Université Paris Cité

École Doctorale 393 Pierre Louis de Santé Publique
Centre de recherche en épidémiologie et statistiques (UMR 1153)
Équipe METHODS

Développement de règles de traitement personnalisé à partir de données d'essais cliniques : focus sur l'estimation des effets individualisés du traitement

Par Florie BOUVIER

Thèse de doctorat de Biostatistique

Sous la direction de Raphaël PORCHER

Et sous la co-direction de Anna CHAIMANI

Présentée et soutenue publiquement le 24/09/2024

Devant un jury composé de :

Roch GIORGI, PU-PH	Aix-Marseille Université	Rapporteur
Silvy LAPORTE, PU-PH	Université Jean Monnet de Saint-Etienne	Rapportrice
David HAJAGE, PU-PH	Sorbonne Université	Examinateur
Delphine MAUCORT-BOULCH, PU-PH	Université Claude Bernard Lyon 1	Examinatrice
Raphaël PORCHER, PU-PH	Université Paris Cité	Directeur de thèse
Anna CHAIMANI, HDR	Université Paris Cité	Co-directrice de thèse

Résumé

Titre : Développement de règles de traitement personnalisé à partir de données d'essais cliniques : focus sur l'estimation des effets individualisés du traitement.

Mots clefs : médecine personnalisée ; effet individualisé du traitement ; hétérogénéité de l'effet traitement ; règle de traitement individualisé ; modèle de prédiction ; apprentissage automatique ; méta-analyse sur données individuelles

Résumé : La médecine personnalisée a pour objectif d'adapter le traitement aux caractéristiques individuelles de chaque patient. L'un des aspects centraux de la médecine personnalisée est d'identifier des sous-groupes de patients qui bénéficient d'un traitement, ou plus d'un traitement que d'un autre. Plusieurs approches existent pour déterminer ces sous-groupes, en particulier l'estimation de l'effet individualisé du traitement (ITE). L'ITE représente l'effet attendu du traitement pour un individu de caractéristiques données. Dans ce travail de thèse, notre objectif était d'étudier le développement de règles de traitement personnalisé en estimant les effets individualisés du traitement, en s'appuyant sur les données issues d'un ou plusieurs essais cliniques randomisés. Nous avons dans un premier temps cherché à estimer les ITE en utilisant une méta-analyse sur données individuelles. Puis, nous avons comparé les règles de traitement produites par diverses méthodes. Enfin, nous avons étudié la discrimination maximale qui pourrait être obtenue pour diverses distributions d'effets de traitement.

Généralement, les modèles prédictifs pour l'ITE se fondent sur les données d'essais contrôlés randomisés uniques, qui ne sont pas toujours assez larges pour atteindre cet objectif, menant à un risque de sur-apprentissage des modèles ou au contraire à l'impossibilité de capturer l'effet de variables pertinentes. Les méta-analyses sur données individuelles (MADI) offrent une solution en rassemblant des données de plusieurs études, améliorant ainsi potentiellement la généralisation des résultats. Notre premier projet a exploré l'estimation des ITE via une MADI, évaluant la performance de différentes méthodes prenant en compte l'hétérogénéité entre les études de la méta-analyse, combinées avec deux méthodes permettant d'estimer les ITE (S-learner et T-learner). Il en ressort que privilégier une approche intégrant des interactions entre le traitement et les variables (S-learner) est bénéfique, sans qu'une méthode spécifique ne se détache en termes de performance.

Identifier des sous-groupes bénéficiant d'une intervention spécifique permet d'établir des règles de traitement individualisé (ITR). De nombreuses techniques d'apprentissage automatique ont été proposées récemment pour générer ces règles. Néanmoins, la concordance des ITR produites par ces méthodes, c'est-à-dire si elles recommandent le même traitement aux mêmes individus, reste incertaine. Notre deuxième projet a comparé les ITR générées par 22 méthodes dans deux essais cliniques, révélant que les méthodes aboutissent à des ITR différentes et donc ne sont pas interchangeables. Le choix de la méthode influence grandement les recommandations de traitement pour les patients, soulevant des préoccupations sur leur usage en pratique.

Pour développer des règles de traitement individualisé efficaces, il est essentiel que ces règles aient une bonne capacité à distinguer les patients qui bénéficient des patients qui ne bénéficient pas de prendre un certain traitement. Autrement dit, il faut que ces règles aient une bonne discrimination. Dans le dernier projet, nous avons exploré le niveau maximal de discrimination pouvant être obtenu pour différentes distributions d'effets de traitement, chacune présentant un niveau d'hétérogénéité différent. Nous avons sélectionné trois métriques de discrimination : la c-statistic for benefit, la concentration of benefit et le population average prescription effect (PAPE). Nos résultats indiquent que la présence d'effets de traitement hétérogènes, n'est pas systématiquement traduite par des résultats favorables en matière de discrimination. La discrimination optimale dépend de la distribution des effets du traitement. De plus, le choix de la métrique utilisée pour évaluer la discrimination a un impact sur les conclusions car elles n'exigent pas les mêmes niveaux d'hétérogénéité de l'effet de traitement pour conduire à une discrimination favorable.

Abstract

Title : Development of individualized treatment rules based on clinical trial data : focus on individualized treatment effects estimation.

Keywords : personalized medicine ; individualized treatment effect ; heterogeneity of treatment effect ; individualized treatment rule ; prediction model ; machine learning ; individual participant data meta-analysis

Abstract : Personalized medicine aims at tailoring treatment to the individual characteristics of each patient. One of the central aspects of personalized medicine is identifying subgroups of patients who benefit from one intervention or more from one intervention than another. Several approaches exist to determine these subgroups, particularly the individualized treatment effect (ITE) estimation. The ITE represents the expected treatment effect for an individual given its characteristics. In this thesis, we aimed to study the development of personalized treatment rules by estimating individualized treatment effects, based on data from one or more randomized clinical trials. We first sought to estimate ITE using an individual participant data meta-analysis. We then compared the treatment rules produced by various methods. Finally, we studied the maximum discrimination that could be obtained for various distributions of treatment effects.

In general, predictive models for ITE are based on data from single randomized controlled trials (RCTs), which are not always large enough to achieve this objective, leading to a risk of overfitting the models or, on the contrary, the impossibility of capturing the effect of relevant variables. Individual participant data meta-analyses (IPD-MA) offer a solution by pooling data from several RCTs, thus potentially improving the generalizability of results. Our first project explored the estimation of ITEs via an IPD-MA, evaluating the performance of different methods that consider heterogeneity between studies in the meta-analysis, combined with two methods for estimating ITE (S-learner and T-learner). The results show that an approach integrating interactions with treatment (S-learner) is beneficial, without any specific method standing out in terms of performance.

Identifying subgroups that benefit from a specific intervention makes it possible to develop individualized treatment rules (ITR). Numerous machine learning techniques have recently been proposed to generate these rules. Nevertheless, the consistency of the ITRs produced by these methods, i.e. whether they recommend the same treatment to the same individuals, remains uncertain. Our second project compared the ITRs generated by 22 methods in two clinical trials, revealing that the methods result in different ITRs and are therefore not interchangeable. The choice of method greatly influences treatment recommendations for patients, raising concerns about their practical use.

To develop effective individualized treatment rules, these rules must have a good capacity to distinguish patients who benefit from taking a certain treatment from patients who do not. In other words, these rules need to have good discrimination. In my last project, we explored the maximum level of discrimination that could be achieved for different distributions of treatment effects, each with a different level of heterogeneity. We selected three discrimination metrics : the c-statistic for benefit, the concentration of benefit, and the population average prescription effect (PAPE). Our results indicate that the presence of heterogeneous treatment effects is not systematically translated into favorable discrimination results. Optimal discrimination depends on the distribution of treatment effects. Furthermore, the choice of metric used to assess discrimination has an impact on the conclusions, as they do not require the same levels of treatment effect heterogeneity to lead to favorable discrimination.

À ma mère, Sylvie

Remerciements

Je remercie Raphaël Porcher pour m'avoir guidé tout au long de cette thèse. Merci pour tous les conseils et toutes les connaissances que tu m'as transmis au cours de ces quatre années.

Je remercie Anna Chaimani pour ses conseils et son expertise en méta-analyse, qui m'ont été précieux pour le premier projet de cette thèse.

Je remercie les Professeurs Silvy Laporte et Roch Giorgi pour avoir accepté de rapporter cette thèse.

Je remercie les Professurs Delphine Maucort-Boulch et David Hajage d'avoir accepté d'examiner le contenu de cette thèse.

Je remercie François Petit pour avoir accepté de participer à mes deux derniers projets de thèse et pour avoir partagé ses connaissances mathématiques.

Je remercie Etienne Peyrot pour son aide sur mes projets et surtout pour m'avoir montré comment optimiser tous mes codes, ce qui m'a fait gagner un temps non négligeable.

Je remercie Alan Balendran pour tous les memes partagés, les morceaux de chocolat et surtout pour son amitié.

Je remercie Isabelle Boutron pour m'avoir acceptée dans le Master CER, me permettant ainsi de développer de précieuses connaissances sur la méthodologie de l'évaluation thérapeutique, et pour m'avoir donné l'opportunité d'effectuer ma thèse dans l'équipe METHODS.

Je remercie tous les autres membres de l'équipe METHODS pour leur formidable accueil. Je n'aurais pas rêvé d'une meilleure équipe pour effectuer ma thèse.

Je remercie tous mes collègues doctorants, qui ont été de formidables compagnons pour ce long voyage qu'est la thèse : Alan, Amira, Anna, les deux Étienne, Lela, Mauricia, Ngan, Ottavio, Peiyu, Tianqi et Tiphaine. Je remercie également les anciens doctorants pour leurs retours d'expérience et leurs conseils, en particulier François et Theodoros.

Je remercie mes amis, en particulier Lina, Laetitia, Florian, Loëssa et Mathieu, pour les soirées et les moments de détente qui m'ont permis de me ressourcer et de tenir le coup.

Je remercie ma famille pour son soutien, et surtout ma mère, Sylvie, pour avoir cru en moi, pour avoir toujours été à mes côtés et pour m'avoir toujours encouragé à développer ma curiosité intellectuelle et à dépasser mes limites.

Valorisation scientifique

Publications issues de cette thèse

— **Bouvier F**, Peyrot E, Balendran A, et al. Do machine learning methods lead to similar individualized treatment rules ? A comparison study on real data. *Statistics in Medicine*. 2024 ; 43(11) : 2043-2061.

— **Bouvier, F.**, Chaimani, A., Peyrot, E., Gueyffier, F., Grenet, G., Porcher, R. (2024). Estimating individualized treatment effects using an individual participant data meta-analysis. *BMC Medical Research Methodology*, 24(1), 74.

Publication annexe

— **Bouvier, F. B.**, Porcher, R. What should be done and what should be avoided when comparing two treatments ? *Best Practice & Research Clinical Haematology* **36,** 2 (2023).

Communications orales et posters

Communications orales

— Symposium sur la médecine de précision, CRESS, 2024 (Paris)

— Causality in practice, séminaire AISSA/CNRS, 2023 (Paris)

Posters

— Présentation du deuxième projet, International society for clinical biostatistics (ISCB), 2023 (Milan)

— Présentation du deuxième projet, EPICLIN/JSCLCC, 2023 (Nancy)

— Présentation du deuxième projet, European Causal Inference Meeting (EuroCIM), 2023 (Oslo)

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

AVC Accident Vasculaire Cérébral. 2

CMC Coefficient de corrélation de Matthews. 41

ECR Essai Clinique Randomisé. 1, 16, 93–95, 98, 99

HTE Hétérogénéité de l'Effet Traitement. 2

IECV Internal-External Cross-Validation (validation croisée interne-externe). 20

ITE Individualized Treatment Effects (effets individualisés du traitement). 7, 8, 10, 13–18, 21, 22, 39–41, 67, 68, 70, 71, 89, 90, 92, 93, 95

ITR Individualized Treatment Rule (règle de traitement individualisé). 10, 11, 14, 22, 39–42, 66–68, 92, 95, 96

MADI Méta-Analyse sur Données Individuelles. ix, 15, 16, 18, 22, 43, 89, 90

PAPE Population Average Prescriptive Effect (effet de prescription moyen de la population). 11, 14, 69, 70, 93

Chapitre 1

Contexte scientifique

1.1 Vers une personnalisation de la médecine

1.1.1 Limites de l'application des résultats d'un essai clinique

Dans la médecine fondée sur les preuves, les données issues de recherches cliniques rigoureuses, telles que les essais cliniques randomisés (ECR) et les méta-analyses d'essais cliniques, sont reconnues comme le plus haut niveau de preuve scientifique pour comparer deux traitements afin de guider les décisions de traitement [1]. Ce sont ces types de données qui constituent le fondement des analyses présentées dans cette thèse.

La randomisation présente dans les ECR garantit que le groupe traitement et le groupe contrôle sont échantillonnés dans la même population, permettant une distribution identiques des facteurs pronostiques dans chacun des groupes. En conséquence, toute variation observée entre les critères de jugement des deux groupes peut être directement attribuée à l'effet du traitement. Des essais contrôlés randomisés bien conçus et bien conduits permettent donc de tirer des conclusions causales sur l'efficacité et la sécurité relatives de traitements.

Cependant, l'application des résultats d'un ECR à un individu peut s'avérer complexe. En effet, l'approche consistant à traiter tous les patients avec le traitement étudié ou à ne traiter aucun patient peut simplifier la notion d'effet du traitement [2, 3, 4, 5, 6, 7].

Rothwell a souligné l'impossibilité d'appliquer les résultats des ECR à l'échelle individuelle

en ré-analysant les données de deux essais [5]. Dans la ré-analyse, les patients de l'*European Carotid Surgery Trial* (ECST) [8] et de l'*UK-TIA Aspirin Trial*[9] ont été divisé en 3 catégories de risque (faible, modéré et fort).
Dans l'essai ECST destiné à évaluer l'efficacité de l'endartériectomie carotidienne pour les patients ayant eu un accident vasculaire cérébral (AVC) et présentant une sténose carotidienne [8], la ré-analyse a révélé des bénéfices uniquement pour les patients à haut risque d'AVC (réduction du risque absolu de 14,1%). Pour les patients à faible risque, la chirurgie n'offrait pas de bénéfice (réduction du risque absolu de -1,4%).
Inversement, la ré-analyse de l'essai UK-TIA évaluant l'efficacité de l'aspirine dans la prévention des AVC chez les patients ayant subi un AVC ischémique mineur [9], a montré que les patients à plus haut risque bénéficiaient moins de l'aspirine que ceux à faible risque.

Ces analyses mettent en lumière l'existence de variations significatives dans les réponses au traitement au sein d'une population. Suivre les directives basées sur l'effet moyen du traitement peut donc être préjudiciable, soit en administrant un traitement inutile à certains patients, soit en privant d'autres patients d'un traitement potentiellement bénéfique [7].

1.1.2 Hétérogénéité de l'effet traitement

L'hétérogénéité de l'effet traitement (HTE) se définit comme la variabilité non-aléatoire et explicable de la direction et de la magnitude des effets traitements au sein d'une population [10, 11]. Il y a présence d'hétérogénéité de l'effet traitement lorsqu'un même traitement entraîne des résultats différents chez les individus d'une population donnée [12].

Plusieurs facteurs influencent l'effet traitement et peuvent entraîner une hétérogénéité : une différence du risque de développer l'événement sans le traitement, une différence de complication du traitement due à une différence dans la vulnérabilité aux effets secondaires, et une différence dans la réponse au traitement [7, 3]. Il est donc nécessaire d'identifier les variables qui modifient l'effet du traitement. De plus, identifier et rapporter cette hétérogénéité lors des essais cliniques randomisés permet une interprétation précise de l'efficacité d'un traitement et permet d'éviter des conclusions trompeuses sur l'efficacité d'un traitement pouvant limiter la

généralisation des résultats de l'essai. [13].

L'effet du traitement pour des critères binaires peut être quantifié par la réduction du risque absolu, qui compare la fréquence des événements entre le groupe traitement et le groupe contrôle, ou par la réduction du risque relatif, qui mesure la proportion de réduction des événements dans le groupe traitement par rapport au groupe contrôle.Il est important de noter que le choix de l'échelle pour rapporter l'effet du traitement influence la perception de l'hétérogénéité de l'effet traitement [14]. Par exemple, l'hétérogénéité de l'effet du traitement variera selon que la mesure utilisée est la différence de risque ou le *odds ratio*.

Traditionnellement, l'analyse des sous-groupes a été utilisée pour évaluer l'hétérogénéité en explorant les effets du traitement au sein de sous-groupes de patients prédéfinis [15].

1.1.3 Limites des analyses de sous-groupes

L'approche traditionnelle d'analyse de sous-groupes est désormais jugée insuffisante pour identifier l'hétérogénéité des effets de traitement en raison de ses nombreuses limitations[16, 17]. En comparant des groupes sur la base d'une seule variable, ces analyses négligent la nature multiforme des différences individuelles entre les patients [18]. Les analyses univariées de sous-groupes ne donnent pas d'indications sur l'effet du traitement pour les patients individuels dont les caractéristiques couvrent plusieurs variables, ce qui limite leur utilité pour guider les décisions thérapeutiques personnalisées [19].

Cette méthode augmente le risque d'identifier incorrectement des interactions positives entre une variable et le traitement en multipliant les analyses sans ajustement adéquat pour le risque d'erreur de première espèce [18]. De plus, ces analyses ont des difficultés à identifier les variations des effets de traitement en raison d'une faible puissance statistique, nécessitant des tailles d'échantillon beaucoup plus importantes pour obtenir une puissance suffisante pour détecter les effets d'interaction [20, 4]. Les analyses de sous-groupes, qu'elles soient spécifiées a priori ou réalisées a posteriori, peuvent aussi accroître le risque de résultats faussement positifs en raison des comparaisons multiples, de la faible puissance statistique et de la possibilité de dés-

équilibres aléatoires entre les sous-groupes.

Ces limites soulignent la nécessité de disposer de méthodes avancées intégrant plusieurs variables pour une estimation plus fine et plus personnalisée des effets du traitement. Pour surmonter ces obstacles, l'adoption de modèles de risque multivariés est recommandée [3, 18], car ils ont démontré une efficacité supérieure pour détecter l'hétérogénéité des réponses au traitement [5]. D'autres méthodes existent et sont présentées ci-dessous.

1.2 Approches prédictives des analyses HTE

Les méthodes prédictives pour évaluer l'hétérogénéité de l'effet traitement cherchent à pallier les limites associées aux analyses de sous-groupes traditionnelles. Ces approches visent à développer des modèles prédictifs capables d'identifier le traitement le plus efficace au niveau individuel, en intégrant diverses variables qui affectent les bénéfices ou les effets nocifs du traitement. Prédire les réponses au traitement pour chaque patient représente un défi, étant donné l'impossibilité d'observer directement les effets d'un traitement au niveau individuel et de comparer simultanément les effets de différentes options de traitement chez un même patient.

Les recommandations PATH identifie deux méthodes clés pour évaluer l'hétérogénéité de l'effet du traitement : la modélisation du risque (*risk modeling*) et la modélisation de l'effet traitement (*treatment effect modeling*) [21].

La modélisation du risque implique l'utilisation d'un modèle multivariable pour prédire le résultat dans différents groupes de risque au sein de la population étudiée. Cette méthode vise à révéler comment les effets du traitement varient selon ces strates de risque, ce qui permet d'adapter les recommandations thérapeutiques. En se concentrant sur le risque en tant que variable clé, la modélisation du risque peut affiner l'estimation des effets du traitement, en reconnaissant que le risque d'un événement est un facteur déterminant de ces effets. Ces modèles révèlent souvent des variations substantielles dans la réduction du risque absolu entre les groupes à risque, ce qui est crucial pour prendre des décisions thérapeutiques pertinentes

sur le plan clinique. Une modélisation efficace du risque nécessite des données complètes sur les facteurs de risque potentiels et peut être limitée par la disponibilité et la qualité de ces données. Les résultats peuvent être généralisés à des populations de patients plus larges que ce qui est approprié si les groupes de risque ne sont pas bien définis.

La modélisation de l'effet du traitement développe un modèle directement sur les données des essais randomisés pour prédire les effets différentiels des traitements sur les sous-groupes de patients. Elle intègre souvent des termes d'interaction entre le traitement et les variables, ce qui permet d'inclure des variables susceptibles de modifier l'effet relatif du traitement. En identifiant et en incorporant les interactions entre le traitement et les variables, la modélisation de l'effet du traitement peut fournir des informations nuancées sur les traitements les plus efficaces pour les différents sous-groupes de patients. Cette approche peut considérablement affiner la prise de décision clinique en fournissant des prédictions détaillées sur l'efficacité du traitement pour les patients individuels en fonction de leurs caractéristiques spécifiques. Cependant, l'inclusion de multiples termes d'interaction peut compliquer le modèle et augmenter le risque de sur-apprentissage, en particulier dans les jeux de données comportant un nombre limité d'observations. Étant donné la nature exploratoire de nombreux modèles de l'effet du traitement, il existe un risque d'identifier des faux positifs.

La modélisation du risque et la modélisation de l'effet du traitement jouent toutes deux un rôle crucial dans l'avancement de la médecine personnalisée en permettant de prédire les réponses individuelles au traitement sur la base des caractéristiques spécifiques du patient. Alors que la modélisation du risque se concentre sur la stratification des patients en fonction de leur risque de base afin d'adapter les effets du traitement, la modélisation de l'effet tente de modéliser directement ces effets en se basant sur les données individuelles des patients. Le défi majeur de ces deux approches réside dans leur mise en oeuvre : assurer une qualité suffisante des données, empêcher le sur-apprentissage et valider les modèles de manière adéquate sont des étapes essentielles pour garantir leur fiabilité et leur utilité clinique. Dans cette thèse, nous avons décidé de nous concentrer sur la modélisation de l'effet du traitement car elle peut potentiellement fournir plus d'informations sur les effets du traitement au niveau individuel.

1.3 Estimation des effets individualisés du traitement

1.3.1 Modèle causal de Neyman-Rubin

La comparaison de deux traitements est par essence une question causale [22]. Cela implique d'examiner comment la prise d'un traitement influence le critère de jugement d'intérêt. Comme indiqué auparavant, il n'est pas possible de mesurer directement les effets d'un traitement au niveau individuel à l'exception des essais "n-of-1" qui mesurent l'effet des deux traitements comparés au sein de chaque patients [23]. Les essais "n-of-1" permettent de déterminer si une intervention est susceptible d'être bénéfique ou de provoquer des effets indésirables chez un unique patient donné. Cette méthodologie est particulièrement adaptée dans les contextes cliniques où la variabilité des réponses des patients est importante, comme dans les maladies chroniques, ou lorsque le patient présente des différences importantes par rapport aux personnes qui ont participé aux essais contrôlés randomisés classiques [24, 25].

Une façon courante de formaliser cet effet causal est le cadre des résultats potentiels ou modèle causal de Neyman-Rubin [26, 27, 28]. Il suppose que chaque individu a deux résultats potentiels, l'un s'il a reçu le traitement à évaluer, que nous notons Y^1, et l'autre s'il a reçu le traitement de comparaison, Y^0. La différence entre ces deux résultats, $Y^1 - Y^0$, définit l'effet causal du traitement pour l'individu concerné.

Pour estimer l'effet des traitements, les méthodes statistiques se fondent sur deux hypothèses :

- Cohérence (*Consistency*) : le résultat observé correspond au résultat potentiel, c'est-à-dire que si un patient a reçu le traitement, son résultat observé sera Y^1 et s'il a reçu le contrôle, son résultat observé sera Y^0. Cela implique, en pratique, que les notions de traitement et contrôle soient clairement définies.

- Absence d'interférence (*No interference*) : le résultat ne dépend que du traitement donné au patient, et non du traitement donné à d'autres patients.

1.3.2 Effets individualisés du traitement

Supposons que nous avons accès à un échantillon d'observations indépendant et identiquement distribué. Soit $X \in \mathcal{X} \subset \mathbb{R}^n$ un vecteur de covariables dans l'espace des covariables $\mathcal{X}$, $A \in \{0, 1\}$ une variable indicatrice du traitement considéré et $Y \in \{0, 1\}$ un critère de jugement. Dénotons par Y^1 et Y^0 les résultats potentiels sous le traitement et sous le contrôle introduits précédemment.

L'ITE d'un individu i peut être défini comme :

$$\tau(x_i) = \mathbb{E}(Y_i^1 = 1|X = x_i) - \mathbb{E}(Y_i^0 = 1|X = x_i) \tag{1.1}$$

En utilisant les hypothèses du modèle causal de Neyman-Rubin, l'ITE de l'individu i peut être exprimé comme :

$$\begin{aligned} \tau(x_i) &= \mathbb{E}(Y_i^1 = 1|X = x_i) - \mathbb{E}(Y_i^0 = 1|X = x_i) \\ &= \mathbb{E}(Y_i^1 = 1|A = 1, X = x_i) - \mathbb{E}(Y_i^0 = 1|A = 0, X = x_i) \\ &= \mathbb{E}(Y_i = 1|A = 1, X = x_i) - \mathbb{E}(Y_i = 1|A = 0, X = x_i) \end{aligned} \tag{1.2}$$

Dans les deux prochaines sous-sections, nous allons présenter certaines méthodes utilisées pour estimer l'ITE ainsi que les métriques utilisées pour évaluer les performances des modèles ITE.

1.3.3 Méthodes pour estimer les ITE

Différentes méthodes pour estimer les ITE ont été développées à partir de données issues d'essais cliniques ces dernières années. Nous présentons ici brièvement ces méthodes. Une description détaillée de certaines de ces méthodes est disponible dans les articles des chapitres 3 et 4.

— Les *meta-learners* : ces méthodes décomposent l'estimation de l'ITE en plusieurs problèmes de sous-régression qui peuvent être résolus avec n'importe quelle méthode de régression ou d'apprentissage supervisé. Les *meta-learners* estiment les résultats poten-

tiels à partir d'un ou plusieurs modèles. Parmi ces méthodes, on retrouve le *T-learner* [29], le *S-learner* [30], le *X-learner* [31], le *DR-learner* [32] et le *R-learner* [33].

— *Causal forests* : méthode qui étend l'algorithme original des forêts aléatoires à l'estimation des ITE en empruntant des idées aux méthodes à noyau et au R-learner[34].

— *Causal Boosting* : l'idée du *causal boosting* est de construire un arbre causal comme dans la méthode *causal forests* en ajustant des *weak learners* aux résidus du modèle pour un certain nombre d'itérations afin d'obtenir une approximation des surfaces de réponses permettant de calculer les ITE [35].

— *Causal BART* : méthode non-paramétrique combinant le cadre bayésien, la méthode *gradient boosting* et la méthode de Monte-Carlo par chaînes de Markov pour construire des arbres causaux qui permettent l'estimation des ITE [36].

— *Virtual twins* : cette méthode consiste à prédire les probabilités de réponse de jumeaux issu du groupe traitement et du groupe contrôle à l'aide de modèles contrefactuels [37].

— *A-learning* et *Modified covariate method* : ce sont deux méthodes qui se concentrent sur les interactions entre le traitement et les covariables et estiment l'ITE via une fonction de contraste et non en estimant chaque résultat potentiel [38, 39].

1.3.4 Métriques pour évaluer les modèles ITE

Deux composantes principales de l'évaluation de la performance d'un modèle de prédiction sont la discrimination et la calibration [40, 41]. Dans le cadre des modèles ITE, la discrimination représente la capacité du modèle à distinguer les individus qui bénéficient du traitement de ceux qui n'en bénéficient pas et la calibration représente la concordance entre le bénéfice estimé et le bénéfice prédit.

Pour évaluer la discrimination, la *c-statistic for benefit* proposée par van Klaveren et al. [42] peut être utilisée. Étant donné que le bénéfice individuel, c'est-à-dire l'obtention d'un résultat plus favorable en prenant le traitement qu'en ne le prenant pas, ne peut pas être observé, van Klaveren et al. ont utilisé des paires d'individus, une dans chaque groupe de traitement, avec des ITE prédits proches pour approcher le bénéfice individuel. La *c-statistic for benefit* est

l'extension de la *c-statistic* pour les effets de traitement individualisés. La *c-statistic for benefit* est définie comme la probabilité que, parmi deux paires appariées choisies au hasard $(p1, p2)$ avec un bénéfice estimé inégal, la paire avec le bénéfice estimé le plus élevé ait également une probabilité prédite plus élevée, où le bénéfice estimé se réfère à la différence de résultats entre deux patients avec le même bénéfice prédit mais avec des assignations de traitement différentes. Pour créer les paires, un patient du groupe contrôle est apparié à un patient du groupe traitement dont le bénéfice prédit du traitement est similaire. Les valeurs élevées de la *c-statistic for benefit* sont meilleures. La *c-statistic for benefit* peut être exprimée comme ci-dessous :

$$C_{fb} = P\Big(\hat{\tau}(x_{p1}) > \hat{\tau}(x_{p2}) \,|\, \tau(x_{p1}) > \tau(x_{p2})\Big)$$

où $\tau(x_{p1})$ et $\tau(x_{p2})$ représentent les bénéfices estimés des paires $p1$ et $p2$ et où $\hat{\tau}(x_{p1})$ et $\hat{\tau}(x_{p2})$ représentent les bénéfices prédits des paires $p1$ et $p2$ respectivement.

Une autre métrique permettant d'évaluer la discrimination a été récemment proposée, il s'agit de la concentration du bénéfice (*concentration of benefit*). La concentration du bénéfice (C_b) évalue la mesure dans laquelle les covariables capturent efficacement la variation des effets du traitement et identifient les individus qui bénéficient le plus du traitement [43]. Des valeurs plus élevées de concentration des bénéfices (proches de 1) indiquent une meilleure discrimination. La concentration du bénéfice peut être exprimée mathématiquement comme suivant

$$C_b = 1 - \frac{\mathbb{E}(\tau_1)}{\mathbb{E}[\max(\tau_2, \tau_3)]}$$

où τ_1, τ_2 et τ_3 représentent deux tirages aléatoires de la distribution des effets individualisés du traitement τ.

La calibration peut être évaluée en extrayant l'intercept et la pente de la ligne de régression. Un intercept proche de 0 et une pente proche de 1 indiquent une bonne calibration. Des courbes de calibration peuvent également être tracées [44]. Les prédictions sont divisées en plusieurs intervalles, afin de s'assurer d'inclure des individus assignés au traitement et des individus assignés au contrôle. Dans chaque intervalle, la moyenne du bénéfice prédit est comparée au bénéfice observé.

1.4 Développement de règles de traitement personnalisé

L'estimation des effets individualisés du traitement permet la construction de règles de traitement individualisé (ITR). Les ITR sont des règles de décision qui recommandent un traitement en fonction des caractéristiques des patients et peuvent être mathématiquement exprimées comme : $r : \mathcal{X} \to \{0, 1\}$.
Pour un ensemble donné de covariables $x \in \mathcal{X}$, $r(x)$ indique si le traitement doit être administré ou non à un patient.

Une règle optimale r^{opt} est obtenue lorsque la valeur $\mathcal{V}(r)$ parmi toutes les $r \in \mathcal{R}$, où $\mathcal{R}$ représentant la classe de toutes les règles de traitement, est maximisée [45] :

$$r^{opt} = \arg\max_{r \in \mathcal{R}} \mathcal{V}(r),$$

où $\mathcal{V}(r) = E[Y(r)]$ avec $Y(r) = Y^1 r(x) + Y^0[1 - r(x)]$ représentant le résultat observé si la règle r est respectée.

A partir des ITE, une règle optimale est obtenue en n'administrant le traitement qu'aux patients dont la valeur de $\tau(x)$ est positive, c'est-à-dire $r^{opt}(x) = \mathbb{1}_{\{\tau > 0\}}(x)$.

Il est important de souligner qu'il est possible de développer une règle de traitement personnalisée sans avoir à estimer les ITE. Dans une telle approche, une règle optimale est obtenue en minimisant une fonction de perte de la valeur de la règle. Deux méthodes développant directement des ITR, *outcome weighted learning* [46] et *contrast weighted learning* [47], sont utilisées dans le deuxième projet détaillé dans le chapitre 3. Ce sont deux méthodes qui utilisent une pondération pour déterminer le traitement entraînant le plus de bénéfice.

Plusieurs règles de traitement personnalisé ont été élaborées à partir de données provenant d'essais contrôlés randomisés. L'un de ces exemples est le score SYNTAX II qui utilise la méthode PATH [48]. Cette ITR, développée à partir des données de l'essai SYNTAX, vise à faciliter les décisions de traitement entre le pontage aorto-coronarien et l'intervention coronarienne percutanée pour des patients souffrant d'une maladie coronarienne complexe. Un autre

exemple est la règle de traitement personnalisé développée à l'aide de la méthode T-learner en utilisant les données de l'*International Stroke Trial*, qui recommande ou non l'allocation d'aspirine pour des patients victimes d'un accident vasculaire cérébral [49]. Par ailleurs, la méthode R-learner a été utilisée pour développer une ITR à partir des données de l'essai SPRINT, ayant pour but de recommander le traitement antihypertenseur menant à la plus grande réduction de pression artérielle systolique [50].

1.4.1 Métriques pour évaluer les ITR

Plusieurs métriques peuvent être utilisées pour évaluer la performance d'une ITR. Nous décrivons ici les métriques pouvant être utilisées en prenant en exemple un critère de jugement binaire favorable.

- La valeur d'une règle : la valeur $\mathcal{V}(r) = E[Y(r)]$ représente le résultat moyen si l'ITR a été correctement suivie. les ITR dont la valeur $\mathcal{V}(r)$ est plus proche de 1 sont plus performantes.
- L'effet de prescription moyen de la population (*population average prescriptive effect*, PAPE) : PAPE compare un ITR à une règle de traitement qui traite aléatoirement la même proportion de patients [51] :

  $$PAPE = E[Y(r) - p_r Y^1 - (1 - p_r)Y^0]$$

  où p_r représente la proportion de patients assignés au traitement évalué dans le cadre de l'ITR r.

  Étant donné que des valeurs plus élevées du critère de jugement sont souhaitables, des valeurs plus élevées du PAPE indiquent une meilleure performance de l'ITR. Une valeur de 0 indique que l'ITR n'est pas plus performant que traiter aléatoirement la même proportion de patients. Des valeurs négatives signifient que l'ITR est moins performant. L'avantage du PAPE est qu'il est facile à interpréter.
- L'avantage de la règle en termes de traitement attribué aux personnes ayant un score positif et négatif est évalué à l'aide de deux métriques : B_{pos} et B_{neg}, où B_{pos} représente l'avantage moyen d'administrer le traitement évalué aux personnes ayant un score po-

sitif, c'est-à-dire $r(x) = 1$, et B_{neg} représente l'avantage moyen de ne pas administrer le traitement évalué aux personnes ayant un score négatif, c'est-à-dire $r(x) = 0$ [52]. Les valeurs sont comprises entre -1 et 1, 1 signifiant qu'il y a un avantage à traiter les personnes ayant un score positif pour B_{pos} et un avantage à ne pas traiter les personnes ayant un score négatif pour B_{neg}.

$$B_{\text{pos}} = P(Y = 1 \mid A = 1, r(x) = 1) - P(Y = 1 \mid A = 0, r(x) = 1),$$

$$B_{\text{neg}} = P(Y = 1 \mid A = 0, r(x) = 0) - P(Y = 1 \mid A = 1, r(x) = 0).$$

— Proportion de personnes avec un score négatif : proportion de personnes considérées comme ne bénéficiant pas du traitement expérimental [52].

$$P_{\text{neg}} = P(r(x) = 0)$$

— Diminution du taux d'événements indésirables dans la population sous la règle [52] :

$$\begin{aligned}\Theta =& P(Y = 1 \mid A = 0) - [P(Y = 1 \mid A = 0, r(x) = 1)P(r(x) = 1) \\ &+ P(Y = 1 \mid A = 1, r(x) = 0)P(r(x) = 0)] \\ =& B_{\text{neg}} \cdot P_{\text{neg}}\end{aligned}$$

1.5 Justification et objectifs de la thèse

L'estimation des effets individualisés du traitement est une approche prometteuse pour le développement de règles de traitement personnalisé. Cependant, les essais cliniques étant conçus pour mesurer un effet moyen, utiliser les données d'un seul essai pour estimer ces effets peut conduire à du sur-apprentissage ou à des estimations inexactes des variables les plus significatives. Une solution potentielle à ce problème est l'utilisation d'une méta-analyse sur données individuelles qui regroupe plusieurs essais cliniques comparant les mêmes traitements. La méta-analyse sur données individuelles permet d'avoir un accès à un échantillon de données plus grand ainsi qu'à une population plus diversifiée ce qui pourrait permettre une meilleure estimation des ITE et ainsi au développement d'une règle de traitement personnalisé plus pertinente. Une étude récente a souligné la nécessité de prendre en compte les

variations entre les études pour éviter les biais écologiques, car les différences inter-études peuvent affecter l'estimation de l'effet traitement et entraîner des valeurs différentes des effets du traitement estimés en utilisant seulement les informations intra-études [53]. Cependant, la manière de combiner les approches tenant compte des potentielles différences inter-études et les méthodes d'estimation des effets individualisés du traitement n'est pas évidente. Dans le premier projet de cette thèse, nous avons estimé les ITE en utilisant une méta-analyse sur données individuelles, avec l'objectif de comparer la performance de différentes approches tenant compte de l'hétérogénéité entre les études (modèle naïf, intercept aléatoire, intercept stratifié, rang-1 et entièrement stratifié), et combinant deux méthodes d'estimation des effets individualisés du traitement (S-learner et T-learner).

Récemment, plusieurs techniques d'apprentissage automatique ont été développées pour créer des règles de traitement personnalisé et sont utilisées en pratique. Ces règles peuvent être dérivées de l'estimation des effets de traitement individualisés (ITE), comme c'est le cas avec les méthodes S-learner et T-learner utilisées dans le premier projet de cette thèse, ou en minimisant directement le risque associé à la règle de traitement à l'aide d'une fonction de perte. Des comparaisons ont déjà été faites entre ces différentes méthodes pour établir des règles personnalisées. Cependant l'uniformité des ITR produites par ces méthodes, c'est-à-dire si elles recommandent le même traitement aux mêmes individus, n'est pas encore claire. Dans le deuxième projet de cette thèse, nous avons comparé les règles de traitement individualisé (ITR) produites par 22 méthodes différentes dans deux essais cliniques : l'International Stroke Trial et l'essai CRASH-3. L'International Stroke Trial a été sélectionné car une ITR y a été précédemment développée en utilisant la méthode T-learner. L'essai CRASH-3 a été choisi en raison de l'hétérogénéité observée dans l'administration précoce du traitement.

Dans les deux premiers projets, nous avons observé des difficultés à obtenir de bons niveaux de discrimination avec les modèles d'effets de traitement individualisés et les règles de traitement individualisées, lesquels peinaient à distinguer les individus bénéficiant du traitement de ceux n'en bénéficiant pas. La discrimination est cruciale pour l'évaluation de la performance, et il semble que l'hétérogénéité des effets du traitement influence les résultats de discrimina-

tion. Bien que la performance des modèles prédictifs ait été largement étudiée, il existe peu de recherches sur les modèles ITE et les ITR, principalement à cause des défis liés à l'évaluation de telles performances, étant donné que les résultats des deux traitements ne sont généralement pas observables chez le même patient. Par conséquent, les mesures conventionnelles de prédiction des risques ne sont pas adaptées pour évaluer la performance des modèles qui prédisent les effets individualisés du traitement. De nouvelles métriques ont été proposées pour évaluer la discrimination des modèles ITE et ITR. Nous avons exploré trois de ces métriques : la c-statistic for benefit, la Concentration of Benefit et le Population Average Prescription Effect (PAPE). Dans le dernier projet de cette thèse, nous avons démontré le potentiel de discrimination maximale que l'on peut atteindre avec différentes distributions de l'effet du traitement, et avons comparé la discrimination mesurée par ces trois métriques. Pour cela, nous avons créé diverses distributions des effets du traitement et calculé la discrimination des modèles ITE par intégration numérique.

Chapitre 2

Estimation des effets individualisés du traitement dans la méta-analyse sur données individuelles

2.1 Résumé du projet

2.1.1 Introduction et objectifs

La plupart des modèles de prédiction pour estimer l'ITE sont développés à partir de données d'un seul essai contrôlé randomisé. Or, ces essais randomisés sont généralement sous-dimensionnés pour ces objectifs (ils sont dimensionnés pour un test de l'effet du traitement dans l'ensemble de la population d'étude), ce qui peut conduite à un sur-apprentissage des modèles ou au contraire à l'impossibilité de capturer l'effet de variables pertinentes.

Une solution est l'utilisation de méta-analyses sur données individuelles (MADI), qui permettent à la fois de disposer de données d'un plus grand nombre de sujets et d'améliorer la généralisation des résultats. La grande majorité des méthodes statistiques pour la MADI s'est intéressée soit à l'estimation de l'effet moyen du traitement, soit, beaucoup moins fréquemment, au développement de modèles de prédiction [54, 55, 56].

Quelques travaux ont envisagé l'estimation de l'ITE à partir d'approches en deux étapes pour la MADI [57, 58], et, récemment, une étude des méthodes pour estimer les interactions entre des covariables individuelles et le traitement à partir de données de MADI a aussi envisagé

une approche en une étape, c'est-à-dire l'utilisation d'un modèle pour l'ensemble des données de la MADI [53]. Ce papier a montré l'importance de tenir compte des variations entre les études mais surtout de n'utiliser que l'information intra-étude pour estimer un effet différentiel du traitement pour éviter des biais écologiques (l'information inter-études pourrait influencer l'estimation de l'effet différentiel du traitement de façon différente que l'information intra-étude).

Dans le contexte d'une étude unique ou d'un ECR, un large éventail d'approches a été proposé pour estimer les ITE [30, 29, 31, 34, 39, 38]. Dans ce travail, nous avons considéré deux stratégies appelées meta-learners, le S-learner et le T-learner. À notre connaissance, la manière de combiner adéquatement ces approches avec celles qui tiennent compte de l'hétérogénéité dans la MADI en une étape n'a pas été étudiée.

L'objectif de ce projet était d'étudier la performance de différentes méthodes prenant en compte l'hétérogénéité entre les études inclues dans la méta-analyse, combinées avec deux méthodes permettant d'estimer les ITE.

2.1.2 Méthodes

Dans cette section, les différentes méthodes sont décrites en considérant un critère de jugement binaire.

2.1.2.1 Estimation des ITE

L'ITE, qui est la différence entre les bénéfices prédits de deux traitements compte tenu d'un ensemble de caractéristiques des patients, est estimé de la manière suivante :

$$\hat{\tau}(x) = \hat{\mu}(x, z = 1) - \hat{\mu}(x, z = 0).$$

où $\hat{\mu}(x, z)$, $z \in \{0, 1\}$ représente le résultat moyen prédit sous le traitement z pour un individu avec les covariables x.

Pour estimer l'ITE, deux méta-learners ont été utilisés, le S-learner et le T-learner, qui dé-

composent l'estimation de ITE en sous-problèmes de régression [31]. A notre connaissance, les méthodes prenant en compte l'hétérogénéité inter-études n'ont été proposées que pour les modèles de régression. Cela a conduit à l'utilisation des meta-learners pour estimer les ITE, avec une préférence pour le S-learner et le T-learner, qui sont les plus appropriés pour les données des ECR avec une randomisation 1 :1.

Le S-learner estime l'ITE à l'aide d'un modèle de régression unique, dans lequel des interactions entre la variable indicatrice du traitement et les covariables pertinentes sont introduites. En considérant par exemple un modèle de régression logistique, le S-learner consiste à estimer le modèle suivant :

$$\text{logit}\mu(x,z) = \alpha + \theta' x + \gamma z + \eta' xz.$$

Nous en déduisons pour tous les individus :

$$\hat{\mu}(x,1) = \text{expit}(\hat{\alpha} + \hat{\theta'} x + \hat{\gamma} + \hat{\eta'} x),$$

et

$$\hat{\mu}(x,0) = \text{expit}(\hat{\alpha} + \hat{\theta'} x).$$

Le T-learner estime l'ITE à l'aide de deux modèles de régression distincts, l'un construit à partir des données du groupe traitement et l'autre à partir des données du groupe contrôle. Les deux modèles suivants :

$$\text{logit}\mu(x,0) = \alpha^0 + \theta'^0 x,$$

pour les individus ayant $z = 0$ et

$$\text{logit}\mu(x,1) = \alpha^1 + \theta'^1 x,$$

pour les individus ayant $z = 1$, sont modélisés et $\hat{\tau}$ est obtenu à partir de $\hat{\mu}(x,1) = \text{expit}(\hat{\alpha}^1 + \hat{\theta'}^1 x)$ et $\hat{\mu}(x,0) = \text{expit}(\hat{\alpha}^0 + \hat{\theta'}^0 x)$ pour tous les individus de la méta-analyse.

2.1.2.2 Modèles de prédiction des risques dans la MADI

Soit $x_{ij} = (x_{ij1}, \ldots, x_{ijN})$ un vecteur de valeurs de covariables pour le sujet $i \in (1, \ldots, N_j)$ dans l'étude $j \in (1, \ldots, J)$. Nous avons considéré les cinq modèles suivants :

— Modèle naïf (NA) : Une première approche considère que toutes les données proviennent d'une seule population et suppose donc qu'il n'y a pas d'hétérogénéité inter-études. Dans ce modèle, un intercept commun et des effets prédictifs communs sont inclus. Cette approche naïve peut entraîner un biais en cas d'hétérogénéité inter-études avérée.
Le modèle peut être exprimé comme cela :

$$\text{logit}(p_{ij}|x_{ij}) = \alpha + \theta' x_{ij} \tag{2.1}$$

où p_{ij} désigne la probabilité que le sujet i de l'essai j développe le critère de jugement.
Lorsque des prédictions individuelles sont faites pour estimer l'ITE à un niveau de covariable x, ces prédictions sont obtenues par $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}' x)$, où $\hat{\alpha}$ et $\hat{\theta}$ sont les estimateurs du maximum de vraisemblance de α et θ respectivement, dans le modèle 2.1.

— Modèle à intercept aléatoire (RI) : une deuxième approche consiste à supposer que l'hétérogénéité dans la MADI ne se produit que sur le risque de base, c'est-à-dire que l'intercept varie entre les études, mais que les effets des prédicteurs sont les mêmes dans chaque étude. Dans ce modèle, nous considérons un effet aléatoire pour modéliser la distribution de l'intercept entre les études.
Le modèle sous-jacent peut être écrit comme cela :

$$\text{logit}(p_{ij}|x_{ij}, \alpha_j) = \alpha_j + \theta' x_{ij} \tag{2.2}$$

avec $\alpha_j \sim \mathcal{N}(\alpha, \tau_\alpha^2)$.
Les prédictions individuelles sont obtenues par $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}' x)$. Les estimateurs α et θ sont obtenus par le maximum de vraisemblance qui est approximé par la quadrature adaptative de Gauss-Hermite.

— Modèle à intercept stratifié (SI) : Une troisième approche consiste à inclure un intercept différent pour chaque étude, en tant qu'effet fixe.

Avec un critère de jugement binaire :

$$\text{logit}(p_{ij}|x_{ij}, \alpha_j) = \sum_{m=1}^{J} \alpha_m I(m = j) + \theta' x_{ij} \tag{2.3}$$

où $I(\cdot)$ désigne la fonction indicatrice.

Pour dériver les prédictions individuelles en $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}' x)$, l'estimateur de θ est obtenu par maximum de vraisemblance. Pour obtenir un seul $\hat{\alpha}$, nous avons utilisé une méta-analyse à effets aléatoires des α_m, avec pondération par variance inverse, comme le suggèrent Debray et al. et Royston et al. [54, 59].

— Modèle entièrement stratifié (FS) : Une quatrième approche consiste à considérer qu'il existe une hétérogénéité entre les études à la fois sur les risques de base et sur les effets des prédicteurs. Dans ce cas, nous calculons des intercepts et des prédicteurs différents pour chaque essai inclus dans la méta-analyse.

Avec un critère de jugement binaire :

$$\text{logit}(p_{ij}|x_{ij}, \alpha_j, \theta_j) = \sum_{m=1}^{J} (\alpha_m I(m = j) + \theta'_m I(m = j) x_{ij}) \tag{2.4}$$

où α_m and θ_m, for $m = 1, .., J$, sont des paramètres réels à estimer. Cela équivaut à appliquer un modèle distinct à chaque étude incluse dans la méta-analyse. Les prédictions individuelles sont alors obtenues comme ceci $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}' x)$, où l'intercept $\hat{\alpha}$ et les prédicteurs $\hat{\theta}$ estimés sont obtenus à l'aide d'une méta-analyse multivariée à effets aléatoires.

$$\begin{pmatrix} \alpha_m \\ \theta_m \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \alpha \\ \theta \end{pmatrix}, V \right)$$

où V est la matrice de covariance inter-études de l'intercept et des prédicteurs.

— Modèle de rang 1 (R1) : Une dernière approche considère que les prédicteurs linéaires partagent une direction commune dans l'espace des covariables mais que la taille de leurs effets peut être systématiquement différente [60]. Ce modèle peut être considéré comme un intermédiaire entre les modèles à effet commun et le modèle entièrement stratifié. Dans ce cadre, les effets spécifiques à l'étude peuvent varier de manière proportionnelle, modélisés par un effet aléatoire ϕ. Avec un critère de jugement binaire :

$$\text{logit}(p_{ij}|x_{ij}, \alpha_j, \phi_j) = \alpha_j + \phi_j \theta x_{ij} \tag{2.5}$$

avec $\alpha_j \sim \mathcal{N}(\alpha, \tau_\alpha^2)$, $\phi_j \sim \mathcal{N}(1, \tau_\phi^2)$.

Avec le modèle de rang 1, les prédictions individuelles sont obtenues par $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}' x)$, où les estimateurs sont directement obtenus comme dans le modèle RI.

2.1.2.3 Validation des modèles

La validation croisée interne-externe (*internal-external cross-validation*, IECV) a été utilisée pour valider les modèles. Dans l'IECV, le modèle est construit avec $J - 1$ études et validé avec l'étude restante pour chaque permutation de $J - 1$ études. Pour tenir compte de l'hétérogénéité du risque de base, le modèle est recalibré dans les ensembles de données de validation. L'intercept est d'abord estimé avec les différents modèles de prédiction du risque présentés précédemment. Le recalibrage est ensuite effectué en estimant un modèle de régression avec les prédicteurs linéaires $\hat{\theta} x$ du modèle original en tant que *offset*, c'est-à-dire que le paramètre de régression est forcé à être égal à un. Ces étapes sont réalisées pour tous les modèles, à l'exception du modèle naïf qui ignore toute hétérogénéité éventuelle. Pour évaluer les performances des modèles, la discrimination et la calibration ont été prises en compte. L'erreur quadratique moyenne a également été calculée.

Pour évaluer la discrimination, c'est-à-dire la capacité du modèle à distinguer les individus qui bénéficient du traitement de ceux qui n'en bénéficient pas, la *c-statistic for benefit* proposée par van Klaveren et al. [42] a été utilisée. Cette métrique est décrite dans le chapitre 1.
La calibration, la concordance entre le bénéfice estimé et le bénéfice prédit, a été évaluée en extrayant l'intercept et la pente de la ligne de régression. Un intercept proche de 0 et une pente proche de 1 indiquent une bonne calibration. Des courbes de calibration ont également été tracées lorsque les méthodes ont été appliquées à la base de données *individual data analysis of antihypertensive intervention trials* (INDANA). Les prédictions ont été divisées en cinq intervalles, afin de s'assurer d'inclure les individus qui ont été assignés au traitement et les individus qui ont été assignés au contrôle. Dans chaque intervalle, la moyenne du bénéfice prédit a été comparée au bénéfice observé.

2.1.3 Résultats

Les performances des modèles et des méta-learners ont été évaluées dans un premier temps par étude de simulation. Nous avons examiné 24 scénarios dans lesquels nous avons modifié le nombre de covariables, le nombre de patients dans chaque essai et le type de critère de jugement. Les différentes approches ont été ensuite illustrées par l'utilisation de données provenant d'INDANA pour évaluer les modèles. L'illustration s'est concentrée sur quatre essais inclus dans la méta-analyse, chacun ayant un âge médian inférieur à 60 ans.

Dans les contextes que nous avons considérés, et avec des critères de jugement binaires, l'utilisation d'interactions avec le traitement (S-learner) ou de deux modèles différents (T-learner) a eu peu d'impact sur la performance des modèles. Pour les critères de jugement binaires, nous déconseillons l'utilisation du modèle FS lorsque plusieurs covariables sont impliquées, car il est sensible au sur-apprentissage. Pour les données de survie, l'approche S-learner a donné des résultats légèrement meilleurs, mais aucune méthode ne s'est révélée significativement supérieure aux autres. De plus, l'inclusion d'une sélection de variables n'a pas modifié les performances des algorithmes. Les modèles R1 et FS, qui englobent une plus grande hétérogénéité, se sont révélés efficaces pour capturer l'incertitude entourant les prédictions des effets individualisés du traitement. Leurs intervalles de prédiction ont englobé plus systématiquement le véritable ITE que les intervalles obtenus par d'autres méthodes.

2.1.4 Discussion

Dans ce travail, nous avons utilisé la régression pour estimer les ITE. Une des limites des modèles de régression est le risque de mauvaise spécification du modèle et son impact sur l'estimation des ITE. Dans le cadre des meta-learners, il est possible d'utiliser des approches d'apprentissage automatique non paramétriques telles que les forêts aléatoires. Par ailleurs, il existe d'autres meta-learners pour estimer les ITE, tels que le R-learner et le DR-learner [33, 32]. Il existe des approches permettant de créer des règles de traitement personnalisé sans estimer les ITE par la prédiction du bénéfice sous les deux traitements, comme la *modified covariate method* [**tian2014**] or *A-learning* [38]. De plus, des approches ne reposant pas explicitement

sur l'estimation des ITE peuvent également être utilisées, comme la *constrained single-index regression* [61] et bien d'autres [45, 62, 63, 64]. Cependant, la manière de prendre en compte l'hétérogénéité qui peut survenir entre les études n'est pas claire. À notre connaissance, la seule proposition d'une approche non basée sur la régression pour l'élaboration d'une ITR avec des données provenant d'une MADI est un article de Mistry et al. utilisant le partitionnement récursif [65]. Etudier la manière d'adapter ces approches moins sensibles à la mauvaise spécification du modèle pour intégrer l'estimation de l'hétérogénéité entre les études devrait faire l'objet de travaux ultérieurs, par exemple en s'appuyant sur des approches pour l'apprentissage fédéré [66, 67].

2.1.5 Conclusion

Les résultats ont indiqué que lors de la sélection d'une stratégie, l'incorporation des interactions avec le traitement (S-learner) est préférable en raison de sa performance efficace pour les critères de jugement binaires et de survie. En termes de sélection des méthodes, aucune des méthodes comparées n'a démontré de supériorité par rapport aux autres.

La description détaillée des méthodes et des résultats de ce travail se trouve dans l'article présent en section 2.2 qui a été publié en mars 2024 dans la revue *BMC Medical Research Methodology*.

2.2 Article

Bouvier *et al.*
BMC Medical Research Methodology (2024) 24:74
https://doi.org/10.1186/s12874-024-02202-9

BMC Medical Research Methodology

RESEARCH

Open Access

Estimating individualized treatment effects using an individual participant data meta-analysis

Florie Bouvier[1*], Anna Chaimani[1,2], Etienne Peyrot[1], François Gueyffier[3], Guillaume Grenet[3] and Raphaël Porcher[1,4]

Abstract

Background One key aspect of personalized medicine is to identify individuals who benefit from an intervention. Some approaches have been developed to estimate individualized treatment effects (ITE) with a single randomized control trial (RCT) or observational data, but they are often underpowered for the ITE estimation. Using individual participant data meta-analyses (IPD-MA) might solve this problem. Few studies have investigated how to develop risk prediction models with IPD-MA, and it remains unclear how to combine those methods with approaches used for ITE estimation. In this article, we compared different approaches using both simulated and real data with binary and time-to-event outcomes to estimate the individualized treatment effects from an IPD-MA in a one-stage approach.

Methods We compared five one-stage models: naive model (NA), random intercept (RI), stratified intercept (SI), rank-1 (R1), and fully stratified (FS), built with two different strategies, the S-learner and the T-learner constructed with a Monte Carlo simulation study in which we explored different scenarios with a binary or a time-to-event outcome. To evaluate the performance of the models, we used the *c*-statistic for benefit, the calibration of predictions, and the mean squared error. The different models were also used on the INDANA IPD-MA, comparing an anti-hypertensive treatment to no treatment or placebo ($N = 40\,237$, 836 events).

Results Simulation results showed that using the S-learner led to better ITE estimation performances for both binary and time-to-event outcomes. None of the risk models stand out and had significantly better results. For the INDANA dataset with a binary outcome, the naive and the random intercept models had the best performances.

Conclusions For the choice of the strategy, using interactions with treatment (the S-learner) is preferable. For the choice of the method, no approach is better than the other.

Keywords Personalized medicine, Individualized treatment effects, Individual patient data, Meta-analysis

*Correspondence:
Florie Bouvier
florie.brion-bouvier@u-paris.fr
[1] Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), Paris, France
[2] Cochrane France, Paris, France
[3] Laboratoire de Biométrie et Biologie Evolutive UMR 5558, CNRS, Université Lyon 1, Université de Lyon, Villeurbanne, France
[4] Centre d'Épidémiologie Clinique, AP-HP, Hôtel-Dieu, Paris, France

Background

Personalized (or stratified) medicine aims at tailoring a treatment strategy to the individual characteristics of each patient. One key aspect of personalized medicine is to identify individuals who benefit from an intervention. Different approaches exist, with a popular one being the estimation of the so-called individualized treatment effect (ITE). Shortly, the ITE on an additive scale is the

predicted benefit under one treatment minus the predicted benefit under the other treatment, given a set of patients' characteristics. It represents what treatment effect is expected for a patient with these characteristics. ITEs are generally estimated by building prediction models or by using machine learning methods such as random forests [1].

In practice, prediction models for ITE are often developed using data from a single randomized controlled trial (RCT) or observational data [2]. RCTs benefit from randomization but are often underpowered for such a task, which may lead to overfitting or the failure of capturing the effects of many relevant variables. A solution to that problem might be to use individual participant data meta-analyses (IPD-MA), which include larger numbers of patients and may also benefit from increased generalizability. Nevertheless, it is necessary to consider the variation between studies in such data to avoid bias. Previous studies have tackled the incorporation of heterogeneity when estimating the average treatment effect i.e. the average difference of the predicted risk between treatments, or have used IPD-MA to develop risk prediction models [3, 4]. However, it is unclear how to deal with heterogeneity in an IPD-MA while using approaches to estimate ITEs. Fisher et al. [5] and, more recently, Chalkou et al. [6] considered a framework to estimate the ITE in IPD-MA with a two-stage approach. More specifically, Chalkou et al. used a network meta-analysis with individual participant data to, first, estimate a prognostic model. Heterogeneity of treatment effects according to baseline risk predicted by this model was then considered using a two-stage approach with treatment by baseline risk interactions estimated within each trial. Seo et al. used one-stage meta-analytic approaches and focused on methods for selecting which treatment-covariate interactions to include in a model where study-specific intercepts and common effects factors were added; they concluded that shrinkage methods performed better than non-shrinkage methods [7].

In the context of a single study or RCT, a wide range of approaches have been proposed to estimate ITEs [8–12]. To our knowledge, how to adequately combine those with the approaches accounting for heterogeneity in IPD-MA has not been investigated. In this work, we considered two strategies called meta-learners, the S-learner and the T-learner [8].

In this study, we aimed to study the performance of strategies that estimate the ITE from an IPD-MA in a one-stage approach and methods focusing on taking into account the heterogeneity in baseline risks to understand which strategy and method should be used in practice. Different methods were compared using both simulated and real data with binary and time-to-event outcomes. We first present the different models and approaches compared in estimating ITEs. Next, we describe the Monte Carlo simulation study and its results, and the models are then applied to the data of the INDANA meta-analysis, a real individual patient data meta-analysis evaluating anti-hypertensive treatments [13]. We conclude with some discussion and paths for future research.

Methods to estimate individualized treatment effects

In this section, we described the different approaches we compared to estimate the ITE from an IPD-MA accounting for the clustering of patients within trials. We first explain the two approaches used to obtain ITEs from risk prediction models and then the different approaches to develop risk prediction models in an IPD-MA we considered.

ITE estimation

Let us consider a binary outcome without loss of generality. The case of time-to-event outcomes, which is similar in essence, is described in Supplementary Material S1. The ITE, which is the difference in predicted benefits of two treatments given a set of patients' characteristics, is estimated as:

$$\hat{\tau}(x) = \hat{\mu}(x,1) - \hat{\mu}(x,0).$$

where $\hat{\mu}(x,z)$, $z \in \{0,1\}$ represents the predicted mean outcome under treatment z for an individual with covariates x.

To estimate the ITE τ many methods exist. In this project, two meta-learners were used, the S-learner and the T-learner, which decompose the estimation of the ITE into sub-regression problems [8]. The meta-learners can be implemented with various prediction techniques such as regression or random forests for instance. In this work, we decided to use regression since methods to handle the heterogeneity in an IPD-MA have been developed with regression in previous works [3, 4].

The S-learner estimates the ITE using a single regression model, where interactions between the indicator variable for the treatment and relevant covariates are introduced.

Considering for instance a logistic regression model, the S-learner consists in estimating the following model:

$$\text{logit}\mu(x,z) = \alpha + \theta' x + \gamma z + \eta' xz.$$

From this, we derive for all individuals:

$$\hat{\mu}(x,1) = \text{expit}(\hat{\alpha} + \hat{\theta}' x + \hat{\gamma} + \hat{\eta}' x),$$

and

$$\hat{\mu}(x,0) = \text{expit}(\hat{\alpha} + \hat{\theta}'x).$$

Different approaches to obtain estimates of $\alpha, \theta, \gamma, \eta$ are described in the next subsection.

The T-learner estimates the ITE using two separate regression models, one built using data from the treatment group and one built using data from the control group. The two following models:

$$\text{logit}\mu(x,0) = \alpha^0 + \theta'^0 x,$$

for individuals with $z = 0$ and

$$\text{logit}\mu(x,1) = \alpha^1 + \theta'^1 x,$$

for individuals with $z = 1$, are fitted and $\hat{\tau}$ is obtained from $\hat{\mu}(x,1) = \text{expit}(\hat{\alpha} + \hat{\theta'}^1 x)$ and $\hat{\mu}(x,0) = \text{expit}(\hat{\alpha} + \hat{\theta'}^0 x)$ for all individuals in the meta-analysis.

The S-learner algorithm may reduce overfitting compared to the T-learner algorithm as it can adjust the number of interactions included in the model and thus can reduce the number of estimates. However, since IPD-MA is used in this work, the potential overfitting of the T-learner might be reduced due to a larger sample size.

In our case, we want to obtain $\hat{\mu}(x,z)$ using data from an IPD-MA. Several approaches exist to estimate this quantity while accounting for the potential heterogeneity that may arise in a meta-analysis. These approaches are detailed in the next subsection.

Risk prediction models in IPD-MA

Let us consider an IPD-MA where data from individual patients from J randomized controlled trials are available, and the outcome of interest is binary. Different methods to develop a single risk prediction model using IPD-MA have been proposed [3, 4]. Four of them were compared in this work and a naive model, that ignores any heterogeneity which may occur between the different studies included in the meta-analysis, was added to the comparison.

Let $x_{ij} = (x_{ij1}, \ldots, x_{ijN})$ be a vector of covariate values for subject $i \in (1, \ldots, N_j)$ in study $j \in (1, \ldots, J)$. For the purpose of describing the different approaches, we do not differentiate the treatment indicator from other covariates and do not specify interactions between covariates, they could be incorporated in the definition of x_{ij}. We considered the following five models:

- Naive model (NA): A first approach considers that all data comes from a single population, and therefore assumes that there is no heterogeneity. In this model, a common intercept and common predictor effects are included. This naive approach can lead to bias when heterogeneity is actually present. The model can be expressed as:

$$\text{logit}(p_{ij}|x_{ij}) = \alpha + \theta' x_{ij} \quad (1)$$

where p_{ij}refers to the probability of subject i in trial j to develop the outcome. When individual predictions are made to estimate the ITE at a covariate level x, these predictions are obtained by $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}'x)$, where $\hat{\alpha}$ and $\hat{\theta}$ are the Maximum likelihood estimators of α and θ respectively, in the model (1).
- Random intercept model (RI): A second approach is to assume that the heterogeneity in the IPD-MA occurs only on the baseline risk i.e. the intercept varies between studies, but the effects of all predictors are the same in each study. In this model, we consider a random study effect to model the distribution of the intercept across studies. The underlying model can be written as:

$$\text{logit}(p_{ij}|x_{ij}, \alpha_j) = \alpha_j + \theta' x_{ij} \quad (2)$$

with $\alpha_j \sim \mathcal{N}(\alpha, \tau_\alpha^2)$. The individual predictions are obtained by $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}'x)$. Estimators of α and θ are obtained via maximum likelihood which is approximated with the adaptive Gauss-Hermite quadrature.
- Stratified intercept model (SI): A third approach is to include a different intercept for each study, as a fixed effect. With a binary outcome:

$$\text{logit}(p_{ij}|x_{ij}, \alpha_j) = \sum_{m=1}^{J} \alpha_m I(m = j) + \theta' x_{ij} \quad (3)$$

where $I(\cdot)$ denotes the indicator function. To derive the individual predictions as $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}'x)$, the estimator of θ is obtained via maximum likelihood. To obtain a single $\hat{\alpha}$, we used a random-effects meta-analysis of the α_m, with inverse variance weighting, as suggested by Debray et al. and Royston et al. [3, 14]. The choice of a random-effects meta-analysis was based on considering that using separate intercepts for each study implied that some heterogeneity would be expected.
- Fully stratified model (FS): A fourth approach is to consider that there is heterogeneity across studies on both the baseline risks and the predictors' effects. In that case, we calculate different intercept and predictor effects for each trial included in the meta-analysis. With a binary outcome:

$$\text{logit}(p_{ij}|x_{ij},\alpha_j,\theta_j) = \sum_{m=1}^{J}(\alpha_m I(m=j) + \theta'_m I(m=j)x_{ij}) \quad (4)$$

where α_m and θ_m, for $m = 1,..,J$, are real valued parameters to be estimated. This is equivalent to fitting a separate model in each study included in the meta-analysis. The individual predictions are then obtained as $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}'x)$, where a single intercept estimate $\hat{\alpha}$ and single predictor estimates$\hat{\theta}$ are obtained with a random-effects multivariate meta-analysis.

$$\begin{pmatrix}\alpha_m \\ \theta_m\end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix}\alpha \\ \theta\end{pmatrix}, V\right)$$

where V is the between-study covariance matrix of the intercept and the predictor effects.

- Rank-1 model (R1): A final approach considers that the linear predictors share a common direction in the covariate space but that the size of their effects might be systematically different [15]. This model can be thought of as an intermediate between the common effect models and the fully stratified model. In this setting, the study-specific effects can vary in a proportional way, modeled by a random effect ϕ. With a binary outcome:

$$\text{logit}(p_{ij}|x_{ij},\alpha_j,\phi_j) = \alpha_j + \phi_j\theta x_{ij} \quad (5)$$

with $\alpha_j \sim \mathcal{N}(\alpha,\tau^2_\alpha)$, $\phi_j \sim \mathcal{N}(1,\tau^2_\phi)$
With the rank-1 model, the individual predictions are acquired by $\hat{\mu}(x) = \text{expit}(\hat{\alpha} + \hat{\theta}'x)$, where both estimators are directly obtained as in the random intercept model.

The risk models using a time-to-event outcome are described in Section 1 of the Supplementary Material.

Model validation

Internal-external cross-validation (IECV) was used to validate the models. In the IECV, the model is constructed with $J-1$ studies and validated with the remaining study for each permutation of $J-1$ studies. To account for the heterogeneity in baseline risk, the model is re-calibrated in the test datasets. We first estimate the intercept with the different risk prediction models presented previously. Recalibration is then performed by estimating a regression model with the linear predictors $\hat{\theta}x$ of the original model as an offset i.e. the regression parameter is forced to be one. These steps are performed for all models except for the naive model which ignores all potential heterogeneity. To assess the models' performance, discrimination and calibration were considered. We also calculated the mean squared error.

To assess the discrimination, which is the ability of the model to distinguish between individuals who benefit and individuals who do not benefit from taking the treatment, the c-statistic for benefit proposed by van Klaveren et al. [16] was used. Since the individual benefit, i.e. obtaining a more favorable outcome when taking the treatment than when not taking it cannot be observed, van Klaveren et al. used pairs of individuals, one in each treatment group, with close predicted ITE to approach the individual benefit. The c-statistic for benefit is the extension of the c-statistic for individualized treatment effects. The c-statistic for benefit is defined as the probability that from two randomly chosen matched pairs ($p1$, $p2$) with unequal estimated benefit, the pair with greater estimated benefit also has a higher predicted probability, where the estimated benefit refers to the difference in outcomes between two patients with the same predicted benefit but with different treatment assignments. To create the pairs, a patient in the control group is matched to one in the treatment group with a similar predicted treatment benefit. Higher values of the c-statistic for benefit are better. The c-statistic for benefit can be expressed as:

$$C_{\text{for-benefit}} = P\Big(\hat{\tau}(x_{p1}) > \hat{\tau}(x_{p2}) \mid \tau(x_{p1}) > \tau(x_{p2})\Big)$$

where $\tau(x_{p1})$ and $\tau(x_{p2})$ represent the observed benefits of pairs $p1$ and $p2$ and where $\hat{\tau}(x_{p1})$ and $\hat{\tau}(x_{p2})$ represent the predicted benefits of pairs $p1$ and $p2$ respectively.

For the calibration, the agreement between the observed and the predicted benefit, was assessed by extracting the intercept and the slope of the regression line. An intercept close to 0 and a slope close to 1 indicate a good calibration. Calibration curves were also plotted when the methods were applied to the INDANA dataset. The predictions were divided into five bins; to make sure to include individuals who were allocated to the treatment and individuals who were allocated to the control. In each bin, the mean of the predicted benefit was compared to the observed benefit.

Addressing aggregation bias

An issue related to the one-stage approach is the way treatment-covariate interactions are included. Indeed, if the model is not correctly specified, it can lead to aggregation bias which occurs when using the information across studies modifies the interactions' estimates obtained when using only within-study information. In order to avoid aggregation bias, only within-trial interaction should be used to estimate the treatment-covariate interactions. To make sure only within-trial information is used, a solution to distinguish within- and across-trial information has been described in Riley et al. [17].

This method consists in centering the covariates to their study-specific mean and adding the covariates' mean as an adjustment term that explains between-study heterogeneity. Since within- and across-trial information are now uncorrelated, we are able to solely use within-trial information. After conducting some simulations (details are given in Section 2 of the Additional file 1) in which we compared the estimates obtained with the models described in the previous section with and without the aforementioned method, we concluded that not centering variables to their study-specific mean and not including a covariate-mean interaction term did not lead to aggregation bias with the proposed models since the estimates obtained were similar. In their paper, Belias et al. find that using this method leads to very small differences [18]. Therefore, we decided to evaluate the performance of the different models without including the method.

Implementation

All the analyses were performed in R version 4.1.2. The random intercept and the stratified intercept models were developed using `glmer` from the `lme4` package for binary outcomes and using `coxme` from the `coxme` package for time-to-event outcomes. For the rank-1 models, we used `rrvglm` in the `VGAM` package and `coxme` in `coxme`. Finally, the fully stratified model was developed using `glm` and `coxph` from `survival`.

Monte Carlo simulation study

Setting

The performance of the models and meta-learners was evaluated in a simulation study. We considered 24 scenarios in which we changed the number of covariates, the number of patients in each trial, and the type of outcome. The scenarios are briefly described below, and more details are given in Section 3 of Additional file 1. We simulated 1000 IPD-MAs composed of 7 trials for each scenario. All the continuous covariates were drawn from a normal distribution and all the binary covariates were drawn from a Bernoulli distribution. For individual i in study j, the treatment allocation t_{ij} was sampled from a Bernoulli distribution of parameter 0.5, the binary outcome y_{ij} was generated from a Bernoulli distribution of parameter p_{ij}, where $\text{logit}(p_{ij}) = \alpha_j + \theta_j x_{ij} + \gamma_j t_{ij}$ and the time-to-event outcome was generated from a Weibull distribution $f(x; k, b) = bkx_{ij}^{k-1} \exp(-bx^k)$, where k represents the shape parameter and b the scale parameter. We chose $k = 1.15$ i.e. the failure rate increases over time and $b = \frac{a}{\exp(\theta_j x_{ij})^k}$, with $a = 50$ to obtain a stretch distribution.

In 12 scenarios, data was generated with a common treatment effect (all $\gamma_j = \gamma$), whereas in the other 12, we included some variation in the predictor effects.

In scenarios 1 to 3, we considered IPD-MAs with a total number of patients equal to 2800, 1400 and 700 respectively (for simplicity, trials were of identical sample size) composed of 3 covariates, 3 treatment-covariate interactions and a binary outcome. Among the covariates, one of them was binary and the other two were continuous.

In scenarios 4 to 6, we computed IPD-MAs with a total number of patients equal to 2800, 1400 and 700 (trials were of identical sample size) composed of 9 covariates (6 binary and 3 continuous) and 4 treatment-covariate interactions.

Scenarios 7 to 12 had the same configuration as scenarios 1 to 6 but the predictor effects varied according to the trial for some variables.

Scenarios 13 to 18 had the same configuration as scenarios 1 to 6 and scenarios 19 to 24 were similar to scenarios 7 to 12 but instead of a binary outcome, we used a time-to-event outcome.

A summary of all scenarios can be found in Supplementary Table 7 of Additional file 1.

We also tackled the impact of variables' selection on the performance of the approaches. We performed variables' selection using a Group lasso [19] for scenarios 4 to 6 and 10 to 12 with the stratified intercept model. The results can be found in Supplementary Material S3.

Results

Results of scenarios 1 to 6 and 13 to 18 are available in Section 4 of Additional file 1. If outliers were found in the results, they were removed from the analysis.

With 3 covariates, all methods had a nearly equal performance, in terms of discrimination and calibration, with both meta-learners (Fig. 1). The mean c-statistic for benefit values were around 0.52 for all models, and although van Klaveren stated that it was difficult to obtain values over 0.6 [16], it still indicates poor discrimination. The calibration was mediocre, the intercept values were close to 0 but the slope values were not close to 1. The rank-1 and the stratified intercept models had higher MSE values and than the other model. With 9 covariates, the fully stratified underperformed the other models with lower discrimination and calibration as well as higher MSE values. The poor performance of FS might be due to overfitting, different intercept and predictor effects are included in the model for all studies (Fig. 2). The other models performed similarly, they had a good calibration with intercept values and slope values close to 0 and 1, respectively, and acceptable discrimination with c-statistic for benefit values around 0.6. Using the S-learner or the T-learner led to equivalent performances for the NA, RI, SI, and R1 methods. However, for the fully stratified method, it was preferable to use the S-learner

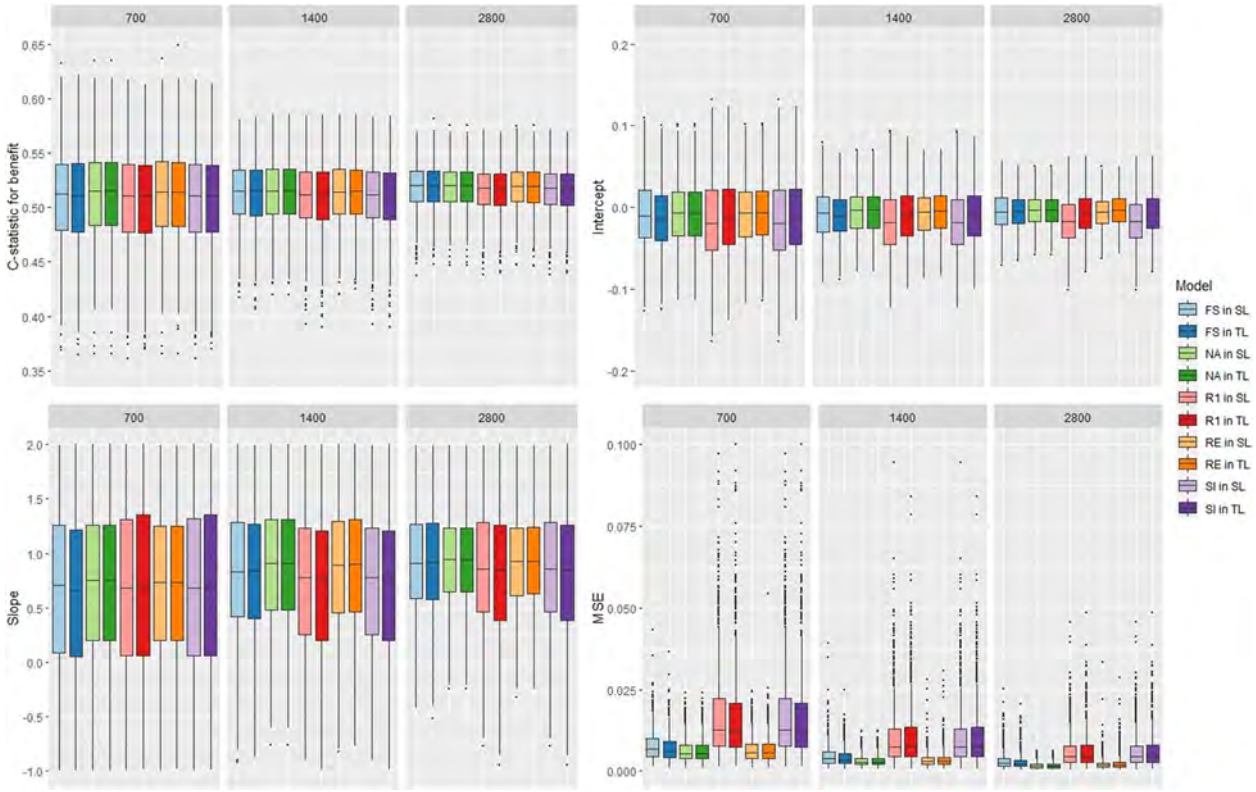

Fig. 1 Boxplot of the models' performance with 3 covariates, a binary outcome, and variation in the predictor effects

approach to obtain better results and lower MSE values. Changing the size of the IPD-MA did not impact the results. When a binary outcome is used and the model includes few covariates, we recommend using the naive, random intercept or fully stratified models which have lower MSE. When the model includes more covariates, we recommend avoiding using the fully stratified model and favoring the other methods. Despite including some heterogeneity in the predictor effects between studies, the naive model, which ignores any potential heterogeneity, did not perform worse than the other methods. The naive model might underestimate some effects and overestimate others, thus leading to a similar performance to the other methods' performances. Similar conclusions were reached in scenarios without variation in the predictor effects (results in Section 4 of Additional file 1).

When a time-to-event outcome was used with 3 covariates, we noticed that using the T-learner led to slightly better discrimination results for all methods, whereas using the S-learner led to better calibration results (Fig. 3). Slope values far from 1 indicating a poor calibration. Higher MSE values were obtained for the fully stratified model and for the rank-1 model when the T-learner was used. The NA, RI, and SI methods' results were similar.

With 9 covariates, using FS led to better calibration results but led to worse discrimination (Fig. 4). The other methods produced analogous discrimination results and had mean c-statistic for benefit values above 0.65. FS had the higher MSE. Overall, using the S-learner led to more stable results and led to lower MSE values.

When a time-to-event outcome is used, we recommend choosing the S-learner approach to estimate ITEs. No methods outperformed the others but with several covariates, the fully stratified model had the best calibration. Similar conclusions were obtained when variation in the predictor effects was not included (results in Section 4 of Additional file 1).

In scenarios where variation in predictor effects was included across studies, the rank-1 and the fully stratified models, which are the models that capture more heterogeneity, did not stand out from the other models and did not lead to better ITE estimation. The fully stratified, which estimates separate intercept and predictor effects for each study included in the meta-analysis, is prone to overfitting. This overfitting was seen in our results, particularly in scenarios with more

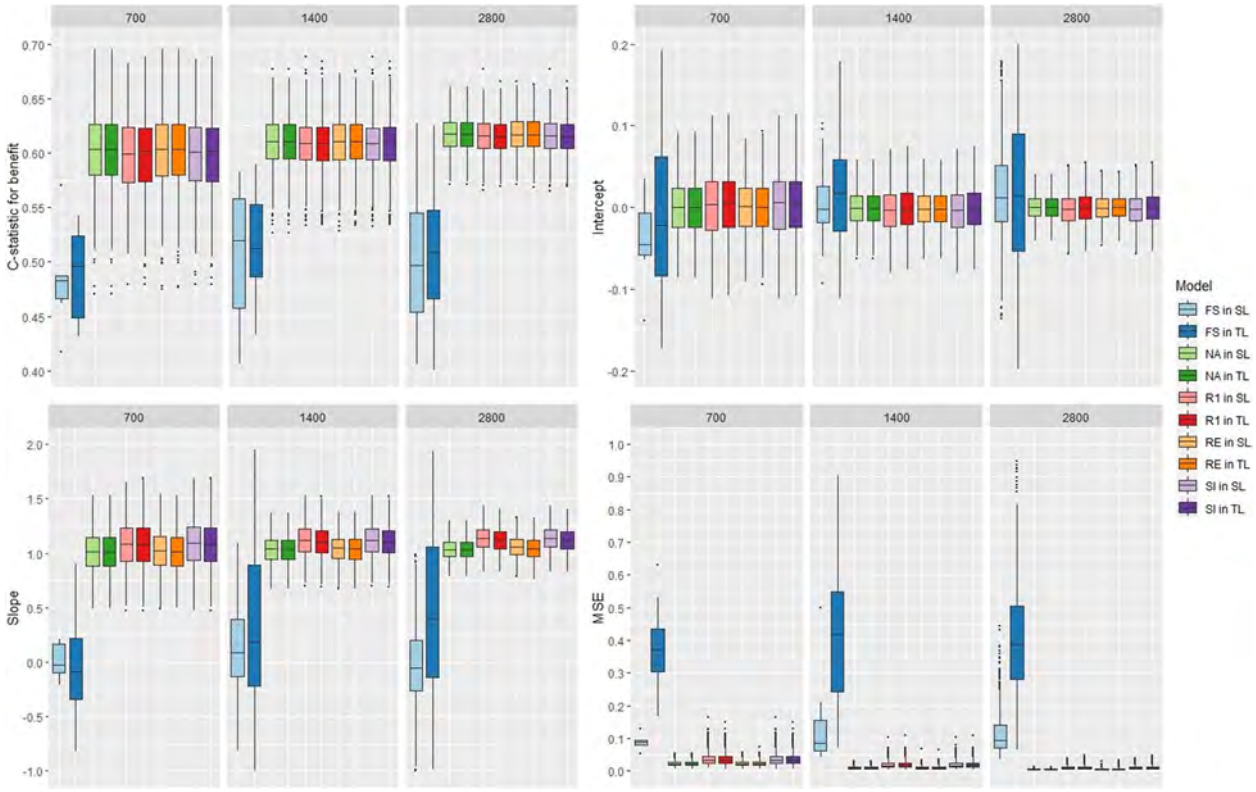

Fig. 2 Boxplot of the models' performance with 10 covariates, a binary outcome, and variation in the predictor effects

covariates. The rank-1 model allows the predictor effects to vary in a proportional way, which might not capture well the type of heterogeneity we considered in the various scenarios (see Supplementary Material S3). In a scenario where heterogeneity was generated in a proportional way, the rank-1 model performed better (Supplementary Material S4)."

To further investigate the performance of the methods, we computed prediction intervals around the predicted ITE of all individuals for each method and calculated the number of times the true ITE was included in the interval (Fig. 5). The predicted ITEs correspond to what would be expected as a treatment effect for an individual with characteristics x_i, had they been assigned to an average trial. In scenarios 7 to 9, FS's prediction intervals were the ones that included the true ITE the most. In scenarios with 9 covariates and a binary outcome (scenarios 10 to 12), the true ITE was more often in the intervals of NA and RI. With a time-to-event, the prediction intervals of FS and R1 with the S-learner included the true ITE more frequently when 3 covariates were used. With 9 covariates, R1's prediction models captured the true ITE more often. Overall, the two methods that included more heterogeneity were the ones that captured the most the true ITE in their prediction intervals.

Illustration on real data

INDANA IPD-MA

To illustrate the different approaches, we used data from the individual data analysis of antihypertensive intervention trials (INDANA) IPD-MA to evaluate the models [20]. This IPD-MA is composed of 9 randomized controlled trials comparing an antihypertensive treatment versus no treatment or a placebo, but given the large disparity between trials, notably for the variable age (See Supplementary Fig. 8, Additional file 1), we decided to compare the different methods on 4 of them for which the median age was under 60 years old. The outcome used in this project was death. The dataset was composed of 40 237 observations and 836 deaths. After comparing the calibration obtained with different combinations of variables, we decided to include the following variables in the final models: age, sex, systolic blood pressure (SBP), serum creatinine and treatment group (Table 1). Since some values were missing, we replaced them using a simple run of a multiple imputation procedure [21]. Considering that the dataset was only used for illustration, we considered that a single imputed dataset would be sufficient. For clinical research, it would be recommended to use several imputed datasets and pool the results [22]. Proper guidance for estimating ITE is lacking but could

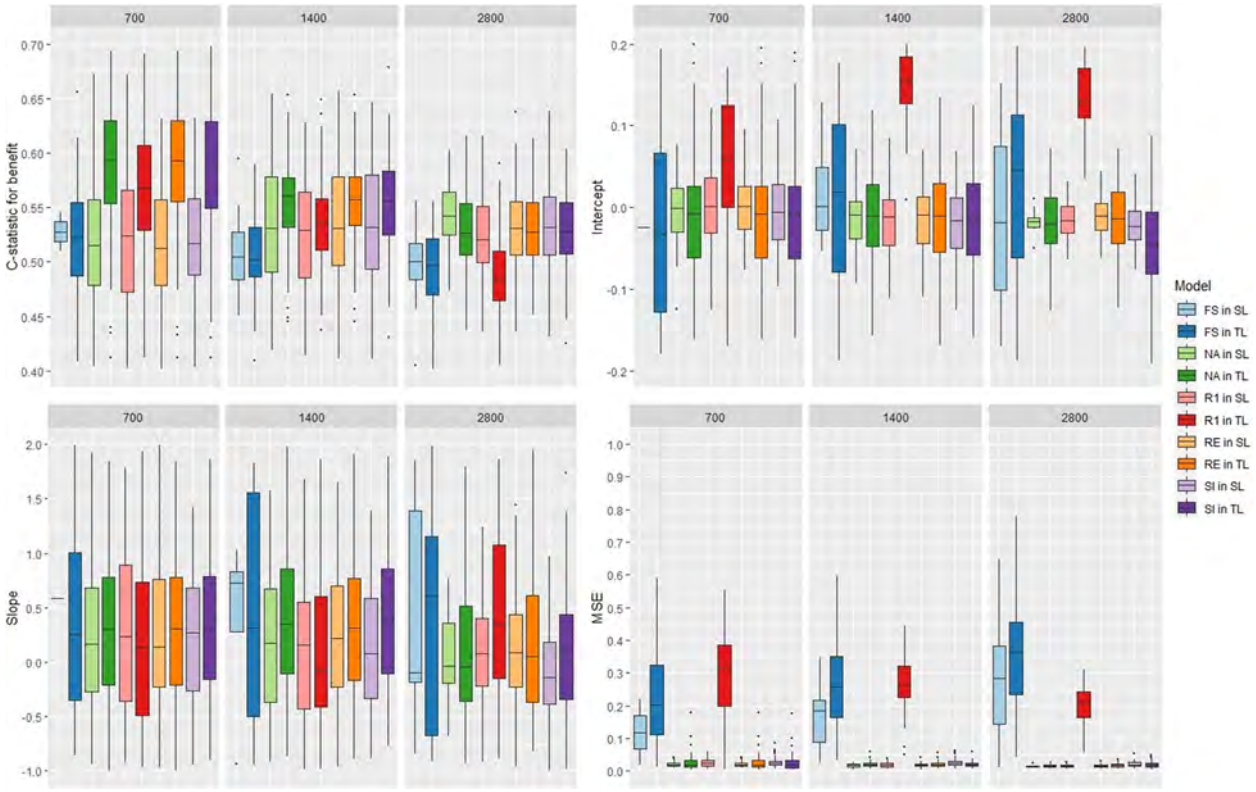

Fig. 3 Boxplot of the models' performance with 3 covariates, a time-to-event outcome, and variation in the predictor effects

be adapted from techniques used for building risk prediction models [23, 24].

Results

Considering death as a binary outcome, a higher c-statistic for benefit was obtained with the S-learner rather than the T-learner, whatever method to handle heterogeneity was used (Table 2). No significant difference was found between the five methods, with c-statistic for benefit values close to 0.5. Even recalling that van Klaveren et al. mentioned it was usual to observe a c-statistic for benefit under 0.6, our results still showed a limited discrimination for the treatment effects [16]. Despite its large sample size (40 237 observations), the dataset only contained 836 events which could also explain why it was difficult to obtain models that discriminated well. In terms of calibration, the median intercept value was close to 0 for every model, with slightly better results when the S-learner was used. With the S-learner, the naive method had a slightly better median slope and the fully stratified method gave the values further from 1. With the T-learner, the RI method had a median slope closer to 1. The SI and R1 methods gave identical median slope values with both approaches. In general, median slope values were not close to 1 which we can visualize in Fig. 6 showing that most points are not close to the diagonal. The MSE values were close to 0 and comparable for every method whatever approach was used. The naive model and the random intercept method built with the S-learner produced the best performances with the INDANA dataset.

The usefulness of adopting a personalized strategy with the INDANA dataset was assessed with three different metrics(Table 3). The individualized treatment rules developed with all methods were compared to a rule treating everyone and to a rule treating no one. The PAPE, which compares the ITR with a treatment rule that randomly treats the same proportion of patients, was also computed [25].

Results showed that there was almost no benefit of using a personalizing strategy with INDANA. PAPE were all around 0 indicating the ITRs did not improve the outcome compared to a rule that randomly treats the same proportion of patients. Similar results were obtained when comparing the ITRs to a rule that treats everyone or to a rule that treats no one. All methods performed similarly with both meta-learners. The

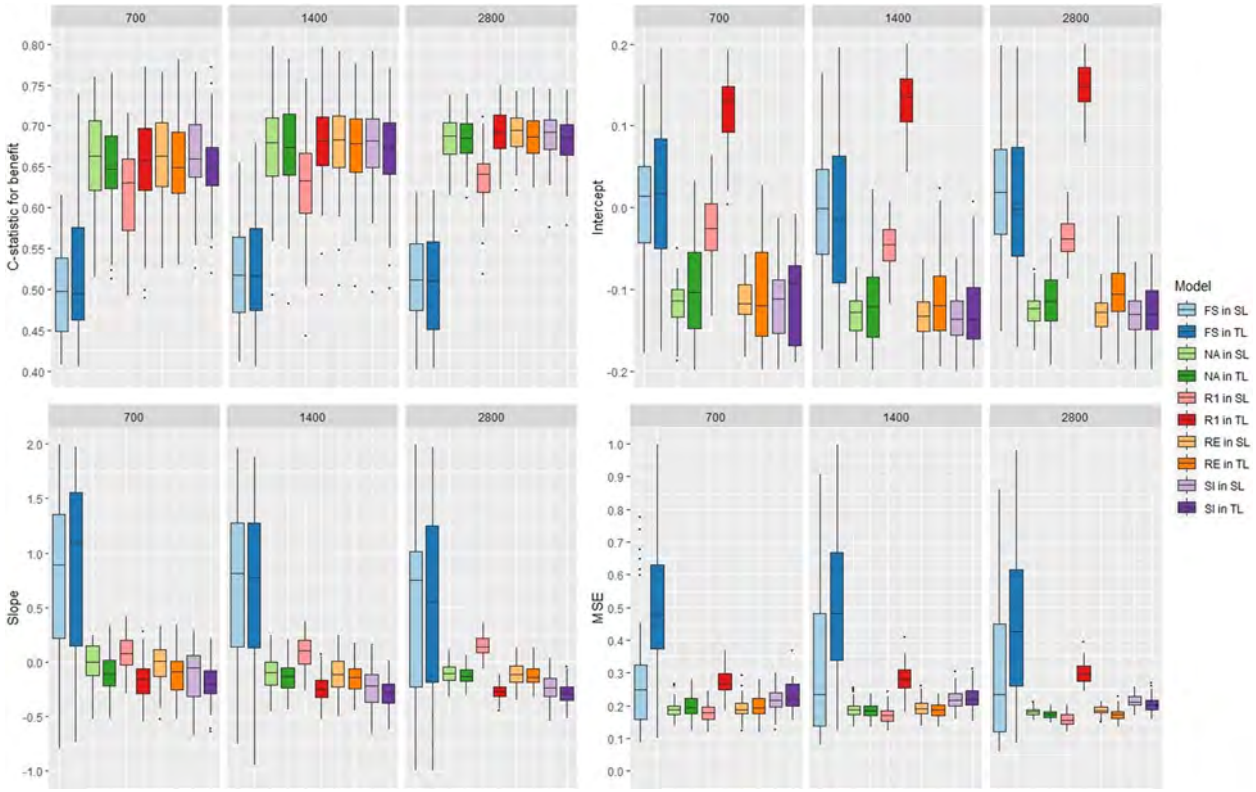

Fig. 4 Boxplot of the models' performance with 10 covariates, a time-to-event outcome, and variation in the predictor effects

limited gain in personalization might be due to the distribution of the treatment effects. Sufficient heterogeneity of treatment effects is needed to develop useful individualized treatment rules.

The distributions of the individualized treatment effects estimated with the different methods were comparable when the same approach, the S-learner or the T-learner, was used (Fig. 7). With the S-learner, all the ITEs were negative. All the ITE estimates were close to 0 which explains the fact that it was difficult to discriminate individuals benefiting from individuals not benefiting from taking the treatment and might indicate a very small treatment effect.

When considering the performance of the methods and the approaches on the train dataset, the discrimination was still low and the calibration improved a bit, especially for the SI and R1 methods (Supplementary Material S5). The c-statistic for benefit values remained close to 0.5 and the ITEs estimated were close to 0 which explains the low discrimination. The fact that the performance did not drastically increase on the train dataset might indicate that the disparity between the trials was too high for them to be meta-analyzed.

Discussion

This paper compared different approaches to estimate individualized treatment effects in an IPD meta-analysis. Using Monte Carlo simulations, the performance of those approaches was compared in terms of calibration and discrimination of the ITE. Eight approaches were considered, combining two strategies for model building (meta-learners), one where interactions between treatment and covariates are added in a regression model (S-learner), and one where two different regression models are fitted for each treatment group (T-learner), with five methods to handle the heterogeneity from the meta-analytic design: naive, random intercept, stratified intercept, fully stratified and reduced rank (rank-1) models. Both binary and time-to-event outcomes were studied. The methods were illustrated in a clinical example.

In the settings we considered, and with binary outcomes, using interactions with treatment (S-learner) or two different models (T-learner) had little impact on the model performance. With a binary outcome, we recommend avoiding using the fully stratified model when several covariates are included, as it is prone to overfitting. For time-to-event data, results were better when the S-learner approach was used but no methods stand

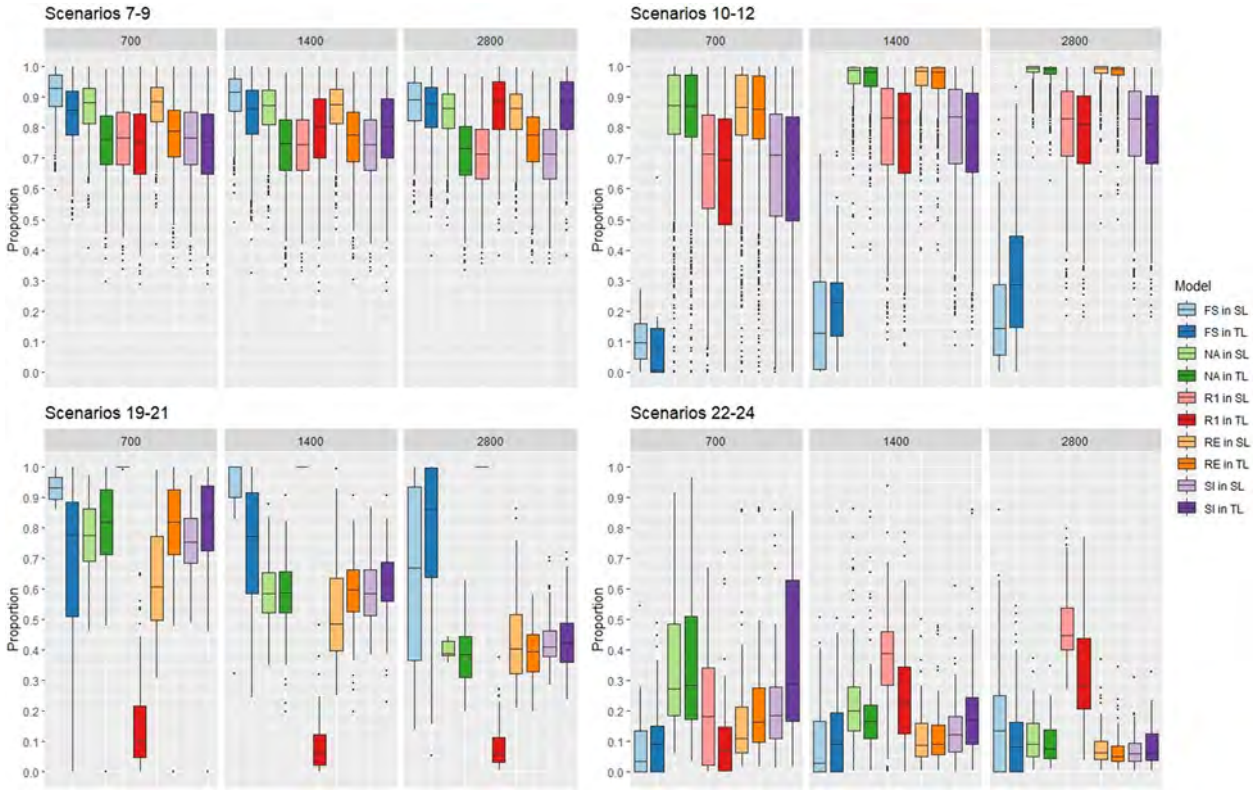

Fig. 5 Number of times the true ITEs was in the prediction intervals of each model

Table 1 Description of the predictors in each trial of the INDANA IPD-MA. The dataset with imputed missing data we analyzed is presented

Variable	ANBP	MRFIT	HDFP	MRC1
Age, mean (SD) years	50.1 (9.0)	46.9 (5.9)	50.8 (9.8)	52.1 (7.5)
Male, no. (%)	2475 (63.0)	8012 (100.0)	5910 (54.0)	9048 (52.1)
SBP, mean (SD) mmHg	154.3 (19.1)	141.1 (14.4)	158.8 (22.8)	161.6 (17.1)
Serum creatinine, mean (SD) μmol/l	87.2 (21.6)	98.0 (13.4)	94.1 (23.2)	84.8 (21.1)
Antihypertensive treatment arm, no. (%)	1988 (50.6)	4019 (50.2)	5485 (50.1)	8700 (50.1)

Table 2 Median results using INDANA with a binary outcome

	S-learner					T-learner				
	NA	RI	SI	R1	FS	NA	RI	SI	R1	FS
C-stat	0.530	0.530	0.530	0.530	0.530	0.507	0.524	0.524	0.524	0.518
Intercept	0.001	0.001	0.002	0.002	0.001	-0.003	-0.003	-0.003	-0.003	-0.003
Slope	1.433	1.453	1.460	1.460	1.961	0.268	0.727	0. 569	0. 569	0.596
MSE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

out and outperformed the others. Additionally, considering variable selection did not change the performance of the algorithms. The rank-1 and the fully stratified models that include more heterogeneity were the methods that captured more of the uncertainty around the ITE prediction, and their prediction intervals included the true

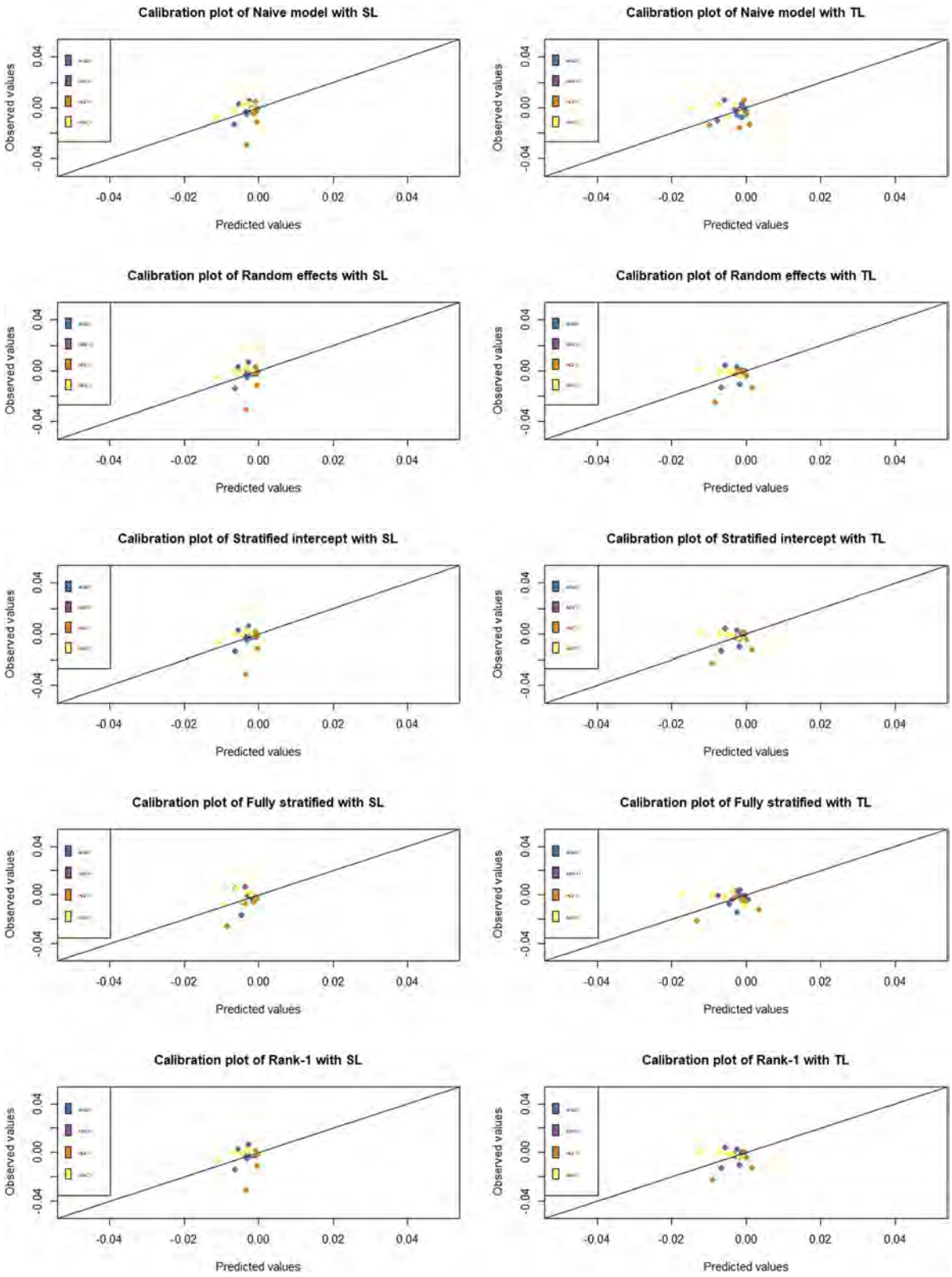

Fig. 6 Calibration plots of the models built with S-learner (left) and T-learner (right) using INDANA with a binary outcome

Table 3 Metrics to assess the usefulness of personalization on the INDANA dataset

	S-learner					T-learner				
	NA	**RI**	**SI**	**R1**	**FS**	**NA**	**RI**	**SI**	**R1**	**FS**
PAPE	0	0	0	0	0	0.001	0	0.001	0.001	0.001
$V(r) - E(Y(0))$	-0.003	-0.003	-0.003	-0.003	-0.003	-0.002	-0.002	-0.002	-0.002	-0.002
$V(r) - E(Y(1))$	0	0	0	0	0	0.001	0.001	0.001	0.001	0.001

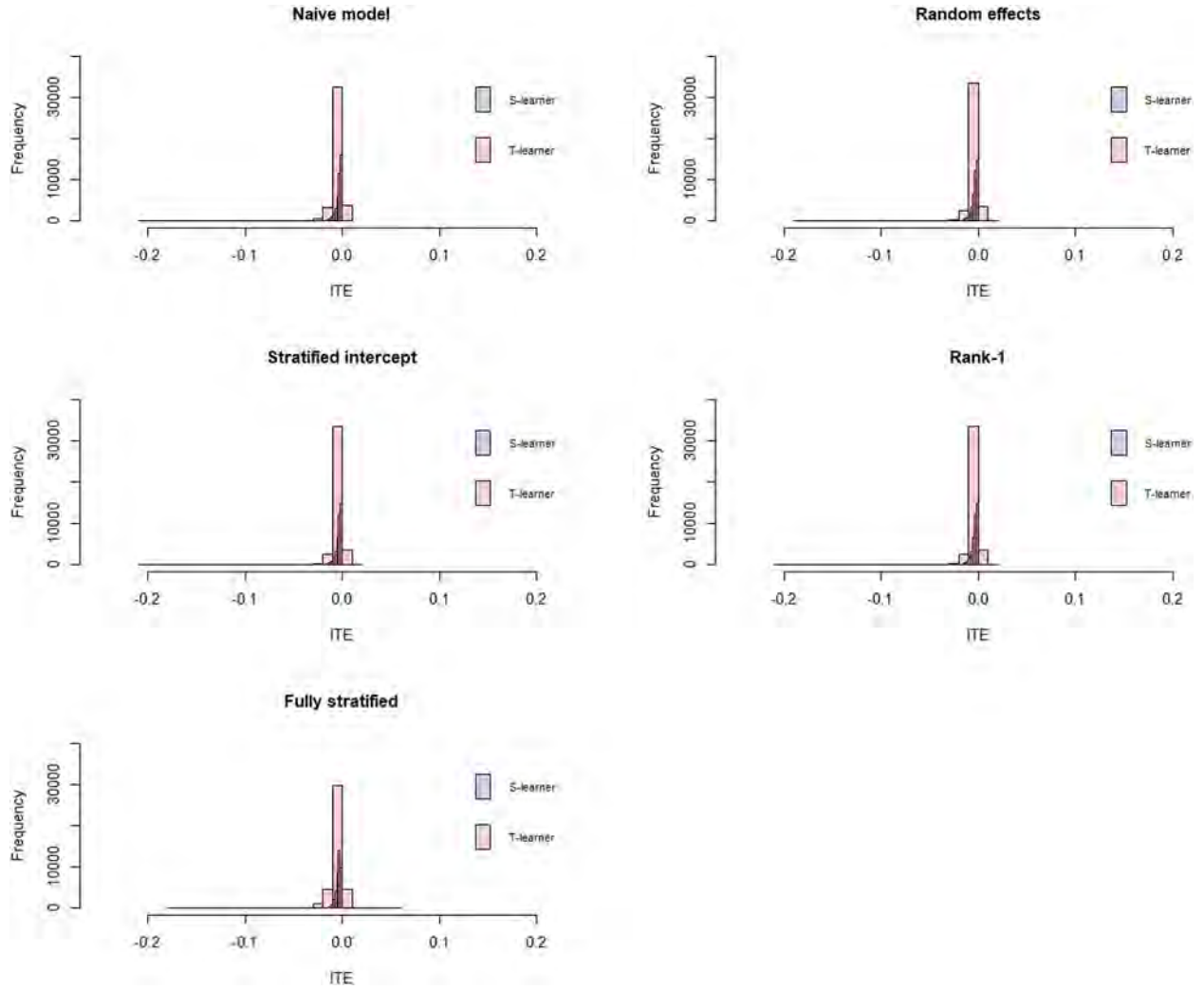

Fig. 7 ITE distribution of the models with the S-learner (blue) and the T-learner (red)

ITE more often than the prediction intervals of the other methods.

In this paper, the ITE was estimated using a one-stage approach. Estimation of the ITE can also be done with a two-stage approach. In a two-stage approach, Fisher et al. [5] advised only considering within-trial interaction i.e. calculating the difference of predicted outcomes in each trial and then comparing the results between trials. Chalkou et al. [6], who also used a two-stage approach to estimate the ITE with IPD-MA within a NMA framework, found that using a pre-specified model (a model with previously identified prognostic factors) rather than a LASSO model yielded better results.

To our knowledge, only one other study investigated the performance of ITE estimation in IPD-MA in a one-stage approach [7] but focusing on variable selection

in models with treatment-by-predictor interactions (S-learner) only. In our work, we also considered the more flexible T-learner approach, whose performance was close to the S-learner when a binary outcome was considered. Since the T-learner may allow non-parametric interactions between treatment and predictors, our results for binary outcomes, slightly differ from recent reports suggesting that models with effect interactions were prone to over-fitting [26]. This may be explained by the use of IPD-MA. Indeed, we had hypothesized that in IPD-MA, where the number of predictors is often limited and the sample size is large, the issues related to over-fitting could be less important. Both simulations and analysis of a real dataset confirmed this hypothesis. We also focused more on the performance of ITE estimation in terms of calibration and discrimination (c-statistic for benefit) of the treatment effect estimate, i.e., on the models' ability to correctly estimate the ITE and separate between individuals benefiting from the treatment and others. Discrimination and calibration are both important to guide treatment decisions.

Steyerberg et al. [4], who compared three of the methods presented in this paper (naive, random intercept, and rank-1) for risk prediction models using an IPD-MA, concluded that rank-1 was the most appropriate method. Here, to estimate the ITE with an IPD-MA, rank-1 did not perform well, especially with a time-to-event outcome.

When deriving predictions for individuals from a new study, we chose to compute a single intercept by performing a meta-analytic approach. Other methods can be employed such as selecting the intercept from the most similar study or estimating the intercept using the outcome prevalence [3]. In our case, without any specific target population in the simulation study and real data analysis, we considered that pooling the intercepts with a meta-analysis would be adequate.

In this work, we decided to use regression to estimate the ITEs. One limitation of regression models is the risk of model misspecification and its impact on estimated ITEs. Within the framework of meta-learners, it is possible to use non-parametric machine-learning approaches such as random forests. Moreover, more robust methods exist to estimate ITEs, such as the R-learner and the DR-learner [9, 27]. Robust approaches for creating individualized treatment rules (ITRs) without estimating the ITEs by predicting the benefit under both treatments exist like the modified covariate method [28] or A-learning [12]. Additionally, approaches without explicitly relying on the ITE estimation can also be used like the constrained single-index regression [29] and many others [30–33]. However, it is unclear how to account for the heterogeneity that may arise between studies. To our knowledge, the only proposal of a non-regression-based approach to ITR development with IPD-MA data is a paper by Mistry et al. using recursive partitioning [34]. Investigating how to adapt those approaches that are less sensitive to model misspecification to incorporate an estimation of heterogeneity between studies should be studied in further work, for instance, building on approaches for federated learning [35, 36]. In our simulation settings, with a large sample size, and no complex interactions or non-linearities between variables, the regression models we used are expected to perform well, and there might be no clear advantage of more complicated approaches. But in more complex situations, this may not be the case, and these remain to be investigated as a follow-up of this work.

In this project, we chose to concentrate on ITE estimation using data from randomized control trials. In practice, personalized strategies have been developed using data from RCTs. The SYNTAX score II which aims to guide decision-making between coronary artery bypass graft surgery (CABG) and percutaneous coronary intervention (PCI) in patients with complex coronary artery disease was developed by Farooq et al. using data from the SYNTAX trial [2]. Other types of data could have been used to estimate ITEs such as observational data, provided that the confounding factors are accounted for. Most approaches for ITE estimation can be used with both RCTs and observational data. Another type of data that is often used in personalized medicine is genomics data. Some penalized methods that can be combined with the meta-learners used in this work have been proposed such as the group LASSO [19]. Studying how to correctly use the risk models that tackle heterogeneity with genomics data might be worth investigating.

We focused on estimating ITEs with a binary treatment. ITE estimation can also be done with multiple treatments. Extension of the meta-learners, that we used in this project, to multiple treatments have been done and compared in previous works [37, 38]. However, it is rare to find a meta-analysis of RCTs that all compare the same set of treatments in practice. The situation would likely be more relevant in the network meta-analysis (NMA) setting. To our knowledge, very few works have tackled the estimation of heterogeneous treatment effects in the NMA context [6]. How to handle heterogeneity between studies within a NMA framework for the ITE estimation may still warrant further study and developments.

IPD-MA benefits from a large sample size, which can facilitate the ITE estimation and can increase generalizability by including trials with not the exact same population. However, some challenges arise in using IPD-MA for estimating individualized treatment effects. Dealing with heterogeneity due to differences

between trials is difficult and it was translated into poor discrimination and calibration in our case study. Another challenge, when a one-stage approach is used, is aggregation bias. Centering variables to their study-specific mean and including the covariate mean as an adjustment term can address this challenge [17].

Extensions of the present work could include the use of observational data instead of randomized control trial data. A further extension with observational data would be to develop methods to estimate this type of prediction models while allowing the datasets to remain located in different data warehouses, similar to the concept of federated learning [35, 36].

Conclusion

In this paper, the performance of different strategies and methods dealing with heterogeneity between trials was evaluated to estimate the ITE with IPD-MA in a simulation study and in a clinical example. Results showed that, for the choice of the strategy, using interactions with treatment (S-learner) is preferable as it performs well with both binary and time-to-event outcomes and that, for the choice of the method, none of the methods we compared outperformed the other methods.

Abbreviations

ITE	Individualized treatment effect
IPD-MA	Individual participant data meta-analysis
RI	Random intercept
SI	Stratified intercept
R1	Rank-1
FS	Fully stratified
SL	S-learner
TL	T-learner

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02202-9.

Supplementary Material 1.

Acknowledgements

The authors wish to thank the INDANA collaboration and the investigators of the individual studies for providing the dataset to illustrate this work.

Authors' contributions

Study concept and design: F.B., R.P., A.C. Analysis and interpretation of data: F.B., A.C., E.P., F.G., G.G., R.P. Drafting of the manuscript: F.B., R.P. Critical revision of the manuscript for important intellectual content: F.B., A.C., E.P., F.G., G.G., R.P. The authors read and approved the final manuscript.

Funding

F.B. and R.P. acknowledge support by the French Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). This work was partially funded by the Agence Nationale de la Recherche, under grant agreement no. ANR-18-CE36-0010-01.

Availability of data and materials

The data that support the findings of this study are available from the INDANA collaboration but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the INDANA collaboration. The code of the simulation study is available at https://github.com/floriebouvier/ITE-estimation-with-IPD-MA.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 13 February 2023 Accepted: 15 March 2024
Published online: 25 March 2024

References

1. Ballarini NM, Rosenkranz GK, Jaki T, Konig F, Posch M. Subgroup identification in clinical trials via the predicted individual treatment effect. PLoS One. 2018;13(10):e0205971. https://doi.org/10.1371/journal.pone.0205971.
2. Farooq V, van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A, et al. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. Lancet. 2013;381(9867):639–50. https://doi.org/10.1016/S0140-6736(13)60108-7.
3. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32(18):3158–80. https://doi.org/10.1002/sim.5732.
4. Steyerberg EW, Nieboer D, Debray TPA, van Houwelingen HC. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Stat Med. 2019;38(22):4290–309. https://doi.org/10.1002/sim.8296.
5. Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? BMJ. 2017;356:j573. https://doi.org/10.1136/bmj.j573.
6. Chalkou K, Steyerberg E, Egger M, Manca A, Pellegrini F, Salanti G. A two-stage prediction model for heterogeneous effects of many treatment options: application to drugs for Multiple Sclerosis. 2020. https://arxiv.org/pdf/2004.13464.pdf. Accessed 3 June 2021.
7. Seo M, White IR, Furukawa TA, Imai H, Valgimigli M, Egger M, et al. Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis. Stat Med. 2020;40(6):1553–73. https://doi.org/10.1002/sim.8859.
8. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A. 2019;116(10):4156–65. https://doi.org/10.1073/pnas.1804597116.
9. Kennedy EH. Optimal doubly robust estimation of heterogeneous causal effects. arXiv. 2020. https://doi.org/10.48550/ARXIV.2004.14497.
10. Susan Athey, Julie Tibshirani, Stefan Wager "Generalized random forests. Ann Stat Ann Statist. 2019;47(2):1148–78.
11. Guo X, Ni A. Contrast weighted learning for robust optimal treatment rule estimation. Stat Med. 2022;sim.9574. https://doi.org/10.1002/sim.9574.
12. Chen S, Tian L, Cai T, Yu M. A general statistical framework for subgroup identification and comparative treatment scoring. Biometrics. 2017;73(4):1199–209. https://doi.org/10.1111/biom.12676.
13. Gueyffier F, Boutitie F, Boissel J, Coope J, Cutler J, Ekbom T, et al. INDANA: a meta-analysis on individual patient data in hypertension. Protocol and preliminary results. Thérapie. 1995;50:353–562.

14. Royston P, Parmar MKB, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Stat Med. 2004;23(6):907–26. https://doi.org/10.1002/sim.1691.
15. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381. https://doi.org/10.1371/journal.pmed.1001381.
16. van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. J Clin Epidemiol. 2018;94:59–68. https://doi.org/10.1016/j.jclinepi.2017.10.021.
17. Riley RD, Debray TPA, Fisher D, Hattle M, Marlin N, Hoogland J, et al. Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: Statistical recommendations for conduct and planning. Stat Med. 2020;39(15):2115–37. https://doi.org/10.1002/sim.8516.
18. Belias M, Rovers MM, Reitsma JB, Debray TPA, IntHout J. Statistical approaches to identify subgroups in meta-analysis of individual participant data: a simulation study. BMC Med Res Methodol. 2019;19(1):183. https://doi.org/10.1186/s12874-019-0817-6.
19. Lim M, Hastie T. Learning interactions through hierarchical group-lasso regularization. 2013. https://arxiv.org/abs/1308.2719. Accessed 3 June 2021.
20. Pocock SJ, McCormack V, Gueyffier F, Boutitie F, Fagard RH, Boissel JP. A score for predicting risk of death from cardiovascular disease in adults with raised blood pressure, based on individual patient data from randomised controlled trials. BMJ. 2001;323(7304):75–81. https://doi.org/10.1136/bmj.323.7304.75.
21. Quartagno M, Carpenter JR. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. Stat Med. 2016;35(17):2938–54. https://doi.org/10.1002/sim.6837.
22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377–99. https://doi.org/10.1002/sim.4067.
23. Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63(2):205–14. https://doi.org/10.1016/j.jclinepi.2009.03.017.
24. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. Biom J. 2015;57(4):614–32. https://doi.org/10.1002/bimj.201400004.
25. Imai K, Li ML. Experimental Evaluation of Individualized Treatment Rules. J Am Stat Assoc. 2021;118(541):242–56. https://doi.org/10.1080/01621459.2021.1923511.
26. van Klaveren D, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. J Clin Epidemiol. 2019;114:72–83. https://doi.org/10.1016/j.jclinepi.2019.05.029.
27. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika. 2020;asaa076. https://doi.org/10.1093/biomet/asaa076.
28. Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. J Am Stat Assoc. 2014;109(508):1517–32. https://doi.org/10.1080/01621459.2014.951443.
29. Park H, Petkova E, Tarpey T, Ogden RT. A Single-Index Model With a Surface-Link for Optimizing Individualized Dose Rules. J Comput Graph Stat. 2022;31(2):553–62. https://doi.org/10.1080/10618600.2021.1923521.
30. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012;68(4):1010–8. https://doi.org/10.1111/j.1541-0420.2012.01763.x.
31. Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR. Doubly Robust Learning for Estimating Individualized Treatment with Censored Data. Biometrika. 2015;102(1):151–68. https://doi.org/10.1093/biomet/asu050.
32. Mo W, Qi Z, Liu Y. Learning Optimal Distributionally Robust Individualized Treatment Rules. J Am Stat Assoc. 2021;116(534):659–74. https://doi.org/10.1080/01621459.2020.1796359.
33. Zhao YQ, Zeng D, Tangen CM, Leblanc ML. Robustifying trial-derived optimal treatment rules for a target population. Electron J Stat. 2019;13(1):1717–43. https://doi.org/10.1214/19-EJS1540.
34. Mistry D, Stallard N, Underwood M. A recursive partitioning approach for subgroup identification in individual patient data meta-analysis. Stat Med. 2018;37(9):1550–61. https://doi.org/10.1002/sim.7609.
35. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc. 2012;19(5):758–64. https://doi.org/10.1136/amiajnl-2012-000862.
36. Lu CL, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc. 2015;22(6):1212–9. https://doi.org/10.1093/jamia/ocv083.
37. Zhao Y, Fang X, Simchi-Levi D. Uplift modeling with multiple treatments and general response types. 2017. http://arxiv.org/abs/1705.08492. Accessed 5 July 2023.
38. Acharki N, Lugo R, Bertoncello A, Garnier J. Comparison of meta-learners for estimating multi-valued treatment heterogeneous effects. 2023. http://arxiv.org/abs/2205.14714. Accessed 11 Apr 2023.

Publisher's Note

Chapitre 3

Comparaison de méthodes de développement de règles individualisées de traitement sur données réelles

3.1 Résumé du projet

3.1.1 Introduction et objectifs

Un aspect fondamental de la médecine personnalisée consiste à identifier les patients qui bénéficieraient de traitements spécifiques, ce qui permet de créer des règles de traitement individualisé (ITR). Par essence, les ITR sont des règles de décision qui recommandent un traitement en fonction des caractéristiques des patients. Les règles de traitement optimales sont particulièrement intéressantes, car elles conduisent au meilleur résultat moyen dans la population si elles sont suivies par tous les individus [45].

Au cours de la dernière décennie, de nombreuses méthodes visant à construire des ITR ont été proposées. Cependant, leurs performances relatives restent floues, et il n'est pas certain que ces ITR dérivées conduisent à recommander le même traitement pour les mêmes individus. Cette question mérite d'être étudiée car si les ITR ne sont pas similaires, il est important de le savoir en amont lors du choix d'une méthode pour dériver un ITR dans la vie réelle.

Des études antérieures ont effectué des comparaisons parmi les méthodes développant une ITR à travers l'estimation des ITE. Plus précisément, les études réalisées par Jacob et Zhang et

al. ont exploré l'efficacité de divers méta-learners (incluant le T-learner, S-learner, X-learner, DR-learner, et R-learner) ainsi que des *causal forests* [68, 69]. L'étude de Jacob a mis en lumière que ces méthodes produisent des estimations hétérogènes des ITE, suggérant l'avantage d'appliquer et de comparer plusieurs approches en pratique. De même, Zhang et al. ont observé des performances différentes parmi ces méthodes. À ce jour, il semble qu'il n'y ait pas de comparaison complète englobant toutes les méthodes que nous avons choisi, en particulier celles qui construisent une ITR sans nécessiter le calcul direct des ITE.

Dans cette étude, nous avons cherché à comparer un large éventail de méthodes pour développer des ITR, à la fois en termes de performance et de concordance. Nous avons comparé 22 méthodes différentes et les avons appliquées à partir des données de deux essais contrôlés randomisés : l'International Stroke Trial (IST) [70] et l'essai CRASH-3 [71].

3.1.2 Méthodes

Nous suivons le cadre des "résultats potentiels" (*potential outcomes framework*) de Rubin [27]. Nous supposons que nous avons accès à un échantillon d'observations indépendantes et identiquement distribuées. Soit $X \in \mathcal{X} \subset \mathbb{R}^n$ un vecteur de covariables dans l'espace des covariables $\mathcal{X}$, $A \in \{0, 1\}$ une variable indicatrice pour le traitement d'intérêt et $Y \in \{0, 1\}$ un résultat binaire. Nous introduisons les résultats potentiels Y^0 et Y^1 qui représentent les résultats binaires qui seraient observés si les patients étaient assignés respectivement au traitement contrôle ou au traitement expérimental. Sans perte de généralité, nous supposons que $Y = 1$ est un événement souhaitable.

3.1.2.1 Métriques

Nous avons d'abord utilisé des métriques permettant d'évaluer la performance de chaque ITR individuellement. Les métriques de performance que nous avons sélectionné sont les suivantes : la valeur de la règle, B_{pos} et B_{neg}, le *population average prescriptive effect* et la *c-statistic for benefit*. Ces métriques sont décrites dans l'introduction ainsi que dans l'article de la section 3.2.

Deux métriques permettant de vérifier si deux ITR ont la même recommandation et s'accordent pour attribuer le traitement aux mêmes patients ont également été utilisées.

— Coefficient de corrélation de Matthews (CMC) : le CMC est utilisé pour mesurer le désaccord, en termes de patients traités, entre deux règles [72]. Les valeurs sont comprises entre -1 et 1, où 1 indique une corrélation positive parfaite, 0 indique l'absence de corrélation et -1 indique une corrélation négative parfaite.

— Coefficient kappa de Cohen : Le coefficient kappa de Cohen mesure l'accord entre deux règles en considérant le nombre d'accords et de désaccords [73]. Il peut varier de -1 à 1. Une valeur inférieure à 0 indique que l'accord entre les deux règles est moins que probable, une valeur de 0 indique qu'il n'y a pas d'accord et une valeur de 1 signifie qu'il y a un accord parfait entre les règles.

3.1.2.2 Approches pour développer des ITR

D'un point de vue conceptuel, on peut distinguer deux catégories de méthodes d'élaboration d'une ITR. Une première classe dans laquelle l'ITR est construite en modélisant d'abord l'ITE, et une règle optimale est établie en donnant le traitement évalué aux individus ayant un ITE positif, et une seconde classe, dans laquelle l'ITR est directement estimée sans calculer les effets de traitement individualisés, et où une règle optimale est établie en minimisant le risque de la valeur de la règle via une fonction de perte. Dans la première classe, deux approches distinctes peuvent être utilisées pour obtenir l'ITE : soit l'estimation de la différence attendue des résultats potentiels entre les traitements, soit l'estimation de l'ITE directement au moyen d'une fonction de contraste. La majorité des méthodes appartiennent à la première catégorie : les méta-learners (T-learner, S-learner, X-learner, DR-learner et R-learner, avec des modèles paramétriques et non paramétriques) [29, 30, 31, 32, 33], PATH [21], *causal forests* [34], *virtual twins* [37], *A-learning* et la *modified covariate method* [38, 39], tandis que *outcome weighted learning* [46] et *contrast weighted learning* [47] appartiennent à la deuxième catégorie. La figure 3.1 présente une classification des méthodes en fonction de la manière dont elles construisent une règle de traitement optimale.

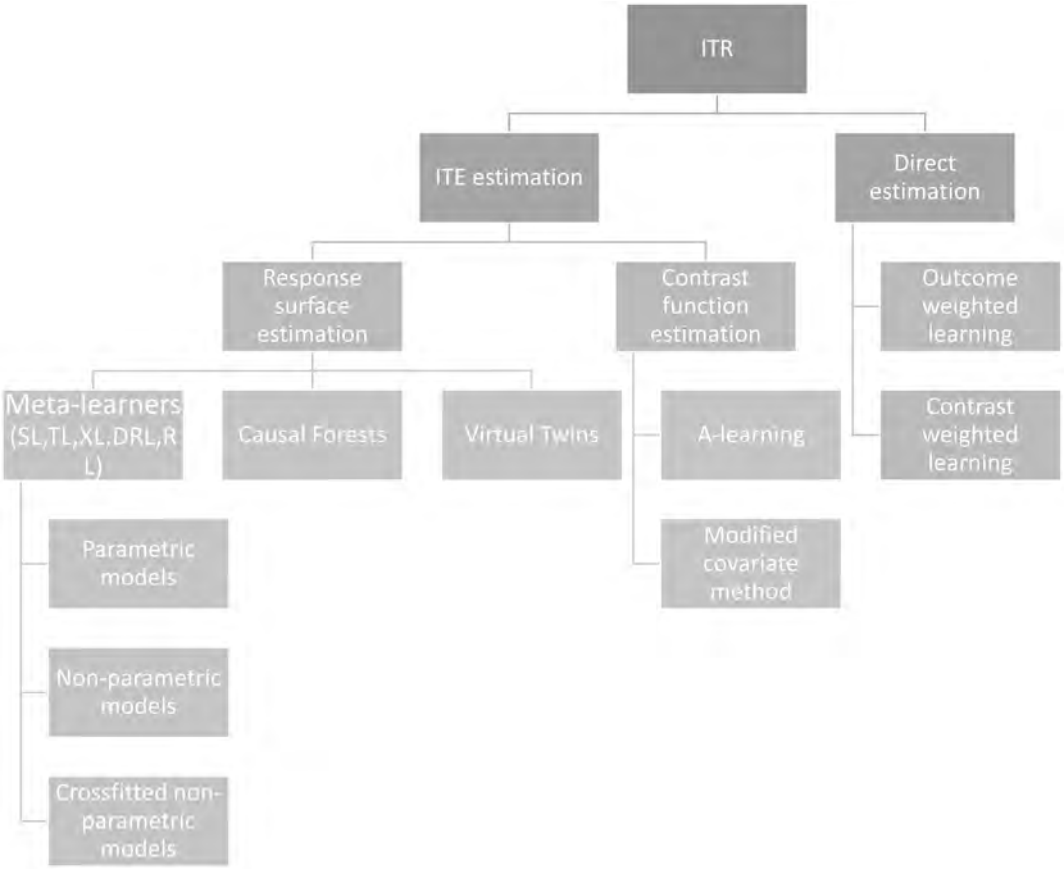

Figure 3.1 – Classification des méthodes.

3.1.3 Résultats

Les résultats ont montré qu'il y avait une grande variabilité dans les recommandations de traitement entre les différentes approches utilisées pour déterminer les ITR dans les deux essais. Les proportions de patients pour lesquels le traitement été recommandé variaient considérablement en fonction de la méthode utilisée pour créer l'ITR et les coefficients de corrélation kappa de Cohen et de Matthews étaient faibles. Cependant, un niveau de concordance plus élevé a été observé entre des méthodes proches (par exemple, entre tous les méta-learners utilisant des modèles paramétriques ou des modèles non paramétriques avec "ajustement croisé" (*cross-fitting*)). Dans l'ensemble, l'évaluation des performances des ITR dans un jeu de validation retenu (33% de l'échantillon original sélectionné au hasard) a révélé que l'efficacité de toutes les ITR était généralement limitée, quelles que soient leurs performances dans le jeu d'entraînement. Cela suggère un fort potentiel d'optimisme pour les algorithmes.

3.1.4 Discussion

Les résultats limités en termes de performance pourraient être dus à la distribution des effets de traitement et au niveau d'hétérogénéité. Bien qu'une certaine hétérogénéité des effets du traitement ait été constatée dans les essais utilisés dans ce travail, en particulier dans l'essai CRASH-3 [71], le niveau d'hétérogénéité pourrait ne pas être suffisant pour élaborer une règle de traitement individualisé bénéfique. Ce résultat a été confirmé par la réalisation de tests du rapport de vraisemblance et le calcul d'indices d'adéquation. Pour les deux essais, les tests du rapport de vraisemblance ont permis de conclure qu'il n'y avait pas de preuve d'hétérogénéité significative et les valeurs des indices d'adéquation ont montré que les interactions entre le traitement et les covariables ne représentaient qu'un faible pourcentage de l'information prédictive. La taille de l'échantillon pourrait également expliquer les performances limitées. Même si les méthodes ont été comparées sur deux grands essais cliniques randomisés, il faudrait peut-être davantage de données pour obtenir de meilleures performances. Une solution pourrait consister à utiliser des MADI, car elles incluent un plus grand nombre de participants.

3.1.5 Conclusion

En résumé, les écarts considérables observés dans la manière dont les différentes méthodes attribuent le traitement indiquent que ces méthodes ne peuvent pas être utilisées de manière interchangeable. Par conséquent, le choix d'une méthode a un impact substantiel sur les patients à qui l'on recommande le traitement expérimental. Cela soulève des questions quant à la capacité d'application de ces méthodes en pratique.

La description détaillée des méthodes et des résultats de ce travail se trouve dans l'article dela section 3.2 qui a été publié en mars 2024 dans la revue *Statistics in Medicine*.

3.2 Article

Received: 29 June 2023 | Revised: 30 January 2024 | Accepted: 1 March 2024

DOI: 10.1002/sim.10059

RESEARCH ARTICLE

Statistics in Medicine WILEY

Do machine learning methods lead to similar individualized treatment rules? A comparison study on real data

Florie Bouvier[1] | **Etienne Peyrot**[1] | **Alan Balendran**[1] | **Corentin Ségalas**[2] | **Ian Roberts**[3] | **François Petit**[1] | **Raphaël Porcher**[1,4]

[1]Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), Université Paris Cité and Université Sorbonne Paris Nord, Paris, France

[2]Bordeaux Population Health Research Center, Université de Bordeaux, Inserm, Bordeaux, France

[3]Clinical Trials Unit, London School of Hygiene & Tropical Medicine, London, UK

[4]Centre d'Épidémiologie Clinique, Assistance Publique-Hôpitaux de Paris, Hôtel-Dieu, Paris, France

Correspondence
Florie Bouvier, Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), Hôpital Hôtel-Dieu, 1 place du Parvis de Notre-Dame, F-75004 Paris, France.
Email: florie.brion-bouvier@u-paris.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-18-CE36-0010-01

Identifying patients who benefit from a treatment is a key aspect of personalized medicine, which allows the development of individualized treatment rules (ITRs). Many machine learning methods have been proposed to create such rules. However, to what extent the methods lead to similar ITRs, that is, recommending the same treatment for the same individuals is unclear. In this work, we compared 22 of the most common approaches in two randomized control trials. Two classes of methods can be distinguished. The first class of methods relies on predicting individualized treatment effects from which an ITR is derived by recommending the treatment evaluated to the individuals with a predicted benefit. In the second class, methods directly estimate the ITR without estimating individualized treatment effects. For each trial, the performance of ITRs was assessed by various metrics, and the pairwise agreement between all ITRs was also calculated. Results showed that the ITRs obtained via the different methods generally had considerable disagreements regarding the patients to be treated. A better concordance was found among akin methods. Overall, when evaluating the performance of ITRs in a validation sample, all methods produced ITRs with limited performance, suggesting a high potential for optimism. For non-parametric methods, this optimism was likely due to overfitting. The different methods do not lead to similar ITRs and are therefore not interchangeable. The choice of the method strongly influences for which patients a certain treatment is recommended, drawing some concerns about their practical use.

KEYWORDS
comparison study, individualized treatment rule, machine learning, personalized medicine

1 | INTRODUCTION

Personalized medicine aims at tailoring a treatment strategy to the individual characteristics of each patient. An essential part of personalized medicine is identifying patients benefiting from a given treatment which allows the construction of individual treatment rules (ITRs). Briefly, ITRs are decision rules that recommend treatment based

on patients' characteristics. Of particular interest are optimal treatment rules, which are rules that would lead to the best average outcome in the population if they were followed by all individuals.[1]

ITRs can be developed using data from randomized controlled trials (RCTs) or observational data. For instance, Farooq et al developed the SYNTAX score II to guide decision-making between coronary artery bypass graft surgery (CABG) and percutaneous coronary intervention (PCI) in patients with complex coronary artery disease using data from the SYNTAX trial.[2] In this paper, we decided to focus on ITRs built from RCTs' data to avoid having to additionally handle confounding factors. However, all the approaches presented here could also be used with observational data.

The PATH statement outlines guidelines for conducting predictive analyses of heterogeneous treatment effects (HTE) in clinical trials. It establishes criteria for predicting HTE and thus developing ITRs, emphasizing the use of a risk modeling approach.[3,4] However, alternative approaches have also been employed for predicting HTE,[5] and a myriad of methods whose goal is to construct an ITR has been proposed in the last decade. Nonetheless, their relative performance is not clearly established and, more importantly, it is not clear whether the derived ITRs would lead to recommending the same treatment for the same individuals. This issue is worth studying because if the ITRs are not similar, it is important to know upstream when choosing a method to derive an ITR in real life.

From a conceptual viewpoint, two classes of methods to develop an ITR can be distinguished. The first class relies on deriving individualized treatment effects (ITE) and then an ITR by recommending treatment to those with a predicted benefit. This class can be further divided into two sub-classes: methods estimating the response surfaces and methods directly estimating the ITE via a contrast function. The second class comprehends methods that directly estimate the ITR without explicitly relying on estimating ITEs or a contrast function.

Some comparisons of methods constructing an ITR via the ITE have been performed in the past. In particular, Jacob and Zhang et al. have both studied the performance of meta-learners (T-learner, S-learner, X-learner, DR-learner and R-learner) and causal forests.[6,7] Jacob found that the methods resulted in differences in terms of ITEs estimates and recommended using multiple methods and comparing their results in practice.[6] In their paper, Zhang et al also found that the methods performed differently.[7] To our knowledge, no comparison has included all the methods we are presenting, particularly methods that directly construct an ITR without calculating ITEs. Furthermore, none of the previous works implemented metrics to assess the agreements between pairs of methods.

In this study, we aimed to compare a wide range of methods to develop ITRs, both in terms of performance and agreement. We compared 22 different methods and applied those using data from two randomized controlled trials: the International Stroke Trial (IST) and the CRASH-3 trial. The remainder of this paper is organized as follows. We start by introducing the statistical setting in Section 2. In Section 3, the different methods are presented. In Section 4, the 22 methods are compared on the two RCTs. Section 5 concludes with a discussion.

2 | STATISTICAL SETTING

In this section, the potential outcomes framework is introduced. Then, we explain how to construct an individualized treatment rule (ITR). Finally, we enumerate the metrics used to compare the ITRs.

2.1 | Causal framework

We follow Rubin's potential outcomes framework.[8] We assume access to an independent and identically distributed sample of observations. Let $X \in \mathcal{X} \subset \mathbb{R}^n$ a vector of covariate in the covariate space $\mathcal{X}$, $A \in \{0,1\}$ be an indicator variable for the treatment of interest and $Y \in \{0,1\}$ be a binary outcome. We introduce potential outcomes Y^0 and Y^1 that represent the binary outcomes that would be observed if patients were assigned to either the control or the evaluated treatment respectively. Without loss of generality, we assume that $Y = 1$ is a desirable event. We make the following assumptions:[9]

- Consistency: the observed outcome corresponds to the potential outcome that is, if a patient received the treatment their observed outcome would be Y^1 and if they received the control, their observed outcome would be Y^0.
- No interference: the outcome only depends on the treatment applied to the patient, and not on the treatment applied to other patients.

- Unconfoundedness: all characteristics associated with both the treatment assignment and the outcome, should have been measured in the study.
- Positivity: all patients have a non-null probability of receiving either treatment.

In the setting of RCTs, Unconfoundedness and Positivity are met by design.

2.2 | Individualized treatment rules

We are interested in constructing individualized treatment rules (ITR) which are decision rules that recommend treatment based on patients' characteristics.

Those rules are modeled as maps $r : \mathcal{X} \rightarrow \{0,1\}$. Accordingly, for a given set of covariates $x \in \mathcal{X}$, $r(x)$ indicates whether or not the treatment should be given to a patient. An optimal rule r^{opt} is obtained when the value $\mathcal{V}(r)$ among all $r \in \mathcal{R}$, with $\mathcal{R}$ being the class of all treatment rules, is maximized[1]:

$$r^{opt} = \arg\max_{r \in \mathcal{R}} \mathcal{V}(r),$$

where $\mathcal{V}(r) = E[Y(r)]$ with $Y(r) = Y^1 r(x) + Y^0[1 - r(x)]$ representing the outcome observed if the rule r was followed.

Constructing an optimal treatment rule can be achieved in two different ways. The first approach involves calculating individual treatment effects (ITE). The ITE τ represents the predicted benefit under one treatment minus the predicted benefit under the other treatment, given a set of patients' characteristics:

$$\tau(x) = E(Y^1 - Y^0 | X = x) = \mu_1(x) - \mu_0(x).$$

An optimal rule is obtained by only giving the evaluated treatment to patients with a positive value of $\tau(x)$ that is, $r^{opt}(x) = \mathbb{1}_{\{\tau > 0\}}(x)$. In this approach, some methods estimate the ITE by estimating the response surfaces whereas others directly estimate the ITE via a contrast function. The second approach consists of directly developing an optimal rule, without estimating the ITEs, by minimizing a loss function of the value of the rule. The methods to develop ITRs considered in this project are described in Section 3.

2.3 | Metrics

Several metrics were used to compare the ITRs developed with different methods described in Section 3. Using several metrics allows us to have a comprehensive view of the performance of the ITRs. Two classes of metrics can be distinguished: metrics whose aim is to estimate the performance of the rules and metrics whose aim is to compare the level of agreement between two rules.

2.3.1 | Performance metrics

First, metrics to assess the quality of a single ITR were used, enabling us to compare the performance of the ITRs.

- The value of a rule: As stated previously, the value $\mathcal{V}(r) = E[Y(r)]$ represents the mean outcome if the ITR was correctly followed. In this project, a desirable binary outcome is considered, thus, ITRs with $\mathcal{V}(r)$ closer to 1 have a better performance.
- The benefit of the rule in terms of assigned treatment among people with a positive and negative score, is assessed with two metrics: B_{pos} and B_{neg}, where B_{pos} represents the average benefit of giving the evaluated treatment among people with a positive score that is, $r(x) = 1$ and B_{neg} represents the average benefit of not giving the evaluated treatment among people with a negative score that is, $r(x) = 0$.[10] The values are between -1 and 1, with 1 meaning there is a benefit in treating people with a positive score for B_{pos} and a benefit in not treating people with a negative score for B_{neg}.

$$B_{pos} = P(Y = 1|A = 1, r(x) = 1) - P(Y = 1|A = 0, r(x) = 1),$$
$$B_{neg} = P(Y = 1|A = 0, r(x) = 0) - P(Y = 1|A = 1, r(x) = 0).$$

- The Population Average Prescription Effect (PAPE): PAPE compares an ITR with a treatment rule that randomly treats the same proportion of patients:[11]

$$\text{PAPE} = E[Y(r) - p_r Y^1 - (1 - p_r)Y^0]$$

where p_r represents the proportion of patients assigned to the evaluated treatment under the ITR r.

The PAPE takes values between −1 and 1. Here, since higher values of the outcome are desirable, higher values of PAPE indicate a better performance of the ITR. A value of 0 indicates that the ITR does not perform better than treating randomly the same proportion of patients. Negative values mean that the ITR performs worse. An advantage of the PAPE is that it is easy to interpret.

- The c-statistic for benefit: it is the probability that from two randomly chosen matched pairs with unequal observed benefit, the pair with greater observed benefit also has a higher predicted probability where the observed benefit refers to the difference in outcomes between two patients with the same predicted benefit but with different treatment assignments.[12] To create the pairs, a patient in the control group is matched to one in the treatment group with a similar predicted treatment benefit. Higher values of the c-statistic for benefit are better. The c-statistic for benefit quantifies how well the rule discriminates patients benefiting from patients not benefiting from taking a given treatment. The c-statistic for benefit can only be calculated for methods using an ITE or a benefit score to derive an ITR.

The standard errors for each metric were calculated through a Bootstrap procedure involving 1000 samples of the original dataset.

2.3.2 | Agreement between two rules

Metrics to see if two ITRs have the same recommendation and agree to allocate the treatment to the same patients were used.

- Matthews correlation coefficient (MCC): Here, the MCC is used to measure the disagreement, in terms of treated patients, between two rules.[13] The values range between −1 and 1, where 1 indicates a perfect positive correlation, 0 indicates no correlation and −1 indicates a perfect negative correlation.
- Cohen's kappa coefficient: Cohen's kappa measures the agreement between two rules by considering the number of agreements and disagreements.[14] It can range from −1 to 1. A value inferior to 0 demonstrates that there is less than chance agreement between the two rules, a value of 0 shows no agreement and a value of 1 means that there is perfect agreement between the rules.

2.4 | Multiple correspondence analysis

A Multiple Correspondence Analysis (MCA) was conducted to see if the ITRs agreed on the treatment decision in the presence of some specific characteristics. All variables included in the different models were put in the MCA, as well as the treatment allocation recommended by each ITR. Continuous variables were categorized and the choice of the categories was motivated by previous works done using the datasets.[15,16]

3 | METHODS TO CONSTRUCT INDIVIDUALIZED TREATMENT RULES

This section presents the different methods that were compared. We selected the most common methods which are either simple to implement with the R software[17] or for which a package is available. As mentioned in Section 2.2, two classes of methods to develop an ITR were distinguished: a first class in which the ITR is constructed by first modeling the ITE,

and an optimal rule is found by giving the treatment evaluated to the individuals with a positive ITE, and a second class, in which the ITR is directly estimated without calculating individualized treatment effects, and where an optimal rule is found by minimizing the risk of the value of the rule via a loss function. In the first class, two distinct approaches can be used to obtain the ITE: either estimating the expected difference of the potential outcomes between treatments or estimating the ITE directly via a contrast function. The majority of the methods fell under the first category: the meta-learners (T-learner, S-learner, X-learner, DR-learner, and R-learner, both with parametric and non-parametric models), PATH, causal forests, virtual twins, A-learning and the modified covariate method, whereas outcome weighted learning and contrast weighted learning fell under the second class. A classification of methods based on how they construct an optimal treatment rule is given in Figure 1. Conceptually, some methods are related and are therefore referred to as belonging to the same family (eg, parametric meta-learners, non-parametric meta-learners, A-learning, and the modified covariate method).

3.1 | Meta-learners

Meta-learners are methods that use sub-regression problems to estimate the ITE via a base learner. In this project, two base learners were implemented: logistic regression and random forest (RF), the latter is selected for its good performance on tabular data.[18] Since meta-learners can use several base learners, they are flexible and can adapt to different types of data.

Meta-learners with a non-parametric model as the base learner can be prone to overfitting. A solution to this potential overfitting is to use cross-fit.[19] Cross-fit consists of splitting the dataset into several folds. Then, the ITEs are learned on every fold and the results are aggregated to derive an ITR (more details in Supplementary Material S1). The meta-learners were compared with and without cross-fit when using random forests as a base learner. When cross-fit was applied, 5 folds and 30 splits were used, because such a choice has been reported as leading to a good performance.[20] However, to our knowledge, there is no standard method or clear guidance on how to perform cross-fit, and other choices exist.[20]

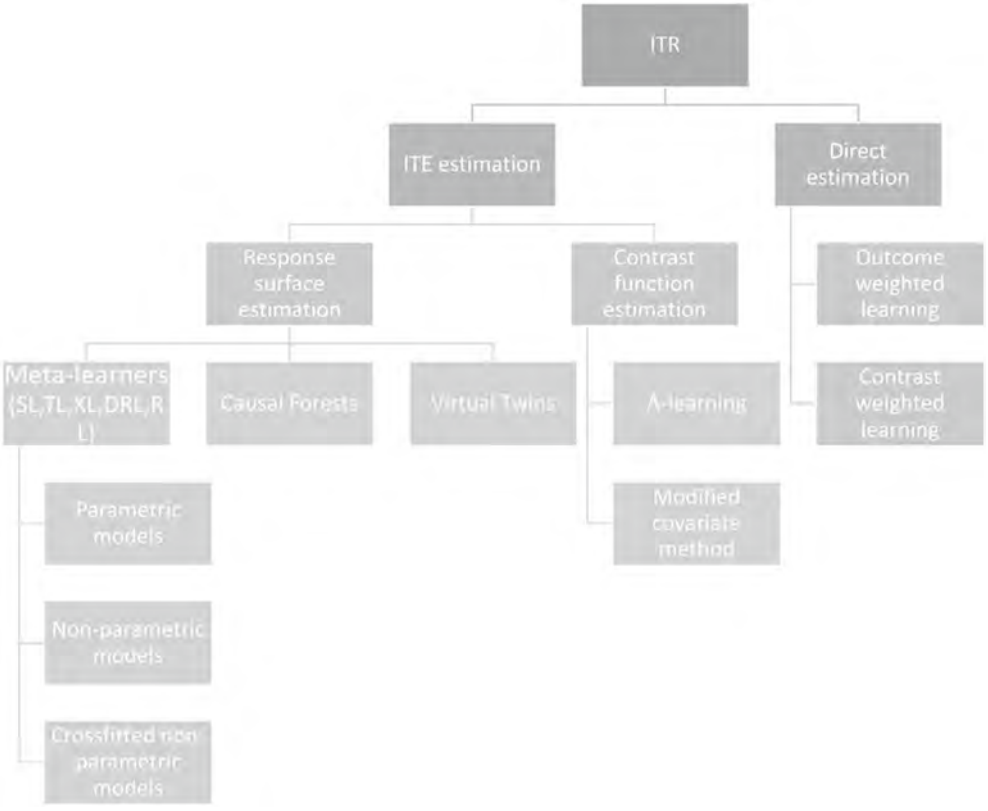

FIGURE 1 Classification of the methods.

In the upcoming segments, we use $\hat{\tau}$ to represent the estimate of τ, and adhere to the same convention for denoting the estimates of other parameters.

3.1.1 | S-learner

The S-learner estimates the treatment effect within a single regression model, where the treatment is included as a feature and where interactions between the treatment and relevant covariates are introduced in the parametric setting.[21] First, use a model to estimate the response function $\mu(x, a)$:

$$\mu(x, a) = \mathbb{E}(Y|X = x, A = a).$$

Then, estimate the individual treatment effect τ:

$$\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0).$$

3.1.2 | T-learner

In the T-learner algorithm, two models are built, one in the treatment group and one in the control group.[21] These models are used to calculate the response functions:

$$\mu_0(x) = \mathbb{E}(Y|X = x, A = 0),$$
$$\mu_1(x) = \mathbb{E}(Y|X = x, A = 1).$$

The ITE is estimated as the difference between the two predicted risks:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x).$$

3.1.3 | X-learner

The X-learner consists of three steps:[21]

1. Estimate the response functions as in the T-learner:

$$\mu_0(x) = \mathbb{E}(Y|X = x, A = 0),$$
$$\mu_1(x) = \mathbb{E}(Y|X = x, A = 1).$$

2. Impute the treatment effects for the individuals in the treated group based on the control-outcome estimator and the treatment effects for the individuals in the control group based on the treatment-outcome estimator and estimate $\hat{\tau}_1(x)$ and $\hat{\tau}_0(x)$:

$$\widetilde{D^1} = Y^1 - \hat{\mu}_0(X^1),$$
$$\widetilde{D^0} = \hat{\mu}_1(X^0) - Y^0,$$
$$\hat{\tau}_1(x) = \mathbb{E}(\widetilde{D^1}|X = x),$$
$$\hat{\tau}_0(x) = \mathbb{E}(\widetilde{D^0}|X = x).$$

3. Define the ITE by a weighted average of the two estimates:

$$\hat{\tau}(x) = w(x)\hat{\tau}_0(x) + (1 - w(x))\hat{\tau}_1(x)$$

where $w(x) \in [0, 1]$ is a weighting function. An estimate of the propensity scores can be chosen as the weighting function, but there is no clear theory on how to choose an optimal weighting function. In the setting of RCTs, it is natural to choose $w(x) = \frac{1}{2}$ if the trial had a 1:1 randomization ratio.

The X-learner has been described as advantageous in an unbalanced design or with sparse treatment effects.[21]

3.1.4 | DR-learner

The DR-learner is a doubly robust estimator that estimates the ITE in two stages.[22] This learner includes double sample splitting to reduce bias.

First, the data $Z_i = (X_i, A_i, Y_i)$ are randomly split into three independent samples D_{1a}, D_{1b}, D_2. Then, the following two steps are applied:

1. Construct propensity score estimates $\hat{\pi}$ of the propensity scores $\pi(X) = \mathbb{P}(A = 1|X = x)$ using D_{1a} and estimate the response functions $\hat{\mu}_0$ and $\hat{\mu}_1$ using D_{1b}.
2. Construct the pseudo-outcome:

$$\hat{\varphi}(Z) = \frac{A - \hat{\pi}(X)}{\hat{\pi}(X)[1 - \hat{\pi}(X)]}[Y - \hat{\mu}_A(X)] + \hat{\mu}_1(X) - \hat{\mu}_0(X).$$

Then, regressing it on covariates X of D_2 to estimate the ITE:

$$\hat{\tau}(x) = \hat{\mathbb{E}}[\hat{\varphi}(Z)|X = x].$$

Cross-fitting can be added as an additional third step:

1. Repeat steps 1 and 2 twice. First, D_{1b} and D_2 are used for step 1 and D_{1a} is used for step 2. Then, D_{1a} and D_2 are used for step 1 and D_{1b} is used for step 2. A final estimate of τ is constructed by averaging the three estimates.

In this work, the propensity score was taken equal to $\frac{1}{2}$ since data from RCTs with a 1:1 randomization ratio were used.

3.1.5 | R-learner

The R-learner estimates the ITEs in two steps:[23]

1. Fit the response function $\hat{\mu}(x)$ and the propensity scores $\hat{\pi}(x)$ with a base learner.
2. Estimate ITEs by minimizing the R-loss, which uses Robinson's decomposition:[24]

$$L_R(\tau(x)) = \frac{1}{n}\sum_{i=1}^{n}[(Y_i - \hat{\mu}(X_i)) - (A_i - \hat{\pi}(X_i))\tau(X_i)]^2 + \Lambda_n(\tau(\cdot))$$

where $\Lambda_n(\tau(\cdot))$ is a regularization term on the complexity of $\tau(\cdot)$.

The response function and the propensity scores can be fitted using a cross-validation procedure and the regularization could be done with a penalized regression such as lasso or ridge, for instance, when logistic regression is used as a base learner.[23] When random forests are used, regularization is achieved by tuning the hyperparameters, particularly by limiting the tree's depth and the number of variables used to build the trees.

3.2 | PATH approach

PATH is a risk modeling approach that has been recommended by the PATH statement.[3,4] This method involves three steps:

1. Fit a regression model (a logistic model with a binary outcome) with the relevant variables to derive the linear predictor
2. Build a model that incorporates the linear predictor, the treatment variable, and the interaction between the linear predictor and treatment to estimate the response functions
3. Derive the ITE based on the response functions

This approach has demonstrated strong performance in diverse scenarios according to previous simulation studies.[25,26]

3.3 | Causal forests

The causal forests algorithm is a special case of generalized random forests (GRF), a flexible and general framework to estimate the ITEs.[27] Causal forests extend the original random forest algorithm by borrowing ideas from kernel-based methods and the R-learner.[23]

In contrast to the standard random forest algorithm in which a prediction for a new observation is obtained by averaging predictions of each tree, here, the trees are used to compute a weighting scheme similar to kernel-based methods. The trees act as weights between training points and any new observations:

$$\alpha_{bi}(x) = \frac{\mathbb{1}_{X_i \in L_b(x)}}{|L_b(x)|}, \quad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B}\alpha_{bi}(x)$$

where X_i corresponds to the covariates of individual i in the training dataset and $L_b(x)$ corresponds to the set of observations in the training set that fall in the same leaf as x for tree b.

Then, the prediction for a new observation is obtained using the adaptive weights by minimizing the R-loss described above.

Another characteristic of causal forests (and more generally of GRF) is the notion of honesty where the training data is split into two parts: one for constructing the tree and the other (the estimation sample) for estimating leaf values for each tree. In doing so, the estimates are less prone to bias and more consistent. The notion of honesty is similar to employing the crossfit in non-parametric meta-learners.

3.4 | Virtual twins

The virtual twins method consists in predicting response probabilities for treatment and control twins for all individuals using counterfactual models.[28] The difference in the probabilities is then used as the outcome in a classification or regression tree. A subgroup of individuals defined by a region S of the covariate space $\mathcal{X}$ for which the treatment effect τ is better than a prespecified threshold can be then identified. The two steps are described below:

1. Fit a random forest in which the covariates, the treatment indicator, and treatment-covariates interactions are included to estimate the response function μ and the ITE τ as in S-learner.
2. Build a regression or a classification tree to find the covariates X that are strongly associated with τ to define region S. Define $\tau*$, a binary variable, as the outcome. When $\tau > c$, $\tau* = 1$ and when $\tau \leq c$, $\tau* = 0$. Develop an ITR based on the value of $\tau*$. Individuals for which $\tau* = 1$ are placed in the estimated region $\hat{S}$. The evaluated treatment is given to individuals in $\hat{S}$

In this work, we built a classification tree and set c equal to 0.
The enhanced treatment effect Q(S) defined as:

$$Q(S) = (P(Y = 1|A = 1, X \in S) - P(Y = 1|A = 0, X \in S)) - (P(Y = 1|A = 1) - P(Y = 1|A = 0))$$

can be estimated by estimating $P(Y = 1|A = 1, X \in \hat{S})$, $P(Y = 1|A = 0, X \in \hat{S})$, $P(Y = 1|A = 1)$ and $P(Y = 1|A = 0)$ using the observed proportions of the data. One of the different approaches that can be used to correct bias is bootstrapping.

Bootstrapping measures the bias of $Q(\hat{S})$, which is then used to adjust $Q(\hat{S})$. In their work, Foster et al compared bootstrapping to other approaches; their conclusion favored bootstrap with 20 samples.[28]

3.5 | A-learning and the modified covariate method

A-learning and the modified covariate method are two methods that focus on treatment-covariates interactions since treatment selection solely depends on the sign of the interactions.[29,30] Given the covariates and the treatment, the estimated outcome can be written as:

$$E(Y|A,X) = m(X) + A\Delta(X)$$

where $m(X), \Delta(X)$ represent respectively the main effect of X and the treatment effect given X. Only the signs of $\Delta(X)$ matter for treatment selection.

In both methods, a personalized benefit score model f is calculated and its sign, which is consistent with the direction of the treatment effect, is used to construct an ITR. An optimal ITR is found for both methods by minimizing a certain loss function ℓ. Details on the loss functions are given below.

3.5.1 | A-learning

In A-learning, the following expected loss function is considered:

$$\ell_A(f) = E(\ell_A(f,x))$$

with

$$\begin{aligned}\ell_A(f,x) = \pi(x)E[M(Y,(1-\pi(x))f(x))|A=1,X=x] \\ + (1-\pi(x))E[M(Y,-\pi(x)f(x))|A=0,X=x].\end{aligned}$$

where $\pi(x)$ represents the propensity scores and M is a positive function, such as the quadratic or cross-entropy (also called logistic loss).

$\ell_A(f,x)$ is then replaced by its empirical version on the observed data:

$$L_A(f) = \frac{1}{n}\sum_{i=1}^{n} M(Y_i,(A_i-\pi(X_i))f(X_i))$$

$\pi(X_i)$ equals $\frac{1}{2}$ in the context of RCTs with 1:1 randomization.

When M is chosen to be the logistic loss, $L_A(f)$ is expressed as:

$$L_A(f) = -\frac{1}{n}\sum_i Y_i(A_i-\pi(X_i))f(X_i) - \log(1+\exp((A_i-\pi(X_i))f(X_i))).$$

3.5.2 | Modified covariate method

Similarly, the expected loss function $\ell_{MCM}(f) = E(\ell_{MCM}(f,x))$ of the modified covariate method where

$$\begin{aligned}\ell_{MCM}(f,x) = E[M(Y,f(X))|A=1,X=x] \\ + E[M(Y,-f(X))|A=0,X=x]\end{aligned}$$

Its empirical version is

$$L_{MCM}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{M(Y_i,(2A_i-1)f(X_i))}{(2A_i-1)\pi(X_i)+1-A_i},$$

which boils down to

$$L_{MCM}(f) = -\frac{2}{n}\sum_{i} Y_i(2A_i-1)f(X_i) - \log(1+\exp((2A_i-1)f(X_i))$$

when the logistic loss function is used and $\pi(X_i) = \frac{1}{2}$. It is worth mentioning that by substituting the benefit score of A-learning in L_A with double the benefit score of the modified covariate method, we obtain $L_{MCM} = 2L_A$.

3.6 | Outcome weighted learning

Outcome weighted learning (OWL) uses a weighted classification framework, in which each patient is weighted based on their outcome, with a hinge loss to estimate an ITR.[31] An optimal treatment rule is obtained by minimizing the following quantity:

$$L(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{(2A_i-1)\pi(X_i)+1-A_i}(1-(2A_i-1)f(X_i))^{+} + \lambda_n\|f\|^2$$

where $x^{+} = \max(x,0)$, $\pi(X_i)$ represents the propensity scores, λ_n is a penalty parameter used to avoid overfitting and $\|f\|$ is a norm for the function $f : \mathcal{X} \to \{0,1\}$. When employing the linear kernel, the Euclidean norm of the coefficients, excluding the intercept, is utilized. The parameter λ_n is chosen by performing a cross-validation. OWL is a consistent estimator and has low variability.[31]

3.7 | Contrast weighted learning

The idea behind contrast weighted learning (CWL) is to use contrasts of the outcome between pairs of patients to build weights used in a weighted classification algorithm to estimate an ITR.[32] A contrast function h is defined for a pair of patients to measure the relative favorability of their outcomes. Several contrast functions exist such as the difference $h(Y_i,Y_j) = Y_i - Y_j$, the log ratio $h(Y_i,Y_j) = \log(Y_i/Y_j)$ or the win indicator $h(Y_i,Y_j) = \text{sgn}(Y_i - Y_j)$, with $\text{sgn}(Y_i - Y_j) = 1$ if $Y_i - Y_j > 0$; $\text{sgn}(Y_i - Y_j) = 0$ if $Y_i - Y_j = 0$ and $\text{sgn}(Y_i - Y_j) = -1$ if $Y_i - Y_j < 0$. In this project, we used the win indicator, considered the most robust contrast function by Guo et al.[32] The optimal ITR is found by minimizing the following function:

$$L(f) = \frac{1}{2}E\left[(\mathbb{1}_{h(Y_i,Y_j)(2A_i-1)f(X_i)<0} + \mathbb{1}_{h(Y_i,Y_j)(2A_j-1)f(X_j)\geq 0})\frac{|h(Y_i,Y_j)|}{\pi(X_i)\pi(X_j)}\right]$$

where $h(Y_i,Y_j)$ is the contrast function between patient i and patient j and $\pi(X_j)$ represents the propensity score. Here $h(Y_i,Y_j) = \text{sgn}(Y_i - Y_j)$ and $\pi(X_i) = \pi(X_j) = \frac{1}{2}$. CWL is a flexible and robust method that only relies on the contrast of outcomes between two patients. However, a correctly specified model is needed to ensure consistency.

3.8 | Implementation

All the analyses were performed in R version 4.1.2. Virtual Twins was implemented using the `aVirtualTwins` package.[33] The package `personalized` was used to develop the modified covariate method and A-learning.[34] For outcome weighted learning and contrast weighted learning, the package `WeightSVM` was used.[35] Causal forests was implemented using the package `grf`.[36] More details about the implementation such as the choice of the hyper-parameters can be found in Supplementary Material S2.

4 | COMPARISON OF THE METHODS ON REAL DATA

In this section, the ITRs obtained when applying the methods described above are compared on two multi-center randomized control trials: the International Stroke Trial and the CRASH-3 trial. The train and test datasets were obtained by splitting the data at the center level using $\frac{2}{3}$ of the data for training.

4.1 | International stroke trial

The 22 methods were first compared on the International Stroke Trial (IST).[15] The IST was chosen because an ITR has been developed on this dataset in the past using the T-learner method and found that 74% of patients would benefit from taking aspirin.[37] The IST is a multi-center randomized control trial that includes 19,435 patients recruited from 466 centers and examines the impact of administrating aspirin, heparin, or both in stroke. For our illustration, we focused on the impact of aspirin on stroke. Nineteen variables were included in the different methods: 16 categorical variables and three continuous variables, similar to what Nguyen et al did.[37] The outcome used was death or dependency at 6 months (1 = no and 0 = yes). The treatment variable was binary (0 = no aspirin and 1 = aspirin). A description of the covariates and the outcome is reported in Supporting Material S3.

Results of the metrics used to evaluate the ITR produced by each method are given in Table 1. The performance in the train dataset of IST can be found in Supplementary Material S4. Higher values of c-statistic for benefit are better but it is rare to obtain values above 0.6.[12] Here, the c-statistic for benefit was close to 0.5 for all the ITRs, indicating poor discrimination. A reason for the poor discrimination could be the lack of strong heterogeneous treatment effects. This hypothesis was confirmed by conducting a likelihood ratio test comparing models with and without treatment-covariate interactions were compared. The test showed that the interactions did not add value meaning no significative heterogeneity was found in the IST dataset. The adequacy index was also computed to see how much predictive information was due to the treatment-covariate interactions. For IST, the adequacy index was equal to 0.993 meaning adding the interactions only accounted for 0.7% of the predictive information. The ITRs had a PAPE close to 0 meaning that the ITRs did not perform better than a rule that treated randomly the same proportion of patients. The PAPE values of most meta-learners were even slightly negative indicating that a non-individualized rule performed slightly better than those individualized rules. The values of B_{pos} and B_{neg} were close to 0, meaning there were not many benefits of giving the evaluated treatment to patients with a positive score or not giving the evaluated treatment to patients with a negative score. The proportion of patients for whom aspirin was recommended by the different ITRs ranged from 0.114 to 0.899, with most methods producing an ITR that recommended the evaluated treatment for more than 50% of patients. Methods belonging to the same family had similar proportions. Despite the significant disparity of proportions, the estimated values of the ITRS were similar showing that giving the evaluated treatment to more or fewer patients did not improve the value. For instance, OWL's ITR recommended treating 0.898 of patients, and CWL's ITR recommended treating 0.114 of patients but their rule's values were 0.399 and 0.400 respectively. The mean outcome when no one was treated (0.380) was close to the mean outcome when everyone was treated (0.396), which further implies that the treatment had a limited impact on the outcome on average. The mean outcome under the individualized rule was above the mean outcome when no one was treated for all methods. However, only five methods (Causal forests, A-learning, modified covariate method, OWL, and CWL) had mean outcome under the rule above the mean outcome when everyone was treated, and even for those methods, the mean outcome did not notably surpass the mean outcome if everyone is treated. Generally, the ITRs developed by the different methods did not drastically improve the mean outcome.

Overall, MCC and kappa's coefficient produced similar values (Figure 2). Most methods had considerable disagreements and thus almost no correlation regarding the people treated with the evaluated treatment in their rules which can indicate that the rules did not consider the same characteristics for the allocation of the treatment. A better concordance was found among methods of the same family. For instance, the ITRs developed with the parametric meta-learners agreed to treat similar patients and had MCC and Cohen's kappa values ranging from 0.77 to 1. Similarly, non-parametric meta-learners had a positive moderate to high correlation with each other and with their crossfitted counterparts. However, they had less correlation with the parametric ones. A-learning and the modified covariate method generated the same ITR and therefore had coefficients of 1. The ITRs obtained with the different methods generally did not recommend the evaluated treatment to the same patients, which draws some concerns for their usability in practice.

A majority of characteristics were located near the origin and were not associated with the treatment allocation of the different ITRs (Figure 3). Virtual twins' ITR recommended not treating patients in a drowsy state and patients with

TABLE 1 Results of the metrics for each method applied to the IST dataset.

	p_r	$\mathcal{V}(r)$ (SE)	$E(Y^0)$ (SE)	$E(Y^1)$ (SE)	B_{pos} (SE)	B_{neg} (SE)	PAPE (SE)	c for benefit (95% CI)
SL	0.567	0.388 (0.009)	0.380 (0.009)	0.396 (0.009)	0.014 (0.017)	−0.019 (0.018)	−0.002 (0.006)	0.495 (0.476; 0.514)
TL	0.567	0.388 (0.009)	0.380 (0.009)	0.396 (0.009)	0.014 (0.017)	−0.019 (0.019)	−0.002 (0.006)	0.495 (0.476; 0.514)
XL	0.622	0.382 (0.009)	0.380 (0.009)	0.396 (0.009)	0.004 (0.016)	−0.034 (0.020)	−0.008 (0.006)	0.495 (0.477; 0.513)
DRL	0.622	0.382 (0.008)	0.380 (0.009)	0.396 (0.009)	0.003 (0.016)	−0.035 (0.020)	−0.008 (0.006)	0.495 (0.477; 0.514)
RL	0.626	0.388 (0.009)	0.380 (0.009)	0.396 (0.009)	0.013 (0.016)	−0.022 (0.020)	−0.003 (0.006)	0.501 (0.483; 0.519)
SL RF	0.526	0.395 (0.009)	0.380 (0.009)	0.396 (0.009)	0.030 (0.017)	−0.004 (0.017)	0.006 (0.006)	0.498 (0.480; 0.516)
TL RF	0.515	0.390 (0.009)	0.380 (0.009)	0.396 (0.009)	0.021 (0.018)	−0.013 (0.018)	0.002 (0.006)	0.498 (0.480; 0.517)
XL RF	0.543	0.393 (0.009)	0.380 (0.009)	0.396 (0.009)	0.025 (0.017)	−0.008 (0.017)	0.004 (0.006)	0.504 (0.486; 0.523)
DRL RF	0.529	0.390 (0.009)	0.380 (0.009)	0.396 (0.009)	0.020 (0.017)	−0.015 (0.018)	0.001 (0.006)	0.500 (0.482; 0.518)
RL RF	0.528	0.387 (0.009)	0.380 (0.009)	0.396 (0.009)	0.018 (0.018)	−0.019 (0.018)	−0.001 (0.006)	0.501 (0.482; 0.519)
SL CF	0.598	0.389 (0.009)	0.380 (0.009)	0.396 (0.009)	0.018 (0.017)	−0.019 (0.019)	0.000 (0.006)	0.497 (0.478; 0.515)
TL CF	0.555	0.385 (0.009)	0.380 (0.009)	0.396 (0.009)	0.009 (0.017)	−0.025 (0.019)	−0.004 (0.006)	0.498 (0.480; 0.516)
XL CF	0.599	0.386 (0.009)	0.380 (0.009)	0.396 (0.009)	0.011 (0.017)	−0.024 (0.018)	−0.004 (0.006)	0.498 (0.480; 0.516)
DRL CF	0.552	0.386 (0.009)	0.380 (0.009)	0.396 (0.009)	0.011 (0.016)	−0.023 (0.019)	−0.003 (0.006)	0.499 (0.481; 0.517)
RL CF	0.548	0.389 (0.009)	0.380 (0.009)	0.396 (0.009)	0.018 (0.018)	−0.016 (0.017)	0.000 (0.006)	0.501 (0.483; 0.519)
PATH	0.723	0.394 (0.009)	0.380 (0.009)	0.396 (0.009)	0.020 (0.015)	−0.003 (0.014)	0.002 (0.005)	0.511 (0.493; 0.529)
Causal Forests	0.899	0.397 (0.008)	0.380 (0.009)	0.396 (0.009)	0.019 (0.013)	0.010 (0.032)	0.002 (0.003)	0.504 (0.486; 0.522)
VT	0.758	0.393 (0.009)	0.380 (0.009)	0.396 (0.009)	0.017 (0.015)	−0.009 (0.019)	0.000 (0.005)	—
AL	0.501	0.419 (0.008)	0.380 (0.009)	0.396 (0.009)	0.089 (0.018)	0.059 (0.018)	0.031 (0.005)	0.566 (0.548; 0.583)
MCM	0.501	0.419 (0.008)	0.380 (0.009)	0.396 (0.009)	0.089 (0.017)	0.059 (0.017)	0.031 (0.005)	0.566 (0.548; 0.583)
OWL	0.898	0.399 (0.009)	0.380 (0.009)	0.396 (0.009)	0.022 (0.013)	0.000 (0.017)	0.004 (0.003)	—
CWL	0.114	0.400 (0.009)	0.380 (0.009)	0.396 (0.009)	0.075 (0.034)	0.007 (0.012)	0.018 (0.004)	—

Note: p_r refers to the proportion of patients for which the rule recommends treatment. $E(Y^0)$ and $E(Y^1)$ refer respectively to the mean outcome when no one is treated and the mean outcome when everyone is treated.

Abbreviations: AL, A-learning; CF, cross-fitted; CWL, contrast weighted learning; DRL, DR-learner; MCM, modified covariate method; OWL, outcome weighted learning; RF, random forests; RL, R-learner; SL, S-learner; TL, T-learner; VT, virtual twins; XL, X-learner.

a TACS stroke and recommended treating fully alert patients, younger patients, and patients with no deficit or disorder. Non-parametric meta-learners produced ITRs that recommended treatment for patients from South America and no treatment for patients from the Middle East, South Asia, and Oceania. CWL's ITR recommended treating patients with a LACS stroke or another type of stroke while OWL's ITR recommended not giving aspirin to unconscious patients. The MCA was concordant with what was found in Figure 2 and reflected well the disagreement in terms of treatment allocation between the ITRs.

4.2 | CRASH-3

In a second stage, we compared the methods on the CRASH-3 dataset.[16] Some heterogeneity in early treatment administration has been found in the CRASH-3 trial, therefore we thought it would be interesting to develop ITRs on this data.[16] CRASH-3 is a multi-center randomized control trial consisting of 9,072 patients from 175 hospitals over 29 countries. The aim of this trial was to examine the effects of tranexamic acid (TXA) in patients with acute traumatic brain injury. This paper used head injury death as the outcome (1 = no and 0 = yes). Six covariates were included in the methods: two categorical variables and four continuous variables. A binary treatment variable (0 = Placebo and 1 = TXA) was used. A description of the covariates and outcome is given in Supporting Material S3.

FIGURE 2 Heatmap representing the MCC and Cohen's Kappa for each combination of two ITRs using the international stroke trial.

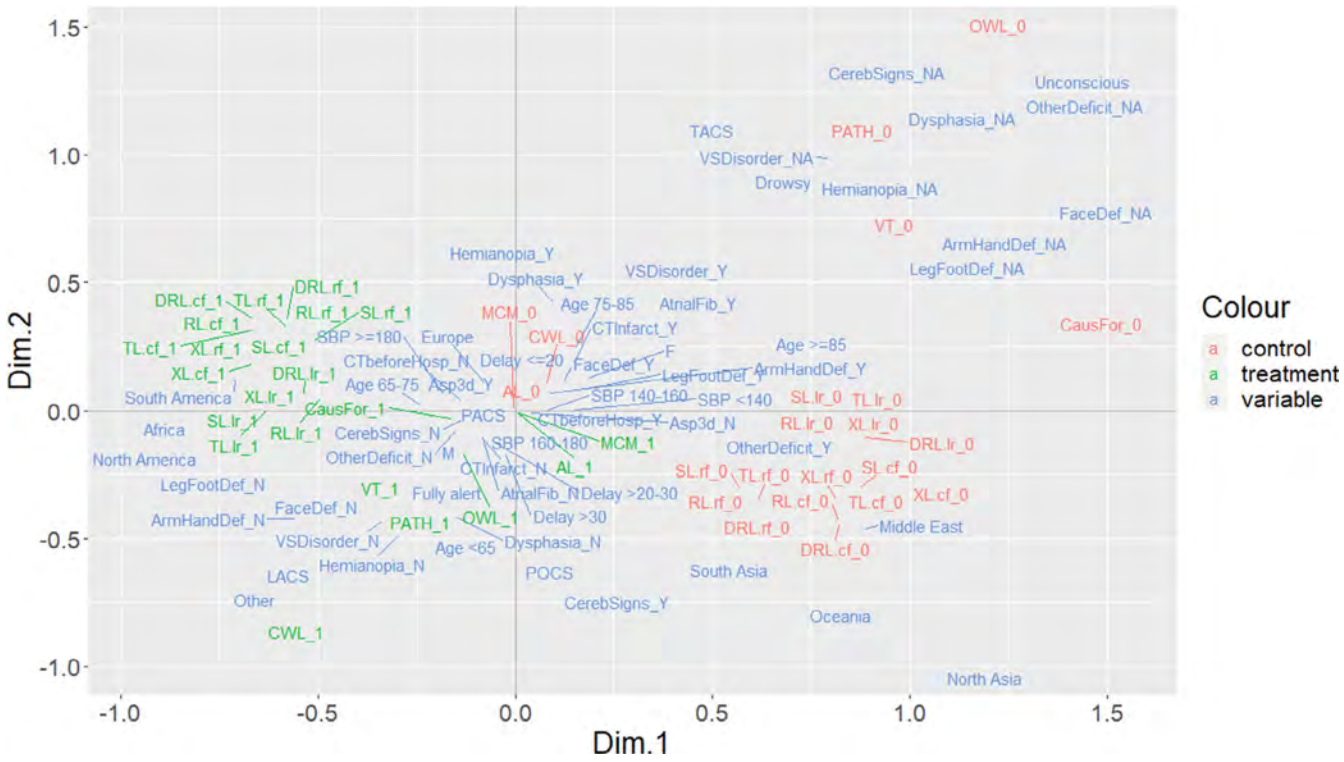

FIGURE 3 Multiple correspondence analysis on the international stroke trial showing all levels of each variable and the treatment recommendation of the individualized treatment rules.

Table 2 shows the values of the metrics obtained with each method. The results of the training set are given in Supplementary Material S3. Recall that higher values of c-statistic for benefit are better but it is rare to obtain values above 0.6.[12] Here, c-statistic for benefit values were all around or under 0.5, indicating poor discrimination except for A-learning and the modified covariate method's ITRs which had higher values (0.721 and 0.720 respectively). Excluding A-learning and the modified covariate method's ITRs, the ITRs were not able to differentiate patients benefiting from taking the evaluated treatment from patients not benefiting. The PAPE values were close to 0 meaning that the ITRs did not perform better than a rule which randomly treated the same proportion of patients. There was a mix of positive and negative values but they all remained close to 0. A-learning and the modified covariate method's ITRs had B_{pos} and B_{neg} values around 0.2, showing some benefits of giving the evaluated treatment to patients with a positive score and not giving the evaluated treatment to patients with a negative score, which was not the case for the other ITRs who had values near 0. A-learning and the modified covariate method outperformed other approaches. This better performance is attributed to the treatment rules they developed, predominantly recommending treatment for patients with moderate Glasgow coma scores and reactive pupils. These patients' profiles align with findings from the CRASH-3 study. The proportions of people for which the treatment was recommended went from 0 to 1 with a majority of methods recommending to give the evaluated treatment to over 60% of patients. Note that CWL's ITR chose to give the evaluated treatment to no one whereas OWL's ITR chose to give the evaluated treatment to everyone. The rules' mean outcomes were almost identical and practically all above 0.8, although the proportion of treated patients differed for

TABLE 2 Results of the metrics for each method applied to the CRASH-3 dataset.

	p_r	$\mathcal{V}(r)$ (SE)	$E(Y^0)$ (SE)	$E(Y^1)$ (SE)	B_{pos} (SE)	B_{neg} (SE)	PAPE (SE)	c for benefit (95% CI)
SL	0.798	0.818 (0.009)	0.802 (0.010)	0.819 (0.010)	0.019 (0.013)	0.007 (0.040)	0.002 (0.007)	0.485 (0.448; 0.522)
TL	0.798	0.818 (0.009)	0.802 (0.010)	0.819 (0.010)	0.019 (0.013)	0.007 (0.037)	0.002 (0.007)	0.485 (0.448; 0.522)
XL	0.711	0.812 (0.010)	0.802 (0.010)	0.819 (0.010)	0.013 (0.014)	−0.021 (0.031)	−0.002 (0.007)	0.480 (0.447; 0.514)
DRL	0.711	0.812 (0.010)	0.802 (0.010)	0.819 (0.010)	0.013 (0.015)	−0.021 (0.030)	−0.002 (0.007)	0.481 (0.448; 0.513)
RL	0.711	0.812 (0.010)	0.802 (0.010)	0.819 (0.010)	0.013 (0.014)	−0.021 (0.030)	−0.002 (0.007)	0.481 (0.448; 0.513)
SL RF	0.496	0.804 (0.010)	0.802 (0.010)	0.819 (0.010)	0.003 (0.019)	−0.030 (0.020)	−0.006 (0.007)	0.489 (0.454; 0.524)
TL RF	0.514	0.805 (0.010)	0.802 (0.010)	0.819 (0.010)	0.007 (0.018)	−0.022 (0.021)	−0.002 (0.007)	0.493 (0.458; 0.528)
XL RF	0.579	0.809 (0.010)	0.802 (0.010)	0.819 (0.010)	0.012 (0.017)	−0.022 (0.021)	−0.002 (0.007)	0.490 (0.457; 0.524)
DRL RF	0.553	0.813 (0.009)	0.802 (0.010)	0.819 (0.010)	0.019 (0.016)	−0.015 (0.022)	0.001 (0.007)	0.518 (0.484; 0.552)
RL RF	0.531	0.812 (0.010)	0.802 (0.010)	0.819 (0.010)	0.018 (0.018)	−0.016 (0.021)	0.001 (0.007)	0.512 (0.479; 0.545)
SL CF	0.641	0.798 (0.010)	0.802 (0.010)	0.819 (0.010)	−0.008 (0.015)	−0.056 (0.027)	0.001 (0.007)	0.469 (0.435; 0.504)
TL CF	0.605	0.803 (0.010)	0.802 (0.010)	0.819 (0.010)	0.000 (0.015)	−0.041 (0.025)	−0.009 (0.007)	0.476 (0.443; 0.510)
XL CF	0.666	0.804 (0.010)	0.802 (0.010)	0.819 (0.010)	0.003 (0.014)	−0.051 (0.028)	−0.010 (0.008)	0.473 (0.440; 0.507)
DRL CF	0.593	0.803 (0.010)	0.802 (0.010)	0.819 (0.010)	0.001 (0.017)	−0.039 (0.024)	−0.010 (0.007)	0.485 (0.451; 0.518)
RL CF	0.592	0.802 (0.010)	0.802 (0.010)	0.819 (0.010)	−0.002 (0.017)	−0.042 (0.023)	−0.011 (0.007)	0.486 (0.454; 0.519)
PATH	0.979	0.820 (0.009)	0.802 (0.010)	0.819 (0.010)	0.019 (0.013)	−0.008 (0.023)	0.001 (0.004)	0.494 (0.462; 0.526)
Causal Forests	0.905	0.817 (0.010)	0.802 (0.010)	0.819 (0.010)	0.016 (0.013)	0.003 (0.058)	−0.001 (0.006)	0.471 (0.437; 0.506)
VT	0.897	0.819 (0.010)	0.802 (0.010)	0.819 (0.010)	0.019 (0.013)	0.018 (0.050)	0.002 (0.007)	—
AL	0.503	0.859 (0.007)	0.802 (0.010)	0.819 (0.010)	0.236 (0.027)	0.212 (0.027)	0.049 (0.004)	0.721 (0.699; 0.743)
MCM	0.503	0.859 (0.007)	0.802 (0.010)	0.819 (0.010)	0.236 (0.027)	0.212 (0.028)	0.049 (0.004)	0.720 (0.698; 0.743)
OWL	1	0.819 (0.010)	0.802 (0.010)	0.819 (0.010)	0.017 (0.014)	—	0.000 (0)	—
CWL	0	0.802 (0.010)	0.802 (0.010)	0.819 (0.010)	—	−0.017 (0.013)	0.000 (0)	—

Note: p_r refers to the proportion of patients for which the rule recommends treatment. $E(Y^0)$ and $E(Y^1)$ refer respectively to the mean outcome when no one is treated and the mean outcome when everyone is treated.

Abbreviations: AL, A-learning; CF, cross-fitted; CWL, contrast weighted learning; DRL, DR-learner; MCM, modified covariate method; OWL, outcome weighted learning; RF, random forests; RL, R-learner; SL, S-learner; TL, T-learner; VT, virtual twins; XL, X-learner.

each method leading us to conclude that there was a negligible treatment effect. This can be emphasized by looking at the mean outcome when no one is treated and the mean outcome when all the individuals are treated. The mean outcome when no one was treated (0.802) was close to the mean outcome when everyone was treated (0.819). Comparing these two mean outcomes to the mean outcome under the rules, we found that all the methods, except the crossfitted S-learner, had a mean outcome higher than the mean outcome when no one was treated, but only three of them (PATH, A-learning, and the modified covariate method) had a better mean outcome than the mean outcome when everyone was treated.

MCC and Cohen's Kappa coefficient were concordant and gave coefficients of similar magnitude (Figure 4). When one of the ITRs recommended treating everyone or no one with the evaluated treatment, it did not make sense to calculate the MCC and Cohen's Kappa coefficient, therefore we put a dashed line in those cases. Parametric meta-learners had a strong concordance with each other with high coefficients. The same thing was observed for non-parametric meta-learners whether crossfit was applied or not, as well as for A-learning and the Modified covariate method. As for the IST, a strong concordance is only found between methods belonging to the same family (eg, parametric meta-learners, non-parametric meta-learners, A-learner and the modified covariate method). Otherwise, the correlation between the ITRs was moderate and most of the time low. The ITRs did not recommend the evaluated treatment to the same patients. The choice of the method had a big impact on the treatment allocation, meaning that in practice two different methods could lead to completely different rules.

The ITRs developed with the parametric meta-learners, virtual twins and causal forests recommended not to treat patients with a low Glasgow Coma Scale score or/and patients with none or only one pupil that reacted, whereas they recommended the treatment to patients with a moderate to high Glasgow Coma Scale score, patients who were female and patients with moderate systolic blood pressure (Figure 5). The non-parametric meta-learners' ITRs recommended treating patients younger patients with relatively high blood pressure and not treating patients with low blood pressure. The MCA reflected well the agreement results that were found in Figure 4. Akin methods' ITRs agreed on the treatment allocation but overall the ITRs did not take into account the same characteristics for the treatment decision.

FIGURE 4 Heatmap representing the MCC and Cohen's Kappa for each combination of two ITRs using the CRASH-3 trial.

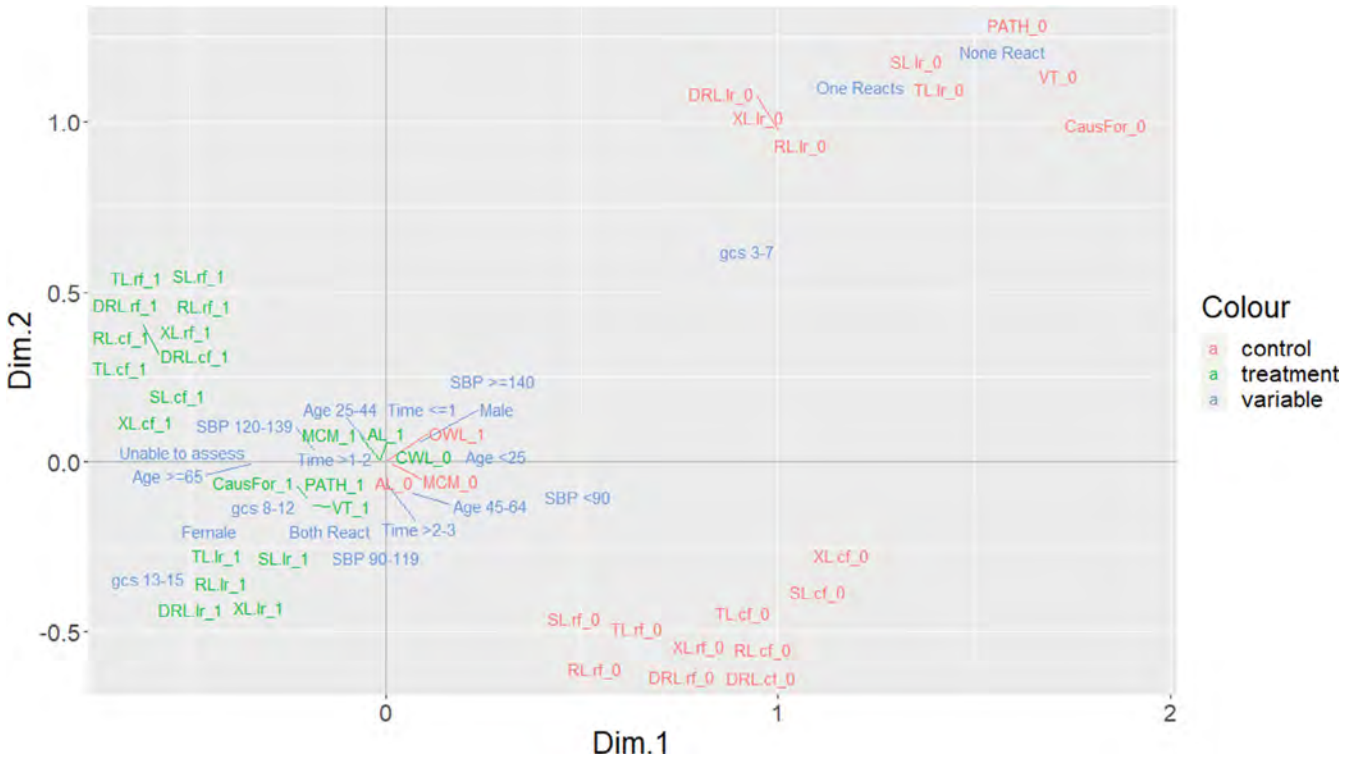

FIGURE 5 Multiple correspondence analysis of the CRASH-3 trial representing the variables' levels and the treatment recommendation of every individualized treatment rule.

5 | DISCUSSION

This paper compared different methods used to construct individualized treatment rules using data from two RCTs: the International Stroke Trial and the CRASH-3 trial. We considered 22 methods belonging to two different classes. The first class included methods that predicted the ITE to derive an ITR: meta-learners (T-learner, S-learner, X-learner, DR-learner and R-learner, both with logistic regression or random forests as a base learner with and without cross-fit), PATH, causal forests, virtual twins, A-learning and modified covariate method. The second class covered methods that directly estimated the ITR without explicitly estimating ITEs: outcome-weighted learning and contrast weighted learning. For each trial, the performance of the ITRs was assessed with various metrics. The pairwise agreement between ITRs was also evaluated.

Results showed that the ITRs obtained by the different methods generally had considerable disagreements regarding the individuals to be treated with the evaluated treatment for both trials. The proportions of patients for whom the evaluated treatment was recommended by the rules were very different depending on which method was employed to build the ITR and the Cohen's kappa and Matthews correlation coefficients were low. A better concordance was found among methods of the same family (eg, among all meta-learners with parametric models, or all meta-learners with non-parametric models and cross-fitting). Overall, when evaluating the performance of ITRs in a hold-out validation sample (33% of the original sample selected at random), results showed that all ITRs had limited performance, whatever the performance in the training set, which suggests a high potential of optimism for the algorithms.

The limited performance results might be due to the distribution of treatment effects and the level of heterogeneity. Although some heterogeneity of treatment effects was found in the trials used in this work, especially in the CRASH-3 trial,[16] the level of heterogeneity might not be sufficient to develop a beneficial individualized treatment rule. This result was reinforced by performing likelihood ratio tests and calculating adequacy indexes. For both trials, the likelihood ratio tests led to the conclusion that there was no evidence of significant heterogeneity and the values of the adequacy indexes showed that the treatment-covariate interactions only accounted for a low percentage of the predictive information. Another explanation for the limited performance might be the sample size. Even if the methods were compared on two large RCTs, perhaps more data is needed to obtain a better performance. A solution might be using individual participant

data meta-analyses (IPD-MA) since they include a larger number of participants. However, one should consider the heterogeneity that may arise between the studies included in the meta-analysis. Different methods to tackle the heterogeneity in IPD-MA have been proposed and compared in previous works.[38-40]

In a previous work, Rekkas et al[25] demonstrated via a simulation study that "complex" methods, which are more flexible, require large sample sizes to perform well and that, when one has access to moderate sample sizes, simpler risk modeling methods recommended by the PATH statement[3] should be preferred to obtain a good performance. Using more parsimonious models with fewer covariates, like what has been done for the SYNTAX II score, might also lead to more robust ITRs with better agreements.[41] Investigating for which distribution of treatment effects, a model can have good discrimination, and thus be able to develop a beneficial ITR, as well as the requirements in effective sample size to allow reliable development of ITRs, is worth studying.

Some comparisons of methods used to construct ITRs have been conducted in the past,[6,7] but to our knowledge, no study has investigated the agreements in terms of the treatment decision with all the methods presented in this project. Both Jacob and Zhang et al have found that the methods had different performances.[6,7] These results were concordant with ours.

Although we compared many methods in this work, we did not include every existing method. Indeed, we decided to focus on methods that are commonly used and that are easily computed or for which an R package was available. We also focused on real data, and a simulation study should be conducted to better delineate the parameters associated with a better performance of the methods. A recent simulation study showed that the sample size and the shape of the distribution of treatments impacted the performance of the methods, particularly the performance of "complex" methods.[25] However, we considered that the illustration on two large RCTs was necessary to study the agreement between the different ITRs in real settings, simulations being often over-simplified. Using real data also allows for tailoring each method, in the sense that each model does not necessarily need to have the same variables. Furthermore, in this paper, we decided to compare the ITRs' decisions on randomized controlled trials. Constructing ITRs can also be done with observational data. Observational databases have the potential to include much more participants, and more diverse participants, than trials, and thus might have both more heterogeneity and larger sample sizes, and could be a better source of data to develop ITRs in practice. Although an effort was dedicated to method optimization, specifically optimizing hyperparameters for tree-based methods through cross-validation to maximize accuracy, it is plausible that further optimization might have led to improved method performance.

In conclusion, the significant disagreements that the methods had regarding the treatment allocation suggest that the methods are not interchangeable. Therefore, the chosen method greatly influences the patients for which the evaluated treatment is recommended. It draws some concerns about their practical use. Some ITRs have been developed in the past using only one method with similar RCTs.[37] Using multiple methods and comparing the obtained ITRs, as suggested by Jacob, might be a solution when one wants to develop an ITR in practice.[6] However, in most cases, more simple approaches such as the risk modeling method advocated in PATH,[3] or carefully adding specific interactions between prespecified treatment-effect modifiers and treatment in the model, as done in the revised SYNTAX score II[41] may be a better strategy than currently available ITR algorithms which may be misleading by overfitting the heterogeneity of treatment effects. Also, methods that allow evaluating the model calibration for benefit may be favored. Evaluating a priori the probability of identifying a beneficial ITR, as suggested by Cain et al,[42] might also be taken under consideration.

AUTHOR CONTRIBUTIONS

Study concept and design: FB, CS, FP and RP. Analysis and interpretation of data: All authors. Drafting of the manuscript: FB, FP, and RP. Critical revision of the manuscript for important intellectual content: All authors.

ACKNOWLEDGEMENTS

The authors wish to thank the investigators of the International Stroke Trial and the CRASH-3 trial for providing the datasets compared in this work. The authors would also like to thank Viet-Thi Tran and François Grolleau for contributing to the comparison using the CRASH-3 dataset. Finally, We thank the two anonymous referees for their comments that greatly improved the paper.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

FUNDING INFORMATION

Florie Bouvier and Raphaël Porcher acknowledge support by the French Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). This work was partially funded by the Agence Nationale de la Recherche, under grant agreement no. ANR-18-CE36-0010-01.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in IST dataset at https://datashare.ed.ac.uk/handle/10283/124 and in CRASH-3 at https://freebird.lshtm.ac.uk/index.php/available-trials/.

ORCID

Florie Bouvier https://orcid.org/0000-0001-6364-6106
Etienne Peyrot https://orcid.org/0009-0006-8520-6201
Alan Balendran https://orcid.org/0000-0002-2779-458X
Corentin Ségalas https://orcid.org/0000-0002-6902-003X
Ian Roberts https://orcid.org/0000-0003-1596-6054
François Petit https://orcid.org/0000-0003-2258-170X
Raphaël Porcher https://orcid.org/0000-0002-5277-4679

REFERENCES

1. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. *Biometrics*. 2012;68:1010-1018.
2. Farooq V, Klaveren D, Steyerberg EW, et al. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. *Lancet*. 2013;381:639-650.
3. Kent David M, Ewout S, David K. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *Bmj*. 2018;363:k4245.
4. Kent DM, Paulus JK, Klaveren D, et al. The predictive approaches to treatment effect heterogeneity (PATH) statement. *Ann Intern Med*. 2020;172:35-45.
5. Alexandros R, Paulus JK, Gowri R, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Med Res Methodol*. 2020;20:264.
6. Daniel J. *CATE Meets ML – the Conditional Average Treatment Effect and Machine Learning*. Springer - Journal: Digital Finance; 2021.
7. Weijia Z, Jiuyong L, Lin L. A unified survey of treatment effect heterogeneity modelling and uplift modelling. *ACM Comput Surv*. 2021;54:1-36.
8. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;6:688-701.
9. Els G, Saskia C, Bianca DS, Moodie Erica EM, Ingeborg W. STRATOS initiative. Formulating causal questions and principled statistical answers. *Stat Med*. 2020;39:4922-4948.
10. Holly J, Brown Marshall D, Margaret P, Ying H. *Statistical Methods for Evaluating and Comparing Biomarkers for Patient Treatment Selection*. UW Biostatistics Working Paper Series. Berlin / Boston: De Gruyter; 2013.
11. Kosuke I, Lingzhi LM. Experimental evaluation of individualized treatment rules. *J Am Stat Assoc*. 2021;118:242-256.
12. Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. *J Clin Epidemiol*. 2018;94:59-68.
13. Pierre B, Søren B, Yves C, Andersen Claus AF, Henrik N. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000;16:412-424.
14. Jacob C. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
15. the International Stroke Trial Collaborative Group,Sandercock Peter AG, Maciej N, Anna C. The international stroke trial database. *Trials*. 2011;12:101.
16. The CRASH-3 trial collaborators. Effects of tranexamic acid on death, disability, vascular occlusive events and other morbidities in patients with acute traumatic brain injury (CRASH-3): a randomised, placebo-controlled trial. *Lancet*. 2019;394:1713-1723.
17. R Core Team. *R: A Language and Environment for Statistical Computing*. Austria: R Foundation for Statistical Computing Vienna; 2022.
18. Léo G, Edouard O, Gaël V. *Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?*; 2022.
19. Victor C, Denis C, Mert D, Esther D, Christian H. *Double/Debiased/Neyman Machine Learning of Treatment Effects*. Nashville, TN: Amercian Economic Association; 2017.
20. Daniel J. *Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects*; 2020.
21. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci USA*. 2019;116:4156-4165.
22. Kennedy EH. Towards optimal doubly robust estimation of heterogeneous causal effects. *Elect J Stat*. 2023;17(2):3008-3049.
23. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*. 2020;108:299-319.
24. Peter R. Root-N-consistent semiparametric regression. *Econometrica*. 1988;56:931-954.

25. Alexandros R, Rijnbeek PR, Kent DM, Steyerberg EW, David VK. Estimating individualized treatment effects from randomized controlled trials: a simulation study to compare risk-based approaches. *BMC Med Res Methodol*. 2023;23:74.
26. David VK, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *J Clin Epidemiol*. 2019;114:72-83.
27. Susan A, Julie T. *Generalized Random Forests*. Ohio: Institute of Mathematical Statistics Beachwood; 2018.
28. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med*. 2011;30:2867-2880.
29. Chen S, Tian L, Cai T, Yu M. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*. 2017;73:1199-1209.
30. Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*. 2014;15:222-233.
31. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. *J Am Stat Assoc*. 2012;107:1106-1118.
32. Xiaohan G, Ai N. Contrast weighted learning for robust optimal treatment rule estimation. *Stat Med*. 2022;41:9574.
33. Francois V, Jared F. aVirtualTwins: Adaptation of Virtual Twins Method from Jared Foster. 2018 R package version 1.0.1.
34. Huling JD, Menggang Y. Subgroup identification using the personalized package. *J Stat Softw*. 2021;98:1-60.
35. Xu T, Chih-Chung C, Chih-Chen L, et al. WeightSVM: Subject Weighted Support Vector Machines. 2022 R package version 1.7-11.
36. Julie T, Susan A, Erik S, Stefan W. grf: Generalized Random Forests. 2022 R Package Version 2.2.1.
37. Nguyen TL, Collins GS, Landais P, Le Manach Y. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials-an illustration with the international stroke trial. *J Clin Epidemiol*. 2020;125:47-56.
38. Florie BB, Anna C, François G, Guillaume G, Raphaël P. Estimating individualized treatment effects using individual participant data meta-analysis working paper or preprint. 2022.
39. Steyerberg EW, Nieboer D, Debray TPA, Houwelingen HC. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: an overview and illustration. *Stat Med*. 2019;38:4290-4309.
40. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med*. 2013;32:3158-3180.
41. Kuniaki T, Serruys Patrick W, Valentin F, et al. Redevelopment and validation of the SYNTAX score II to individualise decision making between percutaneous and surgical revascularisation in patients with complex coronary artery disease: secondary analysis of the multicentre randomised controlled SYNTAXES trial with external cohort validation. *Lancet*. 2020;396:1399-1412.
42. Cain CH, Murray TA, Rudser KD, et al. Design considerations and analytical framework for reliably identifying a beneficial individualized treatment rule. *Contemp Clin Trials*. 2022;123:106951.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bouvier F, Peyrot E, Balendran A, et al. Do machine learning methods lead to similar individualized treatment rules? A comparison study on real data. *Statistics in Medicine*. 2024;1-19. doi: 10.1002/sim.10059

3.3 Analyse de l'agrément des méthodes sur les données de CRASH-2 et CRASH-3

Dans la discussion de l'article, nous avons fait l'hypothèse qu'avoir accès à un jeu de données avec plus d'hétérogénéité pourrait entraîner un meilleur agrément entre les règles développées par les différentes méthodes. Nous avons donc décidé de reproduire l'analyse de concordance en utilisant les données des essais CRASH-2 et CRASH-3. L'essai CRASH-2 a montré comme l'essai CRASH-3 qu'il y avait de l'hétérogénéité dans l'administration précoce du traitement [74]. L'administration précoce du traitement a entraîné un plus grand bénéfice que l'administration tardive chez les patients souffrant d'un traumatisme crânien léger ou modéré alors que le délai de traitement n'a pas eu d'effet chez les patients souffrant d'un traumatisme crânien grave. Combiner CRASH-2 et CRASH-3 nous permet de disposer d'une plus grande hétérogénéité et d'un échantillon de taille plus importante (28 448 observations).

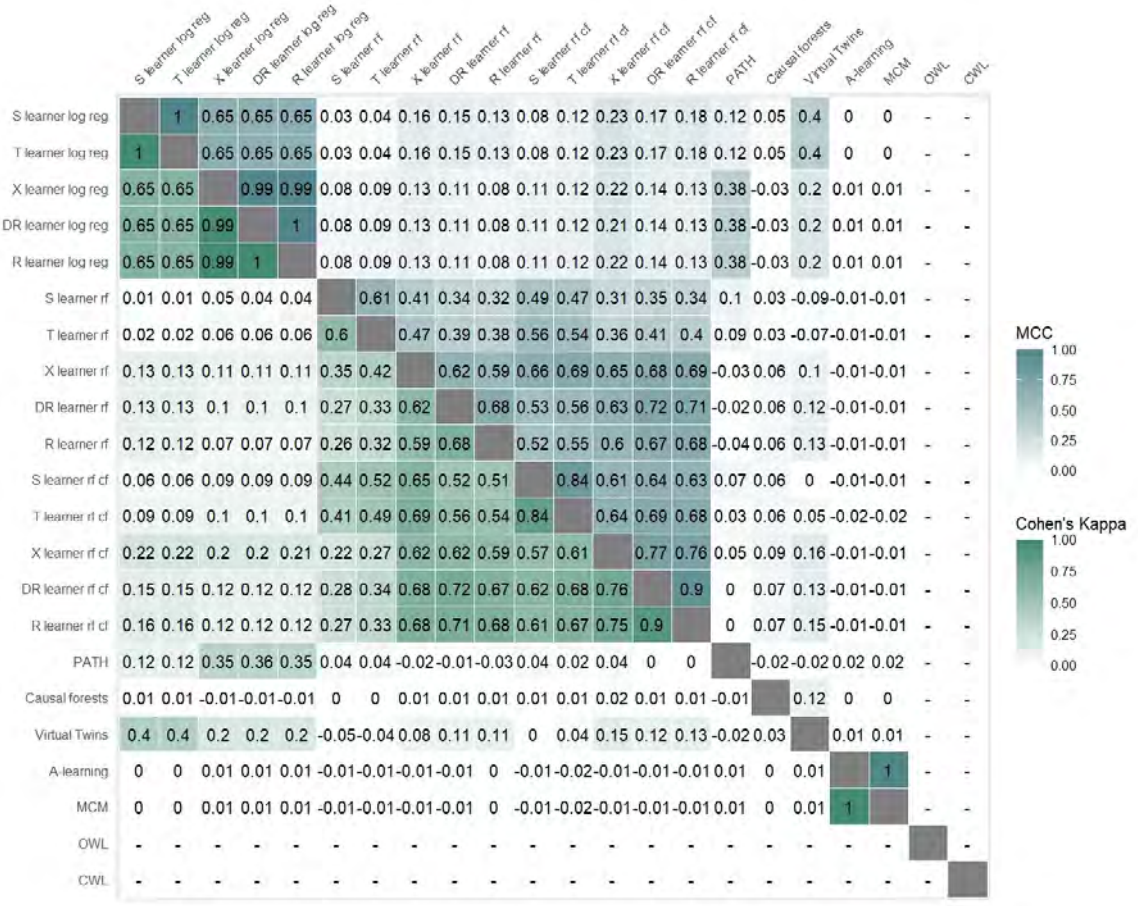

FIGURE 3.2 – Agrément entre les règles de traitement personnalisé développées à partir des données de CRASH-2 et CRASH-3.

Nous avons observé des niveaux de concordance comparables lors de l'utilisation de CRASH-

2 et CRASH-3 qu'en utilisant uniquement CRASH-3. Il existait un fort consensus entre les méthodologies similaires (meta-learners avec régression logistique, meta-learners avec forêt aléatoire, AL et MCM) ; cependant, la concordance était généralement faible entre les autres méthodes. L'augmentation de l'hétérogénéité et une taille plus grande de l'échantillon n'ont pas amélioré l'agrément entre les ITR élaborées à l'aide de ces différentes méthodes.

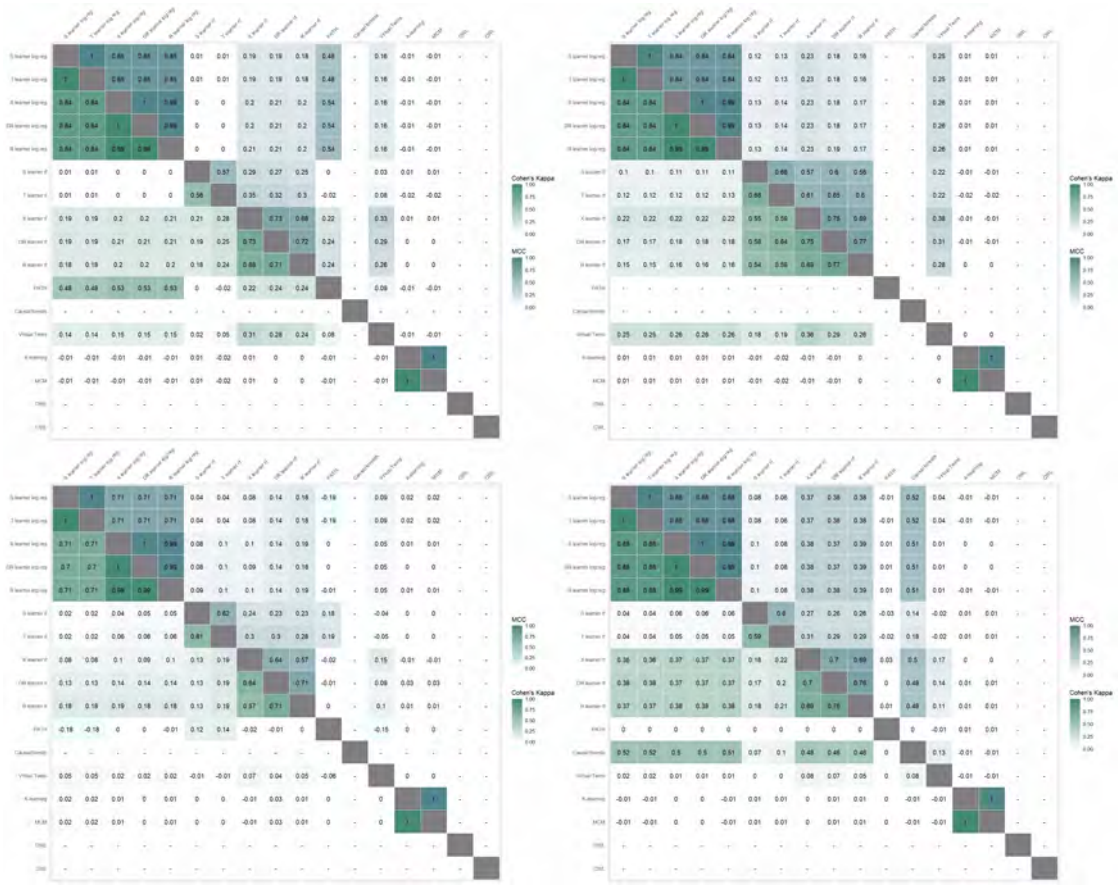

Figure 3.3 – Agrément entre les règles de traitement personnalisé développées à partir des données de CRASH-2 et CRASH-3 après avoir divisé l'ensemble des données en 4 groupes en fonction du risque de base.

Nous avons ensuite cherché à savoir si les méthodes produisaient des règles de traitement similaires lorsque le jeu de données était divisé en sous-groupes selon le risque de base. Diviser le jeu de données selon le risque de base permet d'avoir des patients relativement similaires.

La division de l'ensemble des données en sous-groupes de risque n'a pas conduit à une meilleure concordance entre les ITR. On a trouvé toujours un fort agrément entre les règles développées à partir de méthodes similaires mais l'agrément restait globalement faible.

TABLE 3.1 – Les trois variables les plus importantes utilisées pour recommander le traitement pour chaque méthode.

Méthodes	1	2	3
SL LR	injurytime	gcsE	gcsV
SL RF	gcsV	sbp	injurytime
SL CF	injurytime	gcsV	sbp
TL LR	injurytime	gcsE	gcsV
TL RF	gcsV	sbp	injurytime
TL CF	injurytime	gcsV	sbp
XL LR	injurytime	gcsE	gcsV
XL RF	injurytime	age	gcsM
XL CF	injurytime	age	gcsM
DRL LR	injurytime	gcsV	gcsE
DRL RF	injurytime	age	gcsM
DRL CF	injurytime	age	gcsM
RL LR	injurytime	gcsV	gcsE
RL RF	injurytime	age	gcsM
RL CF	injurytime	age	headinjury
PATH	injurytime	gcsM	headinjury
Causal Forests	injurytime	age	gcsV
Virtual Twins	injurytime	gcsV	gcsM
A-learning	sex	headinjury	gcsE
MCM	sex	headinjury	gcsE
OWL	gcsM	headinjury	gcsV
CWL	—	—	—

La plupart des méthodes ont identifié la durée depuis le traumatisme comme une variable clé pour déterminer s'il convenait de recommander le traitement (Table 3.1). La durée depuis le traumatisme était le facteur prédominant dans les décisions de traitement pour 16 des 22 méthodes. Il convient de noter que les méthodes qui ont utilisé la durée depuis le traumatisme pour l'attribution du traitement utilisent l'estimation des surfaces de réponse.

Les méthodes ont généralement recommandé de prendre en charge les patients ayant subi un traumatisme dans un délai de moins de 2h30. On observe que les approches qui considéraient la durée écoulée depuis le traumatisme comme une variable importante pour l'élaboration d'une règle de traitement personnalisé ont préconisé de soigner les patients le plus rapidement possible.

Table 3.2 – Durée moyenne depuis le traumatisme (en heures) pour recommander le traitement pour chaque méthode.

Méthodes	Temps moyen
SL LR	1.943
SL RF	2.587
SL CF	2.490
TL LR	1.943
TL RF	2.591
TL CF	2.479
XL LR	2.103
XL RF	2.368
XL CF	2.254
DRL LR	2.116
DRL RF	2.375
DRL CF	2.405
RL LR	2.114
RL RF	2.449
RL CF	2.431
PATH	2.856
Causal Forests	2.328
Virtual Twins	2.429
A-learning	2.856
MCM	2.856
OWL	2.856
CWL	—

Si augmenter l'hétérogénéité de l'effet traitement et la taille d'échantillon n'a pas abouti à un meilleur agrément entre les ITR, la plupart des ITR ont considéré la durée depuis le traumatisme comme une variable importante pour développer la règle de traitement qui est la variable où l'hétérogénéité de l'effet traitement a été observée. La majorité des règles préconisaient de donner le traitement aux patients ayant subi le traumatisme moins de 2h30 avant la prise en charge.

Chapitre 4

Discrimination optimale pouvant être obtenue pour plusieurs distributions d'effets du traitement

4.1 Résumé du projet

4.1.1 Introduction et objectifs

Le domaine des modèles de prédiction des risques est bien documenté, offrant diverses méthodologies pour évaluer la performance des modèles [40, 41], en se concentrant sur des aspects tels que la calibration et la discrimination [75, 76], ainsi que le calcul de l'utilité clinique grâce à l'analyse de la courbe de décision [77]. La démonstration d'une forte discrimination est considérée comme un aspect essentiel de la performance d'un modèle, comme l'ont souligné des études précédentes [78, 79].

Contrairement aux modèles de prédiction des risques, il existe peu de travaux sur les modèles d'effets individualisés du traitement et les ITR. L'évaluation des performances des ITR a été décrite par certaines métriques, y compris le concept de la valeur de la règle [80], l'avantage de la règle en termes de traitement attribué [81], et d'autres métriques spécifiques [82]. Ces indicateurs sont axés sur la compréhension des avantages de l'utilisation des ITR par rapport à leur non-utilisation. L'évaluation des ITE est moins courante en raison de la difficulté à mesurer l'efficacité des prédictions sur les effets individualisés du traitement, car il est impossible d'observer les résultats de tous les traitements possibles pour un seul individu. Cette limita-

tion signifie que les métriques traditionnelles de prédiction des risques ne sont pas entièrement applicables. Néanmoins, l'évaluation de la calibration des modèles ITE peut être effectuée en comparant les bénéfices prédits par rapport aux bénéfices estimés à travers des déciles de la population [44]. De nouvelles métriques d'évaluation de la discrimination ont été introduites pour identifier les personnes qui bénéficieront d'un traitement par rapport à celles qui n'en bénéficieront pas. Ce travail se concentre sur trois métriques spécifiques : *c-statistic for benefit* [42], la "concentration du bénéfice" (*Concentration of Benefit*) [43] et "l'effet de prescription moyen de la population" (*Population Average Prescriptive Effect*) [51].

Des études antérieures ont mis en évidence la nécessité d'une différence significative dans le risque d'événement pour distinguer efficacement les patients présentant un risque élevé ou faible de survenue d'un événement [83, 84]. Le degré d'hétérogénéité de l'effet de traitement nécessaire pour créer des ITR efficaces avec une capacité de discrimination optimale reste un sujet à explorer davantage, en particulier en ce qui concerne la discrimination maximale pouvant être obtenue avec différentes distributions de l'effet de traitement.

Nos objectifs consistaient à montrer la capacité de discrimination maximale pouvant être obtenue avec diverses distributions de l'effet du traitement et de comparer la discrimination obtenue par les trois métriques. Nous avons généré différentes distributions d'effets de traitement dans lesquelles la discrimination des modèles ITE a été calculée par intégration numérique.

4.1.2 Méthodes

4.1.2.1 Métriques

Pour toutes les distributions, trois métriques ont été utilisées pour décrire la performance discriminatoire, à savoir la capacité de la règle de traitement individualisé à séparer les individus qui bénéficient du traitement de ceux qui n'en bénéficient pas : la *c-statistic for benefit*, la *concentration of benefit*, et le *population average prescriptive effect*. Ces métriques ont été définies dans le chapite 1.

4.1.2.2 Génération de distributions

Pour produire les distributions des effets individualisés du traitement τ, nous avons commencé par générer les résultats potentiels Y^1 et Y^0. Nous avons opté pour des distributions Beta, qui varient entre 0 et 1, ce qui permet à la distribution des effets individualisés du traitement τ de varier entre -1 et 1. Pour un individu spécifique, ces résultats potentiels ne sont pas indépendants, car les caractéristiques du patient i sont susceptibles de les influencer. Il est donc nécessaire de créer des distributions Beta dépendantes. Une méthode pour créer des distributions Beta dépendantes consiste à générer une distribution de Dirichlet U avec quatre marginales distribuées selon une distribution Beta U_1, U_2, U_3, U_4, ce qui donne une distribution Beta bivariée [85].

Les paramètres de génération des distributions ont été choisis pour créer des distributions unimodales ou bimodales avec des effets de traitement hétérogènes, garantissant que certains individus bénéficient du traitement et d'autres non. L'effet de traitement moyen des distributions générées a été choisi pour être proche de 0 car ces scénarios sont les plus intéressants en médecine personnalisée. Des mélanges de distributions ont été utilisés dans certains scénarios pour obtenir des distributions bimodales.

4.1.3 Résultats

La discrimination maximale réalisable avec les métriques employées pour 20 distributions est présentée dans la figure 4.1.

L'analyse a révélé la capacité de discrimination la plus élevée pouvant être atteinte à travers diverses distributions d'effets de traitement hétérogènes. Elle a également identifié les types de distribution spécifiques et les degrés d'hétérogénéité de l'effet traitement requis pour obtenir une discrimination optimale en ce qui concerne la *c-statistic for benefit*, la *concentration of benefit* et le PAPE. Une *c-statistic for benefit* élevée est obtenue lorsqu'il y a une grande diversité dans les bénéfices. Une *concentration of benefit* élevée est obtenue avec des distributions symétriques qui comprennent un nombre égal d'individus bénéficiant ou non du traitement.

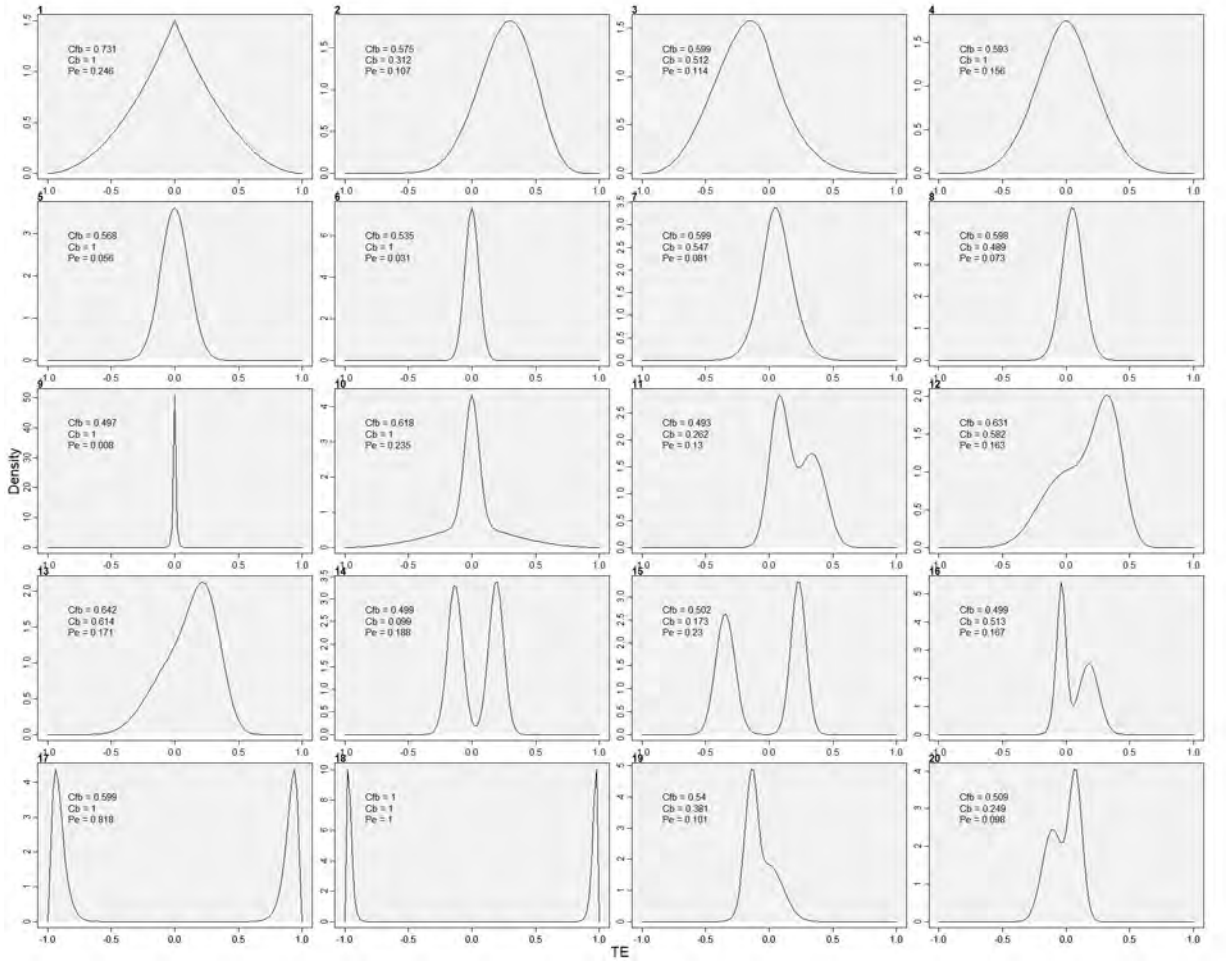

Figure 4.1 – Résultats des distributions sélectionnées pour l'analyse.
C_{fb} = c-statistic for benefit ; Co_b = concentration of benefit ; P_e = 2 × PAPE

Les valeurs de PAPE les plus élevées sont obtenues dans les distributions comprenant des personnes ayant un bénéfice important et des personnes ayant un non-bénéfice important.

4.1.4 Discussion

Les résultats indiquent que des performances favorables pour les trois métriques ont été observées dans les distributions comprenant de nombreux individus avec un ITE proche de 1 et de nombreux individus avec un ITE proche de -1. En revanche, les distributions ne comprenant que des individus ayant connu un petit bénéfice ou un petit non-bénéfice ont mené à une performance limitée pour toutes les métriques. Il est intéressant de noter que les trois métriques utilisées dans ce travail n'ont pas exigé des conditions et des niveaux identiques d'hétérogénéité de l'effet du traitement pour produire des résultats de discrimination élevés. En général, il a été observé que l'obtention d'une bonne *c-statistic for benefit* était associée à des

situations dans lesquelles l'ITE prenait une large gamme de valeurs. Cela peut être attribué au rôle de la *c-statistic for benefit*, qui mesure à quel point il est facile de différencier les individus bénéficiant d'un bénéfice de ceux qui n'en bénéficient pas. Les distributions présentant des bénéfices variés distinguent clairement ces groupes, ce qui permet une meilleure discrimination et donc des valeurs C_{textfb} plus élevées. Les distributions comprenant un nombre égal d'individus bénéficiant et ne bénéficiant pas du traitement ont conduit à une *concentration of benefit* optimale. Rappelons que la *concentration of benefit* est définie comme $Co_{\text{b}} = 1 - \frac{\mathbb{E}(t_1)}{\mathbb{E}[\max(t_2,t_3)]}$, les distributions symétriques ont $\mathbb{E}(t) = 0$, ce qui conduit à $\text{Co}_b = 1$. La performance supérieure d'une règle de traitement individualisé, c'est-à-dire les valeurs optimales de PAPE, a été obtenue lorsque les distributions englobaient des personnes présentant un bénéfice important et des personnes présentant un non-bénéfice important. La plus grande hétérogénéité présente lorsqu'il y a des individus avec de forts bénéfices et de forts non-bénéfices, renforce l'efficacité de la règle de traitement individualisé, conduisant à des valeurs PAPE plus élevées.

4.1.5 Conclusion

En conclusion, la présence d'effets de traitement hétérogènes, qu'ils soient caractérisés par une variabilité dans la direction ou dans la magnitude, ne s'est pas systématiquement traduite par des résultats favorables en matière de discrimination. La discrimination optimale dépend de la distribution des effets du traitement. De plus, le choix de la métrique utilisée pour évaluer la discrimination influence également les résultats. En pratique, l'examen de la distribution des effets du traitement pourrait fournir des indications sur la capacité de discrimination maximale qu'il est possible d'atteindre. Cependant, les ITE sont inobservables et leurs prédictions sont différentes selon la méthode choisie pour développer le modèle, même dans de grands ensembles de données [86].

La description détaillée des méthodes et des résultats de ce travail se trouve dans la section 4.2.

4.2 Article

Evaluating the influence of treatment effects heterogeneity on Discrimination

Florie Bouvier *[1], Etienne Peyrot[1], François Petit[1], and Raphaël Porcher[1,2]

[1]Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), F-75004 Paris, France
[2]Centre d'Épidémiologie Clinique, Assistance Publique-Hôpitaux de Paris, Hôtel-Dieu, Paris, France

Abstract

Analyzing the heterogeneity of treatment effects is crucial in personalized medicine to identify which patients will benefit from specific treatments. The performance of an individualized treatment effects model to guide treatment decisions can be assessed in different ways, with an important one being their ability to effectively discriminate between individuals who benefit from the treatment and those who do not. While many methods and algorithms have been proposed to develop individualized treatment effects models and individualized treatment rules, little is known about the maximum discriminative ability that can be achieved according to the population's underlying distribution of treatment effects. In this work, we calculated the maximum discrimination that can be achieved for a panel of 20 distributions with varying average treatment effects and levels of heterogeneity. The assessment included the following discrimination metrics: the c-statistic for benefit, the concentration of benefit, and the population average prescription effect (PAPE). Results showed the three metrics employed in this study did not require the same levels of treatment effect heterogeneity to lead to high discrimination results. Notably, achieving high c-statistic for benefit and PAPE values required greater heterogeneity than obtaining high concentration of benefit values. Obtaining high discrimination values depends on the distribution of treatment effects and the choice of metric.

Keywords: personalized medicine; individualized treatment effects; discrimination; heterogeneous treatment effects

1 Introduction

The analysis of heterogeneity of treatment effects (HTE), which represents the non-random variability in the direction or magnitude of a treatment effect, is pivotal in personalized medicine for identifying

*Correspondence to: Florie Bouvier (florie.bouvier@u-paris.fr)
Hôpital Hôtel-Dieu, 1 place du Parvis de Notre-Dame, 75004 Paris, France

patients who would benefit from specific treatments and facilitating the development of individualized treatment rules (ITR), which are decision rules recommending treatment based on individual patient characteristics [1–3].

Extensive literature is available for evaluating the performance of risk prediction models [4, 5], as various metrics are used to assess a model's performance comprehensively, focusing on aspects such as calibration [6, 7] and discrimination [8]. Further, the clinical utility of a model can be quantified using methods like net benefit and decision curve analysis [9]. Demonstrating strong discrimination is considered a key aspect of a model's performance [10, 11].

As opposed to risk prediction models, limited works exist on the performance of individualized treatment effects models and ITRs. The assessment of ITR performance has been described by some metrics, including the concept of the ITR's value [12], the benefit of the rule in terms of assigned treatment [13], and other related metrics [14]. These metrics primarily focus on evaluating the benefits of implementing ITRs. Conversely, fewer studies have concentrated on the performance of individualized treatment effects models due to the inherent challenge in measuring such a performance, as the outcomes of both treatments are typically not observable within a single patient. Consequently, conventional risk prediction metrics cannot effectively quantify the performance of models predicting individualized treatment effects. Calibration can be assessed by comparing predicted treatment benefits to estimated treatment benefits across deciles of individualized treatment effects [15]. Recent metrics have also been proposed to evaluate discrimination, which gauges the ability to differentiate individuals who benefit from those who do not. In this work, we were specifically focused on three of them: the c-statistic for benefit [16], the concentration of benefit [17] and the population average prescription effect (PAPE) [18].

Previous works have indicated that a significant disparity in risk of events is necessary to effectively differentiate between patients at high risk of developing the outcome and those at low risk [19, 20]. The extent of heterogeneity in treatment effects required to derive effective ITRs with optimal discriminative ability has yet to be thoroughly investigated. Specifically, there is an interest in understanding the maximum discrimination that can be attained with different distributions of individualized treatment effects.

Our objectives encompassed showing the maximum discriminative capacity achievable with diverse treatment effect distributions and comparing the discrimination obtained by the three aforementioned metrics. We generated different distributions of treatment effects in which the discrimination of the ITE models was calculated using numerical integration. The remaining sections of the paper are structured as follows: Section 2 provides details on the metrics used and how the distributions were generated, Section 3 showcases the results, and Section 4 concludes the paper with a discussion.

2 Methods

This section introduces the metrics employed for the analysis and describes how they were numerically calculated. Twenty distributions featuring heterogeneous treatment effects, with alterations to the direction or magnitude of these effects, were chosen to illustrate the maximum achievable dis-

crimination. Three discrimination metrics were computed in all distributions (see subsection 2.1).

In Rubins' counterfactual framework, the individual benefit for individual i is defined as $B_i = Y_i^1 - Y_i^0$ where Y_i^0 and Y_i^1 refer to the outcomes that would be observed if the individual i was assigned to either the control or the experimental treatment [21]. In practice, B_i is not observable. Hence, we resort to covariates X to estimate the so-called individualized treatment effects $\tau(X) = \mathbb{E}(Y^1 - Y^0|X)$ which represents the difference in predicted benefits of two treatments and can be estimated as:

$$\hat{\tau}(X) = \hat{\mathbb{E}}(Y^1 - Y^0|X).$$

In this project, we focus on binary outcomes. Without loss of generality, we assume that $Y = 1$ is a desirable event so that $\tau > 0$ would lead to recommend the experimental treatment.
In this project, we position ourselves in an oracle situation where the individualized treatment effects are perfectly estimated and correspond to the individual benefits.

2.1 Metrics

For all distributions, three metrics were employed to describe the discriminatory performance i.e. the ability of the individualized treatment rule to separate individuals who benefit from taking the treatment from individuals who do not benefit from taking the treatment: the c-statistic for benefit, the concentration of benefit, and the population average prescriptive effect.

The c-statistic for benefit (C_{fb}) is the probability that, for two patients u and v with an unequal individualized treatment effect, the patient with the greater individualized treatment effect also has a higher individual benefit [16]. Higher values of the c-statistic for benefit indicate a higher performance. It can be mathematically expressed as:

$$C_{\text{fb}} = P\Big(\tau_u > \tau_v \,|\, B_u > B_v\Big)$$

where B_u and B_v represent the individual benefits of patients u and v and where τ_u and τ_v represent the individualized treatment effects of patients u and v respectively.
It's important to note that we have modified the definition here to consider a scenario where individual benefits are accessible, which differs from typical practice since patients usually receive only one treatment. In the original definition, the individual benefit refers to the difference in outcomes between two patients with the same predicted individualized treatment effect but receiving different treatments.
Some concerns have been made regarding the quality of the C_{fb} as a discrimination metric, notably recent studies have shown that the C_{fb} is not a proper scoring rule or that it is sensitive to the way the pairs are matched [22, 23].

The concentration of benefit (Co_b) metric evaluates the extent to which covariates effectively capture the variation in treatment effects and measures how well the individualized treatment effect model

identifies the individuals who experience greater benefits from the treatment [17]. Larger concentration of benefit values, in particular approaching 1, indicate better discrimination. The concentration of benefit can be mathematically denoted as

$$Co_{\text{b}} = 1 - \frac{\mathbb{E}(t_1)}{\mathbb{E}[\max(t_2, t_3)]}$$

where t_1, t_2 and t_3 represent three random draws from the distribution of individualized treatment effects τ.

The population average prescription effect (PAPE) is the difference in mean outcomes when comparing an ITR and a treatment rule that randomly administrates treatment to the same proportion of patients [18]. The PAPE evaluates the efficacy of the ITR in recommending treatment only to the individuals who benefit from it. It is expressed as

$$\text{PAPE} = \mathbb{E}[Y(r) - p_r Y^1 - (1 - p_r)Y^0]$$

where $Y(r) = rY^1 + [1 - r]Y^0$ represents the outcome observed if the rule $r = \mathbb{1}_{\{\tau(X)>0\}}$ was followed and p_r represents the proportion of patients assigned to the experimental treatment under the ITR r. It can be re-expressed as

$$\text{PAPE} = \mathbb{E}[(r - p_r)\tau].$$

The PAPE takes ranges from -0.5 and 0.5, with higher PAPE values implying a better performance of the ITR. A value of 0 indicates that the ITR's performance is no better than random treatment allocation for the same proportion of patients. Further information explaining the reasons for the PAPE falling within the range of -0.5 and 0.5 can be found in Appendix A. To facilitate a more straightforward comparison of the metric results, we used $P_e = 2 \times$ PAPE instead of PAPE to obtain values between -1 and 1 and make it comparable with C_{fb} and Co_b.

2.2 Distribution generation

To produce the distributions of individualized treatment effects τ, we started by generating the potential outcomes Π^1 and Π^0. We opted for Beta distributions, which vary between 0 and 1, allowing the distribution of individualized treatment effects τ to range from -1 to 1. For a specific individual i, these potential outcomes are not independent, as the characteristics of patient i are likely to influence them. This necessitated the generation of dependent Beta distributions. One method for creating dependent Beta distributions involves generating a Dirichlet distribution U with four Beta-distributed marginals U_1, U_2, U_3, U_4, resulting in a bivariate Beta distribution [24].
We now specify such a Dirichlet distribution.
Let $U \sim \text{Dirichlet}(\alpha)$ where $U = (U_1, U_2, U_3, U_4)$ and $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ with $\alpha_i > 0$ for $i = 1, \cdots, 4$ and $U_4 = 1 - U_1 - U_2 - U_3$.

We define the two potential outcome distributions as $\Pi_1 = U_1 + U_2$ and $\Pi_0 = U_1 + U_3$. Given that $(U_1 + U_2, U_3, U_4) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \alpha_4)$ [25] and that a Dirichlet distribution has beta-distributed marginals, we obtain the following [26]:

$$\Pi_1 \sim \text{Beta}(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4) \text{ and } \Pi_0 \sim \text{Beta}(\alpha_1 + \alpha_3, \alpha_2 + \alpha_4)$$

In this study, we consider an oracle scenario where the individualized treatment effects are perfectly estimated i.e. $\hat{\tau} = \tau$. Thus, to produce the individual benefits, we employ the values of the potential outcomes as the probability parameter in a Bernoulli distribution. Specifically:

$$Y^t \sim \text{Bernoulli}(\pi_t), t \in \{0, 1\},$$

where π_t is drawn from the beta distribution Π_t.
The parameters for generating distributions were selected to create unimodal or bimodal distributions with heterogeneous treatment effects, ensuring some individuals benefit from the treatment and others do not. The average treatment effect of the generated distributions was close to 0 as these scenarios hold the most interest in personalized medicine. Mixtures of two distributions were used in some scenarios to obtain distributions of τ with two modes. The parameters selected for each distribution are listed in Appendix B.

2.3 Metrics calculation

The values of the metrics for all distributions were computed via numerical integration, using the R statistical software version 4.1 [27]. The specific formulae are detailed in Appendix A.

To calculate the c-statistic for benefit, we re-expressed its original definition using the Bayes theorem:

$$C_{\text{fb}} = \frac{P(B_u > B_v \mid \tau_u > \tau_v) P(\tau_u > \tau_v)}{P(B_u > B_v)}.$$

Since τ_u and τ_v are continuous and independent, $P(\tau_u > \tau_v) = 0.5$. $P(B_u > B_v)$ can be found using the probability mass function of the Bernoulli distributions. The detailed calculations on how to obtain $P(B_u > B_v \mid \tau_u > \tau_v)$ are presented in Appendix A.

The concentration of benefit was determined by evaluating $\mathbb{E}(t) = \int_{-1}^{1} x f_\tau(x; \alpha_1, \alpha_2, \alpha_3, \alpha_4)\, dx$ and $\mathbb{E}[\max(t_2, t_3)] = 2\int_{-1}^{1} x f_\tau(x) F_\tau(x)\, dx$.

The P_e was calculated by integrating the following integral:

$$2\Big(\int_0^1 x f_\tau(x; \alpha_1, \alpha_2, \alpha_3, \alpha_4)\, dx - \int_0^1 f_\tau(x; \alpha_1, \alpha_2, \alpha_3, \alpha_4)\, dx \int_{-1}^1 x f_\tau(x; \alpha_1, \alpha_2, \alpha_3, \alpha_4)\, dx\Big).$$

3 Results

The illustration of the maximum discrimination achievable with the employed metrics for 20 distributions is presented in Figure 1.

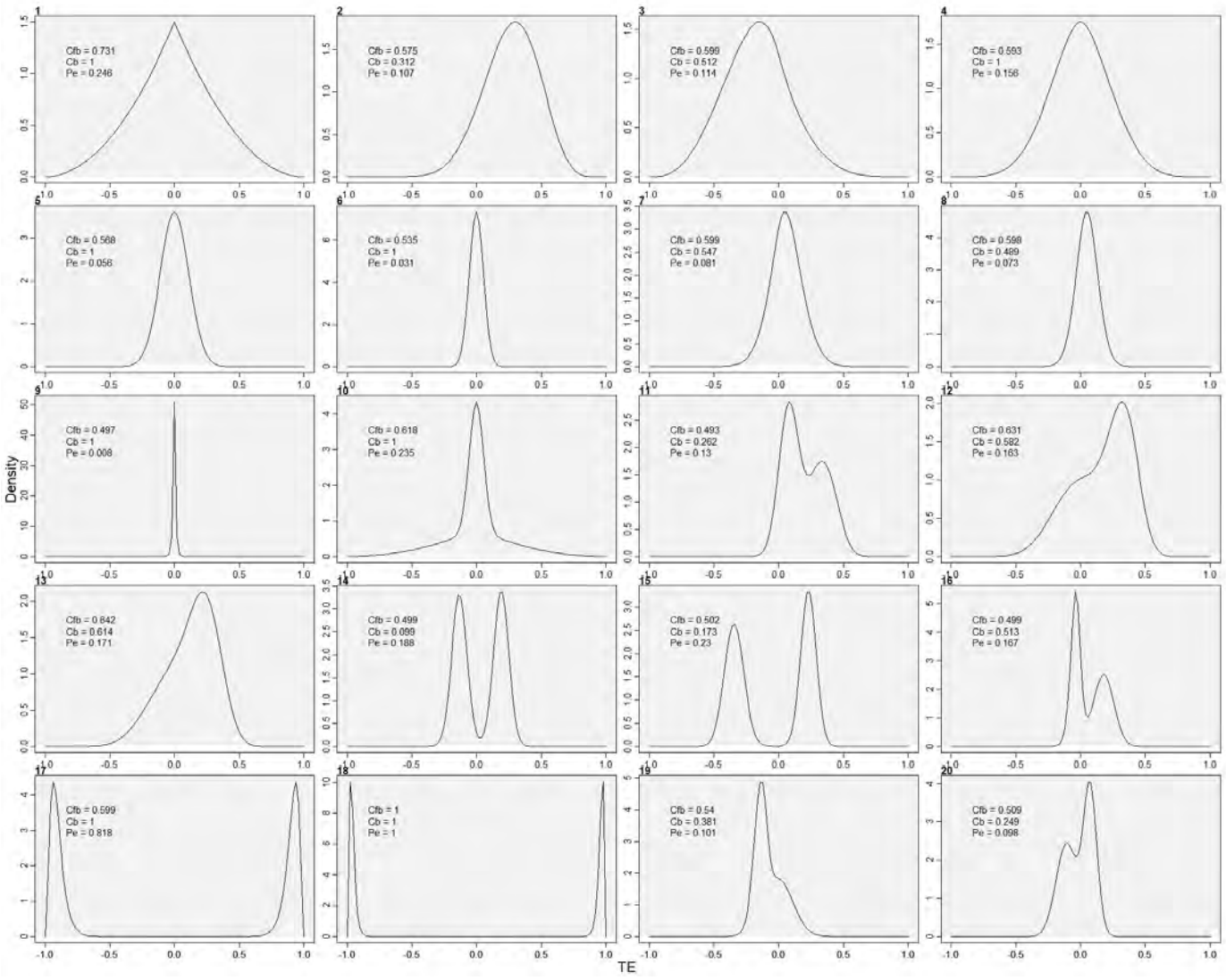

Figure 1: Results of the distributions selected for the analysis.

With some distributions, similar results were observed for all the metrics. For example, a distribution consisting of individuals who either experience a significant benefit (ITE close to 1) or significant non-benefit (ITE close to -1), with both groups equally represented, results in excellent discriminatory power as shown in distribution 18 where all the metrics have a value of 1. Bimodal distributions where individuals experience a small to moderate benefit (ITE between 0 and 0.5) or a small to moderate non-benefit (ITE between 0 and -0.5) lead to limited potential for developing effective personalized treatment strategies, as they result in low metric values. This is observed in distributions 11, 14, 15, 19, and 20, where Co_b values range from 0.099 to 0.381, the c-statistic for benefit values fall between 0.493 and 0.540, and P_e values range from 0.028 to 0.292. Of note, the c-statistic for benefit, which is expected to range from 0.5 to 1, fell below 0.5 for distributions 9, 11, 14, and 16, underscoring the concerns of it not being a proper scoring rule as discussed in previous papers [22, 23].

Disagreement between the three metrics is seen in distributions including an equal number of individuals who either benefited or did not benefit from the treatment, with varying magnitudes of treatment effect. These distributions displayed optimal Co_b values of 1, moderate C_{fb} values, and low P_e values that were close to 0. This is illustrated in distributions 4, 5, and 10. Notably, optimal Co_b values are

expected in such scenarios because symmetrical distributions have an average treatment effect equal to 0, thus $\mathbb{E}(t) = 0$, which leads to $\text{Co}_b = 1$.

A distribution in which the majority of individuals benefit from the treatment, with most patients experiencing a moderate benefit, results in low Co_b and P_e values but a moderate C_{fb} value. On the other hand, patients who experience a significant benefit (ITE around 0.8) or a significant lack of benefit (ITE around -0.8) demonstrate high discriminative ability with Co_b and P_e, while the C_{fb} shows moderate discrimination.

Unimodal distributions with an average treatment effect between -0.2 and 0.2 record low P_e values, suggesting that a personalized treatment strategy did not significantly outperform a non-personalized approach. However, they demonstrate moderate to high values for C_{fb} and Co_b indicating a relatively good discriminative capability.

Distributions where equal numbers of individuals experience benefit and non-benefit from the treatment, with low treatment effects, such as distributions 6 and 9, show low C_{fb} and P_e values but have high Co_b values. Distributions with a small heterogeneity of treatment effects are often seen in real randomized controlled trials, this is the case for the International Stroke Trial, the CRASH-3 trial, or the INDANA meta-analysis [28–30]. When many individuals experience a slight non-benefit, with some experiencing a small to moderate benefit, as seen in distribution 16, this results in low C_{fb} and P_e values but a moderate Co_b value.

The analysis revealed the highest discriminative ability attainable across various distributions of heterogeneous treatment effects. It also identified the specific distribution types and degrees of heterogeneity in treatment effects required for achieving optimal discrimination regarding the c-statistic for benefit, concentration of benefit, and P_e. A high c-statistic for benefit is obtained when many people experience varied benefits. A high concentration of benefit is achieved with symmetric distributions that include equal numbers of individuals benefiting and not benefiting from treatment. The highest P_e values are found in distributions featuring people with a strong benefit and people with a strong non-benefit. A summary of the results can be found in Table 1.

Table 1: Summary table showing the maximum achievable discrimination for the three metrics, categorized into three levels: low, medium, and high.

	C_{fb}	Co_b	PAPE
Distributions with many ITE near 1 or -1	high	high	high
Bimodal distributions with small to moderate ITEs (-0.5 to 0.5)	low	low	low
Symmetrical distributions with varying magnitudes of ITEs	medium	high	low
Distributions with a majority of ITEs between 0 and 0.4	medium	low	low
Distributions with ITEs near 0.8 or -0.8	medium	high	high
Unimodal distributions with an ATE between -0.2 and 0.2	medium/high	medium/high	low
Symmetrical distributions with small values of ITEs	low	high	low
Distributions with many ITEs near -0.1 and others near 0.2	low	medium	low

4 Discussion

This study examined how different distributions of treatment effects, each with varying levels of heterogeneity, impacted the maximum achievable discrimination. The performance of individualized treatment rules derived from these distributions was evaluated through numerical integration, focusing on discrimination measures (c-statistic for benefit, Concentration of Benefit, and Population Average Prescription Effect). Distributions of treatment effects were generated by employing the difference between two dependent Beta distributions and mixtures thereof.

Results indicated that favorable performance across all three metrics was observed in distributions that included many individuals with an ITE near 1 and many with an ITE near -1. Alternatively, distributions consisting solely of individuals who experienced either a minor benefit or a minor non-benefit showed poor performance across all metrics. Such distributions are commonly encountered when working with data from randomized controlled trials. Interestingly, the three metrics employed in this study did not demand identical conditions and levels of treatment effect heterogeneity to yield high discrimination results. In general, it was observed that achieving good c-statistic for benefit values was associated with situations in which the individualized treatment effect takes a large range of values. This can be attributed to the role of the c-statistic for benefit, which measures how well we can differentiate individuals with a benefit from those without a benefit. Distributions with wide benefits clearly distinguish these groups, allowing for better discrimination and thus higher C_{fb} values. Distributions that included equal numbers of individuals benefiting and not benefiting from treatment led to optimal concentration of benefit values. Recall that the concentration of benefit is defined as $Co_{\text{b}} = 1 - \frac{\mathbb{E}(t_1)}{\mathbb{E}[\max(t_2,t_3)]}$, Symmetric distributions have $\mathbb{E}(t) = 0$, which leads to $\text{Co}_b = 1$. The superior performance of an individualized treatment rule i.e. optimal PAPE values, was achieved when distributions encompassed people with a strong benefit and people with a strong non-benefit. The greater heterogeneity present when there are individuals with both strong benefits and strong non-benefits, enhances the effectiveness of the individualized treatment rule, leading to higher PAPE values. The difference between the three metrics suggests that the choice of the metric influences the conclusions drawn from the results. A significantly higher level of heterogeneity is necessary to achieve high c-statistic for benefit values and PAPE compared to what is required for obtaining a high concentration of benefit.

Alternative metrics to evaluate the discrimination of an ITE model include the c-statistic for benefit using 1:1 matching on predicted outcome risk under the control and the model-based c-statistic for benefit [22]. However, these were not included in this work, as they assess the same discrimination aspect as the c-statistic for benefit.

Previous simulation studies have looked into the factors contributing to effective discrimination performance. For instance, Rekkas et al. compared diverse methods for constructing ITRs, revealing that larger sample sizes and a moderate average treatment effect (OR $= 0.8$) were associated with improved c-statistic for benefit values [31]. Hoogland et al. conducted a simulation study assessing the performance of various discrimination metrics, including the c-statistic for benefit, across different sample sizes [22]. They observed that larger sample sizes led to less biased discrimination metrics.

To our knowledge, no study has specifically examined the discrimination and relative performance of different distributions of treatment effects using the three metrics employed in this study. It might be valuable to undertake a simulation study incorporating diverse sample sizes, distribution shapes, and event rates to discern which parameter exerts the most significant impact on discriminative performance.

In conclusion, the presence of heterogeneous treatment effects, whether characterized by variability in the direction or magnitude, did not consistently result in favorable discrimination results. The optimal discrimination depends on the distribution of treatment effects. Furthermore, the choice of metric used to evaluate discrimination also influences the results. In practice, examining the distribution of treatment effects could provide insights into the achievable maximum discriminative ability. However, the ITEs are unobservable and their predictions are different given the method chosen to develop the model, even in large datasets [29].

Declarations

Competing interests

The authors declare that they have no competing interests.

Funding

FB and RP acknowledge support by the French Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

Authors' contributions

Study concept and design: FB, FP, and RP. Analysis and interpretation of data: FB, EP, FP, and RP. Drafting of the manuscript: FB and RP. Critical revision of the manuscript for important intellectual content: FB, EP, FP, and RP.

Acknowledgements

The authors thank François Grolleau and Roch Giorgi for their useful comments and suggestions.

References

[1] Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology* 2013; **66**:818–825. URL https://www.sciencedirect.com/science/article/pii/S0895435613000863.

[2] Kent DM, Steyerberg E, Van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ* 2018; :k4245URL https://www.bmj.com/lookup/doi/10.1136/bmj.k4245.

[3] Kent DM, Paulus JK, Van Klaveren D, D'Agostino R, Goodman S, Hayward R, Ioannidis JP, Patrick-Lake B, Morton S, Pencina M, Raman G, Ross JS, Selker HP, Varadhan R, Vickers A, Wong JB, Steyerberg EW. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. *Annals of Internal Medicine* 2020; **172**:35. URL https://annals.org/aim/fullarticle/2755582/predictive-approaches-treatment-effect-heterogeneity-path-statement.

[4] Harrell F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2015.

[5] Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. volume 19. 2009.

[6] Hosmer DW, Lemesbow S. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods* 1980; **9**:1043–1069.

[7] Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. *Medical Decision Making* 1993; **13**:49–57.

[8] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Jama* 1982; **247**:2543–2546.

[9] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 2006; **26**:565–574.

[10] Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *Journal of Clinical Epidemiology* 1992; **45**:85–89. URL https://linkinghub.elsevier.com/retrieve/pii/089543569290192P.

[11] Gail MH. On criteria for evaluating models of absolute risk. *Biostatistics* 2005; **6**:227–239. URL https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxi005.

[12] Tsiatis AA. *Dynamic treatment regimes: statistical methods for precision medicine*. Chapman & hall/crc monographs on statistics and applied probability. Chapman and Hall/CRC: Boca Raton, 2020.

[13] Janes H, Brown MD, Huang Y, Pepe MS. An approach to evaluating and comparing biomarkers for patient treatment selection. *The International Journal of Biostatistics* 2014; **10**:99–121.

[14] Grolleau F, Petit F, Porcher R. A comprehensive framework for the evaluation of individual treatment rules from observational data 2023.

[15] van Klaveren D, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *Journal of Clinical Epidemiology* 2019; **114**:72–83. URL https://linkinghub.elsevier.com/retrieve/pii/S0895435618310072.

[16] van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. *Journal of Clinical Epidemiology* 2018; **94**:59–68. URL https://linkinghub.elsevier.com/retrieve/pii/S0895435617303037.

[17] Sadatsafavi M, Mansournia MA, Gustafson P. A threshold-free summary index for quantifying the capacity of covariates to yield efficient treatment rules. *Statistics in Medicine* 2020; **39**:1362–1373. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8481.

[18] Imai K, Li ML. Experimental Evaluation of Individualized Treatment Rules. *Journal of the American Statistical Association* 2021; :1–15URL http://arxiv.org/abs/1905.05389. arXiv:1905.05389 [stat].

[19] Pepe MS. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. *American Journal of Epidemiology* 2004; **159**:882–890. URL https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwh101.

[20] Cook NR. Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. *Circulation* 2007; **115**:928–935. URL https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.106.672402.

[21] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0037350.

[22] Hoogland J, Efthimiou O, Nguyen TL, Debray TPA. Evaluating individualized treatment effect predictions: a new perspective on discrimination and calibration assessment 2022. URL http://arxiv.org/abs/2209.06101. arXiv:2209.06101 [stat].

[23] Xia Y, Gustafson P, Sadatsafavi M. Methodological concerns about "concordance-statistic for benefit" as a measure of discrimination in predicting treatment benefit. *Diagnostic and Prognostic Research* 2023; **7**:10. URL https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-023-00147-z.

[24] Olkin I, Trikalinos TA. Constructions for a bivariate beta distribution. *Statistics & Probability Letters* 2015; **96**:54–60.

[25] Ng KW, Tian GL, Tang ML. *Dirichlet and related distributions: Theory, methods and applications*. John Wiley & Sons, 2011.

[26] Moschen LM, Carvalho LM. Bivariate beta distribution: parameter inference and diagnostics 2023.

[27] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2022. URL https://www.R-project.org/.

[28] Nguyen TL, Collins GS, Landais P, Le Manach Y. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials—An illustration with the International Stroke Trial. *Journal of Clinical Epidemiology* 2020; **125**:47–56. URL https://linkinghub.elsevier.com/retrieve/pii/S0895435620300445.

[29] Bouvier F, Peyrot E, Balendran A, Ségalas C, Roberts I, Petit F, Porcher R. Do machine learning methods lead to similar individualized treatment rules? a comparison study on real data. *Statistics in Medicine* 2024; .

[30] Bouvier F, Chaimani A, Peyrot E, Gueyffier F, Grenet G, Porcher R. Estimating individualized treatment effects using an individual participant data meta-analysis. *BMC Medical Research Methodology* 2024; **24**:74.

[31] Rekkas A, Rijnbeek PR, Kent DM, Steyerberg EW, Van Klaveren D. Estimating individualized treatment effects from randomized controlled trials: a simulation study to compare risk-based approaches. *BMC Medical Research Methodology* 2023; **23**:74.

Appendix A

c-statistic for benefit

We have:

$$C_{fb} = \frac{P\left(B_u > B_v \mid \tau_u > \tau_v\right) P\left(\tau_u > \tau_v\right)}{P\left(B_u > B_v\right)}$$

τ_u and τ_v are continuous and independent. Therefore $P\left(\tau_u > \tau_v\right)$ can be calculated as:

$$\begin{aligned}
P\left(\tau_u > \tau_v\right) &= \int_{-1}^{1} \int_{-t_j}^{1} f_\tau(t_i, t_j) dt_i dt_j \\
&= \int_{-1}^{1} \int_{t_j}^{1} f_{\tau_u}(t_i) f_{\tau_v}(t_j) dt_i dt_j \\
&= \int_{-1}^{1} f_{\tau_v}(t_j) \int_{t_j}^{1} f_{\tau_u}(t_i) dt_i dt_j \\
&= \int_{-1}^{1} f_{\tau_v}(t_j) \left(1 - F_{\tau_u}(t_j)\right) dt_j \\
&= \int_{-1}^{1} f_{\tau_v}(t_j) dt_j - \int_{-1}^{1} f_{\tau_u}(t_j) F_{\tau_u}(t_j) dt_j \\
&= 1 - \int_{-1}^{1} f_\tau(t_j) F_\tau(t_j) dt_j \\
&= 0.5
\end{aligned}$$

Where

$$f_\tau(t) = \frac{1}{B(a)} \int_s \int_u u^{\alpha_1 - 1} (t + s - u)^{\alpha_2 - 1} (s - u)^{\alpha_3 - 1} (1 - t - 2s + u)^{\alpha_u - 1} du ds$$

with $u \in \{\max(0, t + 2s - 1), \min(s, t + s)\}$, $s \in \{\max(0, -t), \min(1, 1 - t)\}$ and $t \in [-1, 1]$.

Recall that $B_u = Y_u^1 - Y_u^0$. Thus, we have:

$$\begin{aligned}
P\left(B_u > B_v\right) =& P\left(Y_u^1 - Y_u^0 > Y_v^1 - Y_v^0\right) \\
&= P\left(Y_u^1 = 1, Y_u^0 = 0\right) P\left(Y_v^1 = 1, Y_v^0 = 1\right) + P\left(Y_u^1 = 1, Y_u^0 = 0\right) P\left(Y_v^1 = 0, Y_v^0 = 0\right) \\
&+ P\left(Y_u^1 = 1, Y_u^0 = 0\right) P\left(Y_v^1 = 0, Y_v^0 = 1\right) + P\left(Y_u^1 = 1, Y_u^0 = 1\right) P\left(Y_v^1 = 0, Y_v^0 = 1\right) \\
&+ P\left(Y_u^1 = 0, Y_u^0 = 0\right) P\left(Y_v^1 = 0, Y_v^0 = 1\right)
\end{aligned}$$

$Y^t \sim \text{Bernoulli}\left(\Pi_t\right), t \in \{0, 1\}$ and $\Pi_t \sim \text{Beta}\left(\alpha_t, \beta_t\right), t \in \{0, 1\}$

Thus,

$$\begin{aligned}
P\left(Y^1 = 1, Y^0 = 1\right) &= \int_{[0,1]^2} P\left(Y^1 = 1, Y^0 = 1 \mid \pi_1, \pi_0\right) P\left(\pi_0, \pi_1\right) d\left(\pi_0, \pi_1\right) \\
&= \int_{[0,1]^2} P\left(Y^1 = 1 \mid \pi_1\right) P\left(Y^0 = 1 \mid \pi_0\right) f\left(\pi_0, \pi_1\right) d\left(\pi_0, \pi_1\right)
\end{aligned}$$

$$f_{\pi_0, \pi_1}(x, y) = \frac{1}{B(\alpha)} \int_\Omega u^{\alpha_1 - 1} (x - u)^{\alpha_2 - 1} (y - u)^{\alpha_3 - 1} (1 - x - y + u)^{\alpha_4 - 1} du$$

where $\Omega = (\max(0, x+y-1), \min(x, y))$, $P\left(Y^1 = k \mid \pi_1\right) = p_1^k + (1-p_1)\,(1-k)$ and $P\left(Y^0 = k \mid \pi_0\right) = p_0 k + (1-p_0)\,(1-k)$ with $k \in \{0, 1\}$

The last part $P\left(B_u > B_v \mid \tau_u > \tau_v\right)$ is obtained as follows:

$$P\left(B_u > B_v \mid \tau_u > \tau_v\right) = \sum_{\substack{k,s=-1 \\ k>s}}^{1} P\left(B_u = k \mid \tau_u > \tau_v\right) P\left(B_v = s \mid \tau_u > \tau_v\right)$$

(law of total probability + independence)

$$P\left(B_u = k \mid \tau_u > \tau_v\right) = 2P\left(\tau_u > \tau_v \mid B_u = k\right) P\left(B_u = k\right) \text{(Bayes formula)}$$

$$P\left(B_u = k\right) = \sum_{\substack{m,n=0 \\ m-n=k}}^{1} \int_{[0,1]^2} P\left(Y^1 = m \mid \pi_1, \pi_0\right) P\left(Y^0 = n \mid \pi_1, \pi_0\right) f\left(\pi_1, \pi_0\right) d\pi_1 \pi_0$$

$$P\left(\tau_u > \tau_v \mid B_u = k\right) = \int_{-1}^{1} f_{\tau \mid B_u = k}(t) F_{\tau_v}(t) dt$$

(law of total probability + independance)

$$f_{\tau \mid B_u = k} = \frac{P_{B_u \mid \tau_u}(k) f_\tau}{P\left(B_u = k\right)} \text{(Bayes formula)}$$

$$P\left(B_u = k \mid \tau\right) = \sum_{\substack{u,v=0 \\ u-v=k}}^{1} P\left(Y^1 = u, Y^0 = v \mid \tau_u\right) \text{(law of total probability)}$$

$$P\left(Y^1 = u, Y^0 = v \mid \tau\right) = \frac{f_{\tau \mid Y^1 = u, Y^0 = v} \cdot P\left(Y^1 = u, Y^0 = v\right)}{f_\tau}$$

$$f_{\tau \mid Y^1 = u, Y^0 = v}(t) = \int_0^1 f_{\tau, \pi_0 \mid Y^1 = u, Y^0 = v}(t, s) ds, \forall t \in [-1, 1]$$

$$f_{\tau, \pi_0 \mid Y^1 = u, Y^0 = v}(t, s) = f_{\pi_1, \pi_0 \mid Y^1 = u, Y^0 = v}(t + s, s), \quad \forall (t, s) \in [0, 1]^2$$

$$f_{\pi_1, \pi_0 \mid Y^1 = u, Y^0 = v} = \frac{P\left(Y^1 = u, Y^0 = v \mid \pi_1, \pi_0\right) \cdot f_{\pi_1, \pi_0}}{P\left(Y^1 = u, Y^0 = v\right)}$$

Hypothesis: Y^1 is independent from Y^0 given π_0 and π_1

$$P(Y^1 = u, Y^0 = v \mid \pi_1, \pi_0) = P(Y^1 = u \mid \pi_1) P(Y^0 = v \mid \pi_0)$$
$$P(Y^1 = u \mid \pi_1) = \pi_1^u (1 + \pi_1)^{1-u}$$
$$P(Y^0 = v \mid \pi_0) = \pi_0^v (1 + \pi_6)^{1-v}$$

Concentration of Benefit

$$C_b = 1 - \frac{\mathbb{E}(t_1)}{\mathbb{E}[\max(t_2, t_3)]}$$

$$\mathbb{E}(\tau) = \int_{-1}^{1} x f_\tau(t; \alpha_1, \alpha_2, \alpha_3, \alpha_4)\, dt.$$

where f_τ represents the probability density function of distribution τ.

We recall that

$$\mathbb{E}[\max(t_2, t_3)] = 2\int_{-1}^{1} t f_{\hat{\tau}}(t) F_{\tau}(t)\, dt$$

where F_τ is the cumulative distribution function of distribution τ.

PAPE

Let $r = \mathbb{1}_{\hat{\tau}>0}$ be an optimal rule, $\hat{\tau} = E(Y^1 - Y^0|X)$ be the estimated individualized treatment effects and $p_r = \mathbb{E}(r)$ be the proportion of patients for which treatment is recommended by r. Here, we consider a scenario where the individualized treatment effects are perfectly estimated, thus $\hat{\tau} = \tau$.

$$\text{PAPE} = E((r - pr)\tau$$

|PAPE|$\leq \text{Var}(r)$ in particular:

$$\begin{aligned}
|\text{PAPE}| &\leq \sqrt{E\left[(r-p_r)^2\right]}\sqrt{E\left[\tau^2\right]} \text{(Cauchy-Schwarz)} \\
&\leq \sqrt{E\left[(r-p_r)^2\right]} \text{ since } -1 \leq \tau \leq 1 \\
&\leq \sqrt{E\left[((r-E(r))^2\right]} \\
&= \sqrt{\text{Var}(r)} \\
&= \sqrt{p_r\left(1-p_r\right)}
\end{aligned}$$

$\sqrt{p_r\left(1-p_r\right)}$ reaches its maximal value at $p_r = \frac{1}{2}$ thus |PAPE| $\leq \frac{1}{2}$

Analytical expression

$$\begin{aligned}
\text{PAPE} &= \mathbb{E}[(r - \mathbb{E}(r))\tau] \\
&= \mathbb{E}[\mathbb{1}_{\{\hat{\tau}>0\}}\tau] - \mathbb{E}[\mathbb{1}_{\{\tau>0\}}]\mathbb{E}[\tau]
\end{aligned}$$

We obtain:

$$\text{PAPE} = \int_0^1 t f_\tau(t; \alpha_1, \alpha_2, \alpha_3, \alpha_4)\, dt - \int_0^1 f_\tau(t; \alpha_1, \alpha_2, \alpha_3, \alpha_4)\, dt \int_{-1}^1 t f_\tau(t; \alpha_1, \alpha_2, \alpha_3, \alpha_4)\, dt$$

where f_τ represents the probability density function of distribution τ.

Appendix B

Distribution	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
1	1	1	1	1	—	—	—	—
2	2	8	4	1	—	—	—	—
3	2	1	2	1	—	—	—	—
4	2	2	2	2	—	—	—	—
5	10	10	10	10	—	—	—	—
6	42	42	42	42	—	—	—	—
7	2	3	2	10	—	—	—	—
8	9	7	5	19	—	—	—	—
9	100	1	1	1	—	—	—	—
10	1	1	1	1	40	40	40	40
11	30	28	21	13	12	24	9	1
12	1	10	10	1	12	24	9	1
13	2	20	10	10	1	10	10	1
14	28	39	16	38	35	8	19	21
15	17	4	25	15	26	32	9	33
16	30	3	6	36	24	25	12	13
17	1	1	40	1	1	40	1	1
18	1	1	100	1	1	100	1	1
19	50	1	10	1	10	10	10	10
20	36	16	9	33	31	16	24	5

Chapitre 5

Discussion générale

5.1 Résumé des travaux

5.1.1 Utilisation d'une méta-analyse pour estimer les effets individualisés du traitement

Dans le premier projet de cette thèse, nous avons exploré le potentiel des méta-analyses sur données individuelles dans l'estimation des effets individualisés du traitement.

La MADI utilise des données provenant de plusieurs études, offrant ainsi un ensemble de données qui améliore la puissance statistique et la généralisation des estimations de l'effet du traitement. Cette approche est particulièrement prometteuse parce qu'elle répond au problème courant de la sous-puissance des études utilisées en médecine personnalisée, qui ne parviennent souvent pas à correctement estimer les ITE lorsqu'elles utilisent les données d'un seul essai clinique randomisé. De plus, la MADI permet d'explorer l'hétérogénéité entre les études, ce qui peut être crucial pour adapter les traitements aux caractéristiques individuelles.

Dans ce travail, nous avons comparé différents modèles statistiques pour estimer les ITE à partir d'une MADI en utilisant à la fois des données simulées et des données réelles. L'utilisation du S-learner, un modèle incorporant des termes d'interaction entre le traitement et les variables, s'est avérée efficace pour gérer les complexités de l'estimation de l'ITE à travers divers types de critères de jugement. Aucune méthode statistique n'a été systématiquement plus performante que les autres dans tous les scénarios, ce qui suggère que le choix de la

méthode peut devoir être adapté en fonction des caractéristiques et des résultats spécifiques de l'étude. L'étude a démontré que si la MADI pouvait améliorer l'estimation des ITE, l'application pratique en termes d'amélioration des décisions de traitement n'était pas toujours évidente, comme vu dans l'application aux données INDANA.

Certaines limites à l'utilisation d'une MADI ressortent de ce travail. L'un des principaux défis de la MADI est la gestion de l'hétérogénéité entre les différentes études incluses dans la méta-analyse. Dans le cas étudié, les différents modèles n'ont pas tous réussi à prendre en compte cette hétérogénéité, ce qui peut avoir un impact sur la qualité des estimations de l'ITE. Il existe un risque de sur-apprentissage, en particulier avec les modèles qui tiennent compte d'un degré élevé de variabilité entre les études, comme le modèle entièrement stratifié. La mise en oeuvre des modèles MADI est complexe, car elle fait appel à des techniques statistiques sophistiquées, en utilisant notamment des modèles à effets mixtes, et nécessite d'importantes ressources computationnelles.

Les méta-analyses sur données individuelles ont l'avantage de permettre l'accès à un plus grand nombre de données ce qui entraîne une estimation plus précise des effets individualisés du traitement. Cependant, obtenir les données individuelles de plusieurs études est chronophage et demande une logistique complexe. Une alternative aux méta-analyses pourrait être l'utilisation de l'apprentissage fédéré.

L'apprentissage fédéré est un paradigme d'apprentissage automatique qui permet à plusieurs parties de former un modèle en collaboration sans échanger leurs données sous-jacentes [87].

Contrairement aux modèles traditionnels d'apprentissage automatique centralisé qui nécessitent l'agrégation de données provenant de diverses sources, l'apprentissage fédéré permet à chaque participant d'entraîner un modèle local sur ses données privées, puis de partager les paramètres du modèle avec un serveur central, qui met ensuite à jour les paramètres pour produire un modèle global.

Cette approche permet de relever d'importants défis en matière de protection de la vie privée et de gouvernance des données. Des recherches ont montré que l'apprentissage fédéré peut atteindre des niveaux de performance comparables à ceux des modèles centralisés traditionnels, et même surpasser les modèles formés sur des ensembles de données isolés en raison de la représentation plus large des données.

L'apprentissage fédéré pourrait offrir une alternative prometteuse à la méta-analyse sur données individuelles en améliorant la confidentialité, la conformité, et la représentativité tout en réduisant les contraintes techniques et logistiques associées à l'agrégation centralisée des données. L'apprentissage fédéré est donc particulièrement adapté dans le domaine de la santé où la confidentialité et la sécurité des données sont primordiales.

Des modèles de régression logistique et régression de Cox ont récemment été adaptés au cadre de l'apprentissage fédéré [66, 67]. Adapter ces modèles à l'estimation des effets individualisés du traitement pourrait être un futur projet de recherche.

5.1.2 Comparaison de règles individualisées de traitement développées à partir de méthodes d'apprentissage automatique

Dans le deuxième projet de cette thèse, nous avons examiné dans quelle mesure diverses méthodes d'apprentissage automatique produisent des règles de traitement individualisées similaires lorsqu'elles sont appliquées à des données réelles provenant d'essais cliniques randomisés. L'objectif principal était d'évaluer si ces méthodes recommandent le même traitement pour les mêmes individus. Des métriques de performance et d'agrément ont été utilisées pour comparer les ITRs.

Nous avons comparé 22 approches d'apprentissage automatique pour élaborer des ITR à l'aide de données provenant de deux essais cliniques randomisés. Nous avons classé ces méthodes en deux catégories :

— Les méthodes dérivant les ITR à partir des effets individualisés du traitement, où un bénéfice de traitement est prédit puis utilisé pour décider du traitement

— Les méthodes estimant les ITR directement sans calculer les ITE

Nous avons trouvé que la plupart des ITR issues de différentes méthodes étaient en désaccord considérable sur les recommandations de traitement des patients, ce qui suggère que le choix de la méthode peut influencer de manière significative les décisions cliniques. Les méthodes similaires sur le plan conceptuel (par exemple, celles qui utilisent des modèles statistiques ou des hypothèses similaires) ont montré une plus grande concordance dans leurs recommandations de traitement. D'autre part, en général, les ITR ont démontré une performance limitée, indiquant une marge potentielle d'amélioration et d'optimisation des méthodes en vue d'une application dans le monde réel.

La performance des ITR dépend fortement de la qualité et des caractéristiques des données sous-jacentes, qui peuvent varier considérablement.

En combinant les données des essais CRASH-2 et CRASH-3, nous avons constaté que l'agrément entre les règles développées par les différentes méthodes ne s'est pas amélioré, malgré l'accès à un plus grand échantillon de données et une légère augmentation de l'hétérogénéité de l'effet traitement. Il serait pertinent de mener une étude de simulation pour déterminer si, à partir d'un certain niveau d'hétérogénéité de l'effet du traitement, les méthodes produisent des règles de traitement personnalisé similaires, recommandant ainsi le traitement à des patients similaires. De plus, il serait intéressant d'examiner l'impact du nombre de variables incluses dans les méthodes sur l'agrément des règles. On pourrait s'attendre à un accord plus élevé lorsque peu de variables sont incluses dans le modèle.

5.1.3 Capacité discriminative maximale selon le niveau d'hétérogénéité

Dans le troisième projet de cette thèse, nous avons calculé la discrimination maximale pouvant être obtenue pour différentes distributions d'effets de traitement afin de déterminer l'influence de l'hétérogénéité sur les capacités de discrimination des modèles d'effets de traitement individualisés.

Nous avons utilisé trois métriques spécifiques pour évaluer la discrimination : la *c-statistic for benefit*, la *concentration of benefit*, et le *population average prescription effect* (PAPE). 20 distributions différentes des effets du traitement ont été générées, chacune conçue pour refléter des degrés variables d'hétérogénéité, afin d'observer comment ces métriques se comportent dans différentes conditions. Ces distributions ont été générées à l'aide de distributions Beta dépendantes pour refléter les scénarios du monde réel dans lesquels les caractéristiques individuelles des patients peuvent influencer les résultats du traitement.

Les résultats ont révélé que les performances discriminantes des modèles d'effets individualisés du traitement varient considérablement en fonction de la distribution sous-jacente des effets du traitement. Par exemple, les distributions très hétérogènes dans lesquelles les individus ont un bénéfice important ou un non-bénéfice important ont obtenu des scores de discrimination élevés pour toutes les métriques. En revanche, les distributions plus homogènes avec une variation minimale des effets du traitement ont donné lieu à une discrimination maximale limitée. Cela indique que le niveau d'hétérogénéité a un impact important sur l'efficacité des modèles d'effets de traitement individualisés à prédire quels patients bénéficieront du traitement.

Dans ce projet, nous avons étudié l'impact du niveau d'hétérogénéité dans la distribution des effets du traitement sur la discrimination. Une autre composante importante de la performance d'un modèle ITE est la calibration. Il serait intéressant d'examiner si le niveau d'hétérogénéité de l'effet du traitement influence également de manière significative la calibration maximale pouvant être obtenue.

5.2 Réflexions sur l'utilisation d'essais cliniques randomisés pour développer des règles de traitement individualisé

Les essais cliniques randomisés (ECR) sont largement considérés comme la référence en matière d'évaluation de l'efficacité des traitements et des interventions. Cependant, l'estima-

tion des effets individualisés du traitement à l'aide des données d'essais contrôlés randomisés peut être une tâche complexe et difficile, comme cela a été discuté tout au long de cette thèse. Les essais cliniques randomisés ont généralement des critères d'inclusion stricts qui limitent l'hétérogénéité qui peut être observée. De plus, la population incluse dans un ECR particulier ne représente pas toujours avec précision la population d'intérêt [88, 89, 90]. Les modèles de régression et les méthodes analytiques plus complexes, tout en offrant une plus grande flexibilité pour capturer les effets individualisés, nécessitent des échantillons de plus grande taille et un examen attentif des hypothèses du modèle et des sources potentielles de biais [4].

5.2.1 Utilisation d'autres méthodes

La modélisation de l'effet du traitement permet de développer des règles de traitement personnalisé détaillées. Cependant, cette approche est sensible au sur-apprentissage et nécessite une grande hétérogénéité ainsi qu'un large échantillon pour estimer correctement les effets individualisés du traitement. L'utilisation d'autres méthodes, telles que la modélisation du risque, qui demandent moins de données mais permettent uniquement une stratification large des patients, pourrait être une meilleure solution pour développer des règles de traitement personnalisé à partir de données d'essais cliniques randomisés.

De nouvelles méthodes pour développer des règles de traitement personnalisé ont récemment été proposées, comme l'*Effect Score Analysis* que nous présentons dans la sous-section suivante.

5.2.1.1 Effect Score Analysis

L'*Effect Score Analysis* consiste à créer un modèle qui attribue à chaque patient un score d'effet indiquant le bénéfice attendu d'un traitement [91]. Ce processus comprend plusieurs étapes :

1. Le développement d'un modèle à l'aide d'un ensemble de données pour prédire les effets du traitement en fonction des caractéristiques du patient
2. L'application de ce modèle à un autre ensemble de données pour générer des scores d'effet

3. Le regroupement des patients en strates sur la base de leurs scores
4. L'évaluation des effets du traitement au sein de ces strates afin d'évaluer l'hétérogénéité

Cette méthode permet d'estimer de manière précise et plus stable l'hétérogénéité du traitement par rapport aux méthodes de *machine learning*. Cette méthode peut être vue comme un mélange entre la modélisation du risque et la modélisation de l'effet traitement, en empruntant le cadre de la modélisation du risque pour estimer les effets traitements au niveau de strates de patients.

Il serait intéressant de comparer la performance de l'*Effect Score Analysis* à celle des modèles ITE et de voir si elle parvient à capturer correctement l'hétérogénéité de l'effet traitement.

5.2.2 Identifier a priori la possibilité de développer une règle de traitement individualisé

L'hétérogénéité limitée découlant des critères d'inclusion stricts est l'une des raisons pour lesquelles il est compliqué d'élaborer des règles de traitement individualisé bénéfiques, c'est-à-dire des règles de traitement permettant d'obtenir de meilleurs résultats qu'une règle de traitement classique qui recommande le traitement soit à tout le monde soit à personne. Une étude récente a proposé une méthode pour identifier les ITR bénéfiques qui peuvent être développées en utilisant les données des ECR [92]. La méthode λ_q-LASSO a été introduite pour sélectionner le paramètre de pénalité LASSO permettant de limiter le développement d'ITR non bénéfiques. L'identification d'ITR bénéfiques requiert un certain niveau d'hétérogénéité, notamment, il faut suffisamment de personnes bénéficiant du traitement expérimental ainsi que de personnes bénéficiant du traitement contrôle.

De plus, l'étude a souligné l'importance de développer un ITR en utilisant des variables avec des preuves préalables d'un effet hétérogène du traitement, condition *sine qua non* à la possibilité de développer des ITR efficaces.

Cette étude met également l'accent sur le calcul de la probabilité a priori de développer une ITR bénéfique, permettant aux chercheurs d'éviter le développement d'ITR dont les perfor-

mances sont inférieures à celles de la règle classique optimale.

L'adaptation de cette méthode à la méta-analyse pourrait être une perspective intéressante.

5.2.3 Personomics

Les *personomics*, concept introduit par Roy Ziegelstein, mettent en évidence l'importance de prendre en compte les facteurs psychologiques, sociaux, culturels, comportementaux et économiques qui influencent la santé et les résultats des traitements d'un individu [93].

Les *personomics* aborde les limites de la médecine de précision et de la médecine personnalisée, qui se concentre principalement sur les marqueurs génomiques, protéomiques, et autres marqueurs biologiques. Si ces "-omiques" fournissent des indications précieuses sur les fondements moléculaires de la maladie, elles négligent souvent le rôle essentiel des circonstances de la vie individuelle [94]. Comprendre le contexte personnel et social d'un patient pourrait conduire à des interventions de santé plus efficaces et mieux adaptées [95].

Cette approche vise à garantir que les traitements sont non seulement biologiquement appropriés, mais aussi réalisables dans la pratique et culturellement acceptables pour le patient. Par exemple, des facteurs tels que les connaissances en matière de santé ou les contraintes économiques d'un patient pourraient avoir un impact sur sa capacité à adhérer aux traitements prescrits et à en tirer bénéfice.

L'intégration des *personomics* dans les essais cliniques randomisés pourrait potentiellement faciliter le développement d'ITR plus nuancées et plus efficaces. En tenant compte à la fois des données biologiques et des données *personomics*, les cliniciens pourraient concevoir des stratégies de traitement véritablement personnalisées qui prennent en compte l'ensemble des facteurs affectant les résultats des patients. Par exemple, un patient bénéficiant d'un niveau élevé de soutien social pourrait bénéficier d'une approche thérapeutique différente de celle d'un patient bénéficiant d'un soutien limité, même si leurs marqueurs biologiques sont similaires. Cependant, la manière d'intégrer ces variables dans la conception des essais n'est pas

tout à fait claire. Plusieurs adaptations méthodologiques sont nécessaires, comme l'évaluation de base des variables ou la randomisation des patients.

Par ailleurs, les données *personomics* peuvent aider à identifier les patients qui pourraient avoir besoin de ressources supplémentaires ou d'interventions sur mesure pour obtenir des résultats optimaux. Par exemple, les patients ayant un faible niveau de connaissances en matière de santé pourraient bénéficier d'interventions éducatives plus intensives dans le cadre de leur stratégie de traitement.

Des études supplémentaires sont nécessaires pour déterminer si les variables *personomics* peuvent modifier l'effet du traitement et, par conséquent, être intégrées dans l'élaboration de règles de traitement personnalisé.

5.3 Développement de règles de traitement personnalisé à partir de données observationnelles

L'utilisation des méthodes modélisant l'effet du traitement nécessitent des jeux de données larges pour pouvoir estimer correctement les effets individualisés du traitement et développer des règles de traitement individualisé efficaces. Les essais clinique randomisés ont généralement des tailles d'échantillon limitées en raison du coût et de la logistique intenses. L'utilisation d'une méta-analyse peut pallier à ce problème. En revanche, de part les critères d'inclusion stricts que les essais cliniques possèdent, l'hétérogénéité de l'effet du traitement est limitée. Nous avons vu au cours de cette thèse qu'il est nécessaire d'avoir suffisamment d'hétérogénéité pour pouvoir prédire efficacement les effets individualisés du traitement et développer des règles de traitement individualisés bénéfiques. Une solution pour avoir accès à plus d'hétérogénéité serait d'utiliser des données observationnelles issues de cohortes, de registres ou bien de dossiers médicaux électroniques. Dans les études observationnelles, l'affectation du traitement n'est pas contrôlée par les investigateurs, mais suit simplement les décisions individuelles des médecins et des patients. Différentes sources de données peuvent être utilisées. Étant donné que ces données reflètent la pratique habituelle des soins, les données observationnelles sont souvent appelées "données du monde réel" (*real world evidence*).

L'utilisation de données observationnelles permettent à la fois l'accès à de plus grands jeux de données mais aussi à plus de variables et donc potentiellement à des variables modificatrices de l'effet du traitement qui ne sont pas mesurées au cours d'un essai clinique randomisé. De plus, les données observationnelles ne sont pas soumises à des critères de sélection stricts ce qui permet d'avoir accès aux données de patients représentant la population générale.

Les études observationnelles collectent des données sans intervention du chercheur, reflétant ainsi le monde réel. Cependant, ces études sont souvent sujettes à des facteurs de confusion, car l'affectation des traitements n'est pas randomisée. Cela signifie que les patients recevant des traitements différents peuvent présenter des différences importantes, parfois non mesurées, qui affectent les critères de jugement. Pour relever ces défis et estimer les effets du traitement à partir de données d'observation, les chercheurs utilisent des méthodes statistiques en se basant sur certaines hypothèses causales [27, 26] : cohérence, absence d'interférence, positivité et l'absence de confusion non mesurée.

Diverses techniques, telles que la régression et les méthodes de score de propension, peuvent répondre à ces questions [96, 97, 98, 99]. Les scores de propension estiment la probabilité qu'un patient reçoive un traitement particulier en fonction des caractéristiques observées, en équilibrant les groupes de traitement pour les rendre plus comparables. Ces techniques sont décrites dans l'article "*What should be done and what should be avoided when comparing two treatments*?" [100], disponible en annexe.

L'émulation d'essais est une méthode qui consiste à utiliser des données observationnelles pour imiter un essai contrôlé randomisé [101, 102]. Cette approche est utile lorsque les essais cliniques randomisés ne sont pas faisable en pratique ou lorsqu'il y a un problème d'éthique. L'émulation d'un essai consiste à spécifier les éléments d'un ECR hypothétique, notamment la population, les traitements, les résultats et le moment de la collecte des données [103]. En reproduisant un ECR, les chercheurs visent à réduire les biais qui affectent généralement les études observationnelles, tels que les biais de sélection et les biais liés au temps.

Un aspect essentiel de l'essai émulé consiste à gérer correctement le calendrier des décisions thérapeutiques, surtout lorsque le traitement n'est pas attribué à un moment précis. Pour cela, des stratégies comme le clonage et la censure sont utilisées. Chaque participant est alors hypothétiquement dupliqué : l'un reçoit le traitement et l'autre non. Ils sont ensuite suivis dans le temps pour observer les résultats selon le traitement qui leur a été attribué.

Les études observationnelles, lorsqu'elles sont soigneusement conçues et appliquées avec des méthodes telles que l'ajustement par régression, l'appariement des scores de propension et, en particulier, l'émulation d'essais, peuvent fournir des informations précieuses sur les effets des traitements. Ces méthodes permettent aux chercheurs de se rapprocher des conditions d'un ECR et de produire des estimations utiles pour guider les décisions cliniques.

Conclusion générale

L'objectif de cette thèse était d'étudier le développement de règles de traitement personnalisé en estimant les effets individualisés du traitement à partir de données issues d'un ou de plusieurs essais cliniques randomisés.

Dans le premier projet, nous avons estimé les effets individualisés du traitement en utilisant une méta-analyse sur données individuelles. Les méta-analyses sur données individuelles ont l'avantage de permettre l'accès à un plus grand nombre de données. En revanche cette approche nécessite une grande hétérogénéité de l'effet du traitement pour correctement estimer les effets individualisés du traitement et ainsi développer des règles de traitement individualisé efficaces.

Dans le deuxième projet, nous avons comparé différentes méthodes d'apprentissage automatique pour développer des règles de traitement personnalisé. Les résultats ont révélé une variabilité considérable dans les recommandations de traitement selon les méthodes utilisées, soulignant une limitation pour leur utilisation en pratique.

Dans le troisième projet, nous avons étudié la discrimination maximale pouvant être obtenue pour diverses distributions d'effets de traitement. Nos résultats indiquent que des distributions très hétérogènes permettent une meilleure discrimination, tandis que des distributions plus homogènes limitent cette capacité.

Les résultats de cette thèse apportent des éléments de réflexion importants pour le développement de règles de traitement personnalisé à partir de données d'essais cliniques randomisés. Bien que de nombreuses méthodes aient été proposées ces dernières années pour estimer les

effets individualisés du traitement et développer des règles de traitement personnalisé, il reste encore beaucoup de recherches à effectuer pour identifier l'hétérogénéité de l'effet du traitement présente dans les données et identifier les variables modifiant cet effet.

L'utilisation de données observationnelles offre des perspectives intéressantes pour surmonter certaines limitations des essais cliniques randomisés, en permettant d'accéder à une plus grande hétérogénéité dans l'effet du traitement. Ces approches pourraient aider à mieux estimer les effets individualisés du traitement et à développer des règles de traitement plus bénéfiques.

Bibliographie

[1] M Hassan Murad et al. “New evidence pyramid”. In : *BMJ Evidence-Based Medicine* 21.4 (2016), p. 125-127. doi : 10.1136/ebmed-2016-110401.

[2] Paul P Glasziou et Les M Irwig. “An evidence based approach to individualising treatment”. In : *BMJ* 311.7016 (1995), p. 1356-1359. doi : 10.1136/bmj.311.7016.1356.

[3] Rodney Hayward et al. “Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis”. In : *BMC medical research methodology* 6 (2006), p. 18. doi : 10.1186/1471-2288-6-18.

[4] David M. Kent et Rodney A. Hayward. “Limitations of Applying Summary Results of Clinical Trials to Individual PatientsThe Need for Risk Stratification”. In : *JAMA* 298.10 (2007), p. 1209-1212. doi : 10.1001/jama.298.10.1209.

[5] P.M. Rothwell. “Can overall results of clinical trials be applied to all patients ?” In : *The Lancet* 345.8965 (1995), p. 1616-1619. doi : https://doi.org/10.1016/S0140-6736(95)90120-5.

[6] Andrew J Vickers, Michael W Kattan et Daniel J Sargent. “Method for evaluating prediction models that apply the results of randomized trials to individual patients”. In : *Trials* 8.1 (2007), p. 14. doi : 10.1186/1745-6215-8-14.

[7] Richard Kravitz, Naihua Duan et Joel Braslow. “Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages”. In : *The Milbank quarterly* 82 (2004), p. 661-87. doi : 10.1111/j.0887-378X.2004.00327.x.

[8] Charles Warlow. “MRC European Carotid Surgery Trial : interim results for symptomatic patients with severe (70-99%) or with mild (0-29%) carotid stenosis”. In : *The Lancet* 337.8752 (1991), p. 1235-1243. doi : 10.1016/0140-6736(91)92916-P.

[9] B Farrell et al. “The United Kingdom transient ischaemic attack (UK-TIA) aspirin trial : final results.” In : *Journal of neurology, neurosurgery, and psychiatry* 54.12 (1991), p. 1044. doi : 10.1136/jnnp.54.12.1044.

[10] Ravi Varadhan et al. “A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research”. In : *Journal of Clinical Epidemiology* 66.8 (2013), p. 818-825. doi : https://doi.org/10.1016/j.jclinepi.2013.02.009.

[11] Ravi Varadhan et John D. Seeger. “Estimation and Reporting of Heterogeneity of Treatment Effects”. In : 2013. url : https://api.semanticscholar.org/CorpusID:9040373.

[12] Thorkild IA Sørensen. “Which patients may be harmed by good treatments?” In : *The Lancet* 348.9024 (1996), p. 351-352. doi : https://doi.org/10.1016/S0140-6736(05)64988-4.

[13] Nicholas T. Longford. “Selection bias and treatment heterogeneity in clinical trials”. In : *Statistics in Medicine* 18.12 (1999), p. 1467-1474. doi : https://doi.org/10.1002/(SICI)1097-0258(19990630)18:12<1467::AID-SIM149>3.0.CO;2-H.

[14] Bénédicte Colnet et al. *Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize?* 2023. arXiv : 2303.16008.

[15] Rui Wang et al. “Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials”. In : *New England Journal of Medicine* 357.21 (2007), p. 2189-2194. doi : 10.1056/NEJMsr077003.

[16] Rodney A. Hayward et al. “Reporting Clinical Trial Results To Inform Providers, Payers, And Consumers”. In : *Health Affairs* 24.6 (2005), p. 1571-1581. doi : 10.1377/hlthaff.24.6.1571.

[17] Peter M Rothwell. “Subgroup analysis in randomised controlled trials : importance, indications, and interpretation”. In : *The Lancet* 365.9454 (2005), p. 176-186. doi : https://doi.org/10.1016/S0140-6736(05)17709-5.

[18] David Kent et al. “Assessing and reporting heterogeneity in treatment effects in clinical trials : a proposal.” In : *Trials* 11 (2010), p. 85. doi : 10.1186/1745-6215-11-85.

[19] David M Kent. “Overall average treatment effects from clinical trials, one-variable-at-a-time subgroup analyses and predictive approaches to heterogeneous treatment effects : Toward a more patient-centered evidence-based medicine”. In : *Clinical Trials* 20.4 (2023), p. 328-337. doi : 10.1177/17407745231171897.

[20] Sarah T Brookes et al. “Subgroup analysis in randomised controlled trials : quantifying the risks of false-positives and false-negatives”. In : (2001). doi : 10.3310/hta5330.

[21] David M Kent, Ewout Steyerberg et David Van Klaveren. “Personalized evidence based medicine : predictive approaches to heterogeneous treatment effects”. In : *BMJ* (2018), k4245. doi : 10.1136/bmj.k4245.

[22] Els Goetghebeur et al. “Formulating causal questions and principled statistical answers”. In : *Statistics in Medicine* 39.30 (2020), p. 4922-4948. doi : https://doi.org/10.1002/sim.8741.

[23] Philip Sedgwick. “What is an “n-of-1” trial?” In : *BMJ* 348 (2014). doi : 10.1136/bmj.g2674. eprint : https://www.bmj.com/content/348/bmj.g2674.full.pdf. url : https://www.bmj.com/content/348/bmj.g2674.

[24] Elizabeth O Lillie et al. “The N-Of-1 Clinical Trial : The Ultimate Strategy For Individualizing Medicine ?” In : *Personalized Medicine* 8.2 (2011), p. 161-173. doi : 10.2217/pme.11.7.

[25] RD Mirza et al. “The history and development of N-of-1 trials”. In : *Journal of the Royal Society of Medicine* 110.8 (2017), p. 330-340. doi : 10.1177/0141076817721131.

[26] Jerzy Neyman. “On the application of probability theory to agricultural experiments. Essay on Principles. Section 9 (translation published in 1990)”. In : *Statistical Science* 5 (1923), p. 472-480. doi : 10.1214/ss/1177012031.

[27] D. B. Rubin. “Estimating causal effects of treatments in randomized and nonrandomized studies”. In : *Journal of Educational Psychology* 6 (1974), p. 688-701. doi : 10.1037/h0037350.

[28] D. B. Rubin. “Which ifs have causal answers ? Comment on : “Statistics and causal inference” by P. Holland”. In : *Journal of the American Statistical Association* 81 (1986), p. 961-962. doi : 10.1080/01621459.1986.10478355.

[29] Behram Hansotia et Brad Rukstales. “Incremental value modeling”. In : *Journal of Interactive Marketing* 16.3 (2002), p. 35-46. doi : 10.1002/dir.10035.

[30] Jennifer L Hill. “Bayesian nonparametric modeling for causal inference”. In : *Journal of Computational and Graphical Statistics* 20.1 (2011), p. 217-240. doi : 10.1198/jcgs.2010.08162.

[31] S. R. Künzel et al. “Metalearners for estimating heterogeneous treatment effects using machine learning”. In : *Proceedings of the National Acadademy of Sciences U S A* 116.10 (2019), p. 4156-4165. doi : 10.1073/pnas.1804597116.

[32] Edward H. Kennedy. *Optimal doubly robust estimation of heterogeneous causal effects*. 2020. doi : 10.48550/ARXIV.2004.14497.

[33] X Nie et S Wager. “Quasi-oracle estimation of heterogeneous treatment effects”. In : *Biometrika* 108.2 (2020), p. 299-319. doi : 10.1093/biomet/asaa076.

[34] Susan Athey, Julie Tibshirani et Stefan Wager. *Generalized Random Forests*. 2018.

[35] Scott Powers et al. “Some methods for heterogeneous treatment effect estimation in high dimensions”. In : *Statistics in medicine* 37.11 (2018), p. 1767-1787.

[36] P Richard Hahn, Jared S Murray et Carlos M Carvalho. “Bayesian regression tree models for causal inference : Regularization, confounding, and heterogeneous effects (with discussion)”. In : *Bayesian Analysis* 15.3 (2020), p. 965-1056.

[37] J.C. Foster, J.M.G. Taylor et S.J. Ruberg. “Subgroup identification from randomized clinical trial data”. In : *Statistics in Medicine* 30 (2011), p. 2867-2880. doi : 10.1002/sim.4322.

[38] S. Chen et al. “A general statistical framework for subgroup identification and comparative treatment scoring”. In : *Biometrics* 73.4 (2017), p. 1199-1209. doi : 10.1111/biom.12676.

[39] L. Tian, L. Zhao et L.J. Wei. “Predicting the restricted mean event time with the subject’s baseline covariates in survival analysis”. In : *Biostatistics* 15 (2014), p. 222-233. doi : 10.1093/biostatistics/kxt050.

[40] Frank Harrell. *Regression Modeling Strategies : With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2015. isbn : 978-3-319-19424-0. doi : 10.1007/978-3-319-19425-7.

[41] Ewout Steyerberg. *Clinical Prediction Models : A Practical Approach to Development, Validation, and Updating*. T. 19. 2009. isbn : 978-0-387-77243-1. doi : 10.1007/978-0-387-77244-8.

[42] D. van Klaveren et al. "The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects". In : *Journal of Clinical Epidemiology* 94 (2018), p. 59-68. doi : 10.1016/j.jclinepi.2017.10.021.

[43] Mohsen Sadatsafavi, Mohammad Ali Mansournia et Paul Gustafson. "A threshold-free summary index for quantifying the capacity of covariates to yield efficient treatment rules". In : *Statistics in Medicine* 39.9 (2020), p. 1362-1373. doi : https://doi.org/10.1002/sim.8481.

[44] D. van Klaveren et al. "Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting". In : *Journal of Clinical Epidemiology* 114 (2019), p. 72-83. doi : 10.1016/j.jclinepi.2019.05.029.

[45] B. Zhang et al. "A robust method for estimating optimal treatment regimes". In : *Biometrics* 68.4 (2012), p. 1010-1018. doi : 10.1111/j.1541-0420.2012.01763.x.

[46] Y. Zhao et al. "Estimating Individualized Treatment Rules Using Outcome Weighted Learning". In : *Journal of the American Statistical Association* 107.449 (2012), p. 1106-1118. doi : 10.1080/01621459.2012.695674.

[47] Xiaohan Guo et Ai Ni. "Contrast weighted learning for robust optimal treatment rule estimation". In : *Statistics in Medicine* (2022), sim.9574. doi : 10.1002/sim.9574.

[48] V. Farooq et al. "Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients : development and validation of SYNTAX score II". In : *Lancet* 381.9867 (2013), p. 639-650. doi : 10.1016/S0140-6736(13)60108-7.

[49] T. L. Nguyen et al. "Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials-An illustration with the International Stroke Trial". In : *Journal of Clinical Epidemiology* 125 (2020), p. 47-56. doi : 10.1016/j.jclinepi.2020.05.022.

[50] Junyi Zhou, Ying Zhang et Wanzhu Tu. "A reference-free R-learner for treatment recommendation". In : *Statistical Methods in Medical Research* 32.2 (2023), p. 404-424. doi : 10.1177/09622802221144326. url : https://doi.org/10.1177/09622802221144326.

[51] Kosuke Imai et Michael Lingzhi Li. "Experimental Evaluation of Individualized Treatment Rules". In : *Journal of the American Statistical Association* 118.541 (2021), p. 242-256. doi : 10.1080/01621459.2021.1923511.

[52] H. Janes et al. "An approach to evaluating and comparing biomarkers for patient treatment selection". In : *International Journal of Biostatistics* 10.1 (2014), p. 99-121. doi : 10.1515/ijb-2012-0052.

[53] R. D. Riley et al. “Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates : Statistical recommendations for conduct and planning”. In : *Statistics in Medicine* 39.15 (2020), p. 2115-2137. doi : 10.1002/sim.8516.

[54] T. P. Debray et al. “A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis”. In : *Statistics in Medicine* 32.18 (2013), p. 3158-3180. doi : 10.1002/sim.5732.

[55] T. P. Debray et al. “Get real in individual participant data (IPD) meta-analysis : a review of the methodology”. In : *Research Synthesis Methods* 6.4 (2015), p. 293-309. doi : 10.1002/jrsm.1160.

[56] E. W. Steyerberg et al. “Assessment of heterogeneity in an individual participant data meta-analysis of prediction models : An overview and illustration”. In : *Statistics in Medicine* 38.22 (2019), p. 4290-4309. doi : 10.1002/sim.8296.

[57] D. J. Fisher et al. “Meta-analytical methods to identify who benefits most from treatments : daft, deluded, or deft approach ?” In : *BMJ* 356 (2017), j573. doi : 10.1136/bmj.j573.

[58] Konstantina Chalkou et al. “A two-stage prediction model for heterogeneous effects of treatments”. In : *Statistics in Medicine* (2021), sim.9034. doi : 10.1002/sim.9034.

[59] Patrick Royston, Mahesh K. B. Parmar et Richard Sylvester. “Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer”. In : *Statistics in Medicine* 23.6 (2004), p. 907-926. doi : https://doi.org/10.1002/sim.1691.

[60] E. W. Steyerberg et al. “Prognosis Research Strategy (PROGRESS) 3 : prognostic model research”. In : *PLoS Medicine* 10.2 (2013), e1001381. doi : 10.1371/journal.pmed.1001381.

[61] Hyung Park et al. “A Single-Index Model With a Surface-Link for Optimizing Individualized Dose Rules”. In : *Journal of Computational and Graphical Statistics* 31.2 (2022), p. 553-562. doi : 10.1080/10618600.2021.1923521.

[62] Y. Q. Zhao et al. “Doubly Robust Learning for Estimating Individualized Treatment with Censored Data”. In : *Biometrika* 102.1 (2015), p. 151-168. doi : 10.1093/biomet/asu050.

[63] Weibin Mo, Zhengling Qi et Yufeng Liu. “Learning Optimal Distributionally Robust Individualized Treatment Rules”. In : *Journal of the American Statistical Association* 116.534 (2021), p. 659-674. doi : 10.1080/01621459.2020.1796359.

[64] Ying-Qi Zhao et al. “Robustifying trial-derived optimal treatment rules for a target population”. In : *Electronic Journal of Statistics* 13.1 (2019), p. 1717-1743. doi : 10.1214/19-EJS1540.

[65] Dipesh Mistry, Nigel Stallard et Martin Underwood. “A recursive partitioning approach for subgroup identification in individual patient data meta-analysis”. In : *Statistics in Medicine* 37.9 (2018), p. 1550-1561. doi : 10.1002/sim.7609.

[66] Y. Wu et al. “Grid Binary LOgistic REgression (GLORE) : building shared models without sharing data”. In : *Journal of the American Medical Informatics Association* 19.5 (2012), p. 758-764. DOI : 10.1136/amiajnl-2012-000862.

[67] Chia-Lun Lu et al. “WebDISCO : a web service for distributed cox model learning without patient-level data sharing”. In : *Journal of the American Medical Informatics Association* 22.6 (2015), p. 1212-1219. DOI : 10.1093/jamia/ocv083.

[68] Daniel Jacob. *CATE meets ML – The Conditional Average Treatment Effect and Machine Learning*. 2021.

[69] Weijia Zhang, Jiuyong Li et Lin Liu. “A Unified Survey of Treatment Effect Heterogeneity Modelling and Uplift Modelling”. In : *ACM Computing Surveys* 54.8 (2021). DOI : 10.1145/3466818.

[70] the International Stroke Trial Collaborative Group et al. “The International Stroke Trial database”. In : *Trials* 12.1 (2011), p. 101. DOI : 10.1186/1745-6215-12-101.

[71] The CRASH-3 trial collaborators. “Effects of tranexamic acid on death, disability, vascular occlusive events and other morbidities in patients with acute traumatic brain injury (CRASH-3) : a randomised, placebo-controlled trial”. In : *The Lancet* 394.10210 (2019), p. 1713-1723. DOI : 10.1016/S0140-6736(19)32233-0.

[72] Pierre Baldi et al. “Assessing the accuracy of prediction algorithms for classification : an overview”. In : *Bioinformatics* 16.5 (2000), p. 412-424. DOI : 10.1093/bioinformatics/16.5.412.

[73] Jacob Cohen. “A Coefficient of Agreement for Nominal Scales”. In : *Educational and Psychological Measurement* 20.1 (1960), p. 37-46. DOI : 10.1177/001316446002000104.

[74] Crash-2 Collaborators et al. “The importance of early treatment with tranexamic acid in bleeding trauma patients : an exploratory analysis of the CRASH-2 randomised controlled trial”. In : *The Lancet* 377.9771 (2011), p. 1096-1101.

[75] David W Hosmer et Stanley Lemesbow. “Goodness of fit tests for the multiple logistic regression model”. In : *Communications in statistics-Theory and Methods* 9.10 (1980), p. 1043-1069.

[76] Michael E Miller et al. “Validation of probabilistic predictions”. In : *Medical Decision Making* 13.1 (1993), p. 49-57. DOI : 10.1177/0272989X9301300107.

[77] Andrew J Vickers et Elena B Elkin. “Decision curve analysis : a novel method for evaluating prediction models”. In : *Medical Decision Making* 26.6 (2006), p. 565-574. DOI : 10.1177/0272989X06295361.

[78] George A. Diamond. “What price perfection ? Calibration and discrimination of clinical prediction models”. In : *Journal of Clinical Epidemiology* 45.1 (1992), p. 85-89. DOI : 10.1016/0895-4356(92)90192-P.

[79] M. H. Gail. “On criteria for evaluating models of absolute risk”. In : *Biostatistics* 6.2 (2005), p. 227-239. DOI : 10.1093/biostatistics/kxi005.

[80] Anastasios A. Tsiatis. *Dynamic treatment regimes : statistical methods for precision medicine*. Chapman et Hall/CRC, 2020. ISBN : 978-1-4987-6977-8.

[81] Holly Janes et al. “Statistical Methods for Evaluating and Comparing Biomarkers for Patient Treatment Selection”. In : *UW Biostatistics Working Paper Series* (2013).

[82] François Grolleau, Francois Petit et Raphaël Porcher. *A Comprehensive Framework for the Evaluation of Individual Treatment Rules From Observational Data*. 2023. arXiv : 2207.06275 [stat.ME].

[83] M. S. Pepe. “Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker”. In : *American Journal of Epidemiology* 159.9 (2004), p. 882-890. doi : 10.1093/aje/kwh101.

[84] Nancy R. Cook. “Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction”. In : *Circulation* 115.7 (2007), p. 928-935. doi : 10.1161/CIRCULATIONAHA.106.672402.

[85] Ingram Olkin et Thomas A. Trikalinos. “Constructions for a bivariate beta distribution”. In : *Statistics & Probability Letters* 96 (2015), p. 54-60. issn : 0167-7152. doi : https://doi.org/10.1016/j.spl.2014.09.013.

[86] Florie Bouvier et al. “Do machine learning methods lead to similar individualized treatment rules ? A comparison study on real data”. In : *Statistics in Medicine* (2024).

[87] Nicola Rieke et al. “The future of digital health with federated learning”. In : *NPJ Digital Medicine* 3 (2020), p. 119. doi : 10.1038/s41746-020-00323-1.

[88] C.D. Naylor. “Grey zones of clinical practice : some limits to evidence-based medicine”. In : *The Lancet* 345.8953 (1995), p. 840-842. doi : https://doi.org/10.1016/S0140-6736(95)92969-X.

[89] Cécile Konrat et al. “Underrepresentation of elderly people in randomised controlled trials. The example of trials of 4 widely prescribed drugs”. In : *PloS one* 7.3 (2012), e33559. doi : 10.1371/journal.pone.0033559.

[90] Céline Buffel du Vaure et al. “Exclusion of patients with concomitant chronic conditions in ongoing randomised controlled trials targeting 10 common chronic conditions and registered at ClinicalTrials.gov : a systematic review of registration details”. In : *BMJ Open* 6.9 (2016). doi : 10.1136/bmjopen-2016-012265.

[91] Guanbo Wang, Patrick J Heagerty et Issa J Dahabreh. “Using effect scores to characterize heterogeneity of treatment effects”. In : *JAMA* 331.14 (2024), p. 1225-1226.

[92] Charles H. Cain et al. “Design considerations and analytical framework for reliably identifying a beneficial individualized treatment rule”. In : *Contemporary Clinical Trials* 123 (2022), p. 106951. doi : 10.1016/j.cct.2022.106951.

[93] Roy C. Ziegelstein. “Personomics”. In : *JAMA Internal Medicine* 175.6 (juin 2015), p. 888-889. issn : 2168-6106. doi : 10.1001/jamainternmed.2015.0861. url : https://doi.org/10.1001/jamainternmed.2015.0861.

[94] Roy C. Ziegelstein. “Personomics and Precision Medicine”. English (US). In : *Transactions of the American Clinical and Climatological Association* 128 (jan. 2017), p. 160-168. issn : 0065-7778.

[95] Alexandre Malmartel, Philippe Ravaud et Viet-Thi Tran. “A methodological framework allows the identification of personomic markers to consider when designing personalized interventions”. In : *Journal of Clinical Epidemiology* 159 (2023), p. 235-245.

[96] Robert Griffiths et al. “Addition of rituximab to chemotherapy alone as first-line therapy improves overall survival in elderly patients with mantle cell lymphoma”. In : *Blood, The Journal of the American Society of Hematology* 118.18 (2011), p. 4808-4816.

[97] Daniel E Ho et al. “Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference”. In : *Political analysis* 15.3 (2007), p. 199-236.

[98] Stijn Vansteelandt et Niels Keiding. “Invited Commentary : G-Computation–Lost in Translation ?” In : *American Journal of Epidemiology* 173.7 (mars 2011), p. 739-742. issn : 0002-9262. doi : 10.1093/aje/kwq474. url : https://doi.org/10.1093/aje/kwq474.

[99] Paul R Rosenbaum et Donald B Rubin. “The central role of the propensity score in observational studies for causal effects”. In : *Biometrika* 70.1 (1983), p. 41-55.

[100] Florie Brion Bouvier et Raphaël Porcher. “What should be done and what should be avoided when comparing two treatments ?” In : *Best Practice & Research Clinical Haematology* 36.2 (2023), p. 101473.

[101] Michele Iudici et al. “Time-dependent biases in observational studies of comparative effectiveness research in rheumatology. A methodological review”. In : *Annals of the rheumatic diseases* 78.4 (2019), p. 562-569.

[102] Miguel A. Hernán et James M. Robins. “Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available : Table 1.” In : *American Journal of Epidemiology* 183.8 (2016), p. 758-764. doi : 10.1093/aje/kwv254.

[103] Miguel A Hernán, Wei Wang et David E Leaf. “Target trial emulation : a framework for causal inference from observational data”. In : *Jama* 328.24 (2022), p. 2446-2447.

Annexes

Annexe du chapitre 2

Estimating individualized treatment effects using an individual participant data meta-analysis

Florie Bouvier Anna Chaimani Etienne Peyrot François Gueyffier
Guillaume Grenet Raphaël Porcher

Supplementary material

S1 Methods to estimate individualized treatment effects with a time-to-event outcome

S1.1 ITE estimation

Let $S(t, x, z)$ represent the expected time-to-event at time t under treatment z for an individual with covariates x.
The ITE for a time-to-event outcome is estimated as:

$$\hat{\tau}(x) = \hat{S}(t, x, 1) - \hat{S}(t, x, 0)$$

at a prespecified time-point t.
Considering for instance a Cox regression model, the S-learner consists in estimating the following model:

$$S(t, x, z) = S_0(t) \exp(\theta' x + \gamma z).$$

From this, we derive for all individuals:

$$S(t, x, 1) = S_0(t) \exp(\theta' x + \gamma)$$

and

$$S(t, x, 0) = S_0(t) \exp(\theta' x).$$

With the T-learner, we fit the two following models:

$$S(t, x, 1) = S_0(t) \exp(\theta'^1 x + \gamma^1),$$

$$S(t, x, 0) = S_0(t) \exp(\theta'^0 x).$$

S1.2 Risk prediction models

Using Cox regressions, the different models are expressed as:

- Naive model (NA):

$$\lambda_{ij}(t|x_{ij}) = \lambda_0(t) \exp(\theta x_{ij}) \tag{1}$$

  The individual predictions are obtained by $S(t, x) = \hat{S}_0(t) \exp(\theta x)$, where $\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t))$ and $\hat{\Lambda}_0(t) = \sum_{\tilde{t} \leq t} \hat{\lambda_0}(\tilde{t})$.

- Random intercept model (RI):

$$\lambda_{ij}(t|x_{ij}, \rho_j) = \lambda_0(t) \exp(\theta x_{ij} + \rho_j) \tag{2}$$

  where $\rho_j \sim \mathcal{N}(0, \tau_\rho^2)$. The individual predictions are obtained as $S(t, x|j) = \hat{S}_0(t) \exp(\theta x + \rho_j)$.

- Stratified intercept model (SI):

$$\lambda_{ij}(t|x_{ij}, \lambda_{0j}) = \lambda_{0j}(t)\exp(\theta x_{ij}) \tag{3}$$

The individual predictions are obtained as $S(t, x|j) = \hat{S}_{0j}(t)\exp(\theta x)$.

- Fully stratified (FS):

$$\lambda_{ij}(t|x_{ij}, \lambda_{0j}, \theta_j) = \lambda_{0j}(t)\exp(\theta_j x_{ij}) \tag{4}$$

The individual predictions are obtained as $S(t, x|j) = \hat{S}_{0j}(t)\exp(\theta_j x)$.

- Rank-1 (R1):

$$\lambda_{ij}(t|x_{ij}, \phi_j, \rho_j) = \lambda_0(t)\exp(\phi_j \theta x_{ij} + \rho_j) \tag{5}$$

The individual predictions are obtained as $S(t, x|j) = \hat{S}_0(t)\exp(\theta_x + \rho_j)$.

S2 Potential aggregation bias

We did not simulate the fully stratified model. Since this model consists in stratifying every parameter by trial, aggregation bias is not an issue.

S2.1 Simulations without ecological bias

We simulated a binary outcome following a Bernoulli distribution with parameter P given by:

$$\text{logit}(P) = \beta_0 + \beta_1 x_1 + \beta_2 z + (\beta_3 x_1) \times z,$$

where z denoted the binary treatment indicator, x_1 was a normally distributed variable (see parameterization in table S1). Values for the model parameters were: $\beta_0 = -1.4$, $\beta_1 = 0.02$, $\beta_2 = -0.3$ and $\beta_3 = 0.01$. A total of 1,000 simulations with an IPD-MA sample size of 2800 was performed, and models with and without variable centering as described in Riley et al. [1] were fitted to the data.

Table S1: Distribution of x_1.

	Trial						
Variable	1	2	3	4	5	6	7
x_1, μ (σ)	52 (4)	56 (2)	64 (1)	70 (3)	77 (4)	78 (6)	82 (2)

Table S2: Median parameter estimates (standard errors) over 1000 simulations with the S-learner when ecological bias was not included.

	Model	Random intercept		Stratified intercept		Rank-1	
Parameter	Value	No centering	Centering	No centering	Centering	No centering	Centering
β_0	−1.4	−1.40(0.07)	−1.40(0.10)	−1.27(0.33)	−1.42(0.86)	−1.40	−1.41
β_1	0.02	0.02(0.01)	0.02(0.01)	0.02(0.01)	0.02(0.04)	0.02	0.02
β_2	−0.3	−0.30(0.11)	−0.31(0.11)	−0.30(0.11)	−0.31(0.11)	−0.30	−0.30
β_3	0.01	0.01(0.01)	−0.00(0.01)	0.01(0.01)	0.01(0.01)	0.01	0.00

The parameters obtained with and without centering the variables were similar for all models(Table S2). The intercept obtained with the stratified intercept when centering was included, was closer to the real value than when centering was not included, however, its standard error was larger.

S2.2 Simulations with ecological bias

We simulated a binary outcome in the same way as above but added some ecological bias. Therefore, the values for the model parameters were: $\beta_0 = -1.4$, $\beta_1 = 0.02$, $\beta_2 = -0.3 - ((mean(x_1) - 60)/100)$ and $\beta_3 = 0.01$.

Table S3: Median parameter estimates (standard errors) over 1000 simulations with the S-learner when ecological bias was included.

	Model	Random intercept		Stratified intercept		Rank-1	
Parameter	Value	No centering	Centering	No centering	Centering	No centering	Centering
β_0	−1.4	−1.40(0.08)	−1.40(0.10)	−1.22(0.34)	−1.42(0.88)	−1.34	−1.43
β_1	0.02	0.02(0.01)	0.02(0.01)	0.02(0.01)	0.02(0.04)	0.02	0.02
β_2	−0.40	−0.40(0.12)	−0.40(0.12)	−0.40(0.12)	−0.40(0.12)	−0.40	−0.40
β_3	0.01	0.00(0.01)	−0.00(0.01)	−0.00(0.01)	−0.00(0.01)	0.00	0.00

Centering or not centering the variables led to similar parameters (Table S3). Not centering the variables did not lead to aggregation bias, thus for simplicity, we decided to not add centering in this work.

S3 Simulation settings

S3.1 Covariates generation

In all simulation scenarios, covariates were numbered from x_1 to x_3 or x_1 to x_9, and their distribution varied among the trials j of the meta-analysis, as detailed in the tables S4 and S5. Covariates were drawn either from Gaussian distribution with mean μ and standard deviation σ or from a Bernoulli distribution with parameter π.

Table S4: Distribution parameters for covariates in scenarios with three covariates.

	Trial						
Variable	1	2	3	4	5	6	7
x_1, μ (σ)	52 (4)	56 (2)	64 (1)	70 (3)	77 (4)	78 (6)	82 (2)
x_2, π	0.8	0.4	0.5	0.6	0.5	0.7	0.5
x_3, μ (σ)	186 (13)	182 (16.5)	170 (9.4)	185 (12)	190 (9)	188 (10)	197 (21)

Table S5: Distribution parameters for covariates in scenarios with ten covariates.

	Trial						
Variable	1	2	3	4	5	6	7
x_1, μ (σ)	52 (4)	56 (2)	64 (1)	70 (3)	77 (4)	78 (6)	82 (2)
x_2, π	0.8	0.4	0.5	0.6	0.5	0.7	0.5
x_3, μ (σ)	186 (13)	182 (16.5)	170 (9.4)	185 (12)	190 (9)	188 (10)	197 (21)
x_4, π	0.1	0.005	0.01	0.02	0.05	0.01	0.04
x_5, π	0.002	0.06	0.02	0.02	0.001	0.008	0.04
x_6, π	0.5	0.2	0.3	0.4	0.3	0.25	0.3
x_7, π	0.03	0.001	0.002	0.07	0.003	0.01	0.002
x_8, π	0.13	0.11	0.05	0.25	0.05	0.06	0.04
x_9, μ (σ)	176 (6)	162 (9)	167 (10)	169 (10)	168 (10)	170 (9)	167 (9)

Table S6: Paramaters values for all scenarios.

	No variation				Variation			
	3 covariates		9 covariates		3 covariates		9 covariates	
	Binary	TTE	Binary	TTE	Binary	TTE	Binary	TTE
β_0	-1.4	50	-1.4	50	$-1.4 + ((mean(x_1[j]) - 60)/20)$	$50 + ((mean(x_1[j]) - 60)/20)$	$-1.4 + ((mean(x_1[j]) - 60)/20)$	$50 + ((mean(x_1[j]) - 60)/20)$
β_1	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
β_2	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
β_3	0.02	0.1	0.02	0.1	$0.02 + 0.2 * prop(x_2[j])$	$0.02 + 0.2 * prop(x_2[j])$	$0.02 + 0.2 * prop(x_2)$	$0.02 + 0.2 * prop(x_2[j])$
β_4	-0.3	-0.3	0.82	0.82	$-0.3 + ((mean(x_1[j]) - 60)/100)$	$-0.3 + ((mean(x_1[j]) - 60)/100)$	0.82	0.82
β_5	0.015	0.015	0.8	0.8	0.015	0.015	0.8	0.8
β_6	0.1	0.1	0.7	0.7	0.1	0.1	$0.7 + 0.2 * prop(x_2[j])$	$0.7 + 0.2 * prop(x_2[j])$
β_7	-0.008	-0.008	0.1	0.1	-0.08	-0.008	$0.1 + ((mean(x_1[j]) - 60/100)$	$0.1 + ((mean(x_1[j]) - 60/100)$
β_8	—	—	0.33	0.33	—	—	0.33	0.33
β_9	—	—	-0.02	-0.02	—	—	-0.02	-0.02
β_{10}	—	—	-0.3	-0.3	—	—	$-0.3 + ((mean(x_1[j]) - 60)/100)$	$-0.3 + ((mean(x_1[j]) - 60)/100)$
β_{11}	—	—	0.015	0.015	—	—	0.015	0.015
β_{12}	—	—	0.04	0.04	—	—	0.04	0.04
β_{13}	—	—	0.1	0.1	—	—	0.1	0.1
β_{14}	—	—	-0.008	-0.008	—	—	-0.008	-0.008

S3.2 Parameters generation

The binary outcome was simulated as:

$$\text{logit}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 z + (\beta_5 x_1 + \beta_6 x_2 + \beta_7 x_3) \times z$$

for scenarios with 3 covariates, and as:

$$\text{logit}(P) = \beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6+\beta_7 x_7+\beta_8 x_8+\beta_9 x_9+\beta_1 0z+(\beta_1 1x_1+\beta_1 2x_2+\beta_1 3x_3+\beta_1 4x_6)\times z$$

for scenarios with 9 covariates, where z represents the binary treatment indicator.

The time-to-event outcome was simulated as:

$$\lambda(t) = \beta_0(t) exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 z + (\beta_5 x_1 + \beta_6 x_2 + \beta_7 x_3) \times z)$$

for scenarios with 3 covariates, and as:

$$\lambda(t) = \beta_0(t) exp(\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6+\beta_7 x_7+\beta_8 x_8+\beta_9 x_9+\beta_1 0z+(\beta_1 1x_1+\beta_1 2x_2+\beta_1 3x_3+\beta_1 4x_6)\times z)$$

for scenarios with 9 covariates, where z represents the binary treatment indicator.

S3.3 Simulation scenarios

Table S7: Summary of the 24 simulation scenarios. IPD-MA: individual patients meta-analysis.

Scenario	Outcome	No. covariates	IPD-MA sample size	Heterogeneity
1	Binary	3	2800	No
2	Binary	3	1400	No
3	Binary	3	700	No
4	Binary	9	2800	No
5	Binary	9	1400	No
6	Binary	9	700	No
7	Binary	3	2800	Yes
8	Binary	3	1400	Yes
9	Binary	3	700	Yes
10	Binary	9	2800	Yes
11	Binary	9	1400	Yes
12	Binary	9	700	Yes
13	Time-to-event	3	2800	No
14	Time-to-event	3	1400	No
15	Time-to-event	3	700	No
16	Time-to-event	9	2800	No
17	Time-to-event	9	1400	No
18	Time-to-event	9	700	No
19	Time-to-event	3	2800	Yes
20	Time-to-event	3	1400	Yes
21	Time-to-event	3	700	Yes
22	Time-to-event	9	2800	Yes
23	Time-to-event	9	1400	Yes
24	Time-to-event	9	700	Yes

S4 Simulation results

S4.1 Scenarios 1 to 6 and 13 to 18

Figure S1 shows the results for scenarios 1 to 3 where the predictor effects did not vary across trials. The c-statistic for benefit values, without variation in the predictor effects, were similar between models. We

observed similar performances whether the S-learner or the T-learner was used. Looking at calibration, we found that the models' values are akin for both learners, the intercept values are close to the intended value of 0 but the slope values were far from 1. Concerning the MSE, choosing SI or R1 led to a higher MSE and therefore are not recommended. Increasing the size of the IPD-MA led to lower MSE values. Figure S2 represents the results for scenarios 4 to 6. FS led to worse performance than the other models with bad discrimination and calibration, as well as higher MSE values. The other models had a good performance with mean c-statistic for benefit values above 0.6 and good calibration results

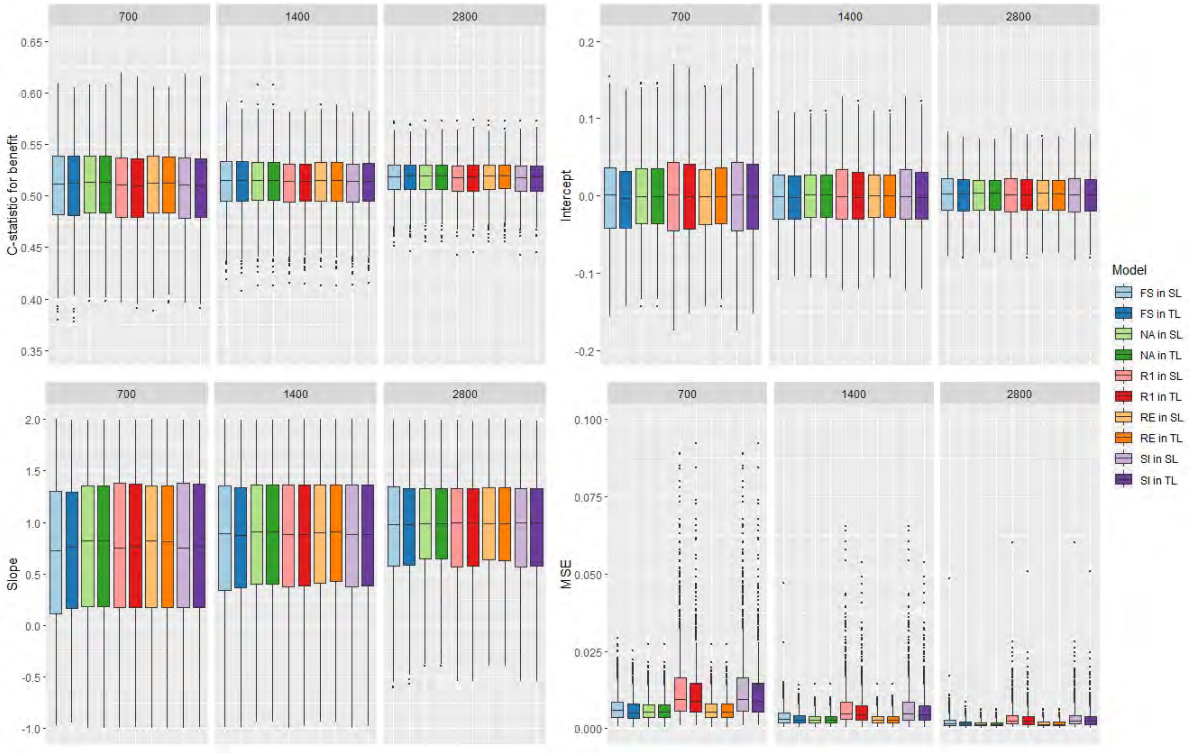

Figure S1: Boxplot of the measures of performance of the models for scenario 1 to 3.

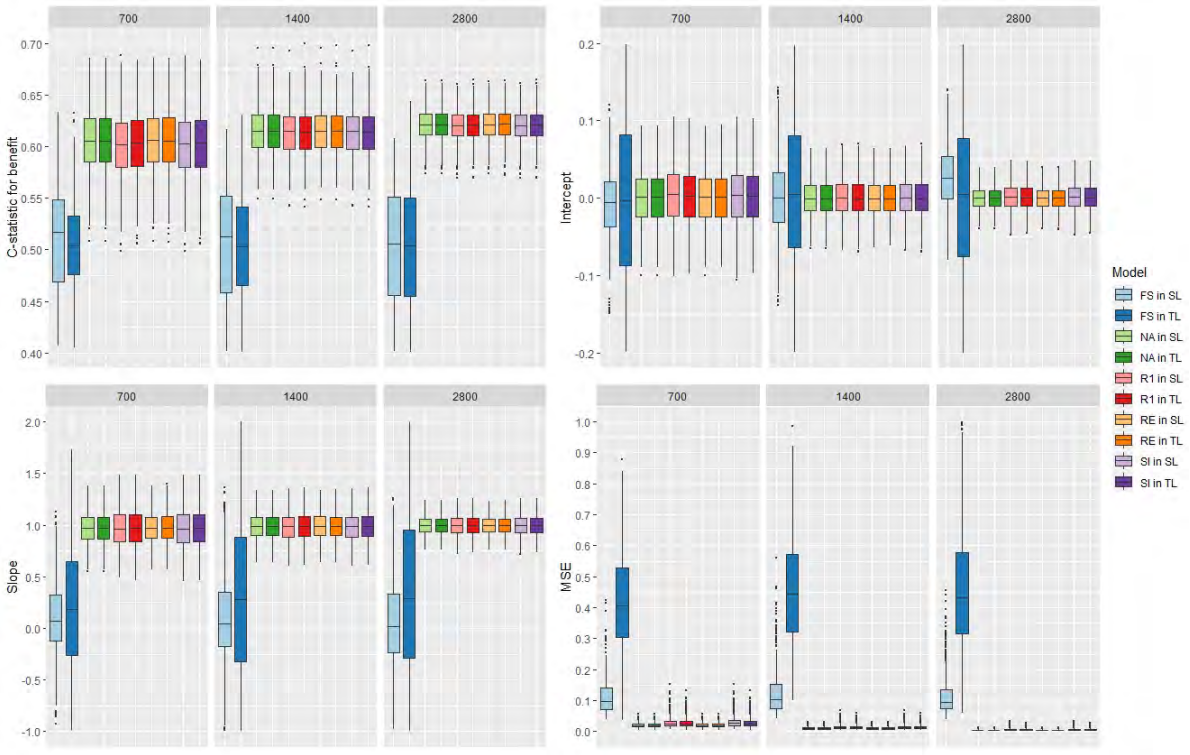

Figure S2: Boxplot of the measures of performance of the models for scenario 4 to 6.

The results of scenarios 13 to 15 are displayed in Figure S3. Using the T-learner led to better discrimination results for all methods, whereas using the S-learner led to better calibration results. Slope values far from 1 indicating a poor calibration. Higher MSE values were obtained for the fully stratified model and for the rank-1 model when the T-learner was used. The NA, RI, and SI methods' results were similar. Results of scenarios 16 to 18 are represented in Figure S5. FS had better calibration results when a time-to-event and 9 covariates were used. However, it had a lower c-statistic for benefit values and a higher MSE.

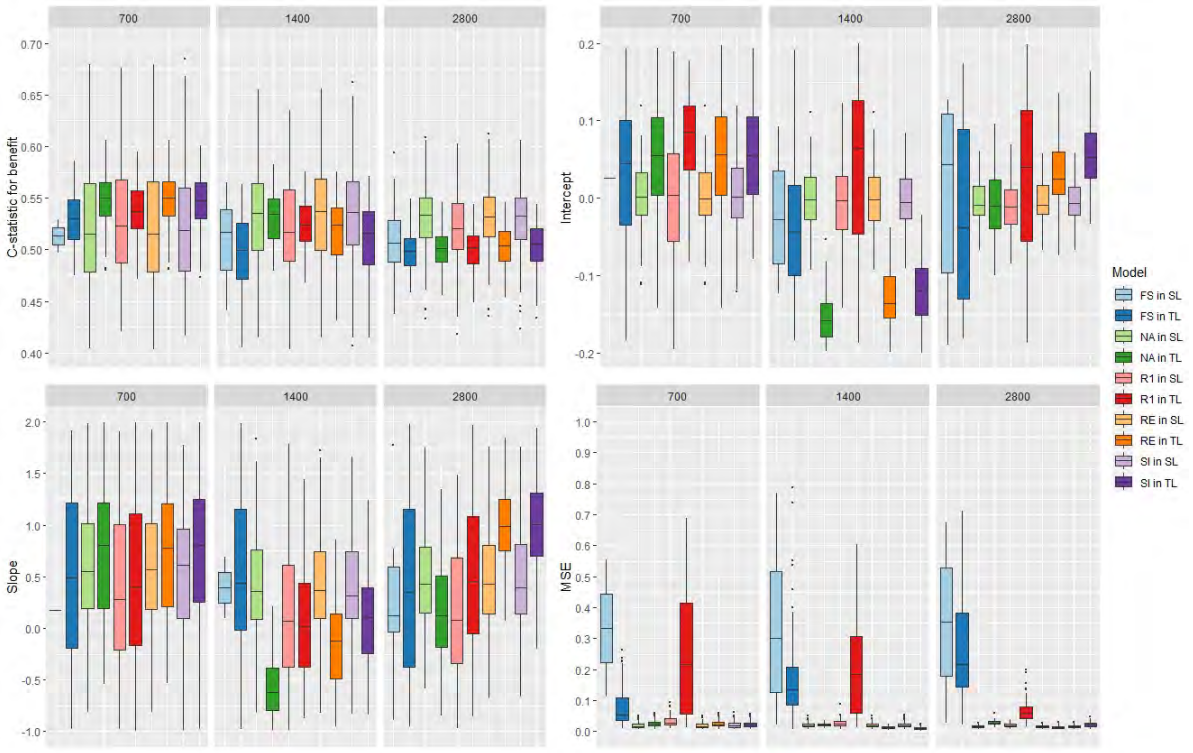

Figure S3: Boxplot of the measures of performance of the models for scenario 13 to 15.

Generally, the rank-1 and the fully stratified models which capture more heterogeneity in the predictor effects, were the two methods that captured more often the true ITE in their prediction intervals (Figure S4).

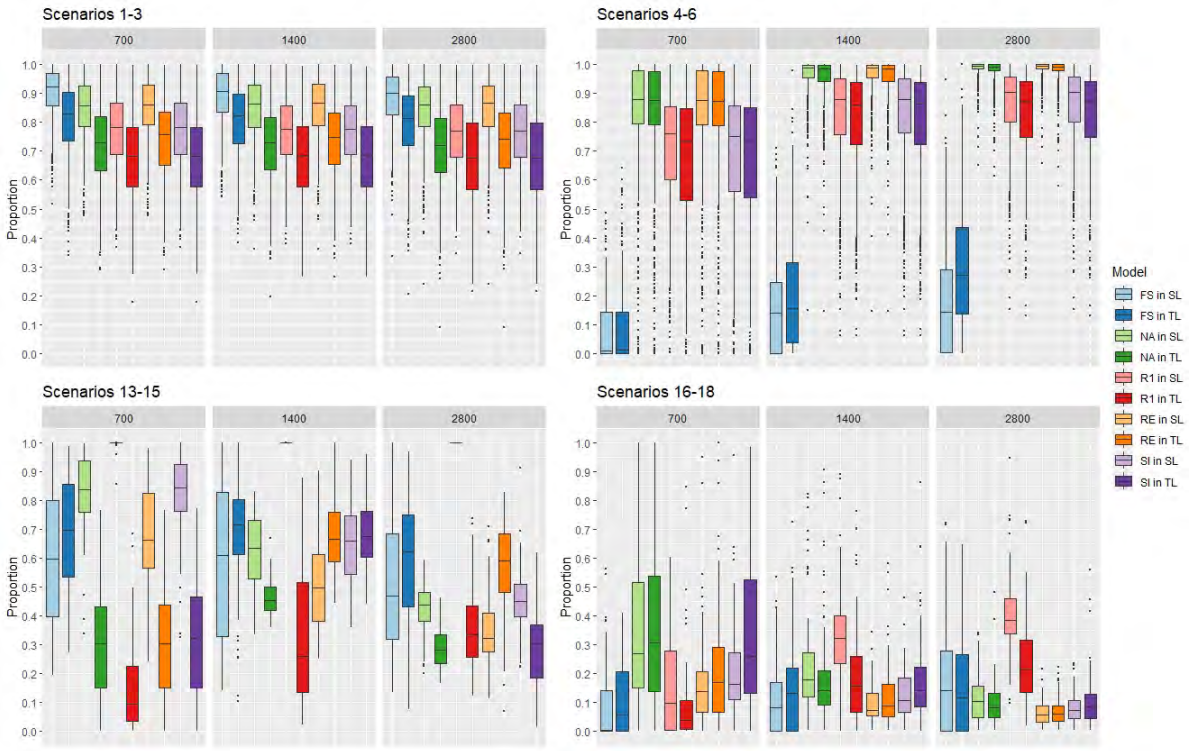

Figure S4: Number of times the true ITEs was in the prediction intervals of each model.

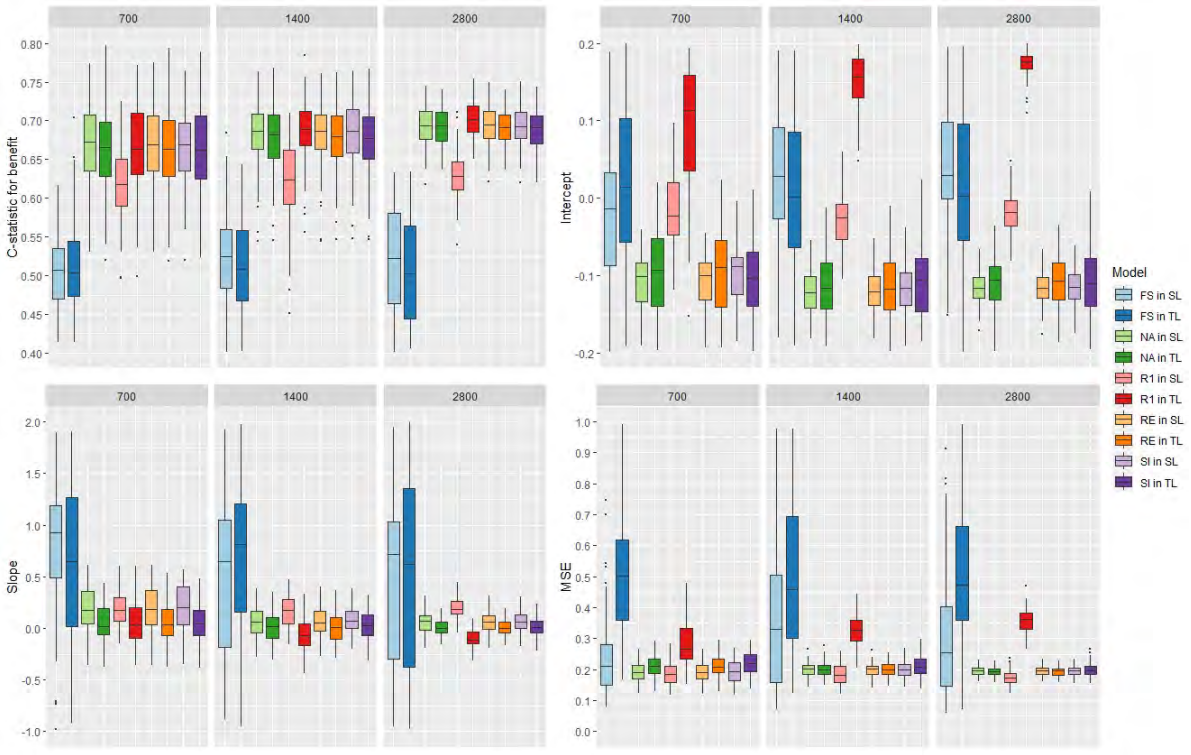

Figure S5: Boxplot of the measures of performance of the models for scenario 16 to 18.

Regarding the approaches with and without variables' selection using the stratified intercept method, the c-statistic for benefit values, intercept values and MSE values obtained were comparable whether variation in predictor effects was included or not (Figure S6 and Figure S7). Slope values differed a bit, but were of equivalent distance to the targeted value of 1. These observations were the same whether the method was used with the S-learner or the T-learner approach. When variation was included, slope values were slightly closer to 1 with variable selection and the T-learner approach. Adding variables' selection did not significantly impact the performances and did not modify our conclusions.

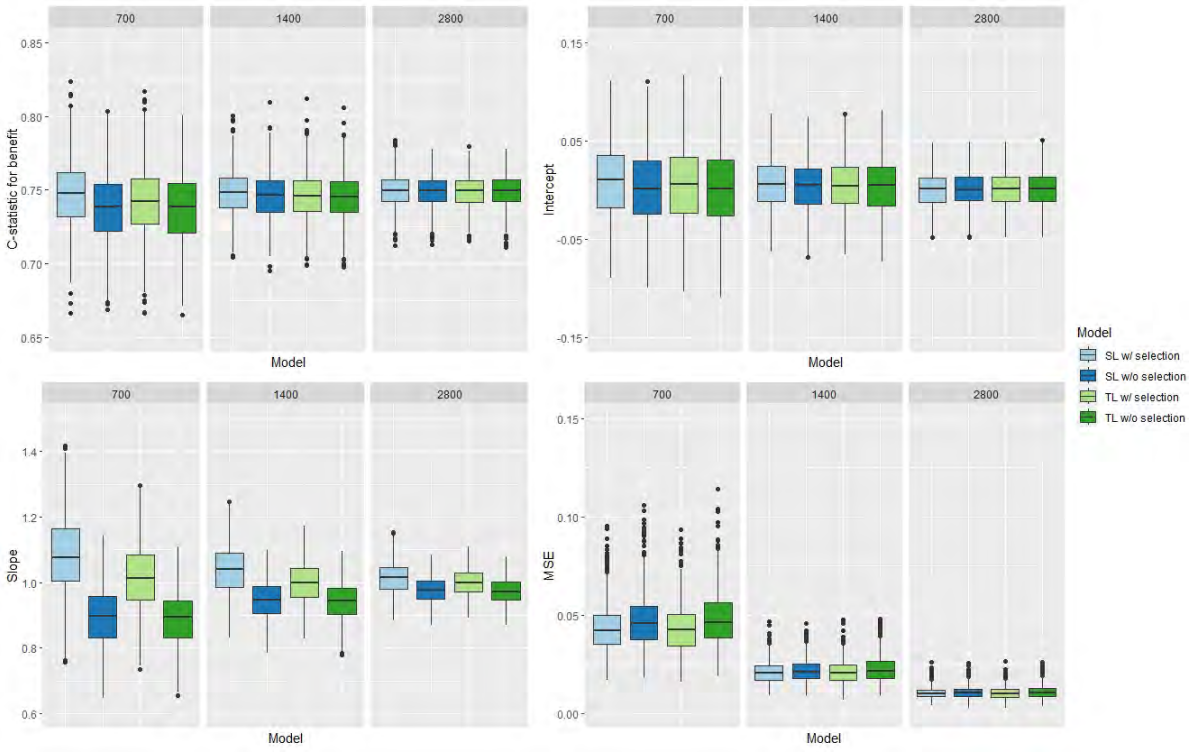

Figure S6: Boxplot of the measures of performance of the models for scenario 4 to 6 with and without variables' selection.

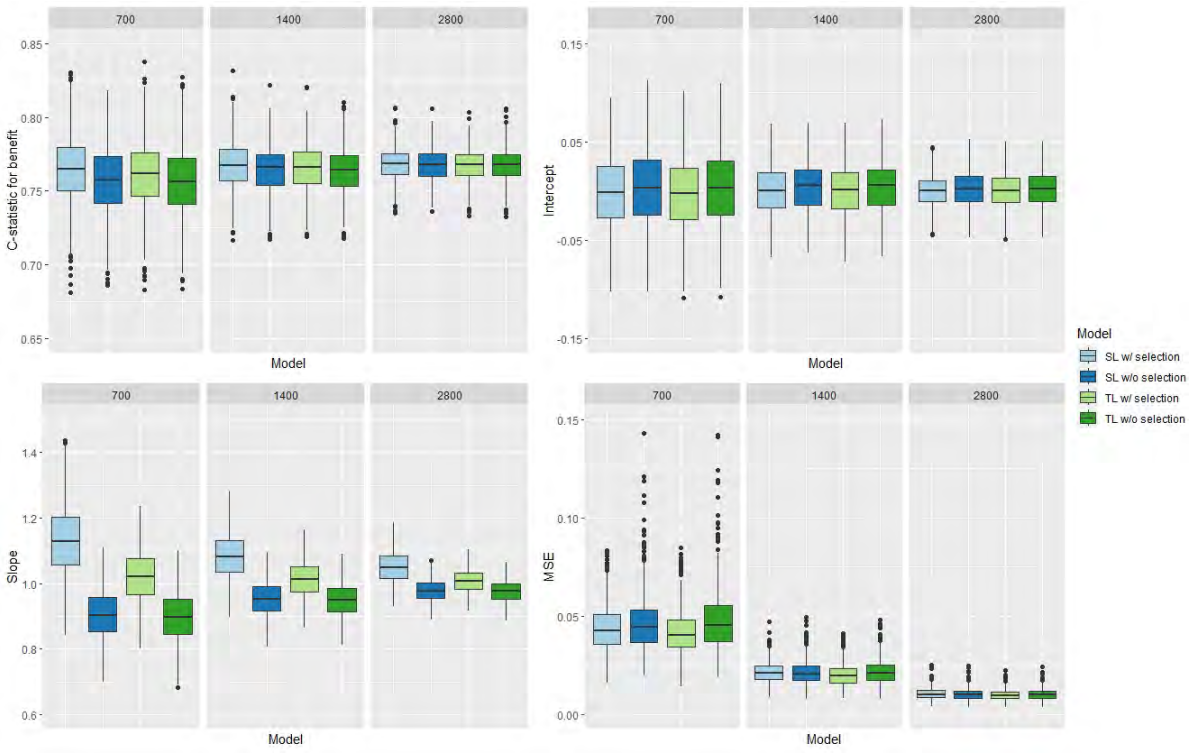

Figure S7: Boxplot of the measures of performance of the models for scenario 10 to 12 with and without variables' selection.

S4.2 Scenario with a proportional heterogeneity

We took scenario 7 but changed the heterogeneity of the treatment variable (β_4) and variable β_3, described in Table 6. In this new scenario, the heterogeneity of β_3 is generated as $\{0.02, -0.02, 0.02, 0.04, 0.01, -0.04, 0.03\}$ and the heterogeneity of β_4 is generated as $\{-0.05, -0.025, 0.025, 0.025, 0.015, 0.05, 0.1\}$.
We obtained better calibration results and a slightly higher median number of times the true ITE was in the prediction interval (Table S8).

Table S8: Median results of the rank-1 model for a scenario in which the heterogeneity was generated in a proportional way.

	S-learner	T-learner
C-statistic for benefit	0.531	0.531
Calibration's intercept	0.001	0.004
Calibration's slope	1.074	1.071
MSE	0.002	0.002
Proportion in prediction interval	0.870	0.799

S5 INDANA IPD-MA

S5.1 Distribution of age

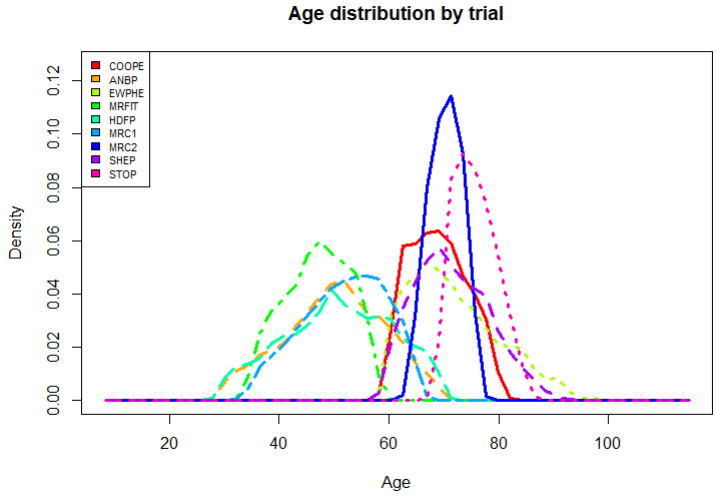

Figure S8: Distribution of age in each trial of INDANA.

S5.2 Results on train dataset

When assessing the performance of the methods and the approaches on the train dataset, the discrimination remained low but the calibration improved, especially for the SI and R1 methods (Table S9). The fact that the performance did not increase a lot on the train dataset might indicate that the disparity between the trials was too high for them to be meta-analyzed.

Table S9: Median results using INDANA with a binary outcome with the training dataset.

	S-learner					T-learner				
	NA	RI	SI	R1	FS	NA	RI	SI	R1	FS
C-stat	0.534	0.532	0.531	0.531	0.532	0.517	0.528	0.529	0.529	0.516
Intercept	-0.001	0.001	0.001	0.001	0.001	-0.001	-0.002	-0.001	-0.001	-0.002
Slope	1.376	1.183	1.133	1.133	0.954	0.868	0.905	0.984	0.984	-0.063
MSE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

References

[1] R. D. Riley, T. P. A. Debray, D. Fisher, M. Hattle, N. Marlin, J. Hoogland, F. Gueyffier, J. A. Staessen, J. Wang, K. G. M. Moons, J. B. Reitsma, and J. Ensor. Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: Statistical recommendations for conduct and planning. *Statistics in Medicine*, 39(15):2115–2137, 2020.

Annexe du chapitre 3

Do machine learning methods lead to similar individualized treatment rules? A comparison study on real data.

Florie Bouvier[1], Etienne Peyrot[1], Alan Balendran[1], Corentin Ségalas[2], Ian Roberts[3], François Petit[1], and Raphaël Porcher[1,4]

[1]Université Paris Cité, Centre de Recherche Épidémiologie et Statistiques–CRESS UMR1153, Inserm, INRAE

[2]Université Bordeaux, Inserm, Bordeaux Population Health Research Center, Bordeaux, France

[3]Clinical Trials Unit, London School of Hygiene & Tropical Medicine, London, UK

[4]Centre d'Épidémiologie Clinique, Assistance Publique-Hôpitaux de Paris, Hôtel-Dieu, Paris, France

Supplementary material

1 Cross-Fit

Non-parametric models requiring nuisance functions, i.e., the sub-models needed to be trained to estimate the value of interest, can be prone to overfitting. For example, there are two nuisance functions in T-learner (the two models used to estimate the response functions) and four in X-learner (two models for the response functions and two models with the imputed treatment effects). Cross-fit (CF) is a technique that can reduce this potential overfitting.[1] First, the dataset is split into equal-sized non-overlapping folds. Then, each nuisance function is estimated on separate folds which gives a first estimate of the value of interest. The folds are swapped until each nuisance function is trained on every single one of them. This gives us as many estimates as there are folds. Then, we average the estimates. The splitting step is repeated to get as many estimates as we want to increase the estimators' stability. Finally, we calculate the median of those estimates. Using the medians, we estimate the ITE and thus the ITR. When cross-fit was applied, 5 folds and 30 splits were used, because such a choice has been reported as leading to a good performance in Jacob's paper.[2] However, to our knowledge, there is no clear guidance on how to perform cross-fit, and other versions exist and can be found in

Jacob's paper.[2] Here, we used cross-fit on S-learner, T-learner, and X-learner, which may be prone to overfitting, but also on the DR-learner and the R-learner.

2 Implementation

2.1 IST

- Restricted cubic splines were used to take care of the non-linearity of continuous covariates in methods using logistic regression. Age and SBP were both concerned by the non-linearity. For *Age*, the chosen knots were: 57,74 and 85. For *SBP*, the knots were: 130, 160 and 200.
- For methods using random forests, a grid search was performed to tune the parameters. The following parameters were used: $ntree = 300$ and $mtry = 4$.
- When using the S-learner approach, all the interactions between the treatment and the covariates were included in the model.

2.2 CRASH-3

- When random forests were used, the tuned parameters were: $ntree = 300$ and $mtry = 2$.

3 Description of variables

3.1 IST

Table S 1 describes the variables input in the models using IST. These variables were chosen to match the work of Nguyen et al.[3] in which an individual treatment rule was developed using the T-learner algorithm, one of the methods compared in this work.

3.2 CRASH-3

Six covariates were included in the different models using the CRASH-3 dataset and are described in Table S 2. These variables were chosen based on their clinical relevance.

4 Results on training data

4.1 IST

Performance results were better when the training data was used rather than the validating data (Table S 3). The discrimination was better with higher c-statistic for benefit values. Particularly, non-parametric meta-learners had c-statistic for benefit values close to 1. The other algorithms had

Table S 1: Description of the variables in the IST dataset.

Variable	Control group 9,715 (50.0%)	Treatment group 9,720 (50.0%)
Female, no. (%)	4567 (47.0)	4461 (45.9)
Age, mean (SD) years	71.7 (11.6)	71.7 (11.6)
Systolic blood pressure, mean (SD) mmHg	160.3 (27.5)	160.0 (27.7)
Delay, mean (SD) h	20.1 (12.5)	20.1 (12.4)
CT before hospitalisation, Yes (%)	6533 (67.2)	6491 (66.8)
Infarct visible at CT, Yes (%)	3239 (33.3)	3176 (32.7)
Atrial fibrillation, Yes (%)	1633 (16.8)	1711 (17.6)
Aspirin within previous 3 days, Yes (%)	2084 (21.4)	2076 (21.4)
Face deficit, No (%)	2522 (26.0)	2567 (26.4)
Face deficit, Not assessable (%)	124 (1.3)	123 (1.3)
Face deficit, Yes (%)	7069 (72.8)	7030 (72.3)
Arm/hand deficit, No (%)	1319 (13.6)	1348 (13.9)
Arm/hand deficit, Not assessable (%)	68 (0.7)	55 (0.6)
Arm/hand deficit, Yes (%)	8328 (85.7)	8317 (85.6)
Leg/foot deficit, No (%)	2230 (23.0)	2272 (23.4)
Leg/foot deficit, Not assessable (%)	122 (1.3)	133 (1.4)
Leg/foot deficit, Yes (%)	7363 (75.8)	7315 (75.3)
Dysphasia, No (%)	5170 (53.2)	5172 (53.2)
Dysphasia, Not assessable (%)	303 (3.1)	281 (2.9)
Dysphasia, Yes (%)	4242 (43.7)	4237 (43.9)
Hemianopia, No (%)	6197 (63.8)	6197 (63.8)
Hemianopia, Not assessable (%)	1958 (20.2)	1987 (20.4)
Hemianopia, Yes (%)	1560 (16.1)	1536 (15.8)
Visuospatial disorder, No (%)	6393 (65.8)	6416 (66.0)
Visuospatial disorder, Not assessable (%)	1733 (17.8)	1715 (17.6)
Visuospatial disorder, Yes (%)	1589 (16.4)	1589 (16.3)
Cerebellar signs, No (%)	7856 (80.9)	7848 (80.7)
Cerebellar signs, Not assessable (%)	795 (8.2)	797 (8.2)
Cerebellar signs, Yes (%)	1064 (11.0)	1075 (11.1)
Other deficit, No (%)	8486 (87.3)	8481 (87.3)
Other deficit, Not assessable (%)	616 (6.3)	633 (6.5)
Other deficit, Yes (%)	613 (6.3)	606 (6.2)
Consciousness, Fully alert (%)	7458 (76.8)	7463 (76.8)
Consciousness, Drowsy (%)	2127 (21.9)	2127 (21.9)
Consciousness, Unconscious (%)	130 (1.3)	130 (1.3)
Stroke type, PACS (%)	3935 (40.5)	3920 (40.3)
Stroke type, TACS (%)	2311 (23.8)	2327 (24.0)
Stroke type, LACS (%)	2331 (24.0)	2326 (24.0)
Stroke type, POCS (%)	1107 (11.4)	1121 (11.5)
Stroke type, Other (%)	31 (0.3)	26 (0.3)
Region, Europe (%)	8228 (84.7)	8234 (84.7)
Region, North America (%)	124 (1.3)	124 (1.3)
Region, South America (%)	346 (3.6)	347 (3.7)
Region, Africa (%)	34 (0.3)	35 (0.4)
Region, Middle East (%)	200 (2.1)	200 (2.1)
Region, North Asia (%)	62 (0.6)	62 (0.6)
Region, South Asia (%)	196 (2.0)	193 (2.0)
Region, Oceania (%)	525 (5.4)	525 (5.4)
Death/dependency at 6 months, Yes (%)	6155 (63.3)	6042 (62.1)

Table S 2: Description of the variables in the CRASH-3 dataset.

Variable	Control group 4,491 (49.5%)	Treatment group 4,581 (50.5%)
Female, no. (%)	884 (19.7)	890 (19.4)
Age, mean (SD) years	41.9 (19.0)	41.8 (19.0)
Systolic blood pressure, mean (SD) mmHg	129.7 (26.4)	130.5 (26.9)
Time since injury, mean (SD) h	1.9 (0.7)	1.9 (0.7)
Glasgow Coma Scale score, mean (SD)	9.6 (3.9)	9.6 (3.9)
Pupil reaction, Both React (%)	3583 (79.8)	3653 (79.7)
Pupil reaction, None React (%)	435 (9.7)	420 (9.2)
Pupil reaction, One Reacts (%)	349 (7.8)	369 (8.1)
Pupil reaction, Unable to assess (%)	124 (2.8)	139 (3.0)
Death, Yes (%)	945 (21.0)	923 (20.1)

moderate discrimination with values around 0.55. PAPE values were close to 0, except for the non-parametric meta-learners, where their ITRs performed better than non-individualized rules which treated the same proportion of patients. B_{pos} and B_{neg} values were close to 0, meaning there were not many benefits of treating patients with a positive score or not treating patients with a negative score, except for the non-parametric meta-learners. The proportion of patients for whom aspirin was recommended by the different ITRs ranged from 0.103 to 0.911, with most methods producing an ITR that recommended treatment for more than 50% of patients, similar to what was obtained in the validating data. Although there were different proportions, the estimated values of the ITRs were similar and close to what was obtained with the validating data for the majority of the methods. The only improvement in estimated values was acquired with the non-parametric meta-learners. Comparing these results to the results obtained on the validating data suggest a high potential for overfitting, especially for non-parametric models.

As for the validating data, the ITRs produced with the training data have significant disagreements except when the algorithms are part of the same family (e.g. parametric meta-learners, non-parametric meta-learners, A-learning and the modified covariate method) (Figure S 1).

4.2 CRASH-3

Performance was better on the CRASH-3 training data than on the validating data (Table S 4). The discrimination was better with higher c-statistic for benefit values, some ITRs even had values approximating 1. As for IST, PAPE values were close to 0, except for the non-parametric meta-learners which had values between 0.090 and 0.185. B_{pos} and B_{neg} values were above 0 showing some benefits of treating patients with a positive score or not treating patients with a negative score, except for CWL. The proportion of patients for whom aspirin was recommended by the different ITRs ranged from 0 to 1 with a majority of methods recommending to treat over 60% of patients, similar to what

Table S 3: Results of the metrics for each method applied to the IST training data.

	p_r	$\mathcal{V}(r)$ (SE)	$E(Y^0)$ (SE)	$E(Y^1)$ (SE)	B_{pos} (SE)	B_{neg} (SE)	PAPE (SE)	cstat (95% CI)
SL	0.602	0.388 (0.006)	0.360 (0.006)	370 (0.006)	0.047 (0.011)	0.044 (0.012)	0.022 (0.004)	0.541 (0.528;0.553)
TL	0.602	0.388 (0.006)	0.360 (0.006)	370 (0.006)	0.047 (0.011)	0.044 (0.012)	0.022 (0.004)	0.541 (0.528;0.553)
XL	0.650	0.387 (0.006)	0.360 (0.006)	370 (0.006)	0.041 (0.011)	0.046 (0.014)	0.020 (0.004)	0.539 (0.526;0.551)
DRL	0.650	0.387 (0.006)	0.360 (0.006)	370 (0.006)	0.041 (0.011)	0.046 (0.013)	0.020 (0.004)	0.539 (0.527;0.551)
RL	0.668	0.385 (0.006)	0.360 (0.006)	370 (0.006)	0.037 (0.010)	0.044 (0.014)	0.018 (0.004)	0.535 (0.523;0.548)
SL RF	0.506	0.806 (0.005)	0.360 (0.006)	370 (0.006)	0.764 (0.008)	0.747 (0.008)	0.441 (0.005)	0.963 (0.960;0.966)
TL RF	0.506	0.845 (0.005)	0.360 (0.006)	370 (0.006)	0.812 (0.007)	0.800 (0.008)	0.480 (0.005)	0.975 (0.972;0.978)
XL RF	0.522	0.692 (0.006)	0.360 (0.006)	370 (0.006)	0.584 (0.010)	0.600 (0.010)	0.327 (0.005)	0.923 (0.917;0.928)
DRL RF	0.510	0.879 (0.004)	0.360 (0.006)	370 (0.006)	0.852 (0.007)	0.860 (0.007)	0.513 (0.005)	0.987 (0.986;0.989)
RL RF	0.507	0.848 (0.005)	0.360 (0.006)	370 (0.006)	0.822 (0.007)	0.819 (0.007)	0.483 (0.005)	0.983 (0.981;0.985)
SL CF	0.589	0.599 (0.007)	0.360 (0.006)	370 (0.006)	0.465 (0.011)	0.384 (0.011)	0.233 (0.004)	0.847 (0.840;0.855)
TL CF	0.554	0.610 (0.006)	0.360 (0.006)	370 (0.006)	0.426 (0.010)	0.494 (0.011)	0.245 (0.005)	0.857 (0.850;0.865)
XL CF	0.599	0.528 (0.006)	0.360 (0.006)	370 (0.006)	0.275 (0.010)	0.381 (0.012)	0.162 (0.004)	0.786 (0.777;0.795)
DRL CF	0.533	0.696 (0.006)	0.360 (0.006)	370 (0.006)	0.575 (0.010)	0.624 (0.010)	0.331 (0.005)	0.919 (0.914;0.924)
RL CF	0.527	0.686 (0.006)	0.360 (0.006)	370 (0.006)	0.569 (0.009)	0.602 (0.010)	0.321 (0.005)	0.915 (0.910;0.920)
PATH	0.643	0.371 (0.006)	0.360 (0.006)	370 (0.006)	0.016 (0.011)	0.002 (0.009)	0.005 (0.004)	0.510 (0.498;0.522)
Causal forests	0.911	0.382 (0.006)	0.360 (0.006)	370 (0.006)	0.024 (0.009)	0.147 (0.022)	0.013 (0.002)	0.579 (0.568;0.591)
VT	0.766	0.374 (0.006)	0.360 (0.006)	370 (0.006)	0.018 (0.010)	0.012 (0.013)	0.006 (0.003)	—
AL	0.498	0.387 (0.005)	0.360 (0.006)	370 (0.006)	0.063 (0.012)	0.055 (0.012)	0.022 (0.003)	0.549 (0.537;0.561)
MCM	0.498	0.387 (0.005)	0.360 (0.006)	370 (0.006)	0.063 (0.012)	0.055 (0.012)	0.022 (0.003)	0.549 (0.537;0.561)
OWL	0.902	0.371 (0.006)	0.360 (0.006)	370 (0.006)	0.012 (0.009)	0.021 (0.011)	0.002 (0.002)	—
CWL	0.103	0.368 (0.006)	0.360 (0.006)	370 (0.006)	0.025 (0.023)	-0.002 (0.009)	0.007 (0.003)	—

p_r refers to the proportion of patients for which treatment is recommended by the rule. SL: S-learner, TL: T-learner, XL: X-learner, DRL: DR-learner, RL: R-learner, RF: random forests, CF: cross-fitted, VT: virtual twins, MCM: modified covariate method, AL: A-learning, OWL: outcome weighted learning and CWL: contrast weighted learning.

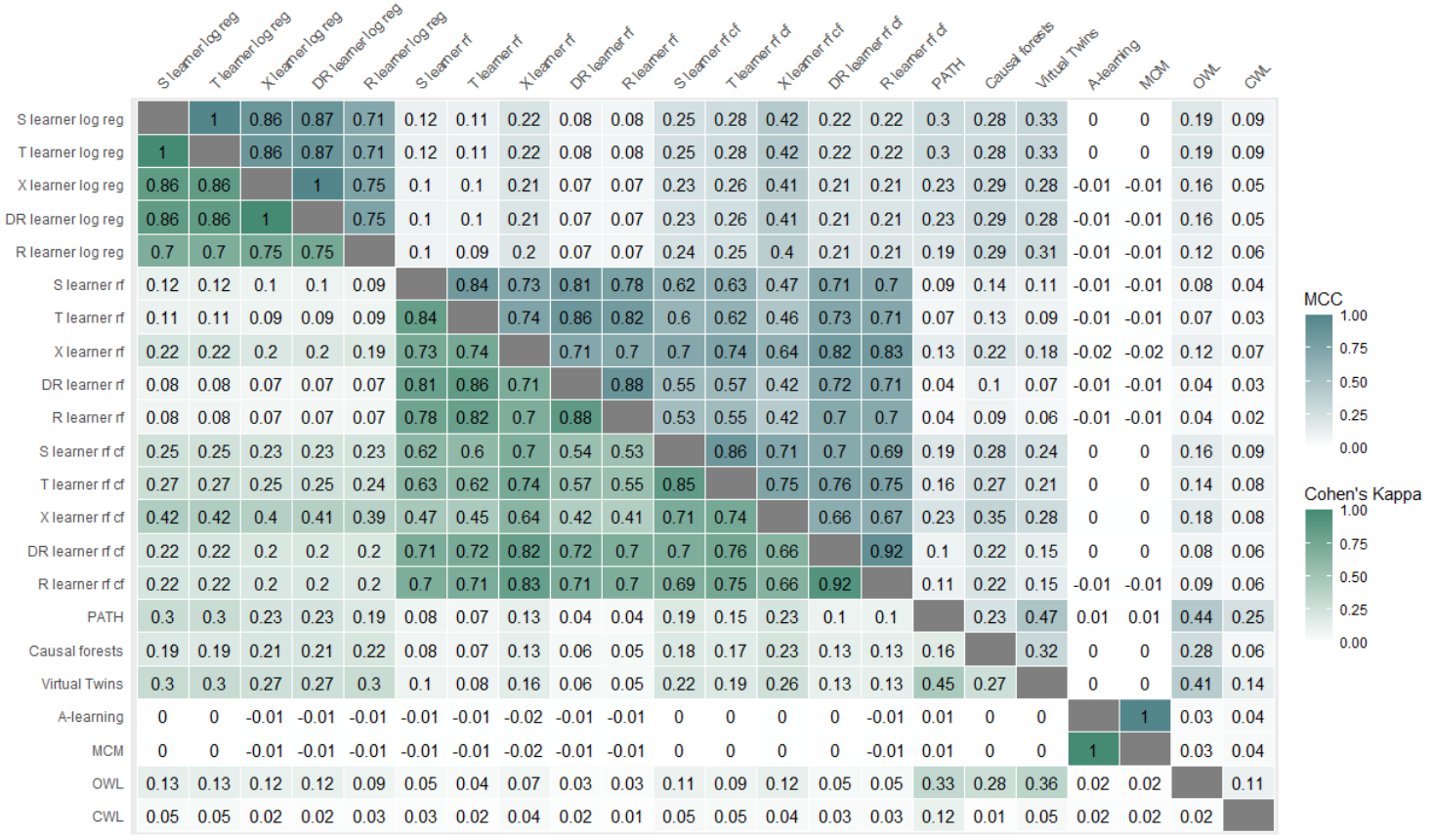

Figure S 1: Heatmap representing the MCC and Cohen's Kappa for each combination of two ITRs using IST training data.

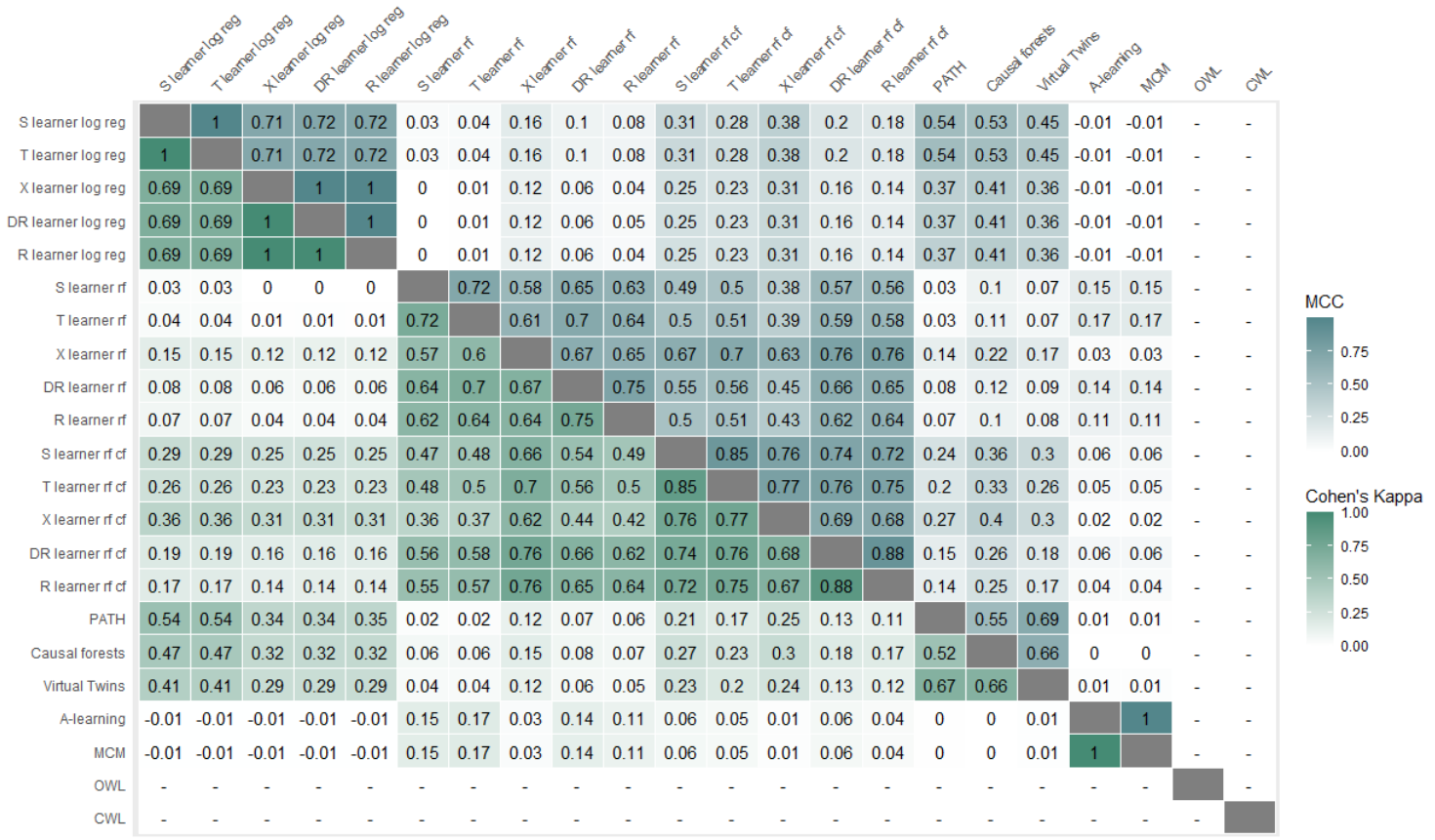

Figure S 2: Heatmap representing the MCC and Cohen's Kappa for each combination of two ITRs using CRASH-3 training data.

was obtained in the validating data. All the ITRs' values were above 0.8, and the best rule's values were obtained with the non-parametric meta-learners.

The ITRs' agreements were similar whether the training or the testing data was used to estimate the correlation coefficients (Figure S 2). A strong concordance is only found between algorithms belonging to the same family (e.g. parametric meta-learners, non-parametric meta-learners, A-learner and the Modified covariate method).

5 ITEs scatterplots

5.1 IST

ITEs were strongly correlated when the methods were akin, but they were badly correlated when the two methods didn't belong to the same family (Figure S 3).

5.2 CRASH-3

The same conclusion is drawn with the CRASH-3 dataset. Similar methods produced similar ITEs (Figure S 4).

Table S 4: Results of the metrics for each method applied to the CRASH-3 training data.

	p_r	$\mathcal{V}(r)$ (SE)	$E(Y^0)$ (SE)	$E(Y^1)$ (SE)	B_{pos} (SE)	B_{neg} (SE)	PAPE (SE)	cstat (95% CI)
SL	0.767	0.820 (0.007)	0.802 (0.007)	0.812 (0.007)	0.022 (0.010)	0.036 (0.027)	0.011 (0.006)	0.528 (0.501;0.555)
TL	0.767	0.820 (0.007)	0.802 (0.007)	0.812 (0.007)	0.022 (0.010)	0.036 (0.028)	0.011 (0.006)	0.528 (0.501;0.555)
XL	0.668	0.815 (0.007)	0.802 (0.007)	0.812 (0.007)	0.017 (0.011)	0.006 (0.021)	0.006 (0.006)	0.523 (0.499;0.548)
DRL	0.668	0.815 (0.007)	0.802 (0.007)	0.812 (0.007)	0.017 (0.011)	0.006 (0.021)	0.006 (0.005)	0.523 (0.499;0.547)
RL	0.669	0.815 (0.007)	0.802 (0.007)	0.812 (0.007)	0.017 (0.011)	0.005 (0.022)	0.006 (0.005)	0.523 (0.499;0.547)
SL RF	0.482	0.983 (0.002)	0.802 (0.007)	0.812 (0.007)	0.474 (0.016)	0.413 (0.015)	0.176 (0.005)	0.963 (0.955;0.971)
TL RF	0.498	0.986 (0.002)	0.802 (0.007)	0.812 (0.007)	0.479 (0.016)	0.453 (0.015)	0.180 (0.005)	0.971 (0.964;0.978)
XL RF	0.580	0.957 (0.004)	0.802 (0.007)	0.812 (0.007)	0.279 (0.012)	0.374 (0.016)	0.150 (0.005)	0.907 (0.894;0.921)
DRL RF	0.555	0.993 (0.001	0.802 (0.007)	0.812 (0.007)	0.420 (0.014)	0.527 (0.015)	0.185 (0.005)	0.982 (0.976;0.987)
RL RF	0.544	0.992 (0.002)	0.802 (0.007)	0.812 (0.007)	0.403 (0.014)	0.472 (0.015)	0.184 (0.005)	0.978 (0.973;0.984)
SL CF	0.633	0.923 (0.005)	0.802 (0.007)	0.812 (0.007)	0.198 (0.012)	0.341 (0.019)	0.115 (0.005)	0.836 (0.819;0.854)
TL CF	0.604	0.926 (0.005)	0.802 (0.007)	0.812 (0.007)	0.212 (0.012)	0.318 (0.018)	0.118 (0.005)	0.843 (0.826;0.860)
XL CF	0.654	0.898 (0.005)	0.802 (0.007)	0.812 (0.007)	0.146 (0.011)	0.254 (0.019)	0.090 (0.005)	0.774 (0.753;0.794)
DRL CF	0.592	0.954 (0.004)	0.802 (0.007)	0.812 (0.007)	0.272 (0.012)	0.386 (0.017)	0.146 (0.005)	0.899 (0.886;0.913)
RL CF	0.580	0.951 (0.004)	0.802 (0.007)	0.812 (0.007)	0.267 (0.012)	0.357 (0.016)	0.143 (0.005)	0.896 (0.882;0.909)
PATH	0.817	0.814 (0.007)	0.802 (0.007)	0.812 (0.007)	0.011 (0.009)	0.024 (0.030)	0.004 (0.006)	0.510 (0.483;0.537)
Causal forests	0.899	0.829 (0.007)	0.802 (0.007)	0.812 (0.007)	0.029 (0.010)	0.170 (0.041)	0.018 (0.005)	0.583 (0.557;0.608)
VT	0.887	0.821 (0.007)	0.802 (0.007)	0.812 (0.007)	0.019 (0.010)	0.069 (0.038)	0.010 (0.005)	—
AL	0.509	0.858 (0.005)	0.802 (0.007)	0.812 (0.007)	0.210 (0.019)	0.189 (0.019)	0.051 (0.004)	0.688 (0.670;0.705)
MCM	0.509	0.858 (0.005)	0.802 (0.007)	0.812 (0.007)	0.210 (0.019)	0.189 (0.019)	0.051 (0.004)	0.688 (0.670;0.706)
OWL	1	0.812 (0.007)	0.802 (0.007)	0.812 (0.007)	0.010 (0.011)	0	0.000 (0.000)	—
CWL	0	0.802 (0.007)	0.802 (0.007)	0.812 (0.007)	0	-0.010 (0.010)	0.000 (0.000)	—

p_r refers to the proportion of patients for which treatment is recommended by the rule. SL: S-learner, TL: T-learner, XL: X-learner, DRL: DR-learner, RL: R-learner, RF: random forests, CF: cross-fitted, VT: virtual twins, MCM: modified covariate method, AL: A-learning, OWL: outcome weighted learning and CWL: contrast weighted learning.

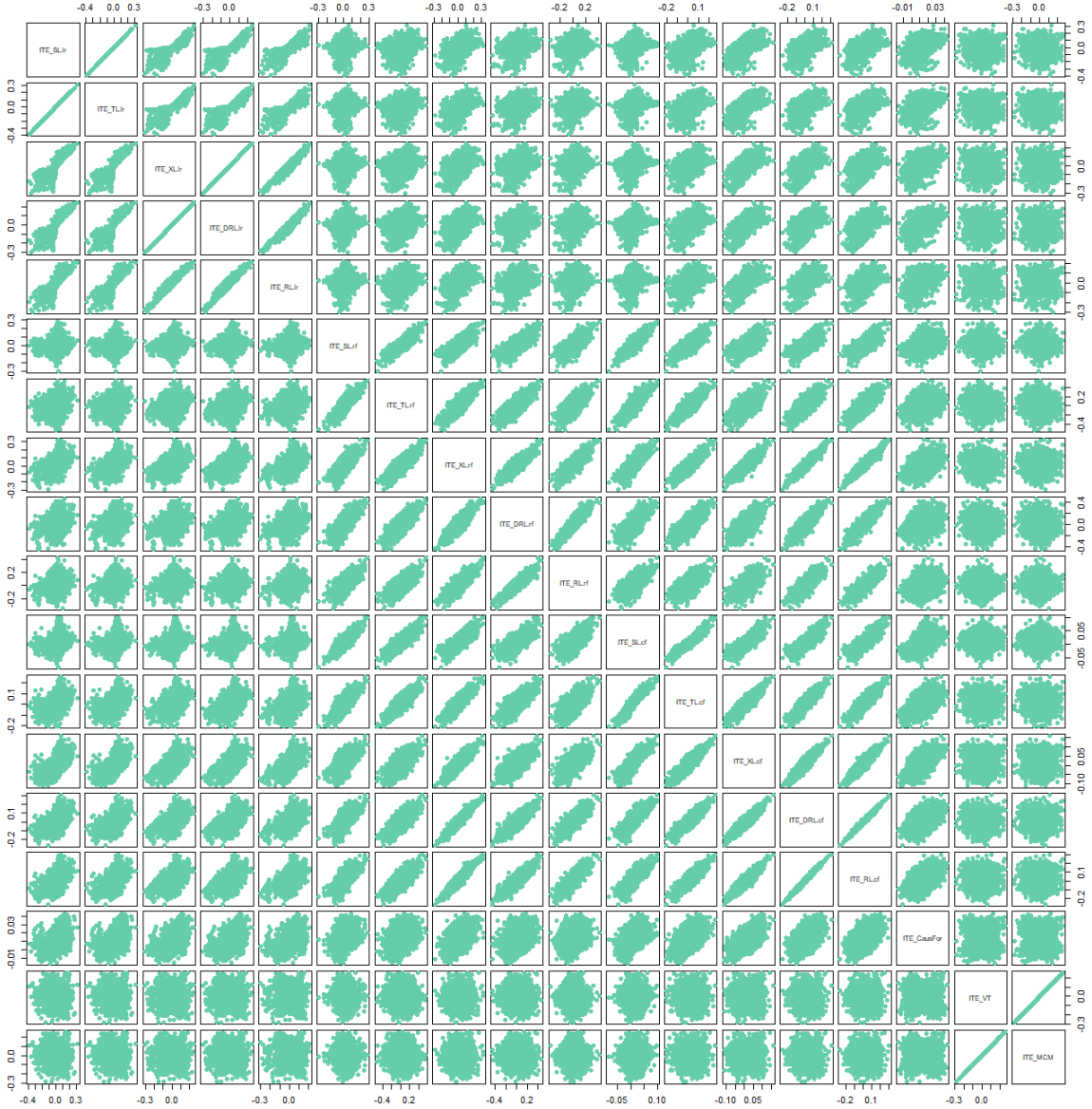

Figure S 3: Matrix of scatterplots representing ITEs using IST validation data.

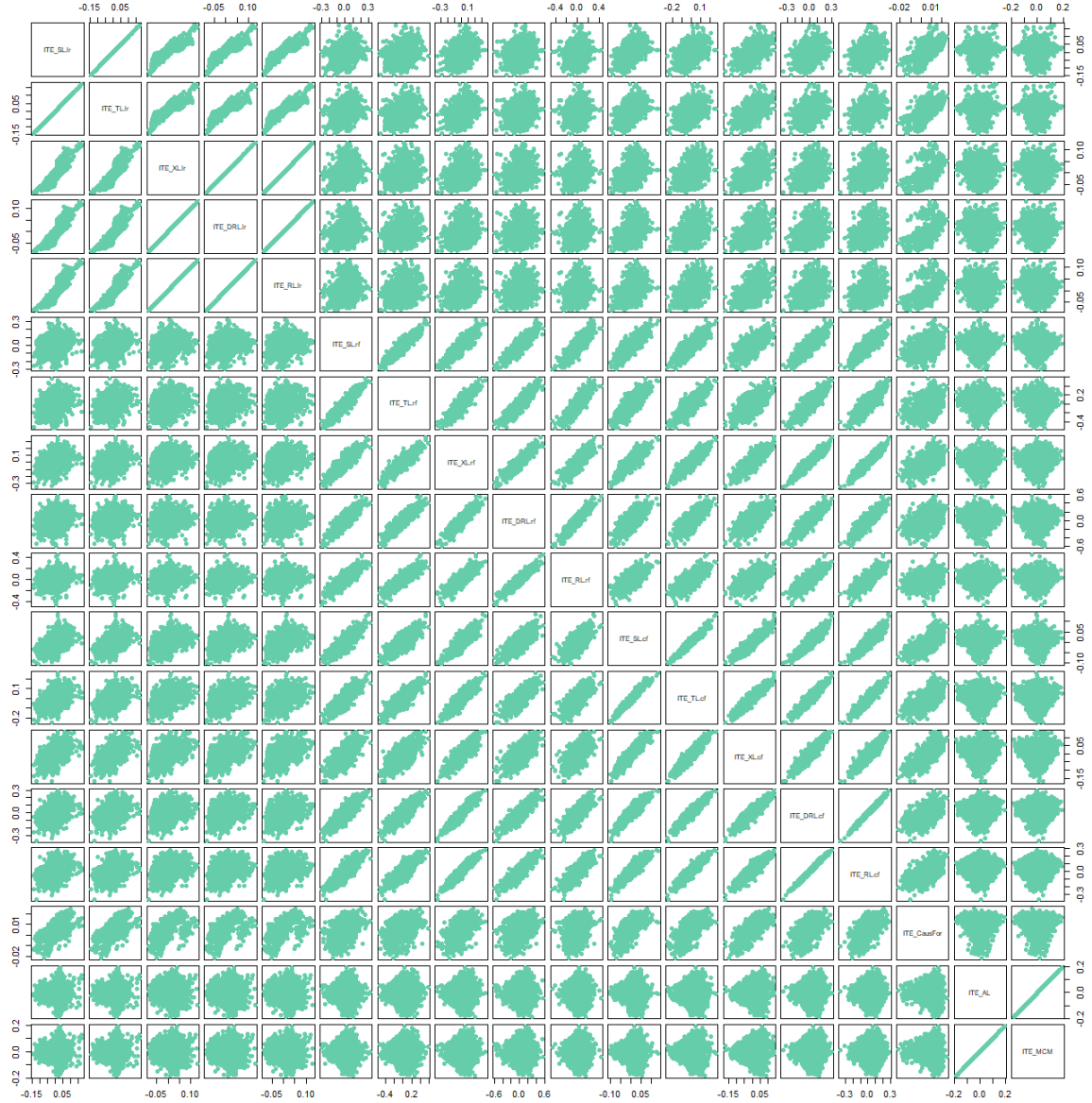

Figure S 4: Matrix of scatterplots representing ITEs using CRASH-3 validation data.

References

1. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. Double/Debiased/Neyman Machine Learning of Treatment Effects. 2017. Number: arXiv:1701.08687 arXiv:1701.08687 [stat].

2. Jacob D. Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects. *arXiv:2007.02852 [stat].* 2020. arXiv: 2007.02852.

3. Nguyen TL, Collins GS, Landais P, Le Manach Y. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials—An illustration with the International Stroke Trial. *Journal of Clinical Epidemiology.* 2020;125:47–56.

What should be done and what should be avoided when comparing two treatments ?

Best Practice & Research Clinical Haematology 36 (2023) 101473

Contents lists available at ScienceDirect

Best Practice & Research Clinical Haematology

journal homepage: www.elsevier.com/locate/issn/15216926

What should be done and what should be avoided when comparing two treatments?

Florie Brion Bouvier [a], Raphaël Porcher [a,b,*]

[a] Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), F-75004, Paris, France
[b] Centre d'Épidémiologie Clinique, AP-HP, Hôpital Hôtel Dieu, F-75004, Paris, France

ARTICLE INFO

Keywords:
Observational studies
Statistics
Causality
Clinical trials
Confounding

ABSTRACT

The preferred approach to compare two treatments is a randomized controlled trial (RCT). Indeed, randomization ensures that the groups compared are similar. Well-designed and well-conducted RCTs thus allow to draw causal conclusions on the relative efficacy and safety of treatments compared. However, it is not always possible to conduct RCTs for all clinical questions of interest, and observational data may also be used to infer on the relative effectiveness of treatments. In this review, we present different approaches that allow statistically valid comparisons of the effectiveness of treatments using observational data under some assumptions. Those are based on regression modelling or the propensity score. We also present the principles of target trial emulation.

1. Introduction

The gold-standard approach to compare two treatments, or more generally to assess the efficacy and safety of an intervention versus a comparator (which could be no intervention at all, a placebo or sham intervention, or another intervention), is a well-designed and well-conducted randomized clinical trial (RCT). There are issues on how to design, conduct and analyze RCTs, from a scientific, ethical or regulatory point-of-view. Those have been summarized in many textbooks [1,2] or documents from regulators, such as guidelines issued by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) (https://www.ich.org/page/search-index-ich-guidelines). Key elements of the validity of an RCT can be classified into internal validity and external validity components [3]. Internal validity relates to selection bias (whether allocation to each group is biased), performance bias (whether other interventions or care differ between groups, apart from the intervention being evaluated), detection bias (whether the outcome is measured the same way in both groups) and attrition bias (whether protocol deviations and loss to follow-up occurred or were handled differently in both groups). External validity refers to the ability of the trial results to be generalized to the target population, and relates to patients' characteristics, treatment regimens, settings in which the trial was conducted (e.g. experience or specialization of centers and care providers), and the definition of outcomes. Assigning the treatment group at random as in RCTs eliminates selection bias, and ensures that the two groups have—on average—the same characteristics, measured or unmeasured. Another key feature is the experimental design. The prospective nature of clinical trials and the requirement to set up a precise protocol allow to limit the other threats to internal validity (sources of bias), for instance using blinding to limit performance and

* Corresponding author. Centre d'Épidémiologie Clinique Hôtel-Dieu 1, place du Parvis Notre-Dame, 75004, Paris, France.
E-mail address: raphael.porcher@aphp.fr (R. Porcher).

https://doi.org/10.1016/j.beha.2023.101473
Received 13 January 2023; Received in revised form 23 February 2023; Accepted 1 May 2023
Available online 6 May 2023

detection bias, specifying precisely how outcomes should be measured, implementing actions to limit attrition, etc. Despite their clear methodological assets, RCTs also have limitations. They are generally very expensive, and require a long time to obtain results, especially when clinical outcomes are mid- or long-term outcomes or rare events. There are also situations where randomization is felt unethical, or will not be accepted by participants or physicians. This can be the case when one wants to compare hematopoietic cell transplant (HCT) to non-HCT; most often there is no randomization, and patients receive HCT or not according to the availability of a donor. It is also not possible, from a financial and logistical point-of-view, to conduct RCTs for pairwise comparisons of all treatments available for a condition, or in all relevant subgroups of patients. Last, there has been evidence that participants to clinical trials generally differ from the target population. This is caused partly by stringent eligibility criteria (e.g. excluding patients vulnerable to adverse effects such as elderly patients or those with comorbidities) [4,5], and also by a selection of participant even within eligibility criteria [6]. Patients in RCTs are also often treated in more experienced centers, and with a better follow-up. This leads to impair the external validity of trials [7,8].

Observational data may provide valuable information to compare treatments [9,10]. In observational studies, the treatment assignment is not controlled by investigators, but simply follows the individual decisions of physicians and patients. Different sources of data can be used. Since those data reflect the usual care practice, observational data are also often referred as real-world evidence. The lack of randomization in observational studies may result in differences in patients' characteristics (observed or unobserved) between the treatment groups. Such differences may bias the estimate of treatment effect. This is called confounding or indication bias. There are however several statistical approaches allowing to estimate treatment effects from observational data. In this article, we review the most popular, and provide some insights to understand how, and under which conditions, they should be used.

2. Potential outcomes and causal assumptions

Comparing two treatments is by essence a causal question [11]. What is of primary interest is the cause-to-consequence relationship between giving one treatment or the other and the outcome, everything else being equal. One common way to formalize this causal effect is the Rubin's potential outcomes or counterfactual outcomes framework [12]. It assumes that each individual has two potential outcomes, one if she/he received the treatment to be evaluated, that we note Y(1), and one if she/he received the comparator treatment (possibly no treatment), Y(0). The causal treatment effect for that individual is Y(1)–Y(0). In practice, there is virtually no way to observe this causal effect, because we cannot observe simultaneously what would have happened if the individual had received the treatment and what would have happened if she/he had received the comparator. We therefore try to estimate the average treatment effect in a population.

The statistical approaches to estimate treatment effects from observational data—also called causal inference methods—rely on several assumptions. Most approaches require four causal assumptions, as detailed below [11].

- *Consistency*: this assumption links the observed outcome under one treatment to the potential outcome. Given the potential outcome framework introduced above, we additionally assume the if an individual received the treatment evaluated her/his observed outcome would be Y(1) and if she/he received the comparator the observed outcome would be Y(0). This assumption seems natural under the potential outcome framework, but it requires in particular that the treatment and comparator are well defined, either at the individual level (e.g. both drugs and regimens are clearly defined), or at the population level (for example the comparator could be "usual care" which may comprise different drugs, but we have to assume that the distribution of those drugs and their dosages is well defined in the population).
- *No interference*: treatment assignment of one individual should not affect the potential outcomes of other individuals. For medical interventions (drugs, surgery, hematopoietic cell transplantation [HCT]) it seems a reasonable assumption. But one may also consider whether patients may discuss side effects of treatments or the relative effectiveness of treatments on forums, for instance, which could influence the perception of the treatment by other patients, and possibly their outcome.
- *Positivity*: all study participants should have a non-null probability of receiving either treatment. More precisely, the study should not include individuals that would always receive one treatment or the other, because of their characteristics. In practice, violations of positivity may also occur by chance, especially in limited samples.
- *No unmeasured confounding* (or *exchangeability*): all confounders, i.e. characteristics associated both with treatment assignment and outcome, should have been measured in the study, so that the analysis can adjust on them.

Methods that do not rely on the no unmeasured confounding exist, like instrumental variables, but are out of the scope of this review.

To adequately estimate treatment effects from observational data, one does assess whether the assumptions above are met, or at least acceptably met. Consistency and no interference are generally assumed, but the investigator should reflect on whether this is reasonable. Positivity is violated when some individuals cannot receive one of the treatments compared, for instance because of contra-indications. Setting clear eligibility criteria, similar to what is done for RCTs, decreases the risk of positivity violations (we will come back to this issue in the section *Target trial emulation*). Positivity can also be checked by examining the distribution of characteristics between groups, and the support of the propensity score (defined below) to ensure it overlaps between groups as illustrated on Fig. 2A.

It is not possible to be certain that all confounders are measured. However, major prognostic factors are usually known, and one can ensure they are used in the analysis. Sensitivity analyses may help assessing the existence of unmeasured confounders and their effect. This can be done by using a negative control, i.e. an outcome that should not be affected by the treatment. A difference between groups on such negative control after adjustment on measured confounders may indicate the existence of unmeasured confounders [13].

Quantitative bias analysis is also increasingly used [14]. One popular approach is to compute the E-value, i.e. the minimum strength of association a confounder should have with treatment assignment and the outcome to fully explain away the treatment effect [15]. For instance, an E-value of 2.5 means that if there was an unmeasured confounder associated with treatment assignment with a relative risk of 2.5 and with the outcome with a relative risk of 2.5, then this unmeasured confounder may cause by itself the treatment effect observed in the study after adjustment for the measured cofounders.

3. What treatment effect? Estimands

An important step when comparing treatments with observational studies is to define the estimand. An estimand describes precisely the treatment effect reflecting the clinical question of interest, summarizing at a certain population-level what the outcomes would be under different treatments in the same patients. The step is important, because when the effect of the treatment is not the same for all individuals, the population-level summary may depend heavily on the target population [16]. For instance, if a treatment is very effective in younger patients but almost not in elderly patients, then the average treatment effect in a younger population would be much higher than in another, older, population. Different estimands exist, with their own target population and interpretation [11]. The following estimands are the most commonly used.

- Average treatment effect (ATE), defined as the average treatment effect on the whole population. It is the difference in mean outcome if everyone had received the evaluated treatment vs. everyone had received the comparator. Mathematically, it is ATE = E [Y(1)–Y(0)].
- Average treatment effect in the treated (ATT), which refers to the average treatment effect among individuals who were assigned to the evaluated treatment, i.e., the difference between the average outcomes observed for the treated patients and the average outcomes they would have obtained if they had not been treated. It can be written as ATT = E[Y(1)–Y(0) | A = 1] where A = 1 denotes assignment to the treatment.

One may also define the average treatment effect in the untreated (ATU), which is symmetrical to ATT. The choice of the estimand depends on the question we aim to answer. For instance, we would select the ATE to know if a treatment should be given to all eligible patients and we would select the ATT to know if the treatment was beneficial for individuals who actually received it.

4. Outcome regression

A common way to correct for confounding is to use a regression model to "adjust" for measured confounders. This is usually done by analysis of covariance for a continuous outcome, logistic regression for a dichotomous outcome, or a Cox model for a time-to-event outcome. In such a model, measured confounders are simply added as covariates in the model. This was for instance the case in a study evaluating the addition of rituximab to chemotherapy as first-line therapy in patients with mantle cell lymphoma aged above 65 [17]. Primary analysis of overall survival (OS) and time to next therapy were based on a multivariable Cox models.

Outcome regression provides a valid causal effect under the aforementioned four causal assumptions, provided the regression model is correctly specified. Those models also assume that the treatment effect is homogeneous (no treatment effect modifiers). A last limitation of this approach is that it provides no warning in case of positivity violation (i.e. the model will provide an estimate), while the risk of bias produced by extrapolation in that case may be important [18].

Another approach based on regression models consists in fitting a regression model for the outcome with confounders as covariates in each of the treatment group, and then to use the model-based predictions to derive average treatment effects [19,20]. This is called *g*-computation. If predictions are made for the whole sample under each model, this is equivalent to predicting Y(1) and Y(0) for each individual of the study, and the difference of the means of predictions of Y(1) and Y(0) estimates the ATE. If predictions are made only for individuals who were assigned to the evaluated treatment, then the ATT is estimated.

5. Propensity score-based approaches

5.1. Propensity score

The other popular approach for causal inference is to rely on the propensity score [21]. The propensity score is the probability of being assigned to the evaluated treatment given the observed covariates. Theoretical results have shown that the propensity score summarizes all relevant information for treatment assignment, and that under the four causal assumptions above, conditioning on the propensity score is sufficient to control for confounding. This is also true for the propensity score estimated on the data.

One can consider that two individuals with the same value of the propensity score only differ by the treatment they were assigned too, so that comparing their outcomes yields a causal treatment effect. In practice, there are different ways to condition on the propensity score that we review below.

After planning the study and predefining the outcome and the set of confounders to adjust for, the first analytic step is to estimate the propensity score from the data. This is usually done by fitting a regression model with the treatment assigned as independent variable, and the confounders as covariates. The most common is logistic regression, but many other approaches, including machine learning, exist [22]. The predicted probabilities from this model are then used for subsequent analysis.

5.2. Propensity score matching

To compare the outcome of individuals conditional on the propensity score, one simple idea is to match study participants from both groups on their propensity score. Participants receiving the treatment evaluated are matched to participants receiving the comparator, with propensity score as close as possible. The most common matching ratio in clinical studies is 1:1 (one control per treated) [23], but sometimes it may be more efficient to use other ratios (e.g., 1:2), depending on the numbers of participants in the evaluated treatment and comparator groups, respectively. Many matching strategies (e.g. with or without replacement) and algorithms exist [24], but are beyond the scope of this review. Other approaches such as optimal full matching, which creates small subsets of individuals with close propensity score values are also valuable [25].

One way to perform matching has been to pair each treated individual with a control with closest propensity score. However, the closest propensity score may not be too close, and the procedure may result in some imbalance on the upper tail of the propensity score distribution. It is therefore more common to only match two individuals if the difference of their propensity score is small enough [24]. This is called *caliper* matching. As a result, not all treated individuals may have a match, especially if matching is done without replacement which is most common. This requires a careful interpretation of the target population, and the final sample size may be markedly lower than the initial one.

The aforementioned strategy was used to compare cord blood (CB) transplantation and matched related donor (MRD) transplantation in non-remission acute myeloid leukemia [26]. In this study, a 1:1 caliper matching was used, based on the nearest neighbor method, and the caliper width was fixed at 0.2 of the standard deviation of the propensity score. Of a total of 2451 patients (1738 with CB and 713 with MRD transplantation), only 918 were matched (459 from each group).

The analysis after matching relies on comparing the outcomes in both matched groups, estimating the treatment effect. Although formally, matching on the propensity score should not induce within-pair correlation on the outcome (the individuals are close in terms of probability of being assigned to the treatment, but not necessarily in terms of prognosis), it has been advised to use variance estimators that account for matching [27,28]. Accordingly, the analysis of OS and other survival outcomes in the study by Shimomura et al. relied on a Cox model, which accounted for matching ("considering parity in a PS-matched cohort").

5.3. Inverse probability of treatment weighting

Another way to use propensity score to balance groups is inverse probability of treatment weighting (IPTW). Each individual is weighted by the inverse of the probability of being assigned the treatment they were actually assigned to, conditional on the individual's characteristics [29,30]. In the treatment group, this corresponds to weighting participants by the inverse of their propensity score value. In the control group, participants are weighted by the inverse of one minus the propensity score value (which is the probability of being assigned to the comparator conditional on covariates). The Fig. 1 illustrates how this works. Of note, the resulting distribution of confounders in both weighted groups resembles the one of the overall study population, so that the estimand here is the ATE. It is also possible to define weights to target ATT. In that case, individuals of the experimental treatment group are given a weight of 1, and those from the control group are given a weight equal to PS/(1 – PS), where PS is used to denote the propensity score. That way, both groups will have a distribution of confounders that matches the one of the treated group, and the treatment effect estimated is the one corresponding to the treated population (ATT). IPTW is valid under the four previous causal assumptions if, in addition, the propensity score model is correctly specified. Another possible issue with IPTW is the influence of extreme weights, which may yield

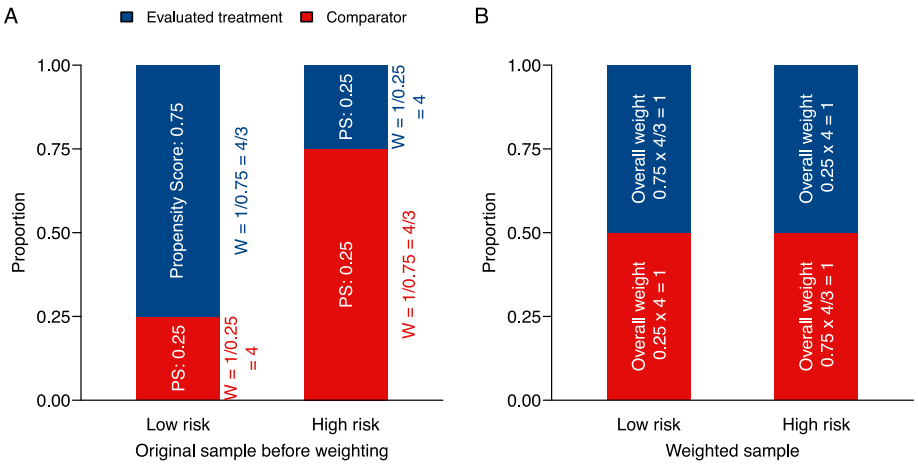

Fig. 1. Illustration of inverse probability of treatment weighting. Assuming risk is the only confounder, given the propensity scores for low risk and high risk patients in the treatment or control group, weights W are obtained (panel A). In the weighted sample (panel B), the overall weight of low risk and high risk patients in each group are balanced. For example, the 25% low risk patients assigned to comparator treatment each receive a weight of 4, so that their overall weight becomes $0.25 \times 4 = 1$. Conversely, the 75% low risk patients assigned to the evaluated treatment only have a weight of 4/3, so that their overall weight is also 1. Within low risk patients, there is the same proportion of treated and control group patients. The same occurs for high risk patients. In the weighted sample, the treated and control groups are thus balanced regarding the confounder.

unstable results. Indeed, since observations are weighted by the inverse of a probability, individuals with propensity score values close to 0 or 1 can have very large weights if they are in the treated or control group, respectively. Two solutions to this exist, one is to trim the sample by excluding individuals with most extreme weights (propensity score trimming) [31], and the other is to truncate extreme weights, so that weights larger than a given threshold value (often the 95th or 99th percentile of the weight distribution) are set at that threshold [32]. While those approaches breach the theoretical properties of IPTW and introduce bias, they may also yield more reliable results in practice.

The analysis of the outcome then simply uses the weighted observations. With time-to-event data, as common in hematology, the analysis often relies on weighted Cox regression. A robust sandwich-type variance estimator has been advocated, but recent work would lead to rather recommend using bootstrapping [33].

The approach was used to assess the effectiveness of COVID-19 convalescent plasma therapy in patients with B-cell lymphoid malignancy and COVID-19 [34]. Individuals with B-cell lymphoid neoplasm included between 1 May 2020 and 1 April 2021 in the convalescent plasma monitored access program implemented in France (n = 81) were compared to patients with the same diseases and hospitalized in French hospitals during the same successive COVID-19 outbreak periods, but who did not receive convalescent plasma (n = 120) using IPTW.

5.4. Other approaches

In addition to the three previously described methods, which are arguably the most commonly found in the medical literature, some other approaches also exist.

One is to use the propensity score in a regression model, instead of all confounders as in section 4. This has the advantage of dimension reduction (only one variable is used instead of many), which may reduce issues related to overfitting, when a large number of confounders are used in a sample with limited effective sample size (e.g. number of events). This approach is valid in a linear model with correctly specified propensity score model, provided the treatment effect is constant across all propensity score values [35]. It may however be problematic in other situations. Such an approach can be found in a study comparing four conditioning regimens for older patients with acute myeloid leukemia receiving allogeneic HCT, where multiple propensity scores were developed using multinomial regression (there were four treatments compared), and used as covariates in a multivariable Cox model [36].

Another method, called subclassification or propensity score stratification, consists in dividing the whole sample into strata according to propensity score values (generally quintiles or deciles of the propensity score distribution), estimate the treatment effect within each stratum, and then pool those different estimates. The method has been popular years ago owing to its computational simplicity, and may work well if there is a reasonable balance of the confounder distribution within each stratum. Otherwise, it is possible to use regression adjustment within strata to correct for the remaining confounding [30].

Last, an approach called double-robust or augmented inverse probability weighting may be used [37]. This approach combines IPTW with regression modelling, where the outcome regression models are used to "augment" the IPTW estimator. It has desirable theoretical properties, and in particular double robustness, which ensures the estimator is consistent if either the propensity score model or the outcome model is correctly specified.

5.5. Assessing propensity score analyses

Propensity scores correct for confounding by achieving balance. Before estimating treatment effects, the "success" of the propensity score in balancing groups should be checked. It is usual to describe matched or weighted groups. To quantify (im)balance, standardized mean differences (SMDs) are often used among other available metrics [38]. This can be found for instance in the table 1 of [26]. The metrics was also used in Ref. [34], although not reported in the main text. Absolute SMDs <0.10 or <0.20 have been described as

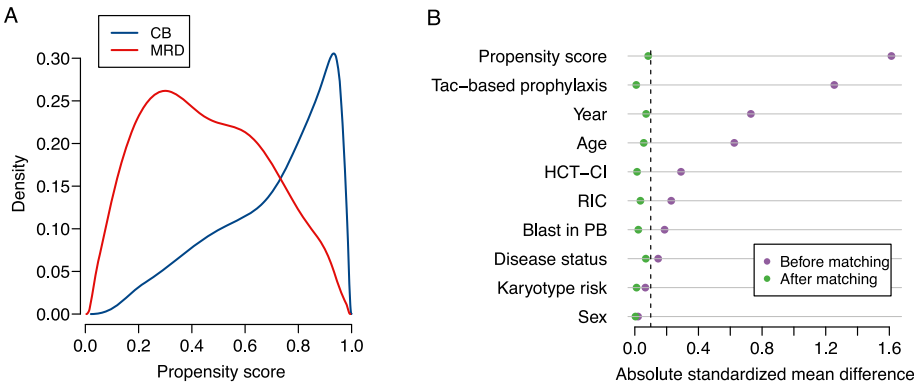

Fig. 2. Illustration of overlap of propensity scores (A) and balance by propensity score matching (B) in a simulated example. The example was simulated to mimic the study by Shimomura et al. [26]. On the panel A, CB denotes cord blood transplantation and MRD matched related donor transplantation. On the panel B, the dashed line showed a standardized mean difference of 0.1, which was chosen by the authors as indicating successful balance.

indicating successful balance [39–41]. An alternative presentation is to display SMDs or absolute SMDs (i.e. their absolute value) in so-called love plots (Fig. 2B).

6. Target trial emulation

Methods described above assume that the assignment to treatment or comparator occurs at a well-defined single timepoint. This may be the case when comparing CB to MRD in the examples previously described. In many situations, however, there is no unique timepoint for decision. This is often the case when comparing a treatment versus no treatment strategy, such as HCT vs. no HCT. Even if it occurs shortly after diagnosis, the time since diagnosis to treatment decision should be handled correctly to avoid time-dependent biases, such as immortal time bias [42–44]. To handle such situations where treatment can be initiated during follow-up, the framework of emulation of a target trial is most useful [45–48]. This consists in first determining the target trial, i.e., a RCT that would ideally be conducted. Then, one emulates this target trial using observational data, using adequate statistical methods to correct for confounding and selection bias, as those described above. In particular, one popular approach is so-called 'cloning and censoring', where participants are cloned, and each clone is allocated to each of the treatment compared. The follow-up of clones is then censored when they deviate from the assigned strategy (Fig. 3). This allows in particular to define a grace period to determine adherence to the assigned treatment group. To correct for selection bias induced by artificial censoring, inverse probability of censoring weighting (IPCW) is used, which is the analogue of IPTW for artificial censoring. This approach also allows to estimate the effect of treatment duration on OS, for instance Ref. [49].

An example of the cloning and censoring approach was used when comparing mortality of hospitalized COVID-19 patients with standard-dose or flexible-dose low-molecular weight heparin thromboprophylaxis [50]. Target trial emulation does not necessarily rely on the cloning and censoring approach and various statistical approaches can be used, as show for instance in studies evaluating the effectiveness of COVID-19 vaccines (e.g. Ref. [51]). Of note, in the example on convalescent plasma therapy in patients with B-cell lymphoid malignancy and COVID-19, the concept of target trial emulation with a 16-day grace period was used, but with a landmark approach [34].

7. Discussion

This review introduced methodological and statistical approaches for comparing treatments using observational data. Of note, similar approaches can be used to derive external control arms for single-arm trials, that are sometimes now the only source of evidence for drug licensing, such as CAR-T cells [52].

We insisted more on what should be done rather than on what should not be done. The most important points are to clearly define

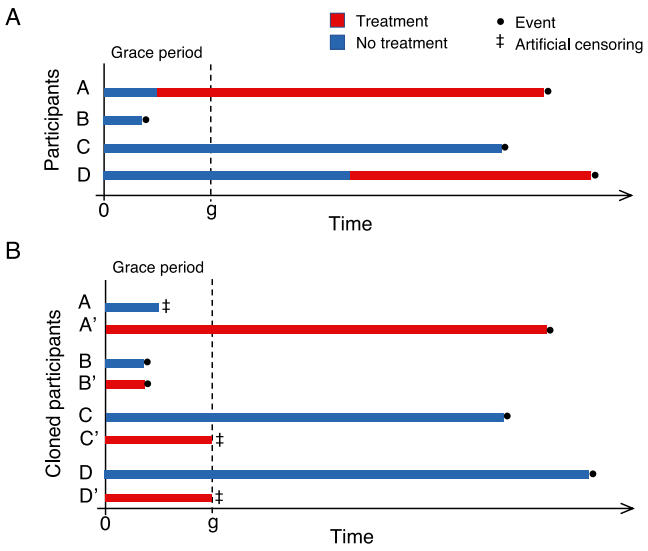

Fig. 3. Illustration of cloning and censoring in an emulated trial. The original data for four fictive participants (A, B, C, D) are given on the panel A. Panel B shows how cloning and censoring is implemented. The grace period (duration g) is used to determine assignment of individuals to one treatment group. All participants are cloned and each cloned is assigned to one treatment group (those denoted by a prime are in the treatment group). Clones are censored when they deviate from the treatment assignment. For instance, clone A deviates from the "no treatment" group when the individual A received treatment during the grace period and is then censored (‡) while the total follow-up of clone A′ is kept because it complied with the treatment group. Conversely the total follow-up of clone C is kept in the no treatment group whereas clone C′ is censored at the end of the grace period, when it becomes certain that the individual will not receive treatment during the grace period. Since individual D was not treated during the grace period, the total follow-up of clone D is kept (no treatment), and the clone D′ is also censored at the end of the grace period, as for clone C'. This would correspond to what would be done in the intention-to-treat analysis of a RCT with a *treatment policy strategy* estimand. The individual B died during the grace period, so we cannot know whether she or he would have received treatment before g. Data are therefore compatible with both treatment strategies; accordingly, both clones B (no treatment) and B' (treatment) are kept.

the clinical question, preferably using a target trial reasoning, and to carefully select potential confounders to be adjusted for. Then, many statistical methods exist for data analysis, each with its pros and cons. One popular approach to compare two treatments in observational studies is to use multivariable regression. Propensity score weighting can improve comparability of the two groups. Of note, confounders should be best selected based on subject matter knowledge, and that data-driven variable selection for the propensity score model using traditional backward procedures or c-statistics is not recommended [53].

The approach outlined above is well adapted to large sample size observational studies.

Difficulties arise when the number of observations is small. In smaller samples, the treatment effect estimates are less precise. When one treatment group is very small, typically less than 50, matching is a good option and can be performed using propensity scores [54]. It has to be noted that matching usually decreases the number of participants and may thereby reduce the power of the study.

The review did not present mathematical details, nor did it mention more complex statistical models, but good tutorials exist for the interested readers [11]. Other approaches than those presented in this manuscript exist, such as instrumental variables, which are in particular valid with unmeasured confounding but necessitate other assumptions. To handle time-varying treatments and confounding, marginal structural models or *g*-estimation have been proposed. These methods are beyond the scope of this review, and we preferred to emphasize the principles of target trial emulation which can be used in the same situations.

Last, issues raised in RCTs, such as performance or attrition bias, may exist in observational studies, although the source of data used may also limit some of those biases.

Summary

Under some assumptions (sometimes referred as *causal assumptions*), it is possible to compare treatments using observational data, by using adequate statistical methods. Most approaches in particular require *no unmeasured confounding*, i.e. that all characteristics associated with both treatment assignment and outcome are measured in the study and adjusted for. Methods to adequately adjust for confounders include the use of regression models and methods based on the *propensity score*. The propensity score is the probability for a particular patient to be assigned to the treatment under study. Balance groups can be achieved either by matching on the propensity score or weighting, using *inverse probability of treatment weighting*.

When conducting such analyses, it is important to check the assumptions, as well as the and success of the propensity score in balancing groups. Sensitivity analyses are also useful.

When planning a causal analysis using observational data, the principles of target trial emulation can be followed, in particular when defining the population, the intervention, the controls and the outcome. It allows to precisely define the estimand, i.e. the treatment effect one wants to estimate. Emulating a target trial also allows to prevent immortal-time bias and other time-dependent biases.

Practice points

- It is possible to compare treatments using observational data, provided adequate statistical methods are used
- Methods rely on several assumptions and, in particular, most require that all confounders are measured and adjusted for in the analysis
- Adequate statistical methods include the use of regression models and methods based on the propensity score, with either matching or weighting
- Principles of target trial emulation are important to carefully frame the clinical question of interest in terms of population, intervention, comparator and outcome, and avoid time-dependent biases

Research agenda

- Treatments, and more generally interventions, should be preferably compared using randomized controlled trials, but observational data may also be used, in particular to assess real-world effectiveness or relevant subgroups of patients
- When using observational data to compare treatments, it is crucial to carefully predefine relevant confounders, and make sure they are available in the data used for analysis

Declaration of competing interest

None.

Acknowledgements

F.B.B. is supported by the French Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

[1] Pocock SJ. Clinical trials: a practical approach. New York: Wiley; 1983.

[2] Senn S. Statistical issues in drug development. Statistics in practice. second ed. Hoboken, NJ: John Wiley & Sons; 2007.
[3] Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001;323:42–6.
[4] Konrat C, Boutron I, Trinquart L, Auleley GR, Ricordeau P, Ravaud P. Underrepresentation of elderly people in randomised controlled trials. The example of trials of 4 widely prescribed drugs. PLoS One 2012;7:e33559.
[5] Buffel du Vaure C, Dechartres A, Battin C, Ravaud P, Boutron I. Exclusion of patients with concomitant chronic conditions in ongoing randomised controlled trials targeting 10 common chronic conditions and registered at ClinicalTrials.gov: a systematic review of registration details. BMJ Open 2016;6:e012265.
[6] Steg PG, Lopez-Sendon J, Lopez de Sa E, Goodman SG, Gore JM, Anderson Jr FA, et al. External validity of clinical trials in acute myocardial infarction. Arch Intern Med 2007;167:68–73.
[7] Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365:82–93.
[8] McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials 2006;7:9.
[9] Concato J, Lawler EV, Lew RA, Gaziano JM, Aslan M, Huang GD. Observational methods in comparative effectiveness research. Am J Med 2010;123:e16–23.
[10] D'Agostino Jr RB, D'Agostino Sr RB. Estimating treatment effects using observational data. JAMA 2007;297:314–6.
[11] Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I. On behalf of" the topic group Causal Inference of the Si. Formulating causal questions and principled statistical answers. Stat Med 2020;39:4922–48.
[12] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66:688–701.
[13] Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 2010;21:383–8.
[14] Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol 2014;43:1969–85.
[15] VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med 2017;167:268–74.
[16] Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol 2006;163:262–70.
[17] Griffiths R, Mikhael J, Gleeson M, Danese M, Dreyling M. Addition of rituximab to chemotherapy alone as first-line therapy improves overall survival in elderly patients with mantle cell lymphoma. Blood 2011;118. 4808-016.
[18] Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 2007;15:199–236.
[19] Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health 2006;60:578–86.
[20] Vansteelandt S, Keiding N. Invited commentary: G-computation–lost in translation? Am J Epidemiol 2011;173:739–42.
[21] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
[22] Alam S, Moodie EEM, Stephens DA. Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. Stat Med 2019;38:1690–702.
[23] Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary JY, Porcher R. Propensity scores in intensive care and anaesthesiology literature: a systematic review. Intensive Care Med 2010;36:1993–2003.
[24] Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. Biom J 2009;51:171–84.
[25] Austin PC, Stuart EA. Optimal full matching for survival outcomes: a method that merits more widespread use. Stat Med 2015;34:3949–67.
[26] Shimomura Y, Sobue T, Hirabayashi S, Kondo T, Mizuno S, Kanda J, et al. Comparing cord blood transplantation and matched related donor transplantation in non-remission acute myeloid leukemia. Leukemia 2022;36:1132–8.
[27] Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Stat Med 2011;30:1292–301.
[28] Gayat E, Resche-Rigon M, Mary JY, Porcher R. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. Pharmaceut Stat 2012;11:222–9.
[29] Rosenbaum PR. Model-based direct adjustment. J Am Stat Assoc 1987;82:387–94.
[30] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med 2004;23:2937–60.
[31] Sturmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution–a simulation study. Am J Epidemiol 2010;172:843–54.
[32] Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol 2008;168:656–64.
[33] Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. Stat Med 2016;35:5642–55.
[34] Hueso T, Godron AS, Lanoy E, Pacanowski J, Levi LI, Gras E, et al. Convalescent plasma improves overall survival in patients with B-cell lymphoid malignancy and COVID-19: a longitudinal cohort and propensity score analysis. Leukemia 2022;36:1025–34.
[35] Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. Stat Med 2014;33:4053–72.
[36] Ciurea SO, Kongtim P, Varma A, Rondon G, Chen J, Srour S, et al. Is there an optimal conditioning for older patients with AML receiving allogeneic hematopoietic cell transplantation? Blood 2020;135:449–52.
[37] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics 2005;61:962–73.
[38] Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. Stat Med 2014;33:1685–99.
[39] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28:3083–107.
[40] Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J Clin Epidemiol 2013;66:S84–90 e1.
[41] Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 2015;34:3661–79.
[42] Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol 2008;167:492–9.
[43] Levesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 2010;340:b5087.
[44] Iudici M, Porcher R, Riveros C, Ravaud P. Time-dependent biases in observational studies of comparative effectiveness research in rheumatology. A methodological review. Ann Rheum Dis 2019;78:562–9.
[45] Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016;183:758–64.
[46] Nguyen VT, Engleton M, Davison M, Ravaud P, Porcher R, Boutron I. Risk of bias in observational studies using routinely collected data of comparative effectiveness research: a meta-research study. BMC Med 2021;19:279.
[47] Hernan MA, Wang W, Leaf DE. Target trial emulation: a framework for causal inference from observational data. JAMA 2022;328:2446–7.
[48] Matthews AA, Danaei G, Islam N, Kurth T. Target trial emulation: applying principles of randomised trials to observational studies. BMJ 2022;378:e071108.
[49] Hernan MA. How to estimate the effect of treatment duration on survival outcomes using observational data. BMJ 2018;360:k182.
[50] Martinez-Ales G, Domingo-Relloso A, Quintana-Diaz M, Fernandez-Capitan C, Hernan MA, Group CH. Thromboprophylaxis with standard-dose vs. flexible-dose heparin for hospitalized COVID-19 patients: a target trial emulation. J Clin Epidemiol 2022;151:96–103.
[51] Horne EMF, Hulme WJ, Keogh RH, Palmer TM, Williamson EJ, Parker EPK, et al. Waning effectiveness of BNT162b2 and ChAdOx1 covid-19 vaccines over six months since second dose: OpenSAFELY cohort study using linked electronic health records. BMJ 2022;378:e071249.

[52] Lambert J, Lengline E, Porcher R, Thiebaut R, Zohar S, Chevret S. Enriching single-arm clinical trials with external controls: possibilities and pitfalls. Blood Adv 2022. **in press**.
[53] Westreich D, Cole SR, Funk MJ, Brookhart MA, Sturmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf 2011;20:317–20.
[54] Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. BMC Med Res Methodol 2012;12:70.