performance entre des modèles simples et des réseaux neuronaux récents à base d'architecture *transformer*.

Nous montrons que dans ces conditions de moyennes tailles d'échantillon, les modèles simples sont plus adaptés que les modèles à base de *transformer*, tant en termes de performance prédictive que d'efficacité en ressources de calcul. Nous constatons une diminution des performances pour les tâches pronostiques avec de faibles prévalences. Pour encourager l'étude plus approfondie de ces méthodes, nous publions les nouveaux modèles introduits avec une API scikit-learn.

Chapitre 4: La prédiction ne suffit pas: nécessité d'un cadre causal pour la prise de décision à partir des données de DPI Ce chapitre exploite le cadre causal pour concevoir des modèles d'aide à la décision utiles. Il montre que des prédictions –même précises comme avec l'apprentissage automatique, peuvent ne pas suffire à fournir des soins adaptés à chaque patient.

En tirant parti des principes de l'inférence causale, nous détaillons les éléments clés nécessaires pour estimer de manière robuste l'effet d'un traitement à partir de données de DPI variant dans le temps. Nous présentons des étapes détaillées permettant de développer des systèmes d'aide à la décision valides à partir des données de DPI grâce à l'émulation d'un essai randomisé. Nous illustrons ce guide par une étude de l'effet de l'albumine sur la mortalité due à la septicémie dans la base de données *Medical Information Mart for Intensive Care database* (MIMIC-IV). Nous étudions l'impact des différents choix d'analyses sur le résultat de l'étude, depuis l'extraction des caractéristiques des patients jusqu'à la sélection de l'estimateur causal. Nous constatons que des biais subtils, tels que le biais du temps immortel, peuvent modifier la conclusion d'une étude. Cependant, nous montrons que ces erreurs peuvent être capturées en émulant avec attention un essai randomisé hypothétique et

en comparant différentes choix de modélisation au sein d'une analyse de vibration. Nous validons notre estimateur de l'effet moyen du traitement à l'aide des résultats d'essais randomisés disponibles dans la littérature. Enfin, nous inspectons l'hétérogénéité de l'effet du traitement dans des sous-populations afin de guider le choix individuel de l'intervention. Dans un esprit didactique, le code et les données sont disponibles publiquement.

Chapitre 5 : Comment sélectionner des modèles prédictifs pour l'inférence causale ? Ce chapitre s'intéresse à la variabilité des résultats constatée dans le chapitre 4 pour différents choix d'estimateurs. Pouvons-nous expliquer pourquoi certains modèles permettent de mieux estimer l'effet du traitement que d'autres ? La théorie de l'apprentissage statistique établit comment sélectionner les modèles pour la prédiction. Ce chapitre montre que les méthodes de sélection utilisées en apprentissage automatique ne permettent pas de choisir les meilleurs modèles pour l'inférence causale. Nous passons en revue des risques plus élaborés présents dans la littérature d'inférence causale. Ces risques reposent sur l'estimation de nuisances qui permettent l'identification de l'effet causal. Cependant, ces risques causaux n'ont pas été évalués empiriquement pour une grande variété de contextes en échantillons finis. Ce chapitre étudie grâce à une étude empirique approfondie les performances de cinq risques causaux pour sélectionner un modèle d'estimation de l'effet de traitement.

Nos résultats montrent que les estimateurs pour l'inférence causale doivent être sélectionnés, validés et ajustés à l'aide de procédures et de mesures d'erreur différentes de celles utilisées classiquement en apprentissage statistique. La sélection du meilleur modèle à l'aide du risque R conduit à de meilleures estimations causales. Malgré le fait qu'il repose sur l'estimation de deux nuisances, ce risque est plus performant que les autres. Nous montrons également de manière théorique que le risque R est une version repondérée du risque non observé oracle entre les deux modèles d'outcomes potentiels. Cette propriété permet une estimation précise de l'hétérogénéité du traitement lorsque la population traitée et la population non traitée diffèrent peu, comme dans les essais randomisés. Pour faciliter la sélection des modèles, nous fournissons un code python mettant en oeuvre notre procédure.