Estimating individualized treatment effects using individual participant data meta-analysis

Florie Brion Bouvier, Anna Chaimani, François Gueyffier, Guillaume Grenet, Raphaël Porcher

Estimating individualized treatment effects using individual participant data meta-analysis

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

[1]Université Paris Cité, Centre of Research Epidemiology and Statistics (CRESS), INSERM U1153, 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, Assistance Publique-Hôpitaux de Paris, Hôtel-Dieu, Paris, France

Abstract

Different approaches can be used to estimate individualized treatment effects with Individual Participant Data Meta-Analyses (IPD-MA). We compared four one-stage models: random effects (RE), stratified intercept (SI), rank-1 (R1) and fully stratified (FS) models, built with two different strategies 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). Simulation results showed that the random effects and the stratified intercept models performed well for both binary and time-to-event outcomes. For the INDANA dataset with a binary outcome, the random effects model had the best performance.

Keywords: Personalized medicine, individualized treatment effects, individual patient data, meta-analysis

1 Introduction

Personalized (or stratified) medicine aims at tailoring a treatment strategy to the individual characteristics of each patient. One key aspect for 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].

Here, we considered meta-algorithms, which are algorithms that decompose the estimation of the ITE into sub-regression problems [2]. We focus on two meta-algorithms: the so-called S-learner and T-learner. Although we did not specifically use machine learning techniques in this work, we kept this terminology, as it fairly reflects the analytical strategy.

The T-learner consists of two steps. First, two regression models are built one in the treatment group and one in the control group. Then, the outcome is predicted for each patient using both models and the ITE is calculated as the difference of both predictions.

The S-learner consists in estimating the treatment effect within a single regression model, where interactions between an indicator variable for the treatment and relevant covariates are introduced. Again, the ITE is calculated as the difference in predictions under both treatments. The S-learner algorithm may reduce overfitting compared to the T-learner algorithm.

In practice, prediction models for ITE are often developed using data from a single randomized controlled trial (RCT) or observational data [3]. RCTs benefit from randomization but are often underpowered for such

a task, which may lead to overfitting or to 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. Previous studies have tackled the incorporation of heterogeneity when estimating the average treatment effect, the average difference of the predicted risk between treatments, or have used IPD-MA to develop prediction models [4, 5]. Fisher et al. [6] and, more recently Chalkou et al. [7], 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 estimate a prognostic model using a one-stage approach. 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 [8]. To our knowledge, no work considered a wider range of methods to estimate ITE or identify individualized treatment rules using IPD meta-analysis data whereas (i) many approaches have been proposed for analysing single study data, and (ii) different meta-analytic models exist for developping risk prediction models [5].

In this study, we aimed to develop a framework that would allow estimating the ITE from an IPD-MA in a one-stage approach with methods focusing on taking into account the heterogeneity in baseline risks. Different methods were compared using both simulated and real data with binary and time-to-event outcomes. We also aimed to compare the S-learner and the T-learner. Section 2 presents the different models and approaches compared in estimating ITEs. In section 3 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 in section 4 [9]. Section 5 concludes with some discussion and paths for future research.

2 Methods to estimate individualized treatment effects

In this section, we describe our framework to develop a prediction model estimating the ITE from an IPD-MA.

To estimate the ITE, we decided to use a one-stage IPD-MA. The one-stage approach consists in analyzing the data from all trials simultaneously.

2.1 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 that the outcome of interest is binary or time-to-event. Different methods to develop a single risk prediction model using IPD-MA have been proposed [4, 5]. Four of them were compared in this work.

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

- Random effects model: A first 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. In the case of a binary outcome, the underlying model can be written as:

$$\text{logit}p_{ij} = \alpha_j + \theta x_{ij},$$

  with $\alpha_j \sim \mathcal{N}(\alpha, \tau_\alpha^2)$ and where p_{ij} refers to the probability of subjet i in trial j to develop the outcome. For survival data, we decided to use a cox regression:

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\theta x_{ij} + \rho_j).$$

  where $\rho_j \sim \mathcal{N}(0, \tau_\rho^2)$.

- Stratified intercept model: A second approach is to include a different intercept for each study, as a fixed effect. With a binary outcome:

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

where $I(\cdot)$ denotes the indicator function. With a time-to-event outcome:

$$\lambda_{ij}(t) = \lambda_{0j}(t) \exp(\theta x_{ij}).$$

- Fully stratified: A third 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:

$$\text{logit}p_{ij} = \sum_{m=1}^{J} (\alpha_m I(m = j) + \theta_m I(m = j) x_{ij}),$$

With a time-to-event outcome:

$$\lambda_{ij}(t) = \lambda_{0j}(t) \exp(\theta_j x_{ij}).$$

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

$$\text{logit}p_{ij} = \alpha_j + \phi_j \theta x_{ij},$$

with $\alpha_j \sim \mathcal{N}(\alpha, \tau_\alpha^2)$, $\phi_j \sim \mathcal{N}(\phi, \tau_\phi^2)$. With a time-to-event outcome:

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\phi_j \theta x_{ij} + \rho_j).$$

2.2 ITE estimation

In this section, we now individualize treatment in the models, represented by a variable z_{ij}. We do not differentiate between the whole set of covariates and effect modifiers. To estimate the ITE with the above models, we use the two meta-learner algorithms. Let us consider the random effects model for instance. Let $\mu(x, z)$ represent the expected binary outcome under treatment z for an individual with covariates x and 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 binary outcome is estimated as:

$$\widehat{ITE} = \hat{\mu}(x_{ij}, 1) - \hat{\mu}(x_{ij}, 0).$$

The ITE for a time-to-event outcome is estimated as:

$$\widehat{ITE} = \hat{S}(t, x_{ij}, 1) - \hat{S}(t, x_{ij}, 0).$$

For S-learner, we use treatment-covariates interactions in the model, such that we obtain the following expression for a binary outcome :

$$logit\mu(x_{ij}, z_{ij}) = \alpha_j + \theta x_{ij} + \gamma_j z_{ij} + \eta x_{ij} z_{ij}.$$

For a time-to-event outcome, we have:

$$S(t, x_{ij}, z_{ij}) = \exp(\theta x_{ij} + \gamma_j z_{ij} + \eta x_{ij} z_{ij} + \rho_j).$$

For T-learner, we build one model using the treatment group and one model using the control group. With a binary outcome:

$$logit\mu(x_{ij}, 0) = \alpha_j^0 + \theta^0 x_{ij},$$

for individuals with $z_{ij} = 0$.

$$logit\mu(x_{ij}, 1) = \alpha_j^1 + \theta^1 x_{ij},$$

for individuals with $z_{ij} = 1$.

With a time-to event outcome:

$$S(t, x_{ij}, 0) = \exp\left(\theta^0 x_{ij} + \gamma_j^0 z_{ij} + \rho_j^0\right),$$

for individuals with $z_{ij} = 0$.

$$S(t, x_{ij}, 1) = \exp\left(\theta^1 x_{ij} + \gamma_j^1 z_{ij} + \rho_j^1\right),$$

for individuals with $z_{ij} = 1$

2.3 Model validation

We used internal-external cross-validation (IECV) to validate the models. In IECV, the model is constructed with $J-1$ studies and validated with the remaining study for each permutation of $J-1$ studies. The intercept of the test dataset is estimated by taking the mean of all intercepts in the train datasets. To assess the performance of the models, both 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, we used the c-statistic for benefit that was proposed by van Klaveren et al. [10]. The c-statistic for benefit corresponds to 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.

For the calibration, the agreement between the observed and the estimated benefit, we divided the predictions into ten bins, a way to make sure we had individuals who were allocated to the treatment and individuals who were allocated to the control. In each bin, we compared the mean of the predicted benefit to the observed benefit and we extracted the intercept and the slope of the regression line. An intercept close to 0 and a slope close to 1 indicates a good calibration. Calibration curves were also plotted when we applied the methods to the INDANA dataset.

2.4 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. 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 proposed by Riley et al. [11]. This methos consists in centering the covariates to their study-specific mean and adding a covariate-mean interaction term that explain 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 the supporting material S1) in which we compared the estimates obtained with the models described in the next section with and without Riley's 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 were similar with and without Riley's method. In their paper, Belias et al. find that using Riley's lead to very small differences [12]. Therefore, we decided to evaluate the performance of the different models without including Riley's method.

2.5 Implementation

All the analyses were performed in R version 3.6.1. The random effects 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 `coxvc` in `coxvc`. Finally, the fully stratified model was developed using `glm` and `coxph` from `survival`.

3 Monte Carlo simulation study

3.1 Setting

The performance of the models and meta-learners were evaluated in a simulation study. We considered 24 scenarios in which we changed the number of covariates, the number of patients in each trial or the type of

outcome. The scenarios are briefly described below, and more details are given in the supporting material S2. 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 $f(x; k, b) = bkx_{ij}^{k-1} \exp(-bx^k)$ with $k = 1.15$ and $b = \frac{50}{\exp(\theta_j x_{ij})^{1.15}}$.

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 scenario 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 and 3 treatment-covariate interactions and a binary outcome. Among the covariates, one of them was binary and the two others were continuous.
In scenario 4 to 6, we computed IPD-MAs with a total number of patients equal to 2800, 1400 and 700 (for simplicity, trials were of identical sample size) composed of 10 covariates (6 binary and 4 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 table S3.

We also tackled the impact of variables' selection on the performance of the meta-algorithms. We performed variables' selection using a Group lasso for scenarios 4 to 6 and 10 to 12 with the stratified intercept model.

3.2 Results

Results of scenarios 7 to 12 and 13 to 18 are available in section 3 of the supporting material. Simulation results for scenarios 1 to 6 are presented in Figure 1 and in Figure 2. In terms of discrimination, for the scenarios with 3 covariates, the models had a similar range of values whether they were built with S-learner or T-learner; FS performed slightly worse than the other models but the difference remained small. The same conclusion can be drawn for the scenarios in which 10 covariates were used. Overall, the choice of the model or the choice of the meta-learner algorithm did not drastically change the results except for R1 and FS. For those models, when the IPD-MA contained 10 covariates, the T-learner algorithm provided no results in 997 simulations when there was 100 patient per trial and in 806 simulations with 200 patients per trial. Regarding calibration, intercept values were more scattered with FS. For the others models, the values were similar. Calibration slopes for RE and SI were more condensed near 1 and the performance of those models was less affected by the change in number of observations. For the mean squared error, we observe similar range of values for RE, SI and R1 whereas the FS has higher MSE values which indicated that it did not perform as well as the other models. Overall, as expected, we obtained a better performance with fewer covariates and a higher number of patients, the comparison between the models was not modified by the number of covariates or the size of the IPD-MA. R1 was unable to produce estimates with the two algorithms for the scenarios in which there were 10 covariates and 1400 or 700 patients. FS was unable to give estimates for the scenarios in which there were 10 covariates and 1400 with T-learner or 700 patients with the two algorithms. The random effects model's results and the stratified intercept model's results are generally more robust given a certain scenario no matter the size of the meta-analysis or the meta-learner algorithm used.

Results of scenarios 19 to 24 with heterogeneity of predictor effects and time-to-event outcomes are given in Figure 3 and Figure 4. RE and SI had the c-statisitic for benefit values closer to 1. The choice of the meta-algorithm did not impact the results. For calibration, we obtained better intercept median values with S-learner in scenarios with 3 or 10 covariates but a more concentrated range of values with T-learner when 10 covariates were included. The slope values were similar with both algorithms but T-learner produced more homogeneous results. Globally, using RE or SI led to better calibration results. The MSE values were comparable for both algorithms except for R1 with 3 covariates where the T-learner algorithm performed better than the S-learner. Regarding the models, R1 and FS gave a lower performance. In conclusion, using the random

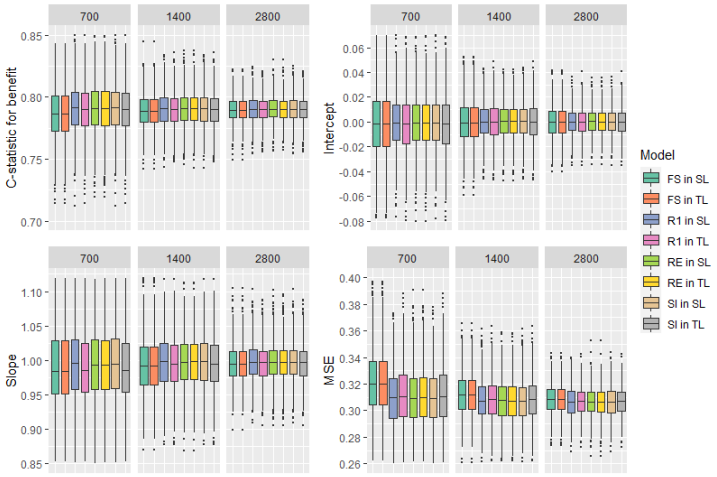

Figure 1: Boxplot of the measures of performance of the models for scenario 1 to 3.

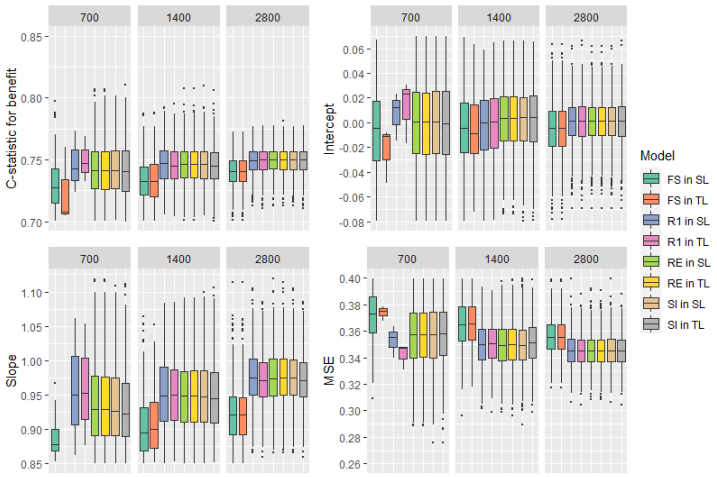

Figure 2: Boxplot of the measures of performance of the models for scenario 4 to 6.

effects or stratified intercept models produced a higher performance. The rank-1 model was unable to generate results in scenarios with 10 covariates.

The performances of the algorithms with and without variables' selection for scenarios 4 to 6 using the stratified intercept model are presented in Figure 5. For discrimination, the c-statistic for benefit values are slightly higher when selection is performed with fewer patients. The algorithms produces similar results. Overall,the intercept values are equivalent with both algorithms. With selection, we obtained better slope results for T-learner. Without selection, the algorithms produced equivalent results. The MSE is lower when selection is done for both algorithms.

The comparison of the algorithms with and without variables' selection for scenarios 10 to 12 using the stratified intercept model is shown in Figure 6. Similar calibration results are obtained for both algorithms. The c-statistic values are higher without selection. For discrimination, the c-statistic for benefit values are slightly higher when selection is performed with fewer patients. The algorithms produces similar results. When the IPD-MA is smaller, using selection leads to better intercept results. Globaly,the intercept values are alike for both algorithms. The range of slope values are smaller without selection. The MSE values are closer to 0 when variables' selection is performed.

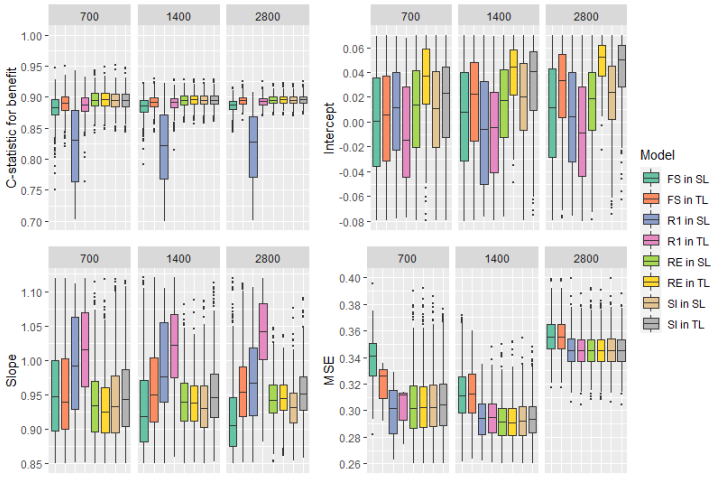

Figure 3: Boxplot of the measures of performance of the models for scenario 19 to 21.

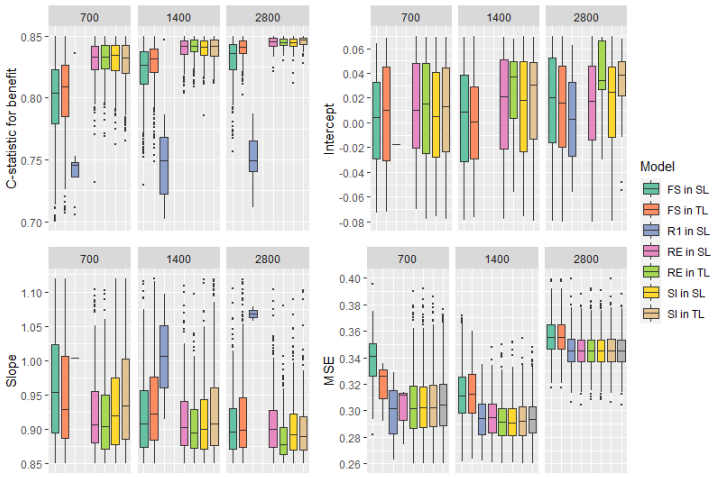

Figure 4: Boxplot of the measures of performance of the models for scenario 22 to 24.

4 Illustration on real data

4.1 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 [13]. This IPD-MA is composed of 9 randomized controlled trials comparing an antihypertensive treatment versus no treatment or a placebo , but given that there is a lot of disparity between trials, notably for the variable age (the figure can be found in in the supporting material S4), we decided to compare the different methods on four of them for which the median age was under 60 years old. The outcome used in this project was death. The dataset has 40 237 observations and 836 deaths. After comparing the calibration obtained with different combinaisons of variables, we decided to include in the final models: the age, the sex, the systolic blood pressure (SBP), the serum creatinine and the treatment group (Table 1). Since some values were missing, we replaced them using a simple run of a multiple imputation procedure [14]. 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 [15]. Proper guidance for estimating ITE is lacking but could be adapted from techniques used for building risk prediction models [16, 17].

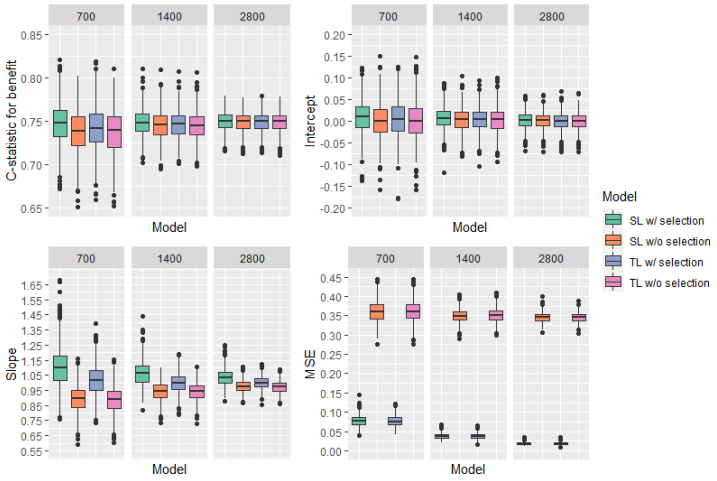

Figure 5: Boxplot of the measures of performance of the models for scenario 4 to 6 with and without variables' selection.

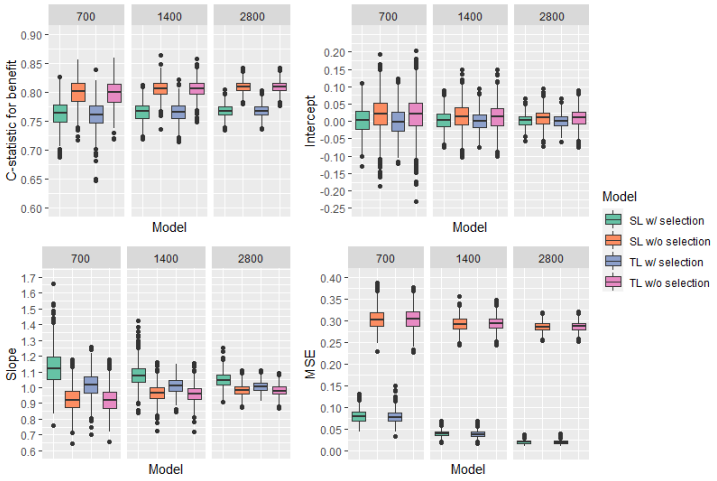

Figure 6: Boxplot of the measures of performance of the models for scenario 10 to 12 with and without variables' selection.

4.2 Results

Results of the models' performance with a binary outcome are displayed in Table 2 and Figure 7. When comparing the discrimination of the different models, we notice that using S-learner gave higher median c-statistic for benefit values than when it was built with the T-learner algorithm. Overall, we observe similar results for the four methods. All the c-statistic for benefit values are around 0.5. van Klaveren et al. mentioned that it is usual to observe a c-statistic for benefit under 0.6 [10]. Moreover, the dataset contains only 836 event for a total of 40 237 observations which could also explains why it was difficlut to obtain models that discriminate well. The median intercept value was close to 0 for every models, with slighlty better results when S-learner was used. With S-learner, RE had a slighlty better median slope and FS gave the values farther from 1. With T-learner, RE had also the median slope closer to 1. SI and R1 gave identical median slope values with both algorithms. In general, median slope values were not really close to 1 which was confirmed by looking at Figure **??** where we can see that some points are not close to the diagonal. The MSE values were close to 0 and comparable for every model whatever meta-learner algorithm was used. Generally speaking, it seems that the random effects model built with S-learner produced the best performance with the INDANA dataset.

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			
	RE	SI	R1	FS	RE	SI	R1	FS
C-stat	0.529	0.530	0.530	0.530	0.507	0.507	0.507	0.510
Intercept	0.001	0.002	0.002	0.001	-0.003	-0.003	-0.003	-0.003
Slope	1.453	1.460	1.460	1.961	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

5 Discussion

Identifying subgroups that benefit from a treatment is an essential part of medical research. Using IPD-MA to build prediction models that estimate the ITE can reduce overfitting and can increase generalizability. In this paper, we compared two meta-learner algorithms and four methods to build prediction models for the ITE with IPD-MA in a one-stage approach. As mentioned before, when a one-stage approach is used, it is preferable to add variables centered to their study-specific mean and a covariate-mean interaction term, as described by Riley et al. [11], to avoid getting aggregation bias; in our simulation study, we found that models without this method did not add bias and hence we decided to not include it for simplicity. Overall, simulation results showed that the random effects and the stratified intercept models give more accurate and harmonious results in terms of discrimination and calibration. When a binary outcome was used, the choice of the meta-algorithm did not have an impact on the results. However, with a time-to-event outcome, the S-learner is preferable in scenarios without heterogeneity or with heterogeneity and few covariates, whereas the T-learner is a better choice in scenarios with heterogeneity and more covariates. The T-learner approach may be considered as allowing nonparametric interactions between the treatment and predictors, and thus these results differ a bit with recent reports noting that effect models with interactions were prone to overfitting [18]. Including variables' selection did not change the performance of the algorithms. Regarding the conclusions that were observed on the INDANA dataset using a binary outcome, the best performance is obtained with the S-learner algorithm and the random effects model. We had hypothesized that in IPD-MA, where the number of predictors is often limited and the sample size large, the issues related to overfitting could be less important. Results confirmed that hypothesis.

The use of IECV allowed to identify the generalizability of the different models. Steyerberg et al. [5], who compared two of the methods present in this paper (random effects and rank-1) to estimate the Average Treatment Effect with an IPD-MA, concluded that rank-1 was the most appropriate method. In this paper, we chose a one-stage approach to estimate the ITE, but a two-stage approach could also have been selected. In a two-stage approach, Fisher et al. [6] advised to only consider within-trial interaction i.e to calculate the difference of predicted outcomes in each trial and then compare the results between trials. Chalkou et al. [7], who used a two-stage approach to estimate the ITE with IPD-MA with 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.

This study has several limitations. We decided to use regression, although other prediction techniques could have been used, among which penalized regression such as the LASSO, random forests, or support vector machine, for instance [1, 19]. 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 elaborate 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. Of note, to estimate heterogeneity in treatment effects, specific LASSO constraints for a support vector machine classifier have been proposed to separate the sparsity constraints between pre-treatment and causal heterogeneity parameters

of interest [19]. Extension of such an approach to IPD-MA may be worth studying. Last, other approaches to treatment personalization exist than those based on ITE prediction. For instance, recursive partitioning methods have been proposed to identify subgroups of patients benefitting from a treatment, that have been extended to IPD-MA [20]. Such methods were however not considered here.

Extensions of the present work could include the use of observational data instead of randomized control trial data. A further extension with observation data would be to develop methods to estimate this type of prediction models (such as rank-1, for instance), whereas allowing the datasets to remain located on different data warehouses, similar to the concept of federated learning [21, 22].

In conclusion, in this paper, we evaluated the performance of different meta-learner algorithms and methods to estimate the ITE with IPD-MA. For the choice of the algorithm, using S-learner, which consists in consist in fitting a single regression and adding the treatment as an indicator variable, leads to slightly better predictions in scenarios where there is a time-to-event outcome with no heterogeneity or heterogeneity and few predictors. For the choice of the method, random effects and stratified intercept are promising approaches.

Acknowledgments

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

Data availability statement

The IPD of the INDANA meta-analysis used for illustration is not routinely available for sharing.

References

[1] N. M. Ballarini, G. K. Rosenkranz, T. Jaki, F. Konig, and M. Posch. Subgroup identification in clinical trials via the predicted individual treatment effect. *PLoS One*, 13(10):e0205971, 2018.

[2] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Acadademy of Sciences U S A*, 116(10):4156–4165, 2019.

[3] V. Farooq, D. van Klaveren, E. W. Steyerberg, E. Meliga, Y. Vergouwe, A. Chieffo, A. P. Kappetein, A. Colombo, Jr. Holmes, D. R., M. Mack, T. Feldman, M. C. Morice, E. Stahle, Y. Onuma, M. A. Morel, H. M. Garcia-Garcia, G. A. van Es, K. D. Dawkins, F. W. Mohr, and P. W. Serruys. 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*, 381(9867):639–650, 2013.

[4] T. P. Debray, K. G. Moons, I. Ahmed, H. Koffijberg, and R. D. Riley. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine*, 32(18):3158–3180, 2013.

[5] E. W. Steyerberg, D. Nieboer, T. P. A. Debray, and H. C. van Houwelingen. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. *Statistics in Medicine*, 38(22):4290–4309, 2019.

[6] D. J. Fisher, J. R. Carpenter, T. P. Morris, S. C. Freeman, and J. F. Tierney. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *BMJ*, 356:j573, 2017.

[7] Konstantina Chalkou, Ewout Steyerberg, Matthias Egger, Andrea Manca, Fabio Pellegrini, and Georgia Salanti. A two-stage prediction model for heterogeneous effects of many treatment options: application to drugs for multiple sclerosis. 04 2020.

[8] Michael Seo, Ian R. White, Toshi A. Furukawa, Hissei Imai, Marco Valgimigli, Matthias Egger, Marcel Zwahlen, and Orestis Efthimiou. Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis. *Statistics in Medicine*, 40(6):1553–1573, dec 2020.

[9] F Gueyffier, F Boutitie, JP Boissel, J Coope, J Cutler, T Ekbom, Robert Fagard, L Friedman, HM Perry, and S Pocock. INDANA: a meta-analysis on individual patient data in hypertension. Protocol and preliminary results. *Thérapie*, 50:353–562, 1995.

[10] D. van Klaveren, E. W. Steyerberg, P. W. Serruys, and D. M. Kent. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. *Journal of Clinical Epidemiology*, 94:59–68, 2018.

[11] 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.

[12] Michail Belias, Maroeska M. Rovers, Johannes B. Reitsma, Thomas P. A. Debray, and Joanna IntHout. Statistical approaches to identify subgroups in meta-analysis of individual participant data: a simulation study. *BMC Medical Research Methodology*, 19(1):183, December 2019.

[13] S. J. Pocock, V. McCormack, F. Gueyffier, F. Boutitie, R. H. Fagard, and J. P. Boissel. 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*, 323(7304):75–81, 2001.

[14] M Quartagno and J R Carpenter. Multiple imputation for ipd meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, 35(17):2938–2954, 07 2016.

[15] Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, February 2011.

[16] Yvonne Vergouwe, Patrick Royston, Karel G.M. Moons, and Douglas G. Altman. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of Clinical Epidemiology*, 63(2):205–214, February 2010.

[17] Angela M. Wood, Patrick Royston, and Ian R. White. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical Journal*, 57(4):614–632, July 2015.

[18] D. van Klaveren, T. A. Balan, E. W. Steyerberg, and D. M. Kent. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *Journal of Clinical Epidemiology*, 114:72–83, 2019.

[19] K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, 7(1):443–470, 2013.

[20] Dipesh Mistry, Nigel Stallard, and Martin Underwood. A recursive partitioning approach for subgroup identification in individual patient data meta-analysis. *Statistics in Medicine*, 37(9):1550–1561, 04 2018.

[21] Y. Wu, X. Jiang, J. Kim, and L. Ohno-Machado. Grid binary logistic regression (GLORE): building shared models without sharing data. *Journal of the American Medical Informatics Association*, 19(5):758–764, 2012.

[22] Chia-Lun Lu, Shuang Wang, Zhanglong Ji, Yuan Wu, Li Xiong, Xiaoqian Jiang, and Lucila Ohno-Machado. Webdisco: a web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association*, 22(6):1212–1219, Nov 2015.

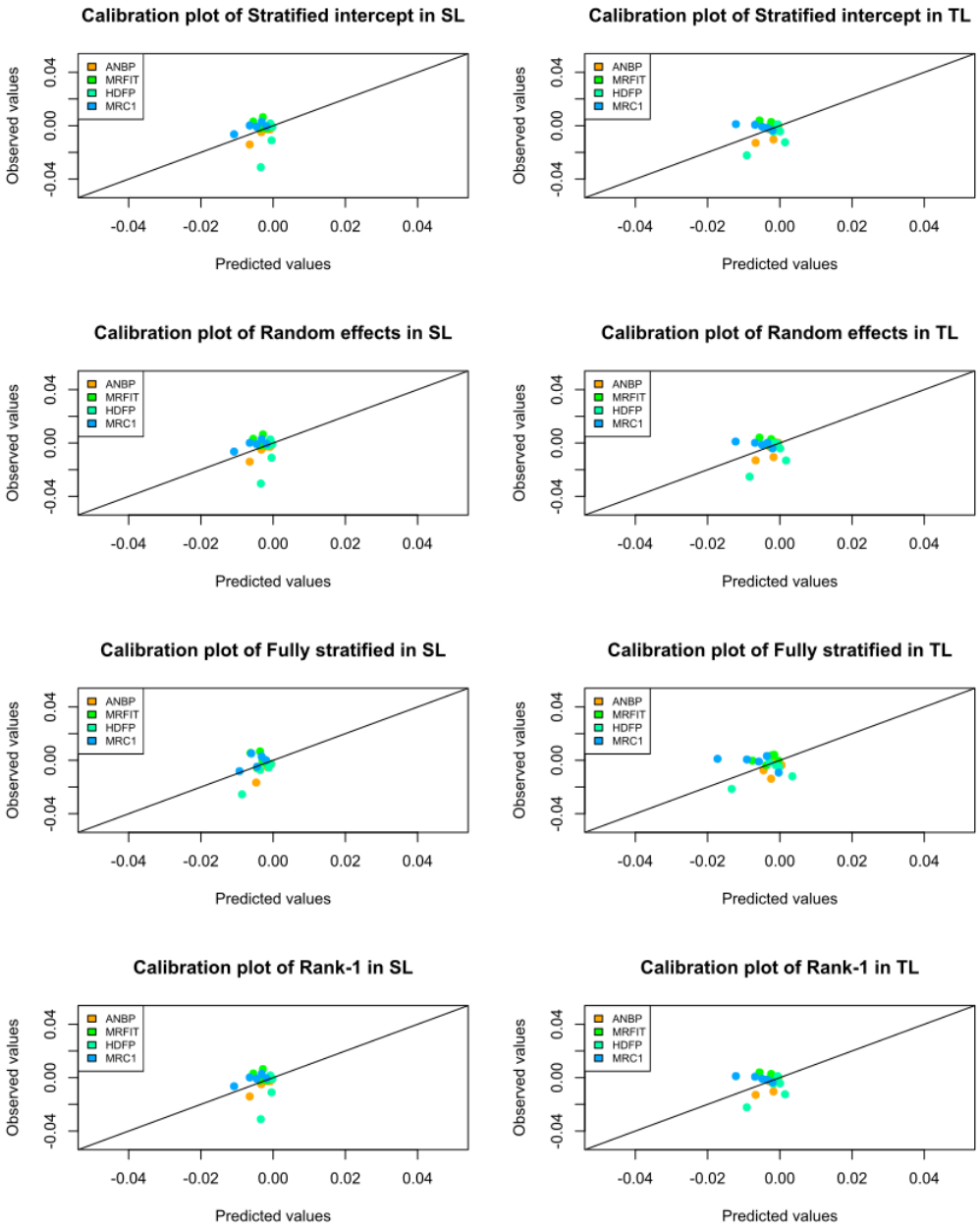

Figure 7: Calibration plots of the models built with S-learner (left) and T-learner (right) using INDANA with a binary outcome.

6 Supplementary material

6.1 Potential aggregation bias

We did not simulated the fully stratified model. Since this model consists in stratifying every parameters by trial, aggregation bias is not an issue.

6.1.1 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 3). 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. [11] were fitted to the data.

Table 3: 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 4: Median parameter estimates (standard errors) over 1000 simulations with the S-learner.

	Model	Random effects		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

6.1.2 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 5: Median parameter estimates (standard errors) over 1000 simulations with the S-learner.

	Model	Random effects		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

6.2 Simulation settings

6.2.1 Covariate generation

In all simulation scenarios, covariates were numbered from x_1 to x_3 or x_1 to x_10, and their distribution varied among the trials of the meta-analysis, as detailed in the tables 6 and 7. Covariates were drawn either from Gaussian distribution with mean μ and standard deviation σ, or from a Bernoulli distributions with parameter π.

Table 6: 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 7: 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)
x_{10}, μ (σ)	6.6 (0.01)	6.5 (0.015)	6.4 (0.011)	6.1 (0.012)	6.4 (0.012)	6 (0.012)	6.4 (0.01)

6.2.2 Simulation scenarios

Table 8: 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	10	2800	No
5	Binary	10	1400	No
6	Binary	10	700	No
7	Binary	3	2800	Yes
8	Binary	3	1400	Yes
9	Binary	3	700	Yes
10	Binary	10	2800	Yes
11	Binary	10	1400	Yes
12	Binary	10	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	10	2800	No
17	Time-to-event	10	1400	No
18	Time-to-event	10	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	10	2800	Yes
23	Time-to-event	10	1400	Yes
24	Time-to-event	10	700	Yes

6.3 Simulation results

Figure 8 and Figure 9 show the results for scenarios 7 to 12 where the predictor effects varied across trials. The c-statistic for benefit values when variation in the predictor effects was included were quite similar between models with a slightly lower range of values for FS. We observed a small increase of performance with S-learner in the scenarios with 3 covariates, especially for SI and R1. No model strategy stood out in terms of discrimination. Looking at calibration, we found that the models' values obtained with S-learner were slightly more gathered around the intended values (0 for the intercept and 1 for the calibration slope). Concerning the MSE, there were no significant differences between the meta-learner algorithms. Choosing FS led to higher MSE and therefore is not recommended. Increasing the size of the IPD-MA led to lower MSE values. Overall, the models showed a good performance. Using the rank-1 model or the fully stratified when the IPD-MA contains a bigger number of variables and less patients procured no estimations. Once again, the scenarios with 10 covariates led to more wide-ranging results. The random effects model and the stratified intercept models provided steadier results across all scenarios.

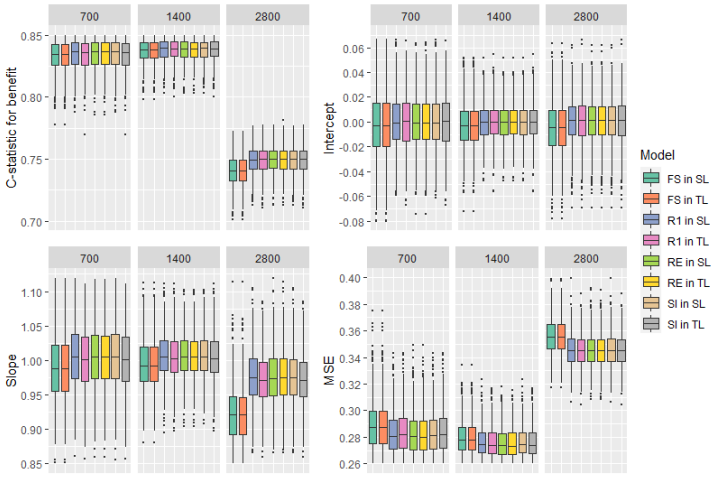

Figure 8: Boxplot of the measures of performance of the models for scenario 7 to 9.

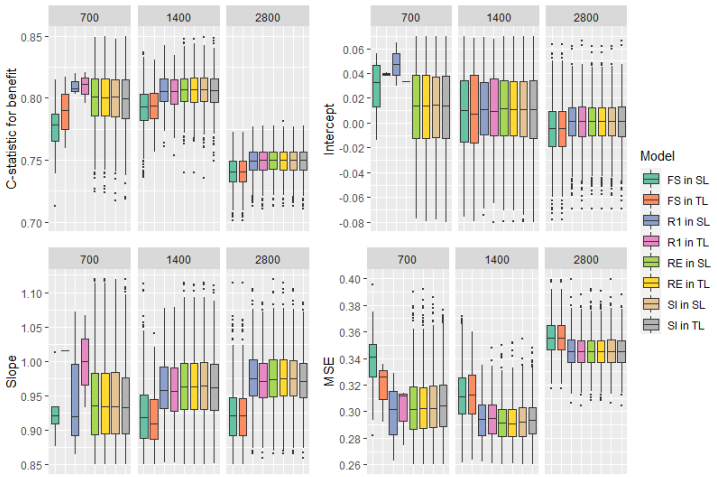

Figure 9: Boxplot of the measures of performance of the models for scenario 10 to 12.

The results of scenarios 13 to 18, which include survival outcomes, are displayed in Figure 10 and Figure 11. RE and SI had the best discrimination with a median c-statistic for benefit value of 0.86 in the scenarios with 3 covariates and a median value between 0.80 and 0.82 in the scenarios with 10 covariates. R1 and FS had a lower and more heterogeneous range of values. For calibration, the values closer to 0 and to 1 ,for the intercept and the slope respectfully, are obtained with SI built with the S-learner algorithm. However, the T-learner algorithm gave more homogeneous calibration results in scenarios with 10 covariates. The best MSE was reached using the RE and SI models with S-learner. In conclusion, with no heterogeneity of predictor effects across trials, the stratified intercept model built with the S-learner algorithm seems to perform better.

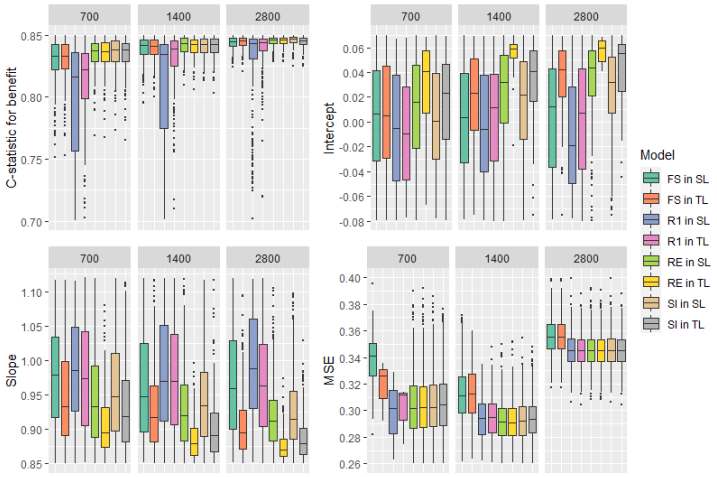

Figure 10: Boxplot of the measures of performance of the models for scenario 13 to 15.

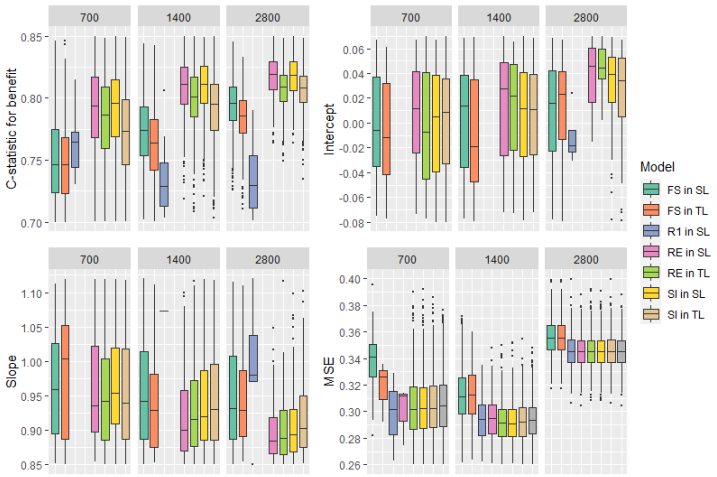

Figure 11: Boxplot of the measures of performance of the models for scenario 16 to 18.

6.4 INDANA IPD-MA

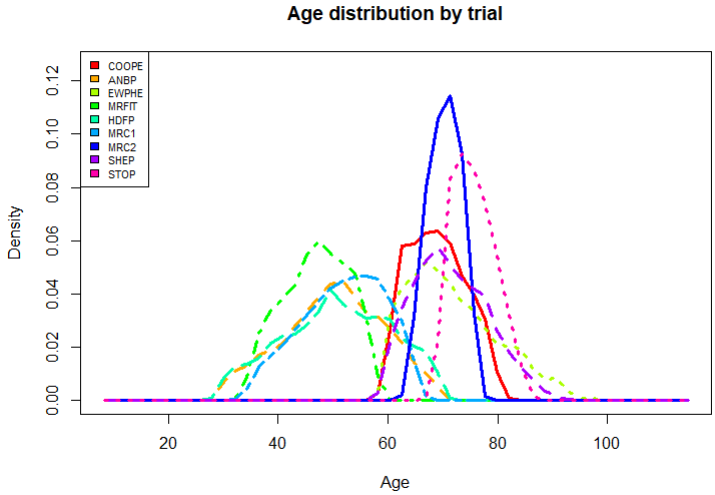

Figure 12: Distribution of age in each trial of INDANA.