Original research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 9: 1–11

DOI: 10.1177/20552076231191967
journals.sagepub.com/home/dhj

Sage

A SuperLearner-enforced approach for the estimation of treatment effect in pediatric trials

Danila Azzolina[1], Rosanna Comoretto[2], Liviana Da Dalt[3], Silvia Bressan[3] and Dario Gregori[4]

Abstract

Background: Randomized Clinical Trials (RCT) represent the gold standard among scientific evidence. RCTs are tailored to control selection bias and the confounding effect of baseline characteristics on the effect of treatment. However, trial conduction and enrolment procedures could be challenging, especially for rare diseases and paediatric research. In these research frameworks, the treatment effect estimation could be compromised. A potential countermeasure is to develop predictive models on the probability of the baseline disease based on previously collected observational data. Machine learning (ML) algorithms have recently become attractive in clinical research because of their flexibility and improved performance compared to standard statistical methods in developing predictive models.

Objective: This manuscript proposes an ML-enforced treatment effect estimation procedure based on an ensemble SuperLearner (SL) approach, trained on historical observational data, to control the confounding effect.

Methods: The REnal SCarring Urinary infEction trial served as a motivating example. Historical observational study data have been simulated through 10,000 Monte Carlo (MC) runs. Hypothetical RCTs have been also simulated, for each MC run, assuming different treatment effects of antibiotics combined with steroids. For each MC simulation, the SL tool has been applied to the simulated observational data. Furthermore, the average treatment effect (ATE), has been estimated on the trial data and adjusted for the SL predicted probability of renal scar.

Results: The simulation results revealed an increased power in ATE estimation for the SL-enforced estimation compared to the unadjusted estimates for all the algorithms composing the ensemble SL.

Keywords

Clinical trials < studies, machine learning < general, paediatrics < medicine, treatment effect, SuperLearner

Submission date: 24 April 2023; Acceptance date: 18 July 2023

Introduction

A Randomized Clinical Trial (RCT) is a study design in which participants are randomly assigned to two or more clinical treatments. RCT is the most rigorously designed hypothesis testing method and is considered the gold standard in clinical research to evaluate the effects of treatments.[1]

In some research settings, the conduct and management of a clinical trial can represent a challenge,[1] due to poor retention or accrual problems, which can negatively impact the quality of study data and increase costs. The literature showed that the leading reason for early failure of

[1]Department of Environmental and Preventive Science, University of Ferrara, Ferrara, Italy
[2]Department of Public Health and Pediatrics, University of Turin, Turin, Italy
[3]Department of Women's and Children's Health, University of Padova, Padova, Italy
[4]Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova, Italy

Corresponding author:
Danila Azzolina, Department of Environmental and Preventive Science, University of Ferrara, Via Fossato di Mortara 64B, 44121, Ferrara, Italy.
Email: danila.azzolina@unife.it

RCT is insufficient enrollment or retention, with a prevalence of the phenomenon ranging from 33.7% to 57%.[2,3]

This problem is evidenced not only in adult RCTs, especially in the oncology[4–6] and cardiology[7] field, but especially in pediatric research due to difficulties in the enrolment process or withdrawal of informed consent. The literature demonstrated that 37% of pediatric RCTs are closed early for inadequate retention.[8]

The management of a pediatric trial is more challenging compared to an adult one[9] due to practical, ethical, and methodological issues.[10] From a practical point of view, this kind of RCT could be initiated only if a favorable benefit-risk balance assessment has been performed for adult experimentation. Furthermore, the regulatory agencies US Food and Drug Administration and European Medicines Agency require that specific plans have to be approved during adult experimentation.[11,12] Ethical issues are also involved because trials are addressing a vulnerable population and oftentimes the focus is on a limited number of patients, as is the case in rare diseases. For this reason, the withdrawal of consent is a very sensitive issue in this research framework.[13] The limited number of patients, together with ethical complications related to the acceptability of the RCT, leads to difficulties in the enrollment of patients.[14] In addition, the unbalanced loss to follow-up in the RCT could involve a systematic attrition bias influencing the statistical power of the study and the balance of confounders between the groups.[15]

All these issues lead to a reduction in the study sample size which can significantly affect the RCT power, compromising the possibility of responding to the primary research question due to a reduction in the likelihood of identifying a treatment effect.[16]

Furthermore, failures in determining the study outcome due to patient dropout in those trials where outcome assessment occurs at follow-up visits can alter the balance of baseline characteristics in patients enrolled in the trial.[15] Attrition bias can occur as a possible result of systematic causes of study withdrawals that affect a certain group of patients. If a cause of withdrawal is prevalent in the intervention arm, the withdrawal imbalance could affect the trial results.[17]

Data missing in a study due to dropouts may cause the traditional statistical analysis of complete or available data to produce a misleading result,[18] especially in the paediatric research setting.[10]

For these reasons, innovative approaches to the analysis of pediatric RCTs have recently been largely supported in the scientific community[11,12] and regulatory agencies.[19–21]

Observational data, for example, can be used as support for experimental research, especially for trials characterized by small sample sizes. The use of observational data that enforces the evaluation of the RCT outcome is proposed in the literature in a Bayesian framework using historical data to define objective priors on the effect of treatment.[22–24] The efficiency of inference can be improved by using external data recovered from historical studies and also from a frequentist point of view.[25] Furthermore, the machine learning (ML) approach has recently been used to combine observational data improving conditional Average Treatment Effect (ATE) estimation and handling possible unmeasured confounding.[26] ML techniques have so far been particularly appealing for their ability to profile the clinical response of a patient and define risk profiles specific to patient characteristics in a precision medicine approach.[27]

Given these premises, this study proposes an ML-enforced treatment effect estimation procedure based on an ensemble SuperLearner (SL) approach to control the confounding effect related to a possible selective dropout mechanism. The method could be applied to adjust the analysis of treatment effect on trial data where the research setting is characterized by observed intercurrent and difficulties in outcome assessment and patient retention.

A paediatric trial candidate for early termination due to under-recruitment, the REnal SCarring Urinary infEction (RESCUE) trial serves as a motivating example.[28] The manifestation and risk profile of renal scars in pediatric patients affected by urinary tract infection (UTI), has been discussed in the literature by evidencing a disease prevalence of 15.6%.[29] In our simulation experiment, the observational data have been generated by assuming a renal scar patient-specific risk profile that follows the disease mechanism explained in the aforementioned research article.[29] The effect of antibiotics combined with steroid treatment has been assumed in simulated RCTs considering different ATE for the treatment arm to prevent renal scarring in pediatric patients affected by UTI. The simulation study proposed in this manuscript reports the performances of the proposed SL-enforced estimation procedure.

Materials and methods

Motivating example RESCUE trial

The RESCUE trial was a randomized controlled double-blind trial.[28] The purpose of this study was to evaluate the effect of adjunctive oral steroids on preventing renal scarring in young children and infants with febrile UTIs compared to antibiotics alone.

The primary outcome was the difference in scarring proportions between treatment arms. The study has been designed in a frequentist setting. By protocol, a sample size of 92 randomized patients per arm was required, also considering 20% of losses to follow-up.

After two years of study conduction, only 18 recruited patients completed the follow-up to determine the study outcome, which resulted in a loss in the final power of 63%.

According to the protocol, some issues raised in trial conduction for the outcome assessment (presence of a

renal scar on renal scintigraphy) occurred at the 6-month follow-up. However, during the study, several patients were lost to follow-up, as parents thought that the scintigraphy at 6 months was useless after the resolution of the acute UTI episode.

The enforced SL-based estimation of ATE

Algorithm description. The SL-enforced estimation algorithm could be applied to enforce ATE estimation in RCT by using an ML technique trained on external observational data (registries, retrospective studies, etc.).

Taking into account a hypothetical RCT aimed at evaluating the treatment effect of a new therapy compared to a placebo or the standard of care, the SL-enforced estimation procedure could be applied to analyze the RCT data, especially in the case where:

1. external observational data defining the disease mechanism are available to train an SL predictive algorithm; and
2. the RCT study setting is characterized, as the RESCUE trial and other pediatric trials, by challenges in patient enrollment and retention. Further details concerning the algorithm have been included in the Supplementary Material. In this research setting, it could be useful to use an algorithm trained on external data to enforce the ATE estimation performance on a new RCT.

Once the feasibility of the procedures for the trial under evaluation has been assessed, the estimation can be initiated; it consists of three different phases (Figure S1, Appendix).

✓ **Step I - SL training.** An SL algorithm is trained and tuned to profile the disease risk profile according to the patient's characteristics on the external observational data even before initiating the planned RCT. This step is useful for estimating disease probability, according to patient characteristics, for new patients to be enrolled in the new RCT.

✓ **Step II - Predict the baseline disease risk profile via SL on RCT patients.** When the new trial begins, the disease probability is predicted on the new patient enrolled in the RCT using the SL tool previously estimated in the observational dataset. A specific patient disease risk profile is stored according to SL prediction. The baseline covariates are considered to estimate the SL model

✓ **Step III - SL adjusted ATE estimation.** Once the new trial is terminated and the data are available for analysis, the estimation procedure is performed by adjusting the ATE estimation in a Logistic regression model treatment by including an adjustment in the final analysis:

(a) The probability of disease is predicted via SL, as reported in step II. The prediction is included as a covariate in the Propensity Score (PS) estimation model together with the dropout indicator. The calculated PS identifies the probability of receiving the treatment according to the patient's specific estimated disease profile and dropout mechanism.

(b) The Inverse of Probability Treatment Weight (IPTW) analysis, to account for a possible imbalance related to the dropout mechanism, is performed in the final analysis.[30] IPTW analysis involves assigning weights to each individual in the dataset based on the inverse of their PS. The idea is to give more weight to individuals with a low probability of receiving the treatment if they are in the treatment group and vice-versa.

(c) The weights w_i are calculated for each i patient by considering this estimation scheme related to the estimated patient-specific propensity PS_i:

$$w_i = 1 / PS_i, \text{ if treated } (T_i = 1)$$
$$w_i = 1 / (1 - PS_i), \text{ if not treated } (T_i = 0).$$

The calculated weights are included in the final treatment effect estimation model which is an IPTW-weighted Logistic regression approach.

All the steps of the analysis were reproduced on synthetic datasets (trial_final.Rdata, observational.Rdata) in an Rmd file (Pseudo_analysis.Rmd) with R code and a report (Pseudo_analysis.html). The files are attached as supplementary files.

Simulation experiment. A Monte Carlo (MC) simulation experiment consisting of 1000 runs has been proposed to evaluate the performance of the SL-enforced estimation method. Each run consists of:

1. Observational and RCT data generation;
2. Unadjusted and SL-enforced ATE estimation; and
3. Calculation of ATE estimation performance measures.

The simulations have been conducted by assuming a Per-Protocol analysis, as a worst-case scenario for analyzing the clinical trial data. Moreover, the simulation results for an Intention To Treat analysis have been also reported in the Appendix.

Data simulation. **Observational data.** A hypothetical historical data set[29] with a sample size of ($n = 1280$) reporting the effects of patients' characteristics on the renal scar probability has been generated in the MC experiment by assuming the covariate effect and disease mechanism as indicated in Shaikh, 2014.[29]

The renal scar event data Y_i for the $i - th$ subject has been simulated from a Binomial random variable $Y_i \sim Binomial(n\ Lp_i)$. Specifically, the probability of the disease has been simulated by assuming a linear predictor

Lp_i defined as follows:

$$Lp_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \beta_7 X_{i7} + \beta_8 X_{i8} + \beta_9 X_{i9} + \beta_{10} X_{i10} + \beta_{11} X_{i11} + \beta_{12} X_{i12} + \beta_{13} X_{i13}.$$

The baseline log odds of renal scar event is $\beta_0 = 0.156/(1\text{-}0.156)$; where 0.156 is the baseline disease probability. The covariates are ($\boldsymbol{X}$) and the ($\boldsymbol{\beta}$) are log OR (Odds Ratio) effects included in the data generating model where:

- $X_1 \sim Binomial\left(n\,;\,\frac{1057}{1057+221}\right)$ is the Age lower than 24 months with $\beta_1 = \log(0.53)$;
- $X_2 \sim Binomial\left(n\,;\,\frac{827}{827+452}\right)$ is the gender female with $\beta_1 = \log(1.13)$;
- $X_3 \sim Binomial\left(n\,;\,\frac{509}{509+549}\right)$ *is the fever* $\geq 39°$*with* $\beta_3 = \log(2, 29)$;
- $X_4 \sim Binomial\left(n\,;\,\frac{376}{376+682}\right)$ *is the* duration of fever (Days) with $\beta_4 = \log(1.11)$;
- $X_5 \sim Binomial\left(n\,;\,\frac{363}{363+494}\right)$ *is the* PMN count >60% with $\beta_5 = \log(1.91)$;
- $X_6 \sim Binomial\left(n\,;\,\frac{512}{512+451}\right)$ is the CRP level >40 mg/L with $\beta_6 = \log(3.06)$;
- $X_7 \sim Binomial\left(n\,;\,\frac{224}{224+902}\right)$ is the presence of an organism other than *Escherichia coli* in the urine examination with $\beta_7 = \log(3.79)$;
- $X_8 \sim Binomial\left(n\,;\,\frac{200}{200+884+112+51}\right)$ is the I and II VUR Grade with $\beta_8 = \log(1.82)$;
- $X_9 \sim Binomial\left(n\,;\,\frac{112}{200+884+112+51}\right)$ is the III VUR Grade with $\beta_9 = \log(3.56)$;
- $X_{10} \sim Binomial\left(n\,;\,\frac{51}{200+884+112+51}\right)$ is the IV and V VUR Grade with $\beta_{10} = \log(22.48)$.

Other latent effects are included in the data-generating model by considering three latent standardized Normal random variables (X_{11}, X_{12}, X_{13}) with log OR effects equal to $\beta_{11} = \beta_{12} = \beta_{13} = \log(1.4)$.

Clinical trial data. The clinical trial data has been generated by assuming the same data generation mechanism as provided in the observational data design stage. Sample sizes per arm n_{RCT} have been assumed to range from 50 to 220. A random treatment allocation T with a balanced assignment has been provided. A dropout mechanism has been assumed to be informative depending on the patient characteristics and treatment assignment as $Binomial(n_{RCT}\,;\,p)$ where $p = 1/(1 + exp(-t\gamma))$ and $t\gamma = log(0.5/(1-0.5)) + log(\gamma_{drop}) * T$. Several dropout effects γ_{drop} have been assumed to range from 1.2 to 1.3.

ATE generation. The ATEs have been simulated. ATE, in an RCT with complete compliance with the treatment, is $ATE = E[Y^1|T_i = 1] - E[Y^0|T_i = 0]$ which is the difference in outcome expectation for each subject i if treated $ATT = E[Y^1|T_i = 1]$ and under the counterfactual scenario if untreated $ATU = E[Y^0|T_i = 0]$.

The loss to follow-up mechanism alters the balance of the experimental conditions among patients; therefore, the ATE is simulated as:

$$logit(ATE) = \underbrace{\log(\pi_{HTE})Lp + \log(\theta)T}_{\text{Heterogeneous Treatment Effect}}.$$

The θ parameter defines the treatment T effect OR assumed as a reduction in the probability of scars ranging from 40% to 20% (OR = 0.6, 0.7, 0.8). The heterogeneity in the treatment response π_{HTE} depends on the disease probability mechanism defined according to the baseline characteristics of the patient Lp. The heterogeneity parameter has been assumed to be equal to $\pi_{HTE} = 1.4,\ 1.5$.

ATE estimation and performance calculation. For each simulation it has been calculated *i*) the Mean Absolute Percent Error (MAPE) defined as , where the true ATE is θ, the estimated ATE is $\hat{\vartheta}$ and n is the number of MC runs; *ii*) the ATE estimation p-value; *iii*) the $\hat{\vartheta}$ estimated ATE.

The ATE within the simulations has been estimated by considering the procedure reported in the algorithm description paragraph.

The weighting system accounts for all the SL algorithms that make up the ensemble SL. The algorithms composing the SL are Gradient Boosting Machine, Improved PREDictive BAGGing classification tree, RANGER Random Forest (RANGER), Generalized Linear Model (GLM), Kernel Support Vector Machine (KSVM), eXtreme Gradient BOOSTting (See Appendix for details). Unadjusted estimates have also been considered for comparison purposes.

Pooling of MC results. After the simulation procedure was completed, the performance of the proposed method was pooled by performing a median and interquartile range (IQR) of the estimated ATE across MC replications; the percentage of trials ensuring a significant ATE, over the MC simulated data have also been calculated.

Calculations have been performed using the R System ver. 3.4.2[31] SuperLearner with the[32] package.

Results

The simulation results revealed a gain in the ability to truly detect the ATE by considering the SL-enforced estimation

compared to the unadjusted estimates for all the algorithms composing the ensemble SL and all the sample sizes.

In all simulation scenarios, the ability to truly detect a significant ATE increases with the sample size; the effect is more evident for an ATE of 0.6 and higher HTE confounding effects and dropout rates. The enforced SL estimation procedure for an ATE of 0.6 rapidly increases the empirical power for a sample size per arm higher than 220 per arm. The unadjusted method is not able to achieve this ATE identification ability in general scenarios. In all cases, an increase in overall power is evidenced, even if minimal, for smaller effects (Figure 1). A similar pattern is evidenced in the Intention To Treat Analysis, ensuring also a slight improvement in the ATE identification ability in comparison with the Per-Protocol Analysis (Figure S2, Appendix).

The decrease in estimation variability defined as the IQR in the MC simulation is evident for all estimation methods considered and all the scenarios of ATE, heterogeneity, and dropout.

The IQR identifies the variability of the point estimates calculated on the simulated trial. A lower variability (IQR) expresses a more efficient treatment effect estimation procedure. The efficiency increase as increases the sample size and is similar for adjusted and unadjusted methods (Figure 2). Similar patterns are reported in the Intention To Treat Analysis are evidenced (Figure S3, Appendix)

The MAPE indicates the, mean across simulated trials of the relative error expressed in percentage. The indicator is a proxy of the bias of the proposed treatment effect estimation. The proposed scenarios showed similar findings for all methods, highlighting certain volatility between results (Figure 3). Another bias indicator is the median estimated ATE across simulations, compared to the assumed true effect. Also, in this case, the indicator evidenced a comparable performance across estimation methods (Figure 4). The results concerning the bias and the ATE estimation are similar in shape for the Intention To Treat approach (Figure 4-5, Appendix)

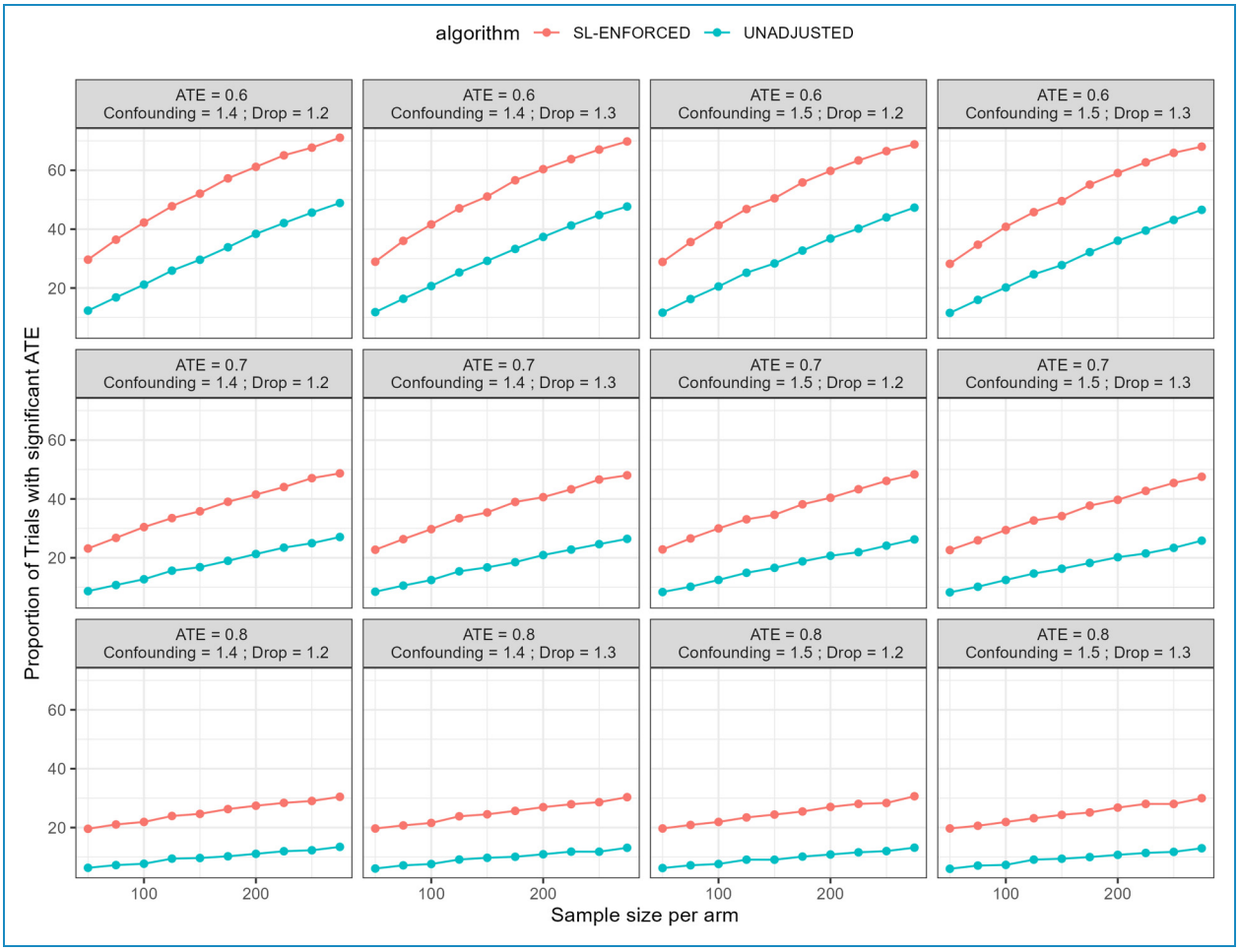

Figure 1. Percentage of trials with a significant ATE, according to the SL algorithm, and the unadjusted estimates according to the sample sizes. Several ATE, drop-out, and heterogeneity of treatment response (confounding) have been assumed.

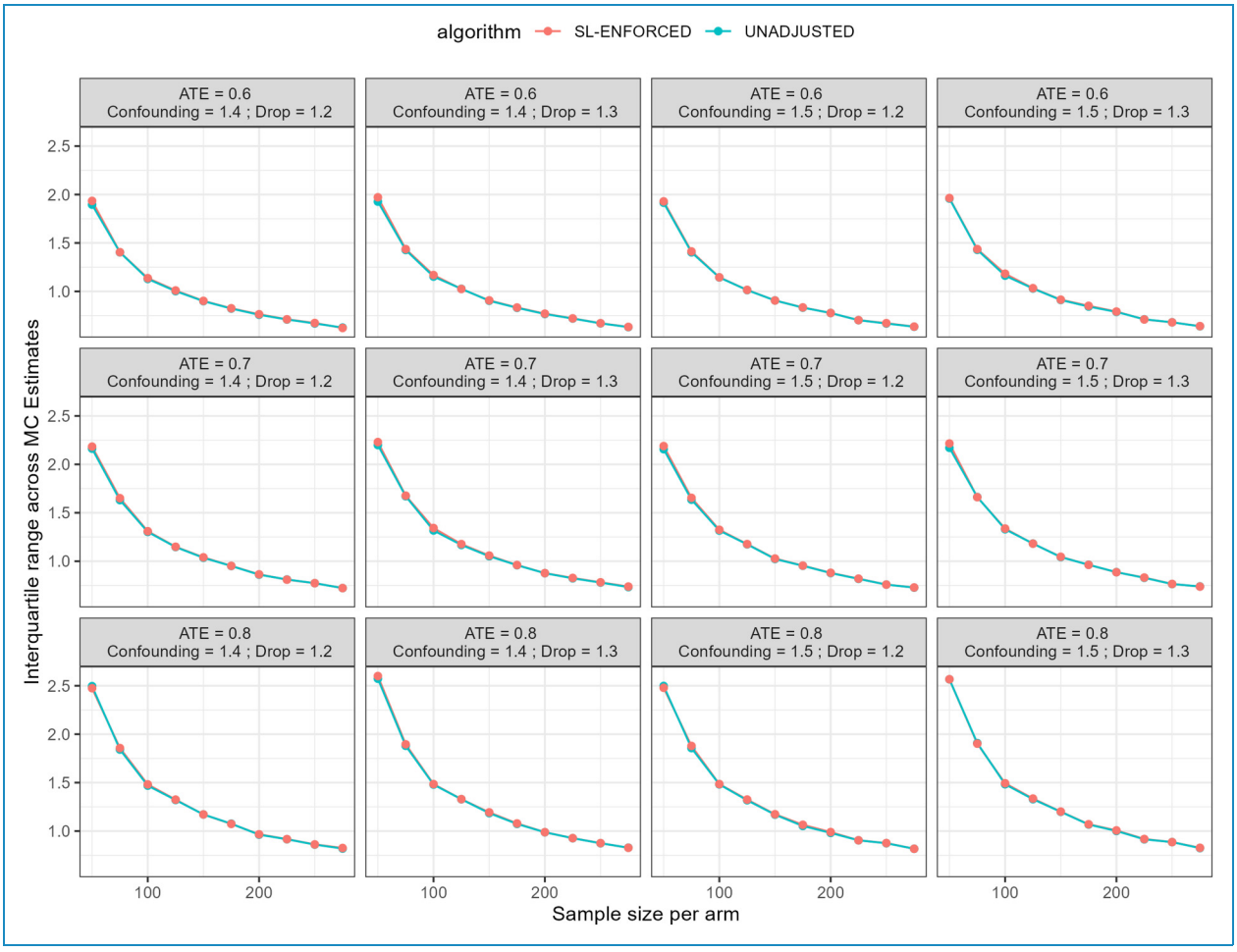

Figure 2. MC Interquartile range for estimated ATE according to machine learning methods composing the ensemble algorithm, the SL algorithm, and for the unadjusted estimates according to the sample sizes. Several ATE, drop-out, and heterogeneity of treatment response (confounding) have been assumed.

Comparison between estimation methods has been also reported considering the single learners composing the ensemble algorithm, the SL algorithm, and the unadjusted estimates. The average across scenarios of simulated trials percentages ensuring a truly significant ATE and MAPE has been computed across simulation scenarios. The results revealed percentages of truly identified ATE across scenarios (Figure 5, Panel A) very similarly higher than 37.8% compared to unadjusted estimates (19.2%) for all ML adjustments. This finding indicates a general improvement in the treatment effect identification ability for ML adjusted method in comparison with the unadjusted ones. Concerning the MAPE, the bias indication is slightly higher for the GLM or unadjusted estimate compared to the lower MAPE observed for KSVM, RANGER, and SL; in general, the performances are very similar across the algorithms considered to define the ensemble SL (Figure 5, Panel B). The results indicate an evident similarity in the performance across the algorithms also for ITT analysis (Figure S6, Appendix).

Discussion

The simulation results revealed a general improvement in the ability to truly detect the ATE by considering the SL-enforced estimation in comparison with the unadjusted estimates for all algorithms composing the ensemble SL and all the sample sizes. The proposed procedure represents the combination of strategies for managing dropout data in RCTs through PS model weighting procedures with the application of ML ensemble algorithms developed on large volumes of observational data useful for controlling the mechanism of outcome development in the RCT characterized by difficulties in the patient retention procedure.

Recently, the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use released a guideline on statistical principles for clinical trials that introduces the general structure to align the objective, conduction, and data analysis of the trial together with an interpretation of the study results.[33] The guideline defines an estimand as the final objective of RCT estimation to

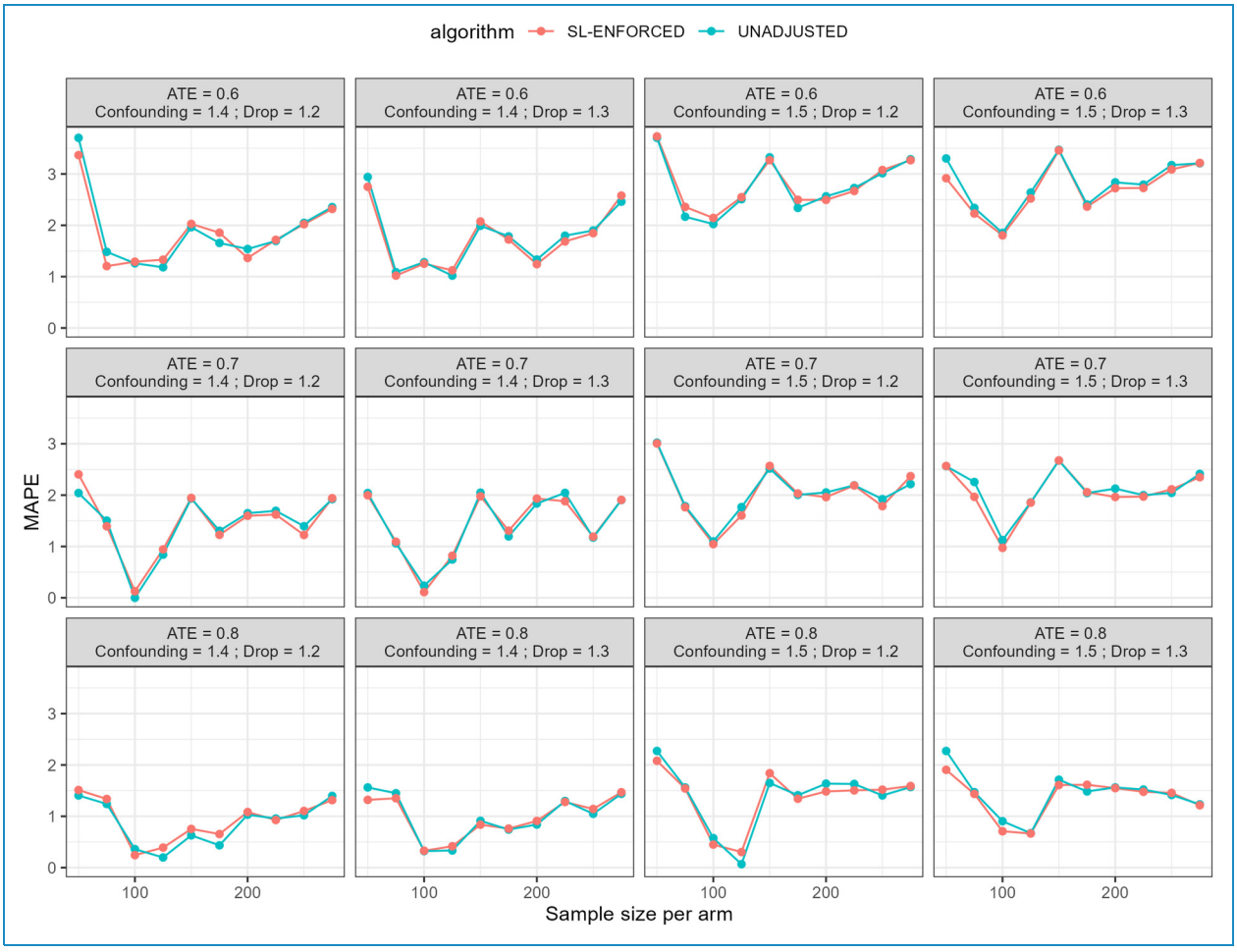

Figure 3. MAPE for ATE estimation according to machine learning methods composing the ensemble algorithm, the SL algorithm, and for the unadjusted estimates according to the sample sizes. Several ATE, drop-out, and heterogeneity of treatment response (confounding) have been assumed.

address the clinical hypothesis behind the trial objective.[34] Furthermore, the document underlines the need to define preliminary procedures to handle dropout events by defining plausible assumptions about the dropout mechanism.[33]

Post-randomization events, including poor compliance with treatment and missing follow-up, could alter the balance of patient characteristics, demystifying the merits of a well-conducted randomization procedure.[35] The gold standard for an RCT analysis is historically represented by the Intention To Treat principle; according to this procedure, all randomized patients should be included in the final analysis.[36] However, conducting an Intention To Treat analysis becomes challenging with a considerable number of dropout events. In several RCT settings, especially in the pediatric field, a considerable fraction of patients enrolled could withdraw their consent for participation in the RCT.[15]

The International Council for Harmonization guideline suggests that the RCT analysis could be tailored to define the outcomes that would have been observed if the patient had continued the trial intervention, hence the hypothetical strategy may be the preferred approach.[33] Within this general context, the literature reports several efforts to handle the dropout mechanism by estimating the PS of IPTW by weighting each patient by the inverse of the probability of receiving treatment given the covariate and treatment history.[37]

Methods do not account for the impact of the mechanism of manifestation of the disease as a possible confounding effect. In this direction, the latest developments in trial biometrics have highlighted how ML techniques can be useful as an aid to manage the mechanism of patient enrollment in RCTs and evaluate possible mechanisms of residual confounding.[38] Common problems of unsuccessful RCTs include difficulties in patient selection, problems in randomization with residual confounders, small trial sizes due to lack of accrual, and missing follow-up.[39]

The scientific literature demonstrated that ML models can be trained to select study participants by predicting

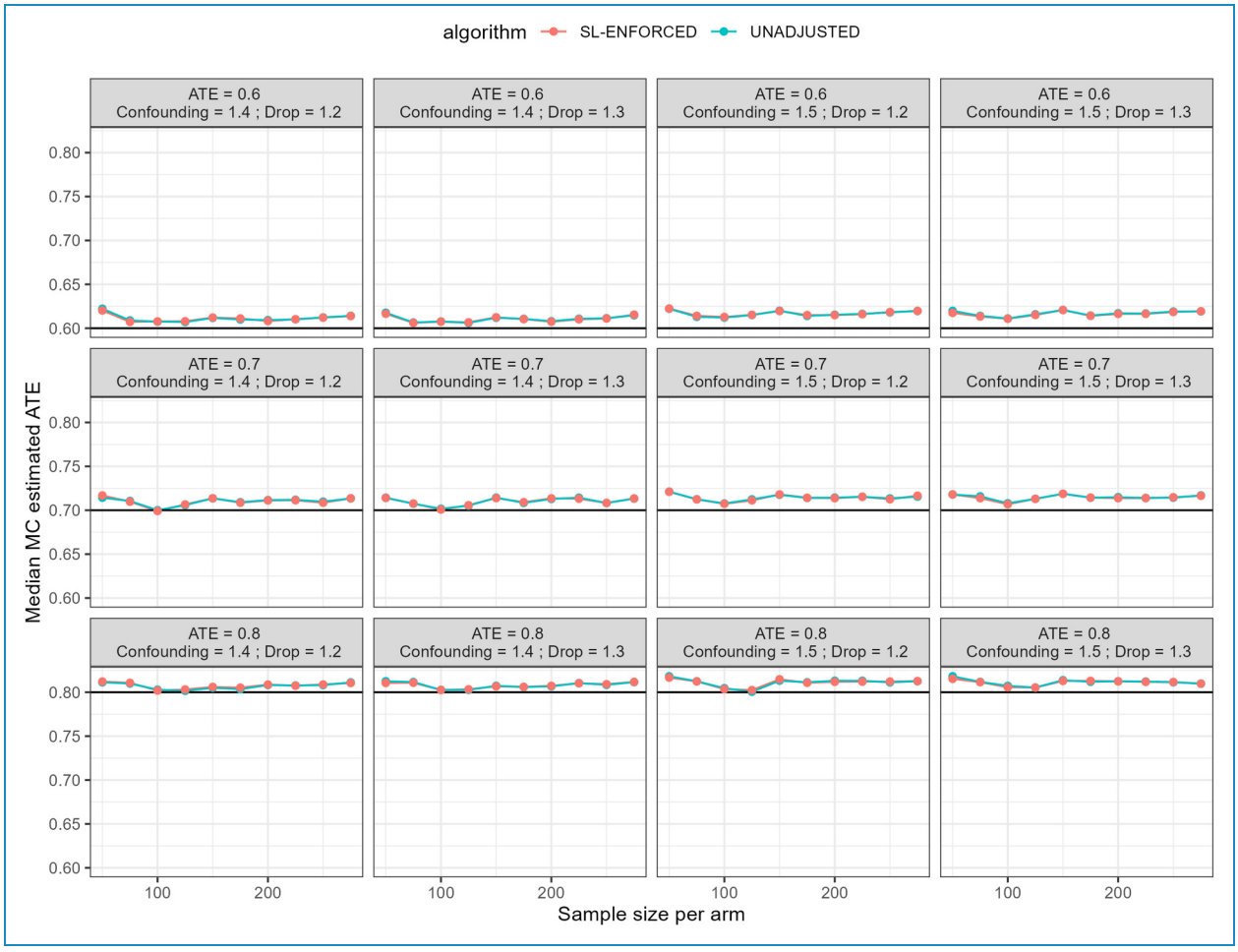

Figure 4. MC median ATE according to machine learning methods composing the ensemble algorithm, the SL algorithm, and the unadjusted estimates according to the sample sizes. Several ATE, drop-out, and heterogeneity of treatment response (confounding) have been assumed. The black line represents the true ATE.

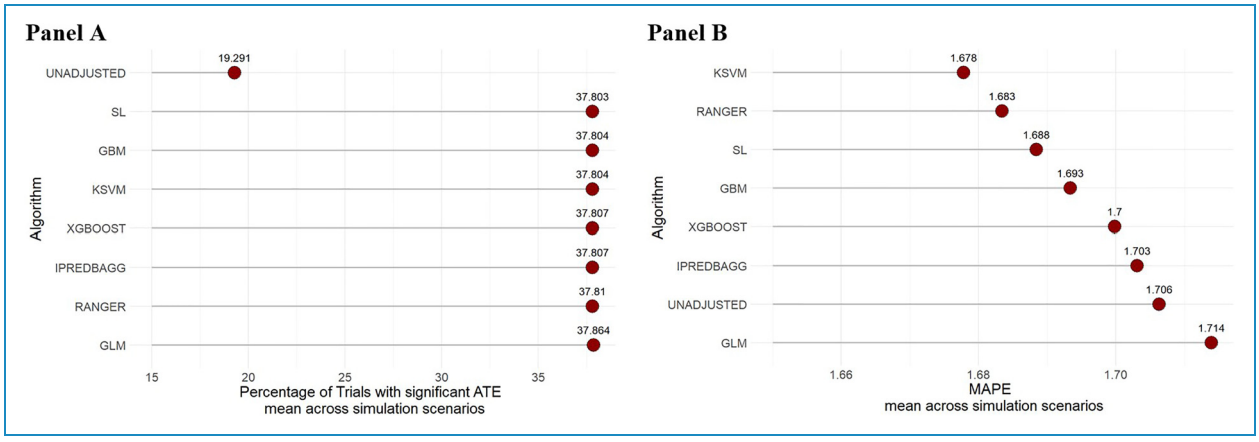

Figure 5. Comparison across estimation methods. Panel A reports the average across scenarios of simulated trials percentages ensuring a truly significant ATE; Panel B reports the MAPE mean across simulation scenarios. The results have been reported by considering the machine learning methods composing the ensemble algorithm, the SL algorithm, and the unadjusted estimates. GBM (Gradient Boosting Machine), IPREDBAGG (Improved PREDictive BAGGing classification tree), RANGER (RANGER Random Forest), UNADJUSTED (Unadjusted Estimate without ML enforcing), GLM (Generalized Linear Model), KSVM (Kernel Support Vector Machine), SL (SuperLearner), XGBOOST (eXtreme Gradient BOOSTing).

the natural history of each RCT participant and to assess study endpoints using a data-driven method. Given these advantages, ML facilitates efficient execution and improvement of statistical power compared to traditional RCTs.[38] In addition, large volumes of observational data can be used to build these predictive algorithms and improve the performance of clinical trials. In the literature, other efforts are also proposed in this regard. For example, in a Bayesian framework, a two-step procedure has been proposed that combines the PS calculation with Bayesian models integrating data from nonrandomized studies with data from RCTs in previous distributions.[40] Other efforts are reported by considering also ML methods to combine RCT and observational studies improving the generalizability of RCTs using the representativeness of observational data.[26] However, the possible benefit of the methods on the study of statistical power in RCTs characterized for problems in patient retention is not demonstrated.

Instead, this research demonstrates an increase in the ability to truly detect the ATE for the SL-enforced procedure. The proposed SL approach adjusts the treatment effect estimates based on patient-specific disease risk profiles predicted by the ML model. This adjustment helps to create balanced treatment and control groups by accounting for individual variations and potential confounders, resulting in improved power to detect the true treatment effect.

Moreover, the similarity in variability between the SL-adjusted estimates and unadjusted methods evidenced in this simulation experiment, indicates that the ensemble approach could capture the inherent variability in the data without introducing additional noise or instability. The ensemble nature of the SL approach incorporates multiple ML algorithms, each with its strengths and weaknesses. By combining the predictions of these algorithms, the SL approach leverages the diversity of the ensemble to capture a wide range of possible treatment effects.[41] This comprehensive approach helps to ensure that the estimation process encompasses the inherent variability in the data, resulting in a similar variability (IQR) to unadjusted methods.

The similarity in bias and point estimation between the SL-adjusted estimates and unadjusted methods has been also evidenced suggesting that the ensemble SL approach successfully could control for confounding bias. It is worth noting that in this simulation study, the SL approach adjusts the treatment effect estimates based on patient-specific disease risk profiles predicted by the ML model. By doing so, the SL could mitigate the bias associated with confounders, resulting in a mean bias similar to unadjusted methods.

Moreover, the comparison between estimation methods, including the ensemble SL algorithm and unadjusted estimates, reveals that the SL-adjusted estimates consistently achieve higher percentages of truly identified ATE compared to unadjusted estimates. This implies that incorporating the SL algorithm and ML adjustments improves the ability to accurately detect and estimate the true treatment effect in a pediatric RCT even for smaller sample sizes. The improved performance in truly identifying significant ATE is a crucial finding, as it enhances the reliability and validity of the study results.

This finding could facilitate the application of the method and is particularly promising especially in pediatric research settings where large volumes of observational data are available, but, at the same time, keeping patients enrolled in the trial is challenging, especially for research settings characterized by procedural and study conduction issues.

Limitations and future research developments. The literature used to simulate the data referred to the prevalence of the disease and the effect size is mainly considered to present the properties of the method rather than the application to a real clinical trial case.

More research developments are needed to evaluate the performance of the SL-enforced method, where the heterogeneity of treatment response does not depend on the mechanism of disease manifestation, but on other confounding issues related to latent causes.

Moreover, the results obtained from the SL-enforced estimation procedure using historical observational data may have some issues in the generalizability to the specific clinical trial setting. The predictive models developed on observational data provide estimates based on the available information and may not account for all factors considered in the trial design. For ensuring better compliance of the ensemble SL model with RCT data, external validation efforts could be needed. Validation studies using independent datasets from similar clinical trial settings can help assess the performance and generalizability of the predictive models. Such validation efforts can provide insights into the potential biases introduced by the heterogeneity between observational data and clinical trials.

Indeed, sensitivity analysis efforts, conducted with varying assumptions and parameters, can explore the impact of heterogeneity and assess the stability of the treatment effect estimates across different scenarios. The sensitivity analysis could be also useful because the SL-enforced estimation procedure relies on the available data, and the presence of unmeasured confounders cannot be completely ruled out. In this general framework, varying assumptions or scenarios in the analysis to assess the robustness of the results could be a valuable effort.

Conclusions

The developed model could be effectively used in a clinical trial to enforce the estimation of the effect of treatment by

adjusting the final estimate for a patient-specific disease risk profile.

Contributorship: Original Draft preparation (DA), Writing review and editing (DA, SB, RC, LDD), Formal Analysis (DA), methodology (DA, DG), supervision (DG).

Declaration of Conflicting Interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical Approval: Not Applicable.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Guarantor: Danila Azzolina.

ORCID iDs: Danila Azzolina https://orcid.org/0000-0002-8185-5742
Dario Gregori https://orcid.org/0000-0001-7906-0580

Supplemental material: Supplementary material for this article is available online.

References

1. Farrell B, Kenyon S and Shakur H. Managing clinical trials. *Trials* 2010; 11: 78.
2. Pak TR, Rodriguez M and Roth FP. Why clinical trials are terminated. *bioRxiv* 2015: 021543.
3. Williams RJ, Tse T, DiPiazza K, et al. Terminated trials in the ClinicalTrials.gov results database: evaluation of availability of primary outcome data and reasons for termination. *PLoS One* 2015; 10: e0127242.
4. Rimel B. Clinical trial accrual: obstacles and opportunities. *Front Oncol* 2016; 6: 103. doi:10.3389/fonc.2016.00103
5. Mannel RS and Moore K. Research: an event or an environment? *Gynecol Oncol* 134: 441–442. doi:10.1016/j.ygyno.2014.08.001
6. Stensland KD, McBride RB, Latif A, et al. Adult cancer clinical trials that fail to complete: an epidemic? *JNCI: J Nat Cancer Inst* 2014; 106.
7. Broadwin C, Azizi Z and Rodriguez F. Clinical trial technologies for improving equity and inclusion in cardiovascular clinical research. *Cardiol Ther* 2023: 1–11.
8. Beasant L, Brigden A, Parslow RM, et al. Treatment preference and recruitment to pediatric RCTs: a systematic review. *Contemp Clin Trials Commun* 2019; 14: 100335.
9. Kern SE. Challenges in conducting clinical trials in children: approaches for improving performance. *Expert Rev Clin Pharmacol* 2009; 2: 609–617.
10. Greenberg RG, Gamel B, Bloom D, et al. Parents' perceived obstacles to pediatric clinical trial participation: findings from the clinical trials transformation initiative. *Contemp Clin Trials Commun* 2018; 9: 33–39.
11. Baiardi P, Giaquinto C, Girotto S, et al. Innovative study design for paediatric clinical trials. *Eur J Clin Pharmacol* 2011; 67: 109–115.
12. Huff RA, Maca JD, Puri M, et al. Enhancing pediatric clinical trial feasibility through the use of Bayesian statistics. *Pediatr Res* 2017; 82: 814.
13. Joseph PD, Craig JC and Caldwell PH. Clinical trials in children. *Br J Clin Pharmacol* 2015; 79: 357–369.
14. McQuillan T, Wilcox-Fogel N, Kraus E, et al. Integrating musculoskeletal education and patient care at medical student–run free clinics. *PM and R* 2017; 9: 1117–1121.
15. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp Clin Trials Commun* 2018; 11: 156–164.
16. Billingham L, Malottki K and Steven N. Small sample sizes in clinical trials: a statistician's perspective. *Clin Investig (Lond)* 2012; 2: 655–657.
17. Kearney A, Rosala- Hallas A, Bacon N, et al. Reducing attrition within clinical trials: the communication of retention and withdrawal within patient information leaflets. *PLoS ONE* 2018; 13: e0204886.
18. Berger VW. Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials. *Biom J* 2005; 47: 119–127.
19. Committee for Medicinal Products for Human. Guideline on the clinical development of medicinal products for the treatment of cystic fibrosis. *London, European Medicines Agency*.
20. European Agency of Medicines. *Guideline on the requirements for clinical documentation for orally inhaled products (OIP) including the requirements for demonstration of therapeutic equivalence between two inhaled products for use in the treatment of asthma and chronic obstructive pulmonary disease (COPD) in adults and for use in the treatment of asthma in children and adolescents*. European Medicines Agency London, 2009.
21. Committee for Proprietary Medicinal. Note for Guidance on Evaluation of Anticancer Medicinal Products in Man. *The European Agency for the Evaluation of Medicinal Products, London*.
22. Ruberg SJ, Beckers F, Hemmings R, et al. Application of Bayesian approaches in drug development: starting a virtuous cycle. *Nat Rev Drug Discov* 2023; 22: 235–250.
23. Azzolina D, Berchialla P, Gregori D, et al. Prior elicitation for use in clinical trial design and analysis: a literature review. *Int J Environ Res Public Health* 2021; 18. DOI: 10.3390/ijerph18041833
24. Azzolina D, Lorenzoni G, Bressan S, et al. Handling poor accrual in pediatric trials: a simulation study using a Bayesian approach. *Int J Environ Res Public Health* 2021; 18: 1–16. doi:10.3390/ijerph18042095
25. Chevret S, Timsit J-F and Biard L. Challenges of using external data in clinical trials- an illustration in patients with COVID-19. *BMC Med Res Methodol* 2022; 22: 321.
26. Colnet B, Mayer I, Chen G, et al. Causal inference methods for combining randomized trials and observational studies: a review. 2020. DOI: 10.48550/ARXIV.2011.08047
27. Peng J, Jury EC, Dönnes P, et al. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. *Front Pharmacol* 2021; 12: 720694.

28. Da Dalt L, Bressan S, Scozzola F, et al. Oral steroids for reducing kidney scarring in young children with febrile urinary tract infections: the contribution of Bayesian analysis to a randomized trial not reaching its intended sample size. *Pediatr Nephrol* 2021; 36: 3681–3692.
29. Shaikh N, Craig JC, Rovers MM, et al. Identification of children and adolescents at risk for renal scarring after a first urinary tract infection: a meta-analysis with individual patient data. *JAMA Pediatr* 2014; 168. doi:10.1001/jamapediatrics.2014.637
30. van der Wal WM and Geskus RB. ipw : an R package for inverse probability weighting. *J Stat Soft* 2011; 43. DOI: 10.18637/jss.v043.i13
31. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/ 2022.
32. Polley E, LeDell E, Kennedy C, et al. *Package 'SuperLearner'*. CRAN, 2019.
33. Guideline IH. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. *E9 (R1) Step* 2019; 4: 20.
34. Gogtay N, Ranganathan P and Aggarwal R. Understanding estimands. *Perspect Clin Res* 2021; 12: 106.
35. Fergusson D. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. *Br Med J* 2002; 325: 652–654.
36. Gupta SK. Intention-to-treat concept: a review. *Perspect Clin Res* 2011; 2: 109–112.
37. Hogan JW, Roy J and Korkontzelou C. Handling drop-out in longitudinal studies. *Statist Med* 2004; 23: 1455–1497.
38. Lee CS and Lee AY. How artificial intelligence can transform randomized controlled trials. *Transl Vis Sci Technol* 2020; 9: 9–9.
39. Nichol AD, Bailey M, Cooper DJ, et al. Challenging issues in randomised controlled trials. *Injury* 2010; 41: S20–23.
40. Zhao H, Hobbs BP, Ma H, et al. Combining non-randomized and randomized data in clinical trials using commensurate priors. *Health Serv Outcomes Res Methodol* 2016; 16: 154–171.
41. Lanera C, Berchialla P, Lorenzoni G, et al. A SuperLearner approach to predict run-in selection in clinical trials. *Comput Math Methods Med* 2022. DOI: 10.1155/2022/4306413