Topics in Heterogeneous Treatment Effect Estimation

by

Jiabei Yang

B.S., B.A., Shanghai Jiao Tong University, 2015

M.S., Harvard T.H. Chan School of Public Health, 2017

A Dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Biostatistics at Brown University

Providence, Rhode Island

October 2022

This dissertation by Jiabei Yang is accepted in its present form
by the Department of Biostatistics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Christopher H. Schmid, Advisor

Date _____ _____

Jon A. Steingrimsson, Advisor

Recommended to the Graduate Council

Date _____ _____

Issa J. Dahabreh, Reader

Approved by the Graduate Council

Date _____ _____

Andrew G. Campbell,
Dean of the Graduate School

Jiabei Yang

Updated Jun 2022

Contact Information

Email: jiabei_yang@brown.edu
LinkedIn: https://www.linkedin.com/in/jiabei-yang
GitHub: https://github.com/jiabei-yang

Education

Ph.D. Biostatistics — Expected Aug 2022
Brown University, Providence, RI
Advised by: Jon A. Steingrimsson and Christopher H. Schmid
Thesis: Topics in Heterogeneous Treatment Effect Estimation

M.S. Biostatistics — May 2017
Harvard T.H. Chan School of Public Health, Boston, MA

B.S. Bioinformatics and Biostatistics — Jun 2015
B.A. French
Shanghai Jiao Tong University, Shanghai, China

Honors and Awards

Student Travel Award — 2020
International Conference on Health Policy Statistics

Best Doctoral or Post-Doctoral Trainee Poster Presentation — 2019
Public Health Research Day, Brown University

Arawana Scholarship (top 5%) — 2014
Shanghai Jiao Tong University

Gold Medal — 2014
International Genetically Engineered Machine Competition

Publications

Yang, J., Dahabreh, I. J., and Steingrimsson, J. A. (2021). Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. *Biometrics*.

Yang, J., Steingrimsson, J. A., and Schmid, C. H. (2021). Sample size calculations for n-of-1 trials. *arXiv preprint*.

Kaplan, H. C., Opipari-Arrigan, L., **Yang, J.**, Schmid, C. H., Schuler, C. L., Saeed, S. A., et al. (2022). Personalized Research on Diet in Ulcerative Colitis and Crohn's Disease (PRODUCE): A Series of N-of-1 Diet Trials. *Official journal of the American College of Gastroenterology | ACG*, 10-14309.

Wei, Y., Coull, B., Koutrakis, P., **Yang, J.**, Li, L., Zanobetti, A., and Schwartz, J. (2021). Assessing additive effects of air pollutants on mortality rate in Massachusetts. *Environmental Health, 20*(1), 1-10.

Marcus, G. M., Modrow, M. F., Schmid, C. H., Sigona, K., Nah, G., **Yang, J.**, et al. (2021). Individualized Studies of Triggers of Paroxysmal Atrial Fibrillation: The I-STOP-AFib Randomized Clinical Trial. *JAMA cardiology*.

Kravitz, R. L., Marois, M., Sim, I., Ward, D., Kanekar, S., Yu, A., ..., **Yang, J.**, et al. (2021). Chronic Pain Treatment Preferences Change Following Participation in N-of-1 Trials, but not Always in the Expected Direction. *Journal of Clinical Epidemiology*.

Kravitz, R. L., Aguilera, A., Chen, E. J., Choi, Y. K., Hekler, E., Karr, C., ..., **Yang, J.**, et al. (2020). Feasibility, acceptability, and influence of mHealth-supported n-of-1 trials for enhanced cognitive and emotional well-being in US volunteers. *Frontiers in Public Health, 8,* 260.

Steingrimsson, J. A., and **Yang, J.** (2019). Subgroup identification using covariate-adjusted interaction trees. *Statistics in medicine, 38*(21), 3974-3984.

Wei, Y., Wang, Y., Lin, C. K., Yin, K., **Yang, J.**, Shi, L., et al. (2019). Associations between seasonal temperature and dementia-associated hospitalizations in New England. *Environment International, 126*, 228-233.

Braun, D., **Yang, J.**, Griffin, M., Parmigiani, G., and Hughes, K. S. (2018). A clinical decision support tool to predict cancer risk for commonly tested cancer-related germline mutations. *Journal of Genetic Counseling, 27*(5), 1187-1199.

Li, J., Jia, J., Li, H., Yu, J., Sun, H., He, Y., ..., **Yang, J.**, et al. (2014). SysPTM 2.0: an updated systematic resource for post-translational modification. *Database*, 2014.

Invited Talks

Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. *Causal Inference and Machine Learning Working Group, Department of Biostatistics, Harvard T.H. Chan School of Public Health and Harvard Data Science Initiative*, 2021.

`nof1`: an R package for analyzing and presenting n-of-1 trials. *The 42nd Annual Meeting of the Society for Clinical Trials*, 2021.

`nof1`: an R package for analyzing and presenting n-of-1 trials. *Western North American Region of the International Biometric Society*, 2021.

Contributed

Incorporating group structure into tree-based algorithms and group selection through importance measures. *Joint Statistical Meetings*, 2020.

`nof1ins`: an R Package for analyzing and presenting individual n-of-1 studies. *Public Health Research Day, Brown University*, 2020.

Sample size calculations in n-of-1 trials. *International Conference on Health Policy Statistics*, 2020.

Sample size calculations in n-of-1 trials. *Joint Statistical Meetings*, 2019.

Sample size calculations in n-of-1 trials. *Public Health Research Day, Brown University*, 2019.

ASK2ME - The All Syndromes Known to Man Evaluator; Clinical decision support tool to predict cancer risk for commonly tested cancer-related germline mutations. *Sloan Healthcare & Bioinnovations Conference*, 2017.

Submitted

Schmid, C. H., and **Yang, J.** Bayesian Models for N-of-1 Trials. *Submitted to Harvard Data Science Review (Special Issue).*

Teaching Assistant

Practical Data Analysis, *Brown University* Fall 2020
Instructed by: Christopher H. Schmid and Alice J. Paul
- Lectured reproducible research, graphics, data collection, data cleaning and management, principles in simulations, and collaborative research.
- Led lab sessions illustrating variable selection using R.

Applied Multilevel Data Analysis, *Brown University* Spring 2019
Instructed by: Joseph W. Hogan
- Lectured multilevel Poisson regression.

Probability, Statistics, and Machine Learning: Advanced Methods, *Brown University* Spring 2018
Instructed by: Christopher H. Schmid and Alice J. Paul

Data Science II, *Harvard T.H. Chan School of Public Health* Spring 2017
Instructed by: Jukka-Pekka Onnela

Principles of Biostatistics II, *Harvard T.H. Chan School of Public Health* Summer 2016
Instructed by: Kerrie Nelson
- Lectured lab sessions reviewing concepts in lecture and applications in STATA.

Community Involvement

Inaugural Member 2021
Epsilon Iota chapter, Delta Omega Honorary Society in Public Health

Student Representative for Brown University School of Public Health 2021
Council on Education for Public Health Re-accreditation Virtual Site Visit

Ph.D. Student Representative 2019 - 2021
Public Health Curriculum Committee, Brown University School of Public Health

Journal Club Moderator 2018
Department of Biostatistics, Brown University School of Public Health

Volunteer 2016
US-India BioPharma & Healthcare Summit

Volunteer 2015
International Genetically Engineered Machine Competition

Reveiwer

Statistics in Medicine, Statistical Methods in Medical Research, Journal of the American Medical Informatics Association, Harvard Data Science Review (Special Issue), Medicine

Professional Memberships

American Statistical Association 2017 - Present
Institute of Mathematical Statistics 2017 - Present

Industry Experience

Data Science Intern 2019
AT&T Labs - Statistics Research, New York City, NY
- Developed tree-based algorithms that account for group structure among input features and defined group importance measures for grouped feature selection.

Biostatistics Intern 2017
dMed Biopharmaceutical, Shanghai, China
- Simulated median progression free survival (mPFS) for ovarian cancer patients based on earlier trial results and germline mutation carrier composition in Chinese population.

Acknowledgements

I would like to express my deepest appreciation to my dissertation committee. I would like to thank my advisors, Dr. Christopher H. Schmid and Dr. Jon A. Steingrimsson, for the guidance, the generous support and the maximum understanding through the program. I thank Chris for bringing me on collaborative projects, for both challenging and helping me in communicating with non-statistical collaborators and for bringing up aspects in research I have not thought of. I thank Jon for guiding me through writing a statistical paper, for being resourceful when exploring a new research area and for the kind flexibility when there were difficult times. Both of them have been essential to my progress and to become an independent researcher through the program. I would like to thank Dr. Issa J. Dahabreh for always being knowledgeable, for consistently offering expertise and insightful comments and for the encouragement every time we interact.

I would like to thank the faculty and staff in the Department of Biostatistics for the continuous support, for helping me navigate the PhD program, and for building and maintaining a warm community for us to grow and thrive. In addition, to the friends, the PhD cohort, colleagues and mentors I interacted during the program, thank for the good time, the motivation and the mentorship in the past few years. I am excited to continue learning and being inspired through our interactions in the future.

Finally, but most importantly, I would like to thank my husband, Yaguang, for the support whenever there is need, and my family for making everyday of my life brighter and more beautiful than I ever imagined it could be.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Heterogeneous treatment effect estimation which aims to determine which treatment works best, for whom and under what circumstances is not only the core mission of comparative effectiveness research and precision medicine [Dahabreh et al., 2016], but is also impacting the studies of how advertising or marketing offers affect consumer purchases, evaluations of the effectiveness of government programs or public policies, and "A/B tests" (large-scale randomized experiments) commonly used by technology firms to select algorithms for ranking search results or making recommendations [Wager and Athey, 2018].

In clinical settings, the estimation of heterogeneous treatment effects is usually performed through subgroup analyses in randomized controlled trials. These analyses are completed by exploring treatment-covariate statistical interactions in outcome regression models [Dahabreh et al., 2017, Lagakos et al., 2006]. However, there are challenges coming with these analyses. These challenges include: firstly, subgroup analyses explores treatment-covariate statistical interactions one variable at a time and when there are multiple variables of interest, patients can belong to multiple subgroups and that makes it hard for apply the subgroup analysis results to individual patients; secondly, the prespcification of subgroup analyses in randomized controlled trials requires subject matter knowledge; lastly, the sample size of clinical trials are usually underpowered to detect such differential treatment effects [Wang et al., 2007, Brookes et al., 2004, Dahabreh et al., 2016].

N-of-1 trials, single participant crossover trials in which multiple treatments are given in sequentially randomized treatment periods [Nikles and Mitchell, 2015], have been recommended as suitable designs for estimating individual-specific treatment effects [Duan et al., 2013]. Previous

work on estimating sample sizes for n-of-1 trials has been limited to a single n-of-1 trial or specific settings in a series of n-of-1 trials. In Chapter 2, we present a procedure for designing n-of-1 trials. We formally define the design components for determining the sample size of a series of n-of-1 trials, present models for analyzing these trials and use them to derive the sample size formula for estimating the population average treatment effect and the standard error of the individual-specific treatment effect estimates. The procedure is implemented and illustrated in the chapter and through a Shiny app.

Extensions of the Classification and Regression Tree algorithms (CART) [Breiman et al., 1984] allow data driven methods for subgroup identification. In randomized controlled trials, the Interaction Tree algorithms were developed which splits the covariate space into subgroups by recursively contrasting treatment effect estimates between groups [Su et al., 2009]. Observational studies allow larger sample size for subgroup analyses; however, subgroup identification in observational studies where the treatment assignment is not randomized requires adjusting for covariates that confound the treatment-outcome relationship. Moreover, if electronic health record database is used, analyses should be able to accommodate 1) repeated and temporally unaligned measures on each participant, 2) time-varying treatment assignment and 3) missing data due to missed clinic visits and dropouts. In Chapter 3, we develop the Causal Interaction Tree (CIT) algorithms for identifying subgroups with heterogeneous treatment effects in cross-sectional observational data. In Chapter 4, we further develop the Longitudinal Identification of Subgroup Algorithm (LISA) for subgroup identification in longitudinal data.

The follow chapters are organized as follows: Chapter 2 presents sample size calculations for a series of n-of-1 trials; Chapter 3 develops the Causal Interaction Tree (CIT) algorithms; Chapter 4 develops the Longitudinal Identification of Subgroup Algorithm (LISA); Chapters A to C are appendices for Chapters A to C respectively.

Chapter 2

Sample Size Calculations for N-of-1 Trials

2.1 Introduction

In the presence of treatment effect heterogeneity, population average treatment effects from parallel-group randomized controlled trials may not accurately represent the risk to individual participants, making individualized treatment recommendations challenging [Kravitz et al., 2004, Greenfield et al., 2007, Olsen et al., 2007]. N-of-1 trials, single participant crossover trials in which multiple treatments are given in sequentially randomized treatment periods [Nikles and Mitchell, 2015], where we refer to a certain length of time where the same treatment is given as a treatment period, have been recommended as suitable designs for estimating individual-specific treatment effects [Duan et al., 2013].

Combining data from a series of n-of-1 trials offers several advantages compared to using data from randomized controlled trials or from a single n-of-1 trial. Firstly, repeated measures on each participant provide more information for estimating population average treatment effects compared to measuring each participant only once in most randomized controlled trials. Secondly, because crossover trials reduce the number of participants needed to achieve the pre-specified power, a series of n-of-1 trials with multiple crossovers is even more efficient at estimating the population average treatment effect. Lastly, the ability to borrow information from other participants increases the

efficiency for estimating individual-specific treatment effects compared to using only data from a single n-of-1 trial [Zucker et al., 2010, Senn, 2019].

Designing a series of n-of-1 trials that achieve the desired power for estimating the population average treatment effect, and that, if of interest, satisfy the standard error requirements for the individual-specific treatment effect estimates requires specifying the sequences with different orders of treatments assigned to the treatment periods, the number of participants (or trials; we can refer to trials as participants because n-of-1 trials are single participant trials and we will use "participants" in later development) assigned to each sequence, and the number of measurements in each period.

Previous work on estimating sample sizes for n-of-1 trials has been limited. Johannessen and Fosstvedt [1991] derived the power for estimating an individual-specific treatment effect using a single n-of-1 trial. Senn [2019] presented sample size formulae for estimating the population average treatment effect and variances of individual-specific treatment effect estimates when combining multiple n-of-1 trials. Both approach assumed specific orders of treatments assigned to treatment periods in the sequences, one measurement in each period, and independent measurements on each participant.

In this paper, we present methods for designing a series of n-of-1 trials. Section 2.2 defines the design components for determining the sample size of a series of n-of-1 trials. Section 2.3 presents models for estimating the population average treatment effect and the individual-specific treatment effects. Section 2.4 and 2.5 derive the sample size formula for estimating the population average treatment effect and the standard error of the individual-specific treatment effect estimates, respectively. Section 2.6 provides details about finding the possible combinations of the design components using the derived formulae. We implement and illustrate the procedure in Section 2.7 and in a Shiny app described in Section 2.8. Section 2.9 summarizes our findings and points to some future extensions.

2.2 Design components for a series of n-of-1 trials

Let Y be the outcome and let A be the treatment assignment. Figure 2.1 presents a schematic for the design of a series of n-of-1 trials where the outcome Y is measured repeatedly at times to which

treatments are assigned. Let i, $i = 1, \ldots, I$, index the sequences with different orders of treatments assigned to treatment periods; let j, $j = 1, \ldots, J_i$, index the J_i participants that are assigned to sequence i; let k, $k = 1, \ldots, K_i$, index the K_i treatment periods in sequence i; let l, $l = 1, \ldots, L_{ijk}$, index the L_{ijk} measurements in the kth treatment period for participant j assigned to sequence i. Therefore, Y_{ijkl} is the lth measurement on the outcome in the kth treatment period for participant j assigned to sequence i with corresponding treatment assignment A_{ijkl}. An n-of-1 trial design then comprises a series of trials with certain sequences where treatments are assigned to a number of treatment periods in which a number of measurements are taken. The design components consist of: 1) the sequences with different orders of treatments assigned to periods in the sequences, which in turn determines the number of sequences I and the number of treatment periods in each sequence K_i, 2) the number of participants assigned to each sequence J_i, and 3) the number of measurements in each period L_{ijk}. Note that determining the sequences in the trials will determine both I and K_i because a different number of treatment periods leads to a different sequence.

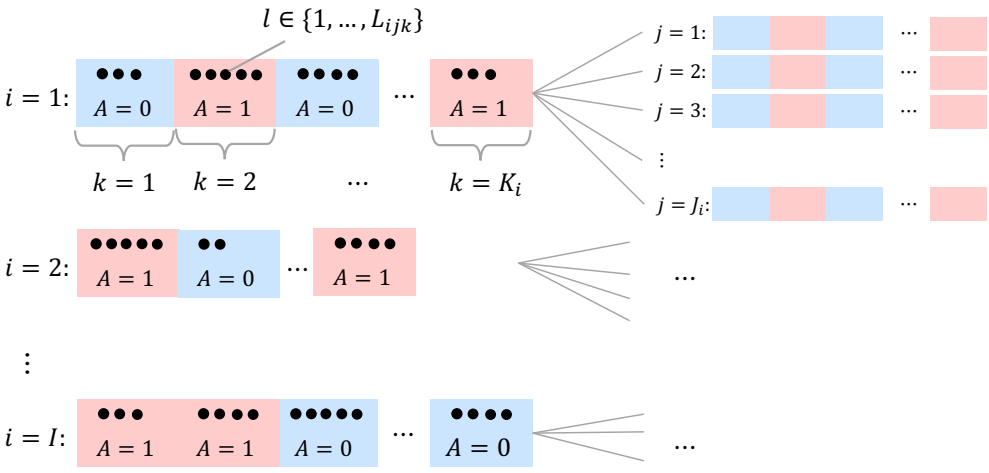

Figure 2.1: Indices for measurements in a series of n-of-1 trials. A is the treatment assignment; we present the scenario where there are two treatments in the trials and A is an indicator for being assigned to the intervention of interest. i, j, k and l are used to index the sequences in trials, participants assigned to each sequence, treatment periods in each sequence and measurements in each treatment period, respectively. J_i, K_i and L_{ijk} are the number of participants assigned to sequence i, the number of treatment periods in sequence i and the number of measurements in the kth treatment period of participant j assigned to sequence i, respectively.

2.3 Models for treatment effect estimation

We make the following assumptions in our derivations. We assume that there are two treatments in the series of n-of-1 trials and refer to the two treatments as intervention and reference treatment; therefore, let treatment assignment A be an indicator for being assigned to the intervention of interest, taking on value of 1 if assigned to the intervention and 0 if assigned to the reference treatment. We assume that the outcome is continuous. The participants are independent and the repeated measurements on each participant, indexed by k and l, can be correlated. There is no underlying trend in the outcome or carryover effects from one treatment to the other throughout the trials.

We now describe the models for estimating treatment effects using n-of-1 trials and start by the naive estimates for estimating the individual-specific treatment effects, which only use the data from a single participant for estimation,

$$Y_{kl} = m + \delta A_{kl} + \epsilon_{kl}, \tag{2.1}$$

where m, the intercept, is the mean of the responses when given the reference treatment, δ, the slope, is the difference between the mean responses of the intervention and reference treatments, and ϵ_{kl} is the residual error. The estimate for δ gives the naive estimate for the individual-specific treatment effect. We specify the variance matrix for the residual error vector to account for the correlation among the repeated measurements on a single participant.

Using data from a series of n-of-1 trials, we can model the intercepts and the slopes for each participant with some structure to estimate the population average treatment effect,

$$Y_{ijkl} = \mu_{ij} + \delta_{ij} A_{ijkl} + \epsilon_{ijkl}. \tag{2.2}$$

The intercept μ_{ij} and the slope δ_{ij} for participant j assigned to sequence i can then be modeled as 1) no pooling (i.e., fixed for each participant), 2) partial pooling (i.e., random from a common distribution), or 3) complete pooling (i.e., common across all participants). We usually do not fit the models with complete pooling of the intercepts because they make the strong assumption that the means of the responses on the reference treatment across the participants are the same, which

usually do not hold; we also do not use the models with no pooling of the slopes for estimating the population average treatment effect because they do not give the estimate directly, but if we use the model with no pooling of both intercepts and slopes, we will have the models for each participant giving the naive estimates for the individual-specific treatment effects. Therefore, four different models span the realm of possibilities for estimating the population average treatment effect, where the intercepts can either be fixed or random and the slopes can either be random or common. The estimate for the common slope in the common slope models or for the average of the random slopes in the random-slope models estimates the population average treatment effect.

Random slopes are more commonly used in order to allow for heterogeneous treatment effects across participants. They also suggest the shrunken estimates for the individual-specific treatment effects by borrowing information from other participants in the series of trials. Assuming a common slope implies that the individual-specific treatment effects are the same in all trials [Jackson et al., 2018]. When the number of measurements for each trial is small, one may choose to model the population average treatment effect using a common slope for more stable estimation.

While it is common to assume random slopes, assuming random intercepts is more controversial. Some recommend fixed intercepts so that estimation of the intercepts does not bias the estimation of the population average treatment effect or of the shrunken estimates for the individual-specific treatment effects. Others have produced simulations showing that the use of fixed intercepts may bias downward the maximum likelihood estimate of the variance for the random slopes [Jackson et al., 2018]. Random intercepts may also be helpful in cases where the information per trial is weak and fixed intercepts are poorly estimated (e.g., cases where there are a small number of measurements per period).

We refer readers to previous studies for details on the use of random or common slopes and the use of fixed or random intercepts [Legha et al., 2018, Riley et al., 2020, White et al., 2019] and will focus on the derivations using all four possible models.

In Model (2.2), participants are assumed to be independent, but repeated measurements on each participant may be correlated. We specify the variance matrix of the residual error vector for each participant to account for the correlation.

2.4 Sample size for population average treatment effect estimation

A general model incorporating fixed or random intercepts, random or common slopes, and a flexible residual error structure on each independent participant is the general linear mixed model

$$\begin{aligned} \boldsymbol{Y}_{ij} =& \mathbf{X}_{ij}\boldsymbol{\theta} + \mathbf{Z}_{ij}\boldsymbol{b}_{ij} + \boldsymbol{\epsilon}_{ij}, \\ \text{Var}(\boldsymbol{b}_{ij}) =& \mathbf{D}, \quad \text{Var}(\boldsymbol{\epsilon}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon} ij}, \quad \boldsymbol{b}_{ij} \perp\!\!\!\perp \boldsymbol{\epsilon}_{ij}. \end{aligned} \tag{2.3}$$

Here, $\boldsymbol{Y}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$ are the outcome and residual error vectors of length $\sum_{k=1}^{K_i} L_{ijk}$ for participant j assigned to sequence i, $\boldsymbol{\theta}$ and $\boldsymbol{b}_{ij}$ are the fixed and random effects, and $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are the corresponding design matrices for fixed and random effects, respectively, where $i = 1, \ldots, I$ and $j = 1, \ldots J_i$; $\mathbf{D}$ is the variance matrix for between-individual errors and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon} ij}$ is the variance matrix for within-individual errors.

Specifications of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon} ij}$ allow heterogeneous residual variance and additional correlation among the repeated measurements on each participant. The illustrations in Section 2.7 and the Shiny app described in Section 2.8 implemented commonly-used variance matrices with homogeneous residual errors and independent, exchangeable and first order autoregressive (AR-1) correlation structures.

Table 2.1 summarizes the elements in Model (2.3) for the four possible models that combine fixed or random intercepts with random or common slopes. We denote the fixed and random intercepts as m_{ij} and $\mu + \gamma_{\mu ij}$, respectively, and the common and random slopes as δ and $\delta + \gamma_{\delta ij}$, respectively, where $\gamma_{\mu ij}$ and $\gamma_{\delta ij}$ are the random components for the random intercepts and slopes, respectively.

The standard error for the generalized least squares estimator of the population average treatment effect [Laird and Ware, 1982], $\hat{\delta}$, is

Model	$\mathbf{X}_{ij}$	$\boldsymbol{\theta}$	$\mathbf{Z}_{ij}$	$\boldsymbol{b}_{ij}$	$\mathbf{D}$	$\boldsymbol{\Sigma}_{ij}$	$\boldsymbol{C}_{\boldsymbol{\theta}}$
Fixed-Common	$\begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 1 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 1 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$ } Treatment periods; $(\sum_{k=1}^{K_i} L_{ijk}) \times (\sum_{i=1}^{I} J_i + 1)$; Indicator column for m_{ij}	$\begin{pmatrix} m_{11} \\ \vdots \\ m_{IJ} \\ \delta \end{pmatrix}$	-	-	-	$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon} ij}$	$(\mathbf{0}^T_{\sum_{i=1}^{I} J_i}, 1)$
Fixed-Random	$\begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 1 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 1 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$ } Treatment periods; $(\sum_{k=1}^{K_i} L_{ijk}) \times (\sum_{i=1}^{I} J_i + 1)$; Indicator column for m_{ij}	$\begin{pmatrix} m_{11} \\ \vdots \\ m_{IJ} \\ \delta \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$ } Treatment periods; $(\sum_{k=1}^{K_i} L_{ijk})$	$\gamma_{\delta ij}$	σ^2_δ	$\mathbf{Z}_{ij}\mathbf{D}\mathbf{Z}_{ij}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon} ij}$	$(\mathbf{0}^T_{\sum_{i=1}^{I} J_i}, 1)$
Random-Common	$\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$ } Treatment periods; $(\sum_{k=1}^{K_i} L_{ijk}) \times 2$	$\begin{pmatrix} \mu \\ \delta \end{pmatrix}$	$\mathbf{1}_{\sum_{k=1}^{K_i} L_{ijk}}$	$\gamma_{\mu ij}$	σ^2_μ	$\mathbf{Z}_{ij}\mathbf{D}\mathbf{Z}_{ij}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon} ij}$	(0, 1)
Random-Random	$\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$ } Treatment periods; $(\sum_{k=1}^{K_i} L_{ijk}) \times 2$	$\begin{pmatrix} \mu \\ \delta \end{pmatrix}$	$\mathbf{X}_{ij}$	$\begin{pmatrix} \gamma_{\mu ij} \\ \gamma_{\delta ij} \end{pmatrix}$	$\begin{pmatrix} \sigma^2_\mu & \sigma_{\mu,\delta} \\ \sigma_{\mu,\delta} & \sigma^2_\delta \end{pmatrix}$	$\mathbf{Z}_{ij}\mathbf{D}\mathbf{Z}_{ij}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon} ij}$	(0, 1)

Table 2.1: Elements in Model (2.3) when we use different forms of intercepts and slopes. We name the models as "form of intercept"-"form of slope"; for example, "Fixed-Common" means the model with fixed intercepts and a common slope. $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are the design matrices for fixed and random effects for participant j assigned to sequence i respectively; $\boldsymbol{\theta}$ and $\boldsymbol{b}_{ij}$ are the fixed effects and random effects respectively; $\mathbf{D}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon} ij}$ are the variance matrices for between-individual and within-individual errors, respectively; $\boldsymbol{\Sigma}_{ij}$ is the variance matrix for the outcome vector $\boldsymbol{Y}_{ij}$; $\boldsymbol{C}_{\boldsymbol{\theta}}$ is the contrast matrix that pulls off the coefficient for treatment effect from the fixed effects vector or the variance for the treatment effect estimate from the corresponding variance matrix, respectively. The dimension of matrices or vectors are labeled to the bottom right corner. $\mathbf{0}$ and $\mathbf{1}$ are column vectors of 0's and 1's with the lengths labeled to the bottom right corner. m_{ij} and μ are the fixed intercepts and the average of the random intercepts respectively; δ is the common slope or the average of the random slopes; $\gamma_{\mu ij}$ and $\gamma_{\delta ij}$ are the random components for the random intercepts and slopes respectively. σ^2_μ and σ^2_δ are the variance of the random intercepts and slopes respectively; $\sigma_{\mu,\delta}$ is the covariance between the random intercepts and slopes.

$$\text{s.e.}(\hat{\delta}) = \sqrt{\boldsymbol{C}_{\boldsymbol{\theta}} \left(\sum_{i=1}^{I} \sum_{j=1}^{J_i} \mathbf{X}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{X}_{ij} \right)^{-1} \boldsymbol{C}_{\boldsymbol{\theta}}^T}, \tag{2.4}$$

where $\boldsymbol{C}_{\boldsymbol{\theta}}$ is the contrast matrix that pulls off the coefficient for treatment effect from the fixed effects vector or the variance for the treatment effect estimate from the corresponding variance matrix; $\boldsymbol{\Sigma}_{ij}$ is the variance matrix for the outcome vector $\boldsymbol{Y}_{ij}$. Table 2.1 presents the forms for $\boldsymbol{C}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{ij}$ in the four possible models.

For a given pre-specified type I error α, a minimal clinically important treatment effect Δ, and variance matrices for the between-individual errors when applicable, $\mathbf{D}$, and for the within-individual errors, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon} ij}$, we can achieve power of at least $1-\beta$ for estimating the population average treatment effect by finding the design components for a series of n-of-1 trials (i.e., the sequences, J_i, and L_{ijk}) that satisfy the following condition,

$$\begin{aligned} 1-\beta \leq & \Phi\left[-Z_{1-\frac{\alpha}{2}} - \frac{\Delta}{\text{s.e.}(\hat{\delta})}\right] + \Phi\left[-Z_{1-\frac{\alpha}{2}} + \frac{\Delta}{\text{s.e.}(\hat{\delta})}\right] \\ = & \Phi\left[-Z_{1-\frac{\alpha}{2}} - \frac{\Delta}{\sqrt{\boldsymbol{C}_{\boldsymbol{\theta}} \left(\sum_{i=1}^{I} \sum_{j=1}^{J_i} \mathbf{X}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{X}_{ij} \right)^{-1} \boldsymbol{C}_{\boldsymbol{\theta}}^T}}\right] \\ & + \Phi\left[-Z_{1-\frac{\alpha}{2}} + \frac{\Delta}{\sqrt{\boldsymbol{C}_{\boldsymbol{\theta}} \left(\sum_{i=1}^{I} \sum_{j=1}^{J_i} \mathbf{X}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{X}_{ij} \right)^{-1} \boldsymbol{C}_{\boldsymbol{\theta}}^T}}\right], \end{aligned} \tag{2.5}$$

where $\Phi(\cdot)$ and Z_p are the cumulative distribution function and the $100 \times p$th percentile of the standard normal distribution, respectively. The first term on the right hand side is small and is commonly ignored in practice.

The model we propose expands upon that of Senn [2019] in deriving sample sizes for n-of-1 trials in several ways. First, Senn [2019] assumed intervention and reference treatments are randomized to each pair of consecutive treatment periods and that limits the sequences to which participants can be assigned. Second, he worked with the differences between the outcomes in each pair of consecutive treatment periods where the outcome on each treatment was measured once. Third, he

assumed no correlation among the repeated measurements on each participant. In our formulation, we allow more flexible sequences with different orders of treatments assigned to the treatment periods, work with outcome measurements directly so the number of measurements in each period can also vary, and allow correlation among the repeated measurements on each participant.

2.5 Standard error of individual-specific treatment effect estimates

We derive the standard error of the individual-specific treatment effect estimates in this section. Section 2.5.1 and 2.5.2 present the standard error of naive and shrunken estimates respectively.

2.5.1 Naive estimates

To specify the heterogeneous residual error variance and correlation among the repeated measurements on the participant of interest, we write Model (2.1) in vector notation for participant j^* assigned to sequence i^* of interest, for whom we will derive the standard error of the naive estimate for the individual-specific treatment effect,

$$\boldsymbol{Y}_{i^*j^*} = \mathbf{X}_{i^*j^*}\boldsymbol{\theta}_{i^*j^*} + \boldsymbol{\epsilon}_{i^*j^*}, \quad \text{Var}(\boldsymbol{\epsilon}_{i^*j^*}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}i^*j^*}.$$

Here, $\boldsymbol{Y}_{i^*j^*}$ and $\boldsymbol{\epsilon}_{i^*j^*}$ are the outcome and the residual error vectors, respectively, of length $\sum_{k=1}^{K_{i^*}} L_{i^*j^*k}$, $\mathbf{X}_{i^*j^*}$ is the design matrix with one column for intercept and the other for treatment indicators corresponding to the treatment periods and $\boldsymbol{\theta}_{i^*j^*} = (\mu_{i^*j^*}, \delta_{i^*j^*})^T$. Specifications of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}i^*j^*}$ allow heterogeneous residual variance and correlation among the repeated measurements on the participant of interest.

The standard error of the resulting naive estimate for the individual-specific treatment effect, $\hat{\delta}_{i^*j^*}$, is

$$\text{s.e.}(\hat{\delta}_{i^*j^*}) = \sqrt{\boldsymbol{C}_{\boldsymbol{\theta}_{i^*j^*}}(\mathbf{X}_{i^*j^*}^T \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}i^*j^*}^{-1} \mathbf{X}_{i^*j^*})^{-1} \boldsymbol{C}_{\boldsymbol{\theta}_{i^*j^*}}^T}, \tag{2.6}$$

where $\boldsymbol{C}_{\boldsymbol{\theta}_{i^*j^*}} = (0, 1)$.

The naive estimates have limitations which show the advantage of the shrunken estimates. At least two treatments are required to estimate the treatment effect and standard error using the

data from a single participant. Additionally, estimating $\mathbf{\Sigma}_{\boldsymbol{\epsilon}_{i^*j^*}}$ using data from only one participant results in considerable uncertainty [Senn, 2019].

2.5.2 Shrunken estimates

Model (2.2) suggests the shrunken estimate for the individual-specific treatment effect for a given participant j^* assigned to sequence i^*, $\hat{\delta}_{i^*j^*} = \hat{\delta} + \hat{\gamma}_{\delta i^*j^*}$. Appendix A.1 shows that the variance of $\hat{\delta}_{i^*j^*} - \delta_{i^*j^*}$, which accounts for the additional variation in random treatment effects [Laird and Ware, 1982], is

$$
\begin{aligned}
&\mathrm{Var}(\hat{\delta}_{i^*j^*} - \delta_{i^*j^*}) \\
=&\boldsymbol{C_\theta}\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i}\mathbf{X}_{ij}^T\mathbf{\Sigma}_{ij}^{-1}\mathbf{X}_{ij}\right)^{-1}\boldsymbol{C_\theta}^T - 2\boldsymbol{C_\theta}\left[\sum_{i=1}^{I}\sum_{j=1}^{J_i}\mathbf{X}_{ij}^T\mathbf{\Sigma}_{ij}^{-1}\mathbf{X}_{ij}\right]^{-1}\mathbf{X}_{i^*j^*}^T\mathbf{\Sigma}_{i^*j^*}^{-1}\mathbf{Z}_{i^*j^*}\mathbf{D}\boldsymbol{C_b}^T + \\
&\boldsymbol{C_b}\left[\mathbf{D} - \mathbf{D}\mathbf{Z}_{i^*j^*}^T\mathbf{\Sigma}_{i^*j^*}^{-1}\mathbf{Z}_{i^*j^*}\mathbf{D} + \mathbf{D}\mathbf{Z}_{i^*j^*}^T\mathbf{\Sigma}_{i^*j^*}^{-1}\mathbf{X}_{i^*j^*}\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i}\mathbf{X}_{ij}^T\mathbf{\Sigma}_{ij}^{-1}\mathbf{X}_{ij}\right)^{-1}\mathbf{X}_{i^*j^*}^T\mathbf{\Sigma}_{i^*j^*}^{-1}\mathbf{Z}_{i^*j^*}\mathbf{D}\right]\boldsymbol{C_b}^T
\end{aligned}
\tag{2.7}
$$

where $\boldsymbol{C_b}$ is 1 for the model with fixed intercepts (i.e., Fixed-Random in Table 2.1) and is $(0, 1)$ for the model with random intercepts (i.e., Random-Random in Table 2.1); the remaining terms are as defined in Table 2.1. The square root of the variance gives the standard error for the shrunken estimate.

2.6 Finding design components in a series of n-of-1 trials

We recommend that practitioners take the following two steps when designing n-of-1 trials. Firstly, use the sample size formula for estimating the population average treatment effect to find the possible combinations of the design components that achieve the pre-specified power requirement; secondly, if they are further interested in estimating the individual-specific treatment effects, they will finalize the design by picking the combinations of the design components that also satisfy the standard error requirements for the individual-specific treatment effect estimates. Illustrations in Section 2.7 and the Shiny app described in Section 2.8 will follow these two steps. In this section, we discuss the ways to find the possible combinations of the design components using the formulae

in Section 2.4 and 2.5.

Among the design components, the sequences with different orders of treatments assigned to periods in the sequences, which in turn determines the number of sequences I and the number of treatment periods in each sequence K_i, the number of participants assigned to each sequence J_i, and the number of measurements in each treatment period L_{ijk}, where $i = 1, \ldots, I$, $j = 1, \ldots, J_i$ and $k = 1, \ldots, K_i$, the first step is to determine the sequences in the series of trials.

Sequences can be specified in two general ways, manually or by randomization. Manual determination usually pre-specifies sequences for a specific purpose. For example, one might wish to start with a placebo, followed by an intervention and then alternate these two treatments a certain number of times. Or one might constrain the number of times the same treatment could be given consecutively when choosing sequences.

We can also follow randomization schemes to determine the sequences. Randomization schemes include but are not limited to: 1) alternating sequences, where the two treatments alternate with the first period assigned at random, 2) pairwise randomization, randomly allocating the order of the two treatments in each consecutive pair of treatment periods, where we refer to the consecutive pairs of treatment periods as blocks, 3) restricted randomization, randomly assigning treatments in the sequence with the restriction that each treatment is assigned to the same number of periods, or 4) unrestricted randomization, completely randomizing treatments to treatment periods in the sequence [Johannessen and Fosstvedt, 1991].

If the number of treatment periods in the sequences are the same (i.e., $K_1 = \cdots = K_I = K$), we can derive the number of sequences I from the number of treatment periods K in each sequence following the randomization schemes (Table 2.2). Except under unrestricted randomization, each treatment is generally assigned to the same number of treatment periods in each sequence so all the sequences in a two treatment design will have an even number of treatment periods. Here, we include odd number of treatment periods for completeness. For odd number of treatment periods in each sequence: 1) under pairwise randomization, we randomly allocate the order of the two treatments in each consecutive pair of periods in the first $K - 1$ periods and randomly assign a treatment to the last period; 2) under restricted randomization, we randomly assign treatments with the restriction that there is only one period difference between the number of periods assigned with the two treatments, regardless of the direction.

Randomization scheme	Odd K	Even K
Alternating sequences	2	
Pairwise randomization	$2^{\frac{K+1}{2}}$	$2^{\frac{K}{2}}$
Restricted randomization	$2\binom{K}{(K-1)/2}$	$\binom{K}{K/2}$
Unrestricted randomization	2^K	

Table 2.2: The number of sequences I under different randomization schemes given the number of treatment periods K in the sequence. "Odd K" and "Even K" refer to scenarios where there are odd and even number of treatment periods in the sequences respectively. When there are odd number of periods in each sequence: 1) under pairwise randomization, we randomly allocate the order of the two treatments in each consecutive pair of crossover periods in the first $K-1$ periods and randomly assign a treatment to the last period; 2) under restricted randomization, we randomly assign treatments with the restriction that there is only one period difference between the number of periods assigned with the two treatments, regardless of the direction.

We will be able to determine the number of sequences I and the number of treatment periods K_i when sequences are specified manually and the relationship between I and K when following randomization schemes for designs with the same number of periods across sequences. The next step is to determine the number of participants assigned to each sequence J_i and the number of measurements in each treatment period L_{ijk}.

We focus on balanced designs in which the number of participants assigned to each sequence, J_i, the number of treatment periods in each sequence, K_i, and the number of measurements in each treatment period, L_{ijk}, are the same so that $J_i = J$, $K_i = K$, and $L_{ijk} = L$. Unbalanced designs where there can be different numbers of treatment periods in the sequences, different numbers of participants assigned to the sequences, or different numbers of measurements in the treatment periods are much more complex and have more moving parts making optimization difficult. Furthermore, if balance is not desired, the nature of the imbalance is often specified. It will be possible to fix one or more of the design components and solve a constrained optimization problem.

Because both the number of participants, IJ and the number of repeated measurements on each participant, KL, affect the cost and practicality of the series of trials [Senn, 2002], determining J and L once I and K are determined trades off between the number of participants and the number of repeated measurements per participant. We can therefore fix either KL or IJ and then find the other.

Specifically, if we choose to first fix KL, I and K are determined when investigators specify sequences manually, and L will be fixed because K is determined; when following randomization

schemes we find the possible combinations of K and L that lead to the fixed product and I will be determined by K following the relationship in Table 2.2. We will then be able to calculate the possible values for the number of participants assigned to each sequence, J, using Equation (2.5) in both scenarios. Alternatively, if we choose to first fix IJ, following a similar procedure, we will also be able to calculate the possible values for the last element in this case, the number of measurements in each treatment period, L. These calculations find the possible combinations of the design components that satisfy the power requirement for estimating the population average treatment effect.

One can then finalize the design using Equations (2.6) or (2.7) to calculate the standard errors of the individual-specific treatment effect estimates if these are of interest and choose the designs that also achieves the required precision.

2.7 Comparison of designs

To illustrate the trade-offs between the number of participants and the number of measurements per participant, we consider a balanced design with pairwise randomization, probably the most common type of n-of-1 design. In balanced designs, the dimensions of within-individual error variance matrices, $\mathbf{\Sigma}_{\boldsymbol{\epsilon}ij}$'s, are the same across participants. We assume that the variance matrices do not vary by participants and equal to $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$. Additionally, the residual errors are homogeneous with a variance of 4 and have an AR-1 correlation structure with a correlation coefficient of 0.4. We set a minimal clinically important treatment effect of $\Delta = 1$, Type I error rate of $\alpha = 0.05$ and Type II error rate of $\beta = 0.2$. When applicable in the model, the variance of the random intercepts (σ^2_μ) and slopes (σ^2_δ) are 4 and 1 respectively, and the covariance of the random intercepts and the random slopes ($\sigma_{\mu,\delta}$) is 1 (i.e., the correlation between the random intercepts and random slopes is 0.5).

Figure 2.2 shows the change in the average required number of measurements across a series of n-of-1 trials ($IJKL$) as a function of the number of measurements per participant (KL, left) and the number of participants (IJ, right) for optimized designs when fixed-intercept models are used for estimating the population average treatment effect. Optimized designs refer to designs that achieve criteria with the smallest possible value of the last element with all the other elements in

I, J, K and L fixed. For example, if given the number of sequences I, the number of participants assigned to each sequence J, and the number of treatment periods in each sequence K, a design with $L \geq 5$ measurements in each treatment period will achieve the pre-specified power, the design with $L = 5$ measurements in each treatment period will be the optimized design and will be presented. Analogous optimization applies if I, J or K is the last component. The shaded area around the lines represent the range of the required number of measurements across trials from different combinations of K and L (left) and of I and J (right) that lead to the same product and the dots on the lines represent the average if there are multiple combinations. As Appendix A.2.1 shows that the form of intercepts has almost no effect on the optimized designs for estimating the population average treatment effect, we show results from models with fixed intercepts and by the form of slopes. Legha et al. [2018] also reported that using restricted maximum likelihood to estimate the population average treatment effect, which gives unbiased estimates of the variances compared with our assumed known variances, the coverage for the effect estimate with fixed or random intercepts are very similar for continuous outcomes.

In general, the required number of measurements across a series of n-of-1 trials is larger when we use the average of random slopes to estimate the population average treatment effect compared to when using a common slope for estimation because the random slope model contains an extra source of variability from the between-individual variance.

Additionally, the required number of measurements across trials stays stable with increasing number of measurements per participant when a common slope is used for estimation, while that increases when we allow random effects around the slope; the required number of measurements across trials decreases with increasing number of participants for models with either form of slopes and the decreasing rate is faster when the random-slope model is used for estimation, which leads to the decreasing discrepancy between the required number of measurements across trials for the common- and random-slope model with more participants in trials. These results are consistent with those in Senn [2019]. There, in a special case of our model described earlier (trials with independent repeated measurements on each participant, one measurement per treatment period and the same number of periods for the two treatments in the sequence), the required number of measurements across trials varied little with the number of measurements per participant for the common effect approach but increased for the random-effect approach. The discrepancy between

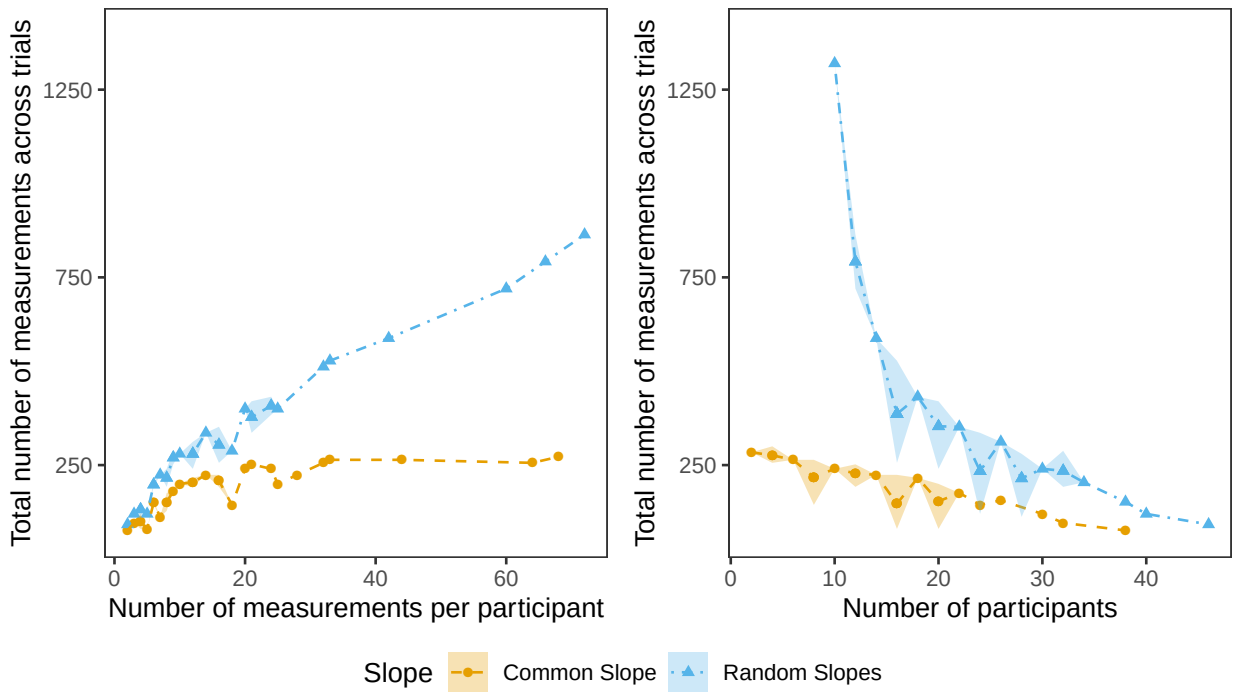

Figure 2.2: Average required number of measurements across a series of n-of-1 trials versus number of measurements per participant (KL, left) and number of participants (IJ, right) for optimized designs when fixed-intercept models are used for estimating the population average treatment effect. "Common Slope" and "Random Slopes" refer to "Fixed-Common" and "Fixed-Random" models in Table 2.1, respectively.

the required number of measurements per participant for the two approaches also decreased with more participants in trials.

Figure 2.3 shows the standard errors of the naive and shrunken estimates for individual-specific treatment effects versus the number of measurements per participant given the number of participants (fixed at 32, left) and versus the number of treatment periods in the sequence further given the number of measurements per participant (fixed at 24, right) for all possible designs that satisfy the power requirement for estimating population average treatment effect. The shaded areas around the lines represent the range of standard errors from different combinations of K and L that lead to the same product (left) and from trials with the same number of treatment periods but with different orders of treatments assigned to the periods (right) and the dots on the lines represent the average if there are multiple combinations. Note that all possible designs for the series of n-of-1 trials are plotted in the figure because as described in Section 2.6 in practice, after using Figure 2.2 to find designs that satisfy the power requirement for estimating the population average treatment effect, we want to further use Figure 2.3 to finalize the design by picking from all the possible designs that satisfy both the power requirement for estimating population average treatment effect and the standard error requirement for estimating the individual-specific treatment effect. For example, if the required total number of measurements is reasonable when we recruit 32 participants in a series of n-of-1 trials in Figure 2.2 and we require the standard error of the shrunken estimates with fixed intercepts for individual-specific treatment effects to be lower than 1, all the designs on the "Shrunken Estimates-Fixed Intercepts" curve in Figure 2.3 will satisfy the requirement. Additionally, if we are able to measure the outcome 24 times on each participant, Figure 2.3 (right) gives all the possible designs. A specific design can be a series of trials with $I = 4$ possible sequences, $J = 8$ participants assigned to each sequence, $K = 4$ treatment periods per sequence, and $L = 6$ measurements per treatment period. Detailed information for all the possible designs is given in the Shiny app (part (e) in Figure 2.4).

As expected, the results in Figure 2.3 show the advantage of shrunken estimates over naive estimates for estimating individual-specific treatment effect. Given fixed number of participants, the standard error of naive estimates is larger than that of shrunken estimates. The difference decreases with increasing number of measurements per participant, but even then the standard errors for naive estimates are larger. Naive estimates benefit more from larger number of treatment

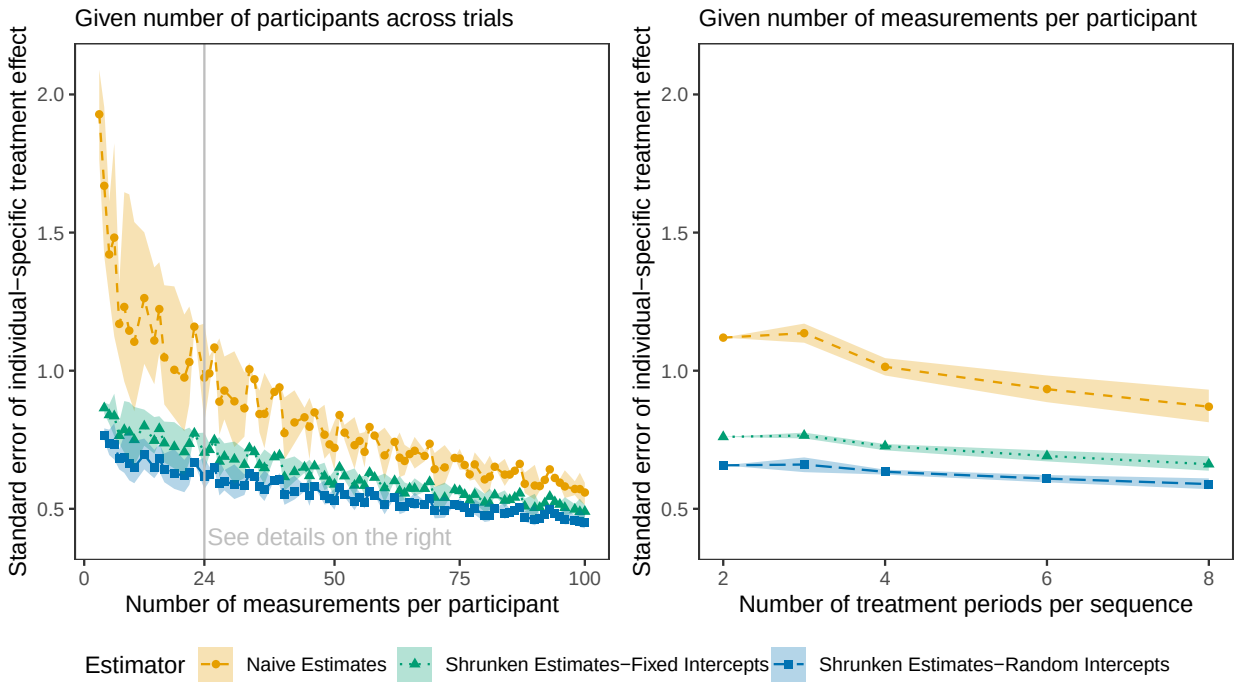

Figure 2.3: Standard error of naive and shrunken estimates for individual-specific treatment effect versus number of measurements per participant given total number of participants across trials (fixed at 32, left) and versus number of treatment periods per sequence further given number of measurements per participant (fixed at 24, right) for all possible designs that satisfy the power requirement for estimating population average treatment effect. "Naive Estimates", "Shrunken Estimates-Fixed Intercepts" and "Shrunken Estimates-Random Intercepts" refer to the standard error of naive estimates, shrunken estimates in the fixed- and random-intercept model, respectively.

periods in the sequence when we further fix the number of measurements on each participant, with a larger drop in standard error when we increase the number of treatment periods in the sequence compared to shrunken estimates. We also found in Figure 2.3 that the standard error of the shrunken estimates in the random-intercept model is slightly smaller than that in the fixed-intercept model.

To evaluate the sensitivity of the results to the varying parameters, Appendix A.2 includes the following additional illustrations:

- Appendix A.2.2 presents results where we use alternative values for the type I and II errors. The average required number of measurements across trials increases with smaller type I and II errors; the rate of increase is higher if 1) we want to reduce smaller type I and II errors, 2) the number of measurements per participant is larger, and 3) the number of participants in trials is smaller.

- Appendix A.2.3 presents results with alternative parameterizations for the residual error variance matrix. We show results assuming 1) independent and exchangeable correlation structure, and different values for 2) the homogeneous residual error variance and 3) residual correlation coefficient under first order autoregressive correlation structure. Trials with exchangeable correlation structure require the fewest measurements across trials, followed by independent correlation structure. Under first order autoregressive correlation structure, the required number of measurements across trials 1) is similar to that under exchangeable correlation structure when the number of measurements per participant is small and the correlation coefficient is large, 2) becomes larger than that under independent correlation structure when the number of measurements per participant is large and the correlation coefficient is small, and 3) increases linearly with homogeneous variance and the rate of increase is higher when the number of measurements per participant is larger and when the number of participants in trials is smaller.

- Appendix A.2.4 presents results with alternative parameterizations for the random-effect variance matrix. Variance of random intercepts and correlation between random intercepts and slopes do not affect the optimized designs that satisfy the power requirement. Holding the number of measurements per participant the same, the average required number of measure-

ments across trials increases linearly with the variance of random slopes; the rate of increase is larger when the number of measurements per participant is larger. Holding the number of participants the same, the average required number of measurements across trials increases more at larger values for the variance of random slopes.

- Appendix A.2.5 presents results with alternative minimal clinically important treatment effects. The required number of measurements across trials increases with smaller minimal clinically important treatment effect; the rate of increase is higher if 1) we want to reduce a smaller minimal clinically important treatment effect, 2) the number of measurements per participant is larger, and 3) the number of participants in trials is smaller.

2.8 Shiny app

We implemented our methods in a Shiny app to allow investigators to design n-of-1 trials interactively without requiring programming knowledge or familiarity with a specific software. The Shiny app is available at `jiabeiyang.shinyapps.io/SampleSizeNof1/`.

Figure 2.4 presents the layout of the Shiny app, where we replicated Figure 2.2 (right) and 2.3 in Section 2.7. In Figure 2.4, part (a) shows the input panel of the app, where investigators are allowed to specify the parameters for designing the series of n-of-1 trials. Part (b)-(e) present information for the possible designs that will satisfy the power and standard error requirements specified by the investigators. Part (b) presents the specified parameters in the input panel and how the average required number of measurements across trials changes as a function of the number of participants for optimized designs (Figure 2.2, right). Part (c) and (d) show the detailed design information for optimized designs and the standard errors of the individual-specific treatment effect estimates for all possible designs that satisfy the power requirement (Figure 2.3), respectively, when one fixes the number of participants across trials by clicking on a specific point in the figure in part (b). Part (e) shows the detailed information for all possible designs that satisfy the power and standard error requirements when one further fixes the number of measurements per participant by clicking on a point in the figure in part (d). If investigators want to know how the average required number of measurements across trials changes as a function of the number of measurements per participant (Figure 2.2, left), they can choose "Total # of measurements vs. # of measurements

per participant" under "Design option" in the input panel. The output in part (d) of the Shiny app will be replaced accordingly by presenting the standard errors versus the number of participants in the trials given the number of measurements per participant.

The Shiny app implements both alternating sequences and pairwise randomization and also allows the user to upload manually specified sequences through "Possible sequences"-"User-specified sequences" in the input panel. Because the number of possible sequences under restricted and unrestricted randomization becomes large as the number of treatment periods in the sequences increases (Table 2.2) and many sequences may be impractical if the same treatment is given in too many consecutive treatment periods, users can complete calculations for these two randomization schemes by picking specific sequences of interest and uploading them through "User-specified sequences".

Additionally, we provide an option to only optimize designs over the scale of the y-axis in part (b) of the Shiny app. Because of the current optimization, even if we allow large number of participants in the trials, the maximum number of participants presented in part (b) of the Shiny app will be small because designs with more participants are not optimized using our definition of optimized designs. Therefore, this option will present all the possible designs within the specified range for the x-axis, only optimized on the scale of the y-axis. Finally, if investigators are not interested in estimating the individual-specific treatment effects, they can clear "Calculate standard error for individual-specific treatment effect estimates" and only part (b) and (c) will be displayed in the output.

2.9 Discussion

We present a procedure for calculating the sample size for n-of-1 trials. We formally define the design components for determining the sample size of a series of n-of-1 trials which include the sequences with different orders of treatments assigned to periods in the sequences, the number of participants assigned to each sequence, and the number of measurements in each treatment period. We present models for analyzing n-of-1 trials and use them to derive the required sample size for estimating population average treatment effect and the standard error of individual-specific treatment effect estimates. We recommend that investigators first use the sample size formula to

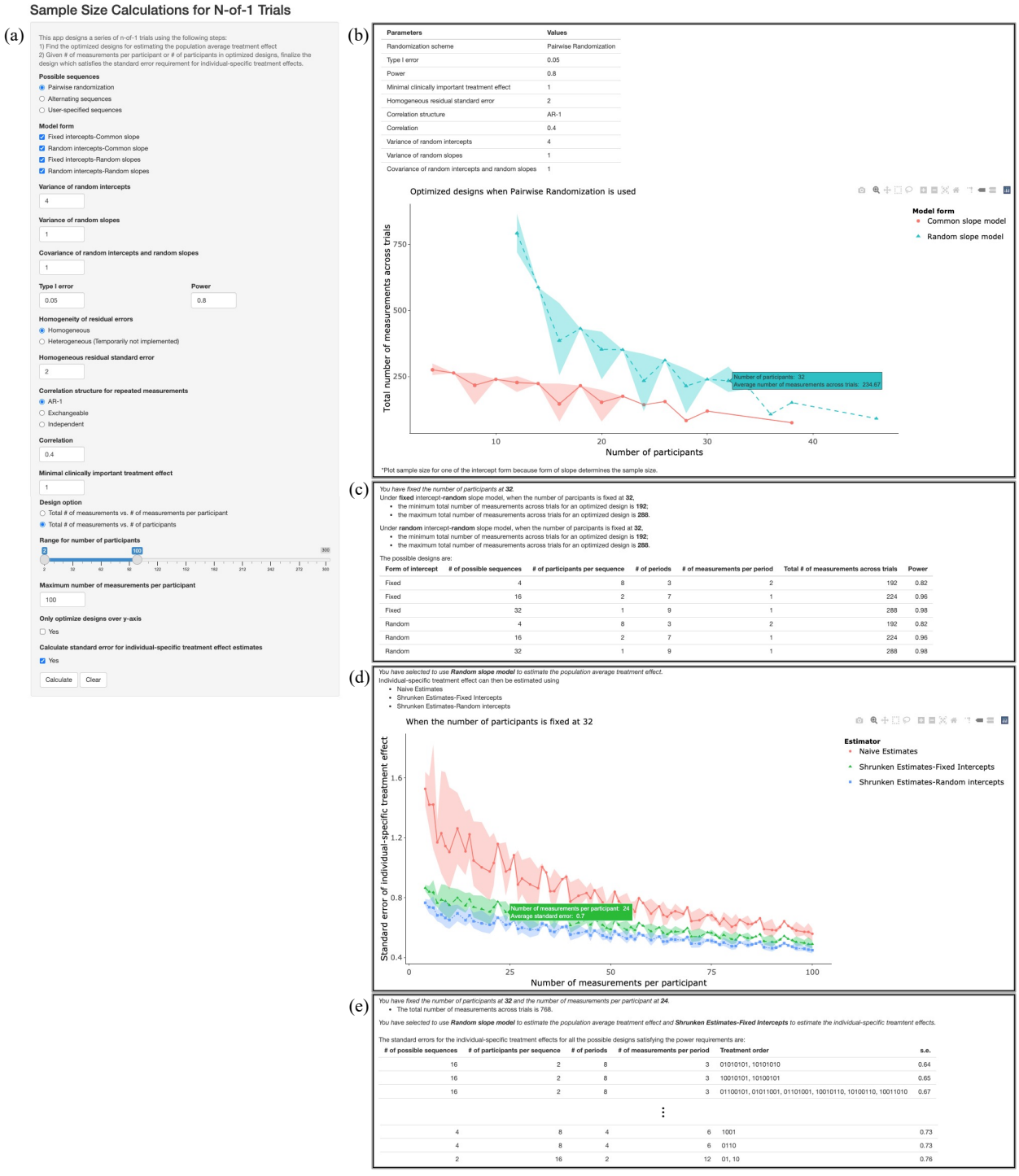

Figure 2.4: Screenshot of Shiny app. Part (a) allows investigators to specify parameters for designing the series of n-of-1 trials; part (b)-(e) present the possible designs that satisfy the power and standard error requirements specified by the investigators.

find the possible combinations of the design components that will satisfy the power requirement for estimating the population average treatment effect, and, if of interest, use the standard error formulae to pick the combinations of design components that will also satisfy the standard error requirements for the individual-specific treatment effect estimates. We implement and illustrate the procedure in the paper and through a Shiny app.

Several directions of future research are practically useful. First, the current derivations assume that the variance components, either for residual errors or for random effects, are known or estimated with adequate precision. Taking into account the fact that the variance components are estimated will involve using a distribution instead of the standard normal distribution when deriving power in Section 2.4 and will facilitate more accurate planning of the n-of-1 trials. Second, it will be helpful to extend the results to discrete outcomes, although correlation among the repeated measurements would then need to be handled differently [Zeger, 1988]. Additionally, it will be valuable to extend the results to trials with more than two treatments [Kravitz et al., 2020] and with multiple outcomes of interest [Barr et al., 2015]. Lastly, to avoid carryover effects, we can introduce washout periods into the derivations when switching from one treatment to another. These periods will limit the number of switches between different treatments in the sequences.

Chapter 3

Causal Interaction Trees: Finding Subgroups with Heterogeneous Treatment Effects in Observational Data

3.1 Introduction

Subgroup identification in randomized trials aims to find subsets of participants that have enhanced treatment effects, in order to support targeted treatment recommendations. This is typically done by performing subgroup analyses or by exploring treatment-covariate statistical interactions in outcome regression models [Dahabreh et al., 2016, 2017]. Non-parametric data-driven approaches for subgroup identification rely on extensions of the Classification and Regression Tree algorithm [Breiman et al., 1984] to explore treatment-covariate interactions [Su et al., 2009, Steingrimsson and Yang, 2019]. Tree-based methods use splitting criteria to recursively partition the covariate space (until some pre-determined stopping criterion is met) in order to create a large and potentially overfit tree that can be used as a prediction model. To reduce overfitting, a subtree of the large tree is selected using pruning criteria. Tree-based methods are appealing for subgroup identification because they can identify treatment-covariate interactions without the need to pre-specify which

interactions to include in the model. An early example of a tree-based algorithm for subgroup identification in randomized trials is the Interaction Tree algorithm [Su et al., 2009]. The algorithm makes splitting decisions by contrasting treatment effect estimates between groups, where the treatment effects are estimated by differences in outcomes between the treatment arms. Lipkovich et al. [2017] review data-driven subgroup identification methods for randomized trials.

In observational data, where treatment is not randomly assigned, confounding of the treatment-outcome relationship complicates subgroup identification. Few studies have assessed tree-based methods for subgroup identification with observational data. Su et al. [2012] and Kang et al. [2014] proposed a likelihood-based splitting statistic with AIC-type pruning criteria, which requires specifying the distribution of the outcome conditional on covariates. Athey and Imbens [2016] proposed the Causal Tree algorithm where splitting decisions minimize an estimator of the mean squared error of the subgroup-specific treatment effect. Wager and Athey [2018] used the Causal Tree algorithm to build a random forest algorithm, an ensemble method that averages multiple fully grown trees. Finally, Powers et al. [2018] proposed ensemble methods to control confounding using inverse probability of treatment weighting. Ensemble methods create black-box individualized prediction models, whereas single tree methods aim to produce clinically interpretable subgroups.

In this paper, we introduce Causal Interaction Tree (CIT) algorithms, a group of novel tree-based algorithms for subgroup identification with observational data. Causal Interaction Trees are extensions of the Interaction Tree algorithm of Su et al. [2009] that use subgroup-specific treatment effect estimators appropriate for observational data for decision-making during tree construction. We have chosen the name "Causal Interaction Tree" for consistency with prior work on Interaction Trees [Su et al., 2009]; the term "causal" alludes to the need to control confounding in order to estimate causal effects and the term "interaction" connects with the literature on methods for detecting effect measure modification [VanderWeele, 2009] by examining statistical interactions [Rothman et al., 1980]. CIT algorithms use inverse probability weighting, g-formula, or doubly robust estimators to estimate subgroup-specific treatment effects. Inverse probability weighting estimators require a correctly specified model for the conditional probability of treatment given covariates (i.e., the propensity score [Rosenbaum and Rubin, 1983]) whereas g-formula estimators require a correctly specified model for the conditional expectation of the outcome given covariates [Robins, 1986]. Doubly robust estimators use a model for the probability of treatment and a model

for the expectation of the outcome and are consistent when either model is correctly specified, but not necessarily both [Bang and Robins, 2005].

In Section 3.2, we begin by defining the Generalized Interaction Tree (GIT) algorithm that includes, as special cases, the Interaction Tree algorithm of Su et al. [2009] and the Covariate Adjusted Interaction Tree algorithm of Steingrimsson and Yang [2019] for randomized trials, as well as the three novel CIT algorithms for observational studies. In Section 3.3, we derive the properties of three subgroup-specific treatment effect estimators that dictate decision-making in the CIT algorithms. In Sections 3.4 and 3.5 we evaluate the performance of the CIT algorithms in simulations and analyses of data from an observational study that evaluated the effectiveness of right heart catheterization on critically ill patients. The Web Appendix contains proofs, additional simulation results, and details regarding the data analyses.

3.2 Generalized Interaction Tree (GIT) Algorithm

Let Y be an outcome measured at the end of followup (binary, continuous, or count); A be an indicator for exposure to treatment that equals 1 when the participant is exposed and equals 0 when not exposed; $\boldsymbol{X}$ be a vector of pre-treatment covariates taking values in $\mathcal{X}$. Any set w that is a subset of $\mathcal{X}$ defines a population subgroup consisting of individuals with $\boldsymbol{X} \in w$. The data collected are assumed to consist of n i.i.d. observations of $(\boldsymbol{X}, A, Y)$. We define $n(w) = \sum_{i=1}^{n} I(\boldsymbol{X}_i \in w)$ to be the number of sampled observations with $\boldsymbol{X} \in w$. Let Y^a be the potential outcome under intervention to set treatment to a, for $a \in \{0, 1\}$ [Rubin, 1974, Robins and Greenland, 2000]. The average treatment effect for the population subgroup with $\boldsymbol{X} \in w$ is defined as $\mathrm{E}(Y^1|\boldsymbol{X} \in w) - \mathrm{E}(Y^0|\boldsymbol{X} \in w) = \mu_1(w) - \mu_0(w)$, where $\mu_a(w) = \mathrm{E}(Y^a|\boldsymbol{X} \in w)$ for $a \in \{0, 1\}$. We are interested in finding a collection of mutually exclusive and exhaustive subsets of $\mathcal{X}$, say, $w_1, \ldots, w_{\widehat{K}}$, that stratify the population into subgroups with heterogeneous treatment effects. The number of identified subgroups, $\widehat{K}$, is estimated from the data (see the algorithm defined below). Our approach for identifying subgroups relies on estimating $\mu_a(w)$ for $a \in \{0, 1\}$ and we use $\widehat{\mu}_a(w)$ to denote a generic estimator of $\mu_a(w)$. In Section 3.3, we describe three specific estimators of $\mu_a(w)$ that can be used with observational data.

We now describe the GIT algorithm (see display for a summary) which unifies prior work on

Algorithm Generalized Interaction Tree (GIT) algorithm

Result:

A tree that partitions the covariate space $\mathcal{X}$ into a set of mutually exclusive and exhaustive subsets, that is, $w_1, \ldots, w_{\widehat{K}}$.

Initialize:

Split the data into an initial tree building dataset and a validation dataset.

1: Create a maximum sized tree $\widehat{\psi}_{\max}$ using the initial tree building dataset:

(a) Define the root node of tree $\widehat{\psi}_{\max}$ as consisting of all observations in the initial tree building dataset. Set the root node as the node of interest.

(b) In the node of interest, identify all permissible $(X^{(j)}, c)$ pairs that split the covariate space into two groups $l = \{X^{(j)} < c\}$ and $r = \{X^{(j)} \geq c\}$.

(c) Consider all such splits in Step 1(b) and divide the node of interest into two mutually exclusive subgroups using the split that gives the largest splitting statistic (as defined in expression (3.1)).

(d) Check pre-determined stopping criteria. If met, denote the current tree as $\widehat{\psi}_{\max}$ and move to step 2 of the algorithm; otherwise, repeat Step 1(b)-1(d) on every node that has not met the stopping criteria (and is not already split into child nodes).

2: Prune $\widehat{\psi}_{\max}$ to create a sequence of candidate trees, $\widehat{\psi}_0, \widehat{\psi}_1, \ldots, \widehat{\psi}_{\widehat{M}}$:

(a) Set $m = 0$ and $\widehat{\psi}_0 = \widehat{\psi}_{\max}$.

(b) Set m to $m + 1$.

(c) Define $\widehat{\psi}_m$ as the subtree of $\widehat{\psi}_{m-1}$ with all the descendants of node h' removed where h' minimizes $g(h)$ among all nodes in tree $\widehat{\psi}_{m-1}$, i.e., $h' = \arg\min_{h \in Q_{\widehat{\psi}_{m-1}}} g(h)$.

(d) Repeat Step 2(b)-2(c) until $\widehat{\psi}_m$ consists only of the root node.

3: Select the final tree from the candidate trees using the validation dataset:

(a) Fix the value of the penalization parameter λ at some quantile of the asymptotic distribution of the splitting statistic defined in expression (3.1).

(b) For each candidate tree $\widehat{\psi}_m$, $\widehat{\psi}_m \in \{\widehat{\psi}_0, \widehat{\psi}_1, \ldots, \widehat{\psi}_{\widehat{M}}\}$, calculate the split complexity $G^{(\lambda)}(\widehat{\psi}_m)$ given by equation (3.2) using the validation set by sending observations down the tree to calculate the splitting statistics defined in (3.1) for each internal node.

(c) Select the final tree as the one that maximizes the validation set split complexity.

subgroup identification using randomized trial data [Su et al., 2009, Steingrimsson and Yang, 2019] and will serve as the basis for three novel algorithms that are appropriate for observational data (introduced in Section 3.3.5). The GIT algorithm begins by splitting the data into an initial tree building dataset and a validation dataset and consists of the following 3 steps:

1. *Creating a maximum sized tree:* We use the initial tree building dataset to create a maximum sized tree. At the beginning of the tree building process, all observations are in a single node, referred to as the root node. Define $X^{(j)}$ as the j-th component of the covariate vector $\boldsymbol{X}$, and let $c \in \mathbb{R}$. The pair $(X^{(j)}, c)$ splits the covariate space into two groups $l = \{X^{(j)} < c\}$ and $r = \{X^{(j)} \geq c\}$, with corresponding treatment effects $\widehat{\mu}_1(l) - \widehat{\mu}_0(l)$ and $\widehat{\mu}_1(r) - \widehat{\mu}_0(r)$, respectively. We define the splitting statistic corresponding to this split as

$$\left[\frac{\{\widehat{\mu}_1(l) - \widehat{\mu}_0(l)\} - \{\widehat{\mu}_1(r) - \widehat{\mu}_0(r)\}}{\sqrt{\widehat{\text{Var}}\left[\{\widehat{\mu}_1(l) - \widehat{\mu}_0(l)\} - \{\widehat{\mu}_1(r) - \widehat{\mu}_0(r)\}\right]}} \right]^2. \tag{3.1}$$

   For this step, all quantities in expression (3.1) are estimated using the initial tree building dataset. The splitting statistic (3.1) measures a standardized difference between the treatment effects in the two groups. When the true treatment effects are identical in the two subgroups, the splitting statistic defined in expression (3.1) converges to a χ^2-distribution with 1 degree of freedom. The node that is being considered for splitting is referred to as parent node and the two nodes that the data is split into are referred to as child nodes.

   To split the node into two subgroups, the GIT algorithm cycles through all permissible $(X^{(j)}, c)$ pairs, and selects the combination that gives the largest splitting statistic in expression (3.1). The algorithm will search through all possible combinations of levels for a categorical covariate $X^{(j)}$ and all possible splits that preserve the ordering for a continuous or an ordinal $X^{(j)}$. We iterate the process within each new subgroup until some pre-determined criteria are met. This procedure results in a large initial tree, $\widehat{\psi}_{\max}$.

2. *Pruning:* The pruning step creates a sequence of subtrees of $\widehat{\psi}_{\max}$ that are candidates for the final tree, reducing the computational complexity of the model selection process. This step was adapted from the Interaction Tree algorithm [Su et al., 2009].

In a given tree, nodes that are split are referred to as internal nodes; nodes that are not split are referred to as terminal nodes. Let a penalization parameter λ be given and define the split complexity for a tree ψ as

$$G^{(\lambda)}(\psi) = \sum_{i \in Q_\psi} G_i(\psi) - \lambda |Q_\psi|. \tag{3.2}$$

$G_i(\psi)$ is the value of the splitting statistic defined in expression (3.1) for internal node i in tree ψ; Q_ψ is the set of internal nodes of ψ; and $|Q_\psi|$ is the number of internal nodes.

Weakest link pruning creates a finite sequence of subtrees of $\widehat{\psi}_{\max}$ built on the initial tree building dataset by sequentially dropping the branch of the tree that has the smallest split complexity. More formally, it follows the steps below:

(a) Set $\widehat{\psi}_0 = \widehat{\psi}_{\max}$ and $m = 0$.

(b) Set m to $m + 1$.

(c) Define $g(h) = \sum_{i \in Q_{\psi_h^*}} G_i(\psi_h^*)/|Q_{\psi_h^*}|$ if $h \in Q_\psi$ and $g(h) = +\infty$ otherwise, where ψ_h^* is the subtree consisting of node h and all descendants of node h. The weakest link, defined in terms of split complexity, of tree $\widehat{\psi}_{m-1}$ is the node $h' = \arg\min_{h \in Q_{\widehat{\psi}_{m-1}}} g(h)$. Define $\widehat{\psi}_m$ as the subtree of $\widehat{\psi}_{m-1}$ with all descendants of h' removed.

(d) Repeat Step (b) and (c) until $\widehat{\psi}_m$ consists only of the root node.

Executing the pruning step results in a sequence of trees $\widehat{\psi}_0 = \widehat{\psi}_{\max}, \widehat{\psi}_1, \ldots, \widehat{\psi}_{\widehat{M}}$.

3. *Final tree selection:* The last step is to select a final tree from the sequence of trees generated in the pruning step, using the validation dataset. For a candidate tree $\widehat{\psi}_m$, $m \in \{0, \ldots, \widehat{M}\}$, we calculate the validation split complexity, as defined in equation (3.2), by sending each observation in the validation dataset down the candidate tree to re-calculate the splitting statistics defined in expression (3.1) for each internal node. This involves re-estimating $\widehat{\mu}_a(w)$ in expression (3.1) using the validation dataset. The final tree is selected as the one that maximizes the validation split complexity for a fixed penalization parameter λ. A common criterion for selecting the penalization parameter is some quantile of the asymptotic distribution of the splitting statistic (3.1) when the treatment effects are identical in the two

subgroups l and r.

The GIT algorithm partitions the covariate space $\mathcal{X}$ into $\widehat{K}$ mutually exclusive and exhaustive subsets, $w_1, \ldots, w_{\widehat{K}}$, each of which defines a subgroup of observations. The next step is to estimate subgroup-specific treatment effects and associated measures of uncertainty. Data-splitting can be used to obtain asymptotically valid confidence intervals for the final tree-based subgroup-specific treatment effect estimators. Data-splitting separates the data into two disjoint parts, with one part used for fitting the GIT algorithms and the other for subgroup-specific treatment effect estimation and confidence interval construction. As the data used for tree building and confidence interval construction are independent, it follows from general results on post-selection inference [Fithian et al., 2014] that the confidence intervals for subgroup-specific treatment effect estimators are asymptotically valid. Thus, if the goal of the analysis is to draw inferences about the subgroup-specific treatment effect estimators, data-splitting is recommended.

3.3 Subgroup-specific treatment effect estimators with observational data

Implementation of the GIT algorithm requires specifying an estimator $\widehat{\mu}_a(w)$ for $\mu_a(w) = \mathrm{E}(Y^a|\boldsymbol{X} \in w)$. In Sections 3.3.2-3.3.4, we define and examine properties of three estimators for $\mu_a(w)$ that can be used with observational data and in Section 3.3.5, we discuss the use of these estimators in connection with the GIT Algorithm.

3.3.1 Identifiability of subgroup-specific treatment effects

The following conditions are sufficient to identify $\mu_a(w)$ using the observed data.

1. Consistency of potential outcomes: for every individual exposed to treatment $A = a$, the observed outcome Y is equal to their potential outcome under the same treatment, Y^a. For binary A, this condition means that $Y = Y^1 A + Y^0(1 - A)$.

2. Mean exchangeability: $\mathrm{E}[Y^a|A = a, \boldsymbol{X} = \boldsymbol{x}] = \mathrm{E}[Y^a|\boldsymbol{X} = \boldsymbol{x}]$ for all covariate patterns $\boldsymbol{x} \in w$ that have a positive density.

3. Positivity: $0 < \mathrm{P}(A = 1|\boldsymbol{X} = \boldsymbol{x}) < 1$ for all covariate patterns $\boldsymbol{x} \in w$ that have a positive density.

The following theorem shows that the subgroup-specific potential outcome means $\mathrm{E}(Y^a|\boldsymbol{X} \in w)$ are identifiable using the observed data (see, e.g., Robertson et al. [2020]); for completeness, we provide a proof in Web Appendix B.1.1.

Theorem 3.3.1 *Under identifiability conditions 1-3, the subgroup-specific potential outcome mean* $\mathrm{E}[Y^a|\boldsymbol{X} \in w]$ *can be written as the observed data functional*

$$\mu_a(w) = \mathrm{E}[\mathrm{E}(Y|\boldsymbol{X}, A = a)|\boldsymbol{X} \in w], \tag{3.3}$$

or equivalently using the inverse probability weighting representation

$$\mu_a(w) = \frac{1}{\mathrm{P}(\boldsymbol{X} \in w)} \mathrm{E}\left[\frac{I(\boldsymbol{X} \in w, A = a)}{\mathrm{P}(A = a|\boldsymbol{X})} Y\right]. \tag{3.4}$$

3.3.2 Inverse probability weighting estimator

The identifiability result (3.4) suggests the inverse probability weighting (IPW) estimator,

$$\widehat{\mu}_{\mathrm{IPW},a}(w) = \frac{1}{n\widehat{\gamma}_w} \sum_{i=1}^{n} \frac{I(\boldsymbol{X}_i \in w, A_i = a)Y_i}{e_a(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}})} = \frac{1}{n(w)} \sum_{i=1}^{n} \frac{I(\boldsymbol{X}_i \in w, A_i = a)Y_i}{e_a(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}})}. \tag{3.5}$$

Here, $\widehat{\gamma}_w = \frac{n(w)}{n}$ is a nonparametric estimator for $\mathrm{P}(\boldsymbol{X} \in w)$, the probability of being in the subgroup defined by w; $e_a(\boldsymbol{X}; \widehat{\boldsymbol{\beta}})$ is an estimator for $\mathrm{P}(A = a|\boldsymbol{X})$, the probability of receiving treatment a given covariates $\boldsymbol{X}$, indexed by parameter $\boldsymbol{\beta}$. When implementing the GIT algorithm, we can estimate the conditional probability of treatment using the whole dataset (initial tree building dataset for steps 1 and 2 of the algorithm and validation dataset for step 3), the data in the parent node being considered for splitting, or the data in each of the child nodes for each possible split. Using subsets of the data allows the models to differ between subgroups which can lead to lower bias; however, it utilizes less information compared to using the whole dataset and is more computationally intensive. In simulations we evaluate the impact of using different subsets of the data to estimate $\boldsymbol{\beta}$. Throughout, we make the positivity assumption $e_a(\boldsymbol{X}; \widehat{\boldsymbol{\beta}}) \geq \varepsilon > 0$, for $a \in \{0, 1\}$. It follows from identifiability result (3.4) that if $e_a(\boldsymbol{X}; \widehat{\boldsymbol{\beta}})$ is correctly specified (i.e., $e_a(\boldsymbol{X}; \widehat{\boldsymbol{\beta}}) \xrightarrow{p} \mathrm{P}(A = a|\boldsymbol{X})$,

where "$\xrightarrow{p}$" denotes convergence in probability), $\widehat{\mu}_{\text{IPW},a}(w)$ is a consistent estimator for $\mu_a(w)$.

Define the true treatment effect difference between two disjoint subgroups defined by l and r as $T(l,r) = \{\mu_1(l) - \mu_0(l)\} - \{\mu_1(r) - \mu_0(r)\}$. The parameter $T(l,r)$ is the population analog of the numerator of the splitting statistic (3.1). Define $\widehat{T}_{\text{IPW}}(l,r)$ as the estimator of $T(l,r)$ obtained by replacing, in the definition of this parameter, the potential outcome means, $\mu_a(w)$, $a \in \{0,1\}$, $w \in \{l,r\}$, by the corresponding inverse probability weighting estimators.

The inverse probability weighting splitting statistic depends on an estimator of the variance of $\widehat{T}_{\text{IPW}}(l,r)$. The following theorem provides the asymptotic variance of $\widehat{T}_{\text{IPW}}(l,r)$ when the model for the conditional probability of treatment is a logistic regression model. Although not explicit in the notation, the model can also include interactions and higher-order terms.

Theorem 3.3.2 *Assume that the conditional probability of treatment is estimated using a correctly specified logistic regression model fit using the data falling in the union of two disjoint subgroups defined by $l \cup r$. Denote the estimated logistic regression coefficients as $\widehat{\boldsymbol{\beta}}_{LR}$ and the corresponding asymptotic limit as $\boldsymbol{\beta}_{LR}$. For a split into two subgroups defined by l and r and $w \in \{l,r\}$, define $\boldsymbol{H}_{\boldsymbol{\beta}_{LR},w} = \mathrm{E}\left[\left\{\frac{AY}{e_1(\boldsymbol{X};\boldsymbol{\beta}_{LR})} + \frac{(1-A)Y}{e_0(\boldsymbol{X};\boldsymbol{\beta}_{LR})}\right\}\frac{\partial}{\partial\boldsymbol{\beta}}e_1(\boldsymbol{X};\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{LR}}\Big|\boldsymbol{X} \in w\right]$ and $\mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \mathrm{E}\left[\frac{\frac{\partial}{\partial\boldsymbol{\beta}}e_1(\boldsymbol{X};\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{LR}}\left\{\frac{\partial}{\partial\boldsymbol{\beta}}e_1(\boldsymbol{X};\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{LR}}\right\}^T}{e_1(\boldsymbol{X};\boldsymbol{\beta}_{LR})e_0(\boldsymbol{X};\boldsymbol{\beta}_{LR})}\Big|\boldsymbol{X} \in (l \cup r)\right]$. The asymptotic variance of $\widehat{T}_{IPW}(l,r)$ is given by*

$$
\begin{aligned}
&\frac{1}{\mathrm{P}\{\boldsymbol{X} \in l | \boldsymbol{X} \in (l \cup r)\}}\,\mathrm{E}\left[\frac{(YA)^2}{e_1(\boldsymbol{X};\boldsymbol{\beta}_{LR})} + \frac{\{Y(1-A)\}^2}{e_0(\boldsymbol{X};\boldsymbol{\beta}_{LR})}\Big|\boldsymbol{X} \in l\right] \\
&+ \frac{1}{\mathrm{P}\{\boldsymbol{X} \in r | \boldsymbol{X} \in (l \cup r)\}}\,\mathrm{E}\left[\frac{(YA)^2}{e_1(\boldsymbol{X};\boldsymbol{\beta}_{LR})} + \frac{\{Y(1-A)\}^2}{e_0(\boldsymbol{X};\boldsymbol{\beta}_{LR})}\Big|\boldsymbol{X} \in r\right] - T(l,r)^2 \\
&- (\boldsymbol{H}_{\boldsymbol{\beta}_{LR},l} - \boldsymbol{H}_{\boldsymbol{\beta}_{LR},r})^T \mathbf{E}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}} (\boldsymbol{H}_{\boldsymbol{\beta}_{LR},l} - \boldsymbol{H}_{\boldsymbol{\beta}_{LR},r}) \\
&- \frac{[\mathrm{P}\{\boldsymbol{X} \in r | \boldsymbol{X} \in (l \cup r)\}\{\mu_1(l) - \mu_0(l)\} + \mathrm{P}\{\boldsymbol{X} \in l | \boldsymbol{X} \in (l \cup r)\}\{\mu_1(r) - \mu_0(r)\}]^2}{\mathrm{P}\{\boldsymbol{X} \in l | \boldsymbol{X} \in (l \cup r)\}\,\mathrm{P}\{\boldsymbol{X} \in r | \boldsymbol{X} \in (l \cup r)\}}. \qquad (3.6)
\end{aligned}
$$

We present a proof and give a consistent variance estimator in Web Appendix B.1.2.

The asymptotic variance result in Theorem 3.3.2 is an extension of the results on marginal treatment effect estimators from Lunceford and Davidian [2004] to subgroup-specific treatment

effect estimators. The variance (3.6) consists of five terms. The first three terms represent the variance where both the conditional probability of treatment and $\mathrm{P}\{\boldsymbol{X} \in l | \boldsymbol{X} \in (l \cup r)\}$ are known. The fourth term reflects the variance reduction when we use the estimated rather than the true conditional probability of treatment. A similar observation was made in Lunceford and Davidian [2004] for the marginal case. Interestingly, the last term in (3.6) shows that using a non-parametric estimator for $\mathrm{P}\{\boldsymbol{X} \in l | \boldsymbol{X} \in (l \cup r)\}$ results in further variance reduction compared to using the true unknown conditional probability.

3.3.3 g-formula estimator

The identifiability result (3.3) suggests the g-formula estimator
$\widehat{\mu}_{\mathrm{g},a}(w) = \frac{1}{n(w)} \sum_{i=1}^{n} I(\boldsymbol{X}_i \in w) g_a(\boldsymbol{X}_i; \widehat{\boldsymbol{\eta}}_a)$, where $g_a(\boldsymbol{X}; \widehat{\boldsymbol{\eta}}_a)$ is an estimator for $\mathrm{E}(Y|\boldsymbol{X}, A = a)$ indexed through a parameter $\boldsymbol{\eta}_a$. We can estimate the conditional expectation of outcome using either the whole dataset or only a subset of the data analogously to the conditional probability of treatment in the inverse probability weighting estimator. It follows from identifiability result (3.3) that if the outcome model is correctly specified (i.e., $g_a(\boldsymbol{X}; \widehat{\boldsymbol{\eta}}_a) \xrightarrow{p} \mathrm{E}(Y|\boldsymbol{X}, A = a)$), $\widehat{\mu}_{\mathrm{g},a}(w)$ is a consistent estimator for $\mu_a(w)$. A variance estimator for $\widehat{\mu}_{\mathrm{g},a}(w)$ is given by equation (13) in Steingrimsson and Yang [2019].

3.3.4 Doubly robust estimator

Web Appendix B.1.4 shows that the first order influence function of $\mu_a(w)$ under the non-parametric model is

$$\frac{1}{\mathrm{P}(\boldsymbol{X} \in w)} \mathrm{E}\left[I(\boldsymbol{X} \in w)\{\mathrm{E}(Y|\boldsymbol{X}, A = a) - \mu_a(w)\} + \frac{I(\boldsymbol{X} \in w, A = a)}{\mathrm{P}(A = a|\boldsymbol{X})}\{Y - \mathrm{E}(Y|\boldsymbol{X}, A = a)\} \right].$$

This influence function suggests the doubly robust (DR) estimator

$$\widehat{\mu}_{\mathrm{DR},a}(w) = \frac{1}{n(w)} \sum_{i=1}^{n} \left[I(\boldsymbol{X}_i \in w) g_a(\boldsymbol{X}_i; \widehat{\boldsymbol{\eta}}_a) + \frac{I(\boldsymbol{X}_i \in w, A_i = a)}{e_a(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}})} \{Y_i - g_a(\boldsymbol{X}_i; \widehat{\boldsymbol{\eta}}_a)\} \right]. \tag{3.7}$$

Let $H_{a,\boldsymbol{\beta},\boldsymbol{\eta}_a,\gamma_w}(e_a(\boldsymbol{X}; \boldsymbol{\beta}), g_a(\boldsymbol{X}; \boldsymbol{\eta}_a), A, Y, \gamma_w) = \frac{I(\boldsymbol{X} \in w)}{\gamma_w} \left[g_a(\boldsymbol{X}; \boldsymbol{\eta}_a) + \frac{I(A=a)}{e_a(\boldsymbol{X}; \boldsymbol{\beta})}\{Y - g_a(\boldsymbol{X}; \boldsymbol{\eta}_a)\} \right]$ be a function indexed by $(a, \boldsymbol{\beta}, \boldsymbol{\eta}_a, \gamma_w)$ and $\mathcal{F} = \{H_{a,\boldsymbol{\beta},\boldsymbol{\eta}_a,\gamma_w} : a \in \{0, 1\}, (\boldsymbol{\beta}, \boldsymbol{\eta}_a, \gamma_w) \in \Theta\}$ where Θ is

the joint parameter space. Following Van Der Vaart and Wellner [1996], for a function $f(\boldsymbol{O})$ define $\mathbb{P}_n\{f(\boldsymbol{O})\} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{O}_i)$ and $\mathbb{G}_n\{f(\boldsymbol{O})\} = \sqrt{n}\left[\mathbb{P}_n\{f(\boldsymbol{O})\} - \mathrm{E}\{f(\boldsymbol{O})\}\right]$. Using this notation, $\mathbb{P}_n\left\{H\left(e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}}), g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a), A, Y, \widehat{\gamma}_w\right)\right\} = \widehat{\mu}_{\mathrm{DR},a}(w)$.

Let $\boldsymbol{\beta}^*$ and $\boldsymbol{\eta}_a^*$ be the asymptotic limits of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\eta}}_a$ (assumed to exist). To derive the large sample properties of $\widehat{\mu}_{\mathrm{DR},a}(w)$, we make the following assumptions:

A.1 The class $\mathcal{F}$ is a Donsker class and the limiting parameters $(\boldsymbol{\beta}^*, \boldsymbol{\eta}_a^*, \mathrm{P}(\boldsymbol{X} \in w))$ are in the interior of Θ [Van Der Vaart and Wellner, 1996].

A.2 $||H(e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}}), g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a), A, Y, \widehat{\gamma}_w) - H(e_a(\boldsymbol{X};\boldsymbol{\beta}^*), g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*), A, Y, \mathrm{P}(\boldsymbol{X} \in w))||_2 \xrightarrow{p} 0$.

A.3 $\mathrm{E}\{H(e_a(\boldsymbol{X};\boldsymbol{\beta}^*), g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*), A, Y, \mathrm{P}(\boldsymbol{X} \in w))^2\} < \infty$.

A.4 At least one of: $g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a) \xrightarrow{p} \mathrm{E}(Y|\boldsymbol{X}, A=a)$ or $e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}}) \xrightarrow{p} \mathrm{P}(A=a|\boldsymbol{X})$ holds.

Theorem 3.3.3 *Under Assumptions A.1-A.4, we have that:*

1. *The doubly robust estimator is consistent, that is* $\widehat{\mu}_{DR,a}(w) \xrightarrow{p} \mu_a(w)$.

2. *The doubly robust estimator has rate of convergence*

$$||\widehat{\mu}_{DR,a}(w) - \mu_a(w)||_2 = O_P\left\{\frac{1}{\sqrt{n}} + \left|\left|e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}}) - \mathrm{P}(A=a|\boldsymbol{X})\right|\right|_2 \left|\left|g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a) - \mathrm{E}(Y|\boldsymbol{X}, A=a)\right|\right|_2\right\}. \tag{3.8}$$

We provide a proof in Web Appendix B.1.5. Theorem 3.3.3 gives useful insights into the asymptotic behaviour of the doubly robust estimator. Assumption A.4 implies that the estimator is doubly robust in that it is consistent if at least one of the models $g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a)$ or $e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}})$ are consistent, but not necessarily both. The rate of convergence result in equation (3.8) shows that if the combined rate of convergence of $g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a)$ and $e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}})$ to the true parameters $\mathrm{E}(Y|\boldsymbol{X}, A=a)$ and $\mathrm{P}(A=a|\boldsymbol{X})$ is at least $\sqrt{n}$, $\widehat{\mu}_{\mathrm{DR},a}(w)$ is $\sqrt{n}$-consistent. This implies that the doubly robust estimator has $\sqrt{n}$ rate of convergence even if more flexible models such as generalized additive models or the highly adaptive lasso [Benkeser and Van Der Laan, 2016], both of which are estimation procedures that have slower than $\sqrt{n}$ but faster than $n^{\frac{1}{4}}$ rate of convergence, are used to estimate the models for

the conditional expectation of the outcome or the conditional probability of treatment. In contrast, the inverse probability weighting and g-formula estimators require $\sqrt{n}$ rate of convergence for the estimation of the probability of treatment or the expectation of the outcome, respectively.

By Donsker preservation theorems (Van Der Vaart and Wellner [1996]), Assumption $A.1$ follows if $e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}})$, $g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a)$, $e_a(\boldsymbol{X};\boldsymbol{\beta}^*)$, and $g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*)$ are Donsker. The Donsker requirements restrict the entropy of the estimators $e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}})$ and $g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a)$ and their respective limits. Popular modeling approaches, including generalized linear models and generalized additive models, satisfy Assumption $A.1$, but many modern machine learning methods do not. The requirements can be avoided by using cross-fitting methods [Chernozhukov et al., 2018]. Assumptions A.2 and A.3 are technical conditions on the convergence and the second moment of $H(e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}}), g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a), A, Y, \widehat{\gamma}_w)$. They hold, for example, if all of $e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}})^{-1}$, $g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a)$, and $\widehat{\gamma}_w^{-1}$ converge in probability and are uniformly bounded and $\text{Var}(Y|\boldsymbol{X}, A=a)$ is finite.

3.3.5 Causal Interaction Tree (CIT) algorithms

We define the inverse probability weighting, g-formula, and doubly robust CIT algorithms – IPW-CIT, g-CIT, and DR-CIT – as the GIT algorithm implemented using $\widehat{\mu}_{\text{IPW},a}(w)$, $\widehat{\mu}_{\text{g},a}(w)$, $\widehat{\mu}_{\text{DR},a}(w)$, respectively. Collectively, we refer to these three algorithms as the Causal Interaction Tree (CIT) algorithms.

All the theory developed in Section 3.3.2-3.3.4 is for a fixed partitioning of $\mathcal{X}$, but the CIT algorithms are based on a data-dependent partitioning. Because all decision-making in our implementation of the CIT algorithms – including splitting decisions, pruning, and final tree selection – is driven by the estimators of $\mu_a(w)$, it is reasonable to expect that more robust and efficient estimators of $\mu_a(w)$ will lead to better performance. For instance, in the analysis of censored observations, more efficient and robust estimators for the statistics used for decision-making in tree-based algorithms have been shown to improve empirical performance compared to naive estimators [Steingrimsson et al., 2016, Sun et al., 2019]. In Web Appendix B.1.6, we show that when the number of terminal nodes grows at a rate of $o\left(\frac{n}{\log(n)}\right)$ the data-adaptive CIT algorithms are L_2 consistent both when the models for the conditional probability of treatment and/or the conditional expectation of the outcome are fit using the whole dataset and when the models are fit separately in each terminal node.

3.4 Simulations

3.4.1 Simulation setup

The covariate vector was simulated from a 6-dimensional mean zero multivariate normal distribution, where $\text{cov}(X^{(j)}, X^{(k)}) = 0.3$ when $j \neq k$, and $\text{Var}(X^{(j)}) = 1$. The treatment indicator A was simulated from a Bernoulli($p(\boldsymbol{X})$) distribution, with parameter
$p(\boldsymbol{X}) = \text{expit}\left(0.6X^{(1)} - 0.6X^{(2)} + 0.6X^{(3)}\right)$. The outcome was simulated using two different settings:

- $Y = 2 + 2A + 2I\left(X^{(1)} < 0\right) + \exp\left(X^{(2)}\right) + 3I\left(X^{(4)} > 0\right) + \left(X^{(5)}\right)^3 + \epsilon$, where $\epsilon \sim N(0, 1)$. For this setting, the treatment effect is the same for all covariate values and the correct tree consists only of the root node. We refer to this setting as the homogeneous setting.

- $Y = 2 + 2A + 2I\left(X^{(1)} < 0\right) + \exp\left(X^{(2)}\right) + 3AI\left(X^{(4)} > 0\right) + \left(X^{(5)}\right)^3 + \epsilon$, where $\epsilon \sim \text{N}(0, 1)$. For this setting, the treatment effect differs depending on whether $X^{(4)} > 0$ or not and the correct tree splits on $X^{(4)}$ at 0. We refer to this setting as the heterogeneous setting.

For both simulation settings, a training and a test set were generated by drawing 1000 independent samples from the joint distribution of $(\boldsymbol{X}, A, Y)$.

3.4.2 Evaluation measures

We used the following measures to evaluate the performance of different methods:

- Mean Squared Error (MSE): Denote the subgroups identified by the tree built on the training set as $w_1, \ldots, w_{\widehat{K}}$. We define the mean squared error as $\frac{1}{1000}\sum_{i=1}^{1000}\sum_{k=1}^{\widehat{K}} I(\boldsymbol{X}_i \in w_k)$ $[\{\widehat{\mu}_1(w_k) - \widehat{\mu}_0(w_k)\} - \{\text{E}(Y|A = 1, \boldsymbol{X}_i) - \text{E}(Y|A = 0, \boldsymbol{X}_i)\}]^2$, where $\boldsymbol{X}_i, i = 1, \ldots, 1000$ are the covariates from the test set.

- Correct tree: Splitting on a continuous covariate at the correct split point has a probability of zero. Therefore, a tree is defined to be correct if it splits on all the continuous variables the correct number of times independently of the splitting point and if it splits on all the categorical or ordinal variables at the correct split points.

- Number of noise variables: The average number of times the tree splits on one of the noise variables (i.e., $\{X^{(1)}, \ldots, X^{(6)}\}$ for the homogeneous treatment effect setting and $\{X^{(1)}, X^{(2)}, X^{(3)}, X^{(5)}, X^{(6)}\}$ for the heterogeneous treatment effect setting).

- Pairwise prediction similarity: Let $I_T(i,j)$ and $I_M(i,j)$ be indicators if participants i and j fall in the same terminal node running down the true tree and the fitted tree, respectively. Pairwise prediction similarity is defined as $1 - \sum_{i=1}^{1000} \sum_{j>i}^{1000} \frac{|I_T(i,j) - I_M(i,j)|}{\binom{1000}{2}}$, and it measures the ability of the tree-based algorithms to stratify observations into different groups.

- Correct first split: The first split is correct if the fully grown tree $\widehat{\psi}_{\max}$ makes the first split correctly (only applicable to the heterogeneous treatment effect setting).

3.4.3 Implementation of the Causal Interaction Tree (CIT) algorithms

We implemented the large tree $\widehat{\psi}_{\max}$ for the CIT algorithms using `rpart's` capability to build fully grown trees using user specified splitting rules and node-specific estimators.

Implementation of the CIT algorithms requires choosing which part of the data to be used to fit the models for the conditional probability of treatment and the conditional expectation of the outcome that are needed for the subgroup-specific treatment effect estimators. The results presented in this section use models that are fit using the data in the parent node being considered for splitting. Alternatives include fitting the models using the whole dataset prior to the tree building process or fitting separate models in each of the child nodes for each possible split, for which the results are presented in Web Appendix B.3.5.

To evaluate the impact of model specification for the conditional probability of treatment and the conditional expectation of the outcome, we implemented three versions of each tree-based algorithm: one that used correctly specified models, one that used misspecified functional forms of covariates, and one that had an unmeasured common cause of the outcome and treatment.

In the first version, the correctly specified logistic regression model for estimating the conditional probability of treatment in $\widehat{\mu}_{\text{IPW},a}(w)$ and $\widehat{\mu}_{\text{DR},a}(w)$ includes main effects of $X^{(1)}, X^{(2)}$ and $X^{(3)}$; the correctly specified linear regression outcome model used to implement $\widehat{\mu}_{\text{g},a}(w)$ and $\widehat{\mu}_{\text{DR},a}(w)$ includes A, $I\left(X^{(1)} < 0\right)$, $\exp\left(X^{(2)}\right)$, $I\left(X^{(4)} > 0\right)$ and $\left(X^{(5)}\right)^3$ for the homogeneous setting and a product term $A \times I\left(X^{(4)} > 0\right)$, instead of $I\left(X^{(4)} > 0\right)$, for the heterogeneous setting. In the

second version of implementation that used misspecified functional forms of covariates, the model for the conditional probability of treatment includes exponentiated forms of all covariates; the outcome model includes main effects of treatment and covariates in their original form and all two-way treatment-covariate product terms. The last version of implementation excluded $X^{(2)}$ from the data to mimic the scenario where there is an unmeasured common cause of the outcome and treatment. The model for the conditional probability of treatment includes main effects of all the remaining covariates. The outcome model includes main effects of treatment and all the remaining covariates and all the remaining two-way treatment-covariate interactions. When there are unmeasured common causes of outcome and treatment, the mean exchangeabilty condition is expected to be violated and all estimators are expected to be biased. This represents use of the methods in observational studies with residual confounding by unmeasured variables, where the target parameter is unidentifiable and the statistical model is misspecified.

The training dataset was split into an initial tree building dataset of size 800 and a validation set of the remaining 200 observations. The penalization parameter λ was chosen to be the 95th percentile of the χ^2-distribution with 1 degree of freedom (i.e., the limiting distribution of the splitting statistic (3.1) under treatment effect homogeneity).

We compared the performance of the CIT algorithms with the Causal Tree (CT) algorithm proposed by Athey and Imbens [2016]. To apply the CT algorithm to datasets where the treatment is not randomly assigned, we follow the documentation of the `causalTree` package and set the `weights` parameter to the inverse of the estimated probability of being exposed to the treatment actually received for each observation. The model for the conditional probability of treatment was implemented in the same way as for the CIT algorithms. Implementation of the CT algorithm using the `causalTree` package requires selecting several tuning parameters. We used the default settings in Athey and Imbens [2016] (referred to as the "Original CT") and a version using the combination of tuning parameters with the highest average rank in terms of minimizing MSE across the two simulation settings (referred to as the "Optimized CT"). For further details on implementation of the CT algorithm we refer to Web Appendices B.3.1 and B.3.2.

Code implementing the simulations is available from the Biometrics website and any updates will be posted to GitHub (link available in the Supporting Information section).

3.4.4 Simulation results

We used 10,000 simulations to compare the different algorithms in settings described in Section 3.4.1. Figure 3.1 shows boxplots of MSE and Table 3.1 shows the proportion of correct trees, average number of noise variables, and pairwise prediction similarity for both simulation settings, and the proportion of trees making a correct first split in the heterogeneous setting.

Table 3.1: Proportion of correct trees (higher is better), average number of noise variables used for splitting (lower is better), pairwise prediction similarity (higher is better), and proportion of trees making the first split correctly (only for heterogeneous setting, higher is better). Columns 3, 4 and 5 show simulation results when the treatment effect is homogeneous, and columns 6, 7, 8 and 9 show simulation results when the treatment effect is heterogeneous. Original CT refers to the original Causal Tree algorithm by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Section 3.4.3, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

		Homogeneous Effect			Heterogeneous Effect			
Algorithm	Model	Correct Trees	Number Noise	PPS	Correct Trees	Number Noise	PPS	Correct First Split
Original CT	Unmeasured Cov	0.00	28.34	0.05	0.00	21.49	0.55	0.63
	Mis Func	0.00	25.28	0.07	0.00	19.87	0.55	0.30
	True	0.02	25.75	0.08	0.02	19.31	0.57	0.58
Optimized CT	Unmeasured Cov	0.99	0.03	0.99	0.52	0.15	0.77	0.87
	Mis Func	0.99	0.01	1.00	0.27	0.27	0.66	0.47
	True	0.99	0.01	1.00	0.59	0.19	0.81	0.84
IPW-CIT	Unmeasured Cov	0.90	0.60	0.94	0.10	0.61	0.57	0.71
	Mis Func	0.76	1.11	0.88	0.03	1.20	0.53	0.20
	True	0.90	0.59	0.95	0.05	0.70	0.53	0.47
g-CIT	Unmeasured Cov	0.95	0.12	0.97	0.27	0.15	0.62	0.96
	Mis Func	0.94	0.12	0.97	0.30	0.12	0.63	0.96
	True	1.00	0.00	1.00	0.99	0.00	1.00	1.00
DR-CIT	Both Unmeasured Cov	0.89	0.65	0.94	0.54	0.60	0.79	0.97
	Both Mis Func	0.90	0.56	0.95	0.67	0.65	0.86	0.99
	True Treat Mis Func Out	0.91	0.58	0.95	0.71	0.58	0.87	1.00
	True Out Mis Func Treat	0.95	0.14	0.98	0.93	0.14	0.99	1.00
	Both True	0.95	0.14	0.98	0.93	0.12	0.99	1.00

The results in Figure 3.1 and Table 3.1 are consistent with the properties of the subgroup-specific treatment effect estimators described in Section 3.3. When the models for the conditional probability of treatment and/or the conditional expectation of the outcome are correctly specified,

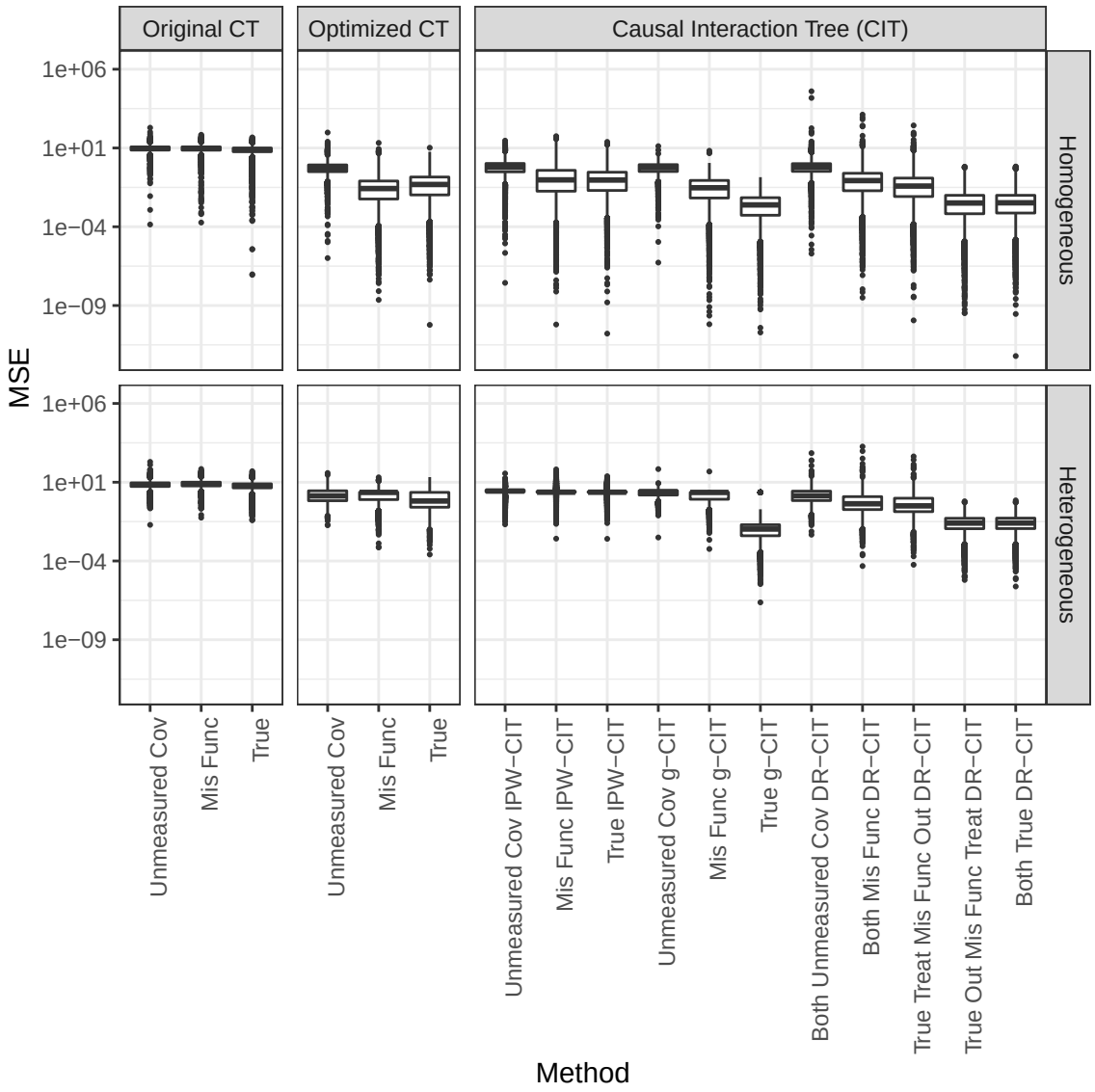

Figure 3.1: Mean squared error (MSE) for the different tree building algorithms for the homogeneous (top) and the heterogeneous (bottom) settings described in Section 3.4.1. Lower values indicate better performance. Original CT refers to the original Causal Tree algorithm by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Section 3.4.3, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

the g-CITs show the best overall performance, closely followed by the DR-CITs, and the IPW-CITs show the worst performance. All CIT algorithms show the best performance when the models required for implementation are correctly specified. When the models are misspecified, the robustness of the doubly robust estimator is also confirmed: a) the DR-CITs with a correctly specified model for the conditional expectation of the outcome but a misspecified model for the conditional probability of treatment perform substantially better than the IPW-CITs with a misspecified model for the conditional probability of treatment, and b) the DR-CITs with a correctly specified model for the conditional probability of treatment but a misspecified model for the conditional expectation of the outcome perform similarly to the g-CITs with a misspecified model for the conditional expectation of the outcome in the homogeneous setting but substantially better in the heterogeneous simulation setting. Overall, the DR-CITs have the best performance, followed by the g-CITs; the IPW-CITs show the worst overall performance among the CIT algorithms. Finally, the results in Figure 3.1 and Table 3.1 show that g-CITs and DR-CITs perform substantially better than both versions of the CT algorithm.

To further evaluate the performance of the CIT algorithms, B.3 includes additional simulation results:

- Web Appendix B.3.3 compares the running times of the CT and CIT algorithms. DR-CITs run on average 8-20 times faster than IPW-CITs and g-CITs. The reason is that, unlike the variance estimators of the other two algorithms, the variance estimator of DR-CIT does not involve calculating the inverse of the quadratic form of the design matrix, substantially reducing computation time.

- Web Appendix B.3.4 presents results when an alternative method for final tree selection proposed in Steingrimsson and Yang [2019] is used in connection with the CIT algorithms. A description of this final tree selection method is given in B.2. The results show that the CIT algorithms are more likely to overfit when the alternative final tree selection method is used.

- Web Appendix B.3.5 presents simulation results for the CIT algorithms when the models for the conditional probability of treatment and the conditional expectation of the outcome are fitted a) prior to the tree building process using the whole dataset and b) separately within each child node for each possible split. The results where the models are fitted using

the whole dataset are similar to those presented in this section. When the models are fitted separately in each child node, the CIT algorithms are more likely to overfit compared to when the models are fitted using the data in the parent node being considered for splitting.

- Web Appendix B.3.6 presents results when the models for the conditional probability of treatment and the expectation of outcome for DR-CITs are fitted using random forests. The performance is similar to that when the models are fitted by generalized linear models.

- Web Appendix B.3.7 presents simulations when the outcome is binary and the covariate vector includes both continuous and categorical variables. The results show similar trends to the simulations presented in this section.

- We used simulations to evaluate data-splitting for confidence interval construction; the results in Web Appendix B.3.7 show good coverage of the resulting confidence intervals.

- Web Appendix B.3.8 presents simulations under both a lower signal-to-noise ratio and a different correlation structure among the covariates. The results are similar to those presented in this section.

- Web Appendix B.3.9 presents simulations when the true underlying subgroups do not follow a tree structure. The results show that the CIT algorithms are able to identify the signal variables that explain treatment effect heterogeneity (in the heterogeneous setting) and produce root-only trees in the homogeneous setting. The CITs also outperform the CTs; specifically, the Original CTs overfit the data and the Optimized CTs fail to identify any signal in the heterogeneous setting.

3.5 Analysis of the SUPPORT Study

We used the CIT and the CT algorithms to analyze data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT), an observational study that evaluated the effectiveness of right heart catheterization (RHC) on critically ill patients [Connors et al., 1996]. The original analysis used matching and a follow-up analysis by Hirano and Imbens [2001] used inverse probability weighting to estimate the average treatment effect.

The SUPPORT dataset consists of 5735 participants; the treatment group (2184 participants) received RHC during the first 24 hours in the intensive care unit; the control group (3551 participants) did not receive RHC in the same time window. Participants who experienced death within 48 hours were excluded from the original study. We focus on 30-day survival as the outcome of interest; there was no censoring so we treat the outcome as binary. There are 51 covariates available for analysis; the covariates are listed in Web Appendix B.4.1. For further details on the study, we refer to Connors et al. [1996].

For the CIT algorithms, we randomly selected 80% of the dataset (4588 participants) for initial tree building and the remaining 20% of the dataset (1147 participants) as the validation set for final tree selection. Following Connors et al. [1996], we modeled the conditional probability of treatment using a logistic regression model that included all covariates as main effects. The outcome model included the main effects of treatment and all covariates and all two-way treatment-covariate interactions. Other implementation choices were the same as those described in Section 3.4.

The final tree from all the three CIT algorithms and the Optimized CT algorithm consists only of a root node. That is, none of the algorithms identified any subgroups with differential treatment effects. The root-only tree is consistent with the results in Connors et al. [1996], where none of the pre-specified subgroup analyses identified heterogeneous treatment effects. On the contrary, the Original CT algorithm produced an unreasonably large tree with 168 terminal nodes. For algorithms producing root-only trees, the data-splitting based doubly robust estimator for population average treatment effect is 0.067 with a confidence interval of $[0.027, 0.108]$, which is similar to the estimates in Hirano and Imbens [2001]. Web Appendix B.4.3 provides further details on the data-splitting procedure.

A common criticism of tree-based algorithms is the instability of the tree building process to minor modifications of the data. To evaluate the stability of the CIT algorithms, we fit the algorithms on 1000 subsamples of the SUPPORT data where each subsample randomly sampled half of the observations (2868 participants) without replacement. The results are presented in Web Appendix B.4.2 and show that the IPW-CIT, g-CIT and DR-CIT algorithms produce root-only trees for 82.8%, 99.7%, and 87.9% of the subsamples, respectively.

The large tree $\widehat{\psi}_{\max}$ (from step 1 of the GIT algorithm) built by the DR-CIT first splits on if the estimated probability of surviving 2 months at study entry is greater than or equal to 0.85.

The data-split treatment effect for those with survival probability lower than 0.85 was estimated to be 0.066 with a 95% confidence interval of $[0.022, 0.108]$; for those with survival probability greater than 0.85, the data-split treatment effect was estimated to be -0.007 with a confidence interval $[-0.091, 0.081]$. The treatment effect difference between the two subgroups was 0.076 with a confidence interval $[-0.014, 0.169]$. Web Appendix B.4.3 includes further details on the data-splitting procedure. In agreement with our results, a pre-defined subgroup analysis in Connors et al. [1996] found that the effect of RHC on death is greatest when the estimated 2-month survival probability is around 0.7 and the effect decreases when the survival probability exceeds 0.8, although the difference was not statistically significant.

Consistent with the results in Connors et al. [1996], all the CITs and the Optimized CTs identify no treatment effect heterogeneity in the SUPPORT data. To evaluate if CITs can identify subgroups in the presence of treatment effect heterogeneity, we used the SUPPORT data to simulate such a setting and the results are presented in Web Appendix B.4.4. The results show that the IPW-CIT, g-CIT and DR-CITs are able to correctly identify the variable associated with the treatment effect heterogeneity the majority of the times (95%, 78% and 94%, respectively).

3.6 Discussion

The CIT algorithms are novel approaches for subgroup identification in observational data. The algorithms are based on the GIT algorithm, which also encompasses the Interaction Tree algorithm of Su et al. [2009] and the Covariate Adjusted Interaction Tree algorithm of Steingrimsson and Yang [2019] for randomized trials, as special cases. The CIT algorithms use inverse probability weighting, g-formula, or doubly robust estimators of subgroup-specific treatment effect for decision-making during tree construction. We evaluated the finite sample properties of these algorithms in simulations and in analyses of data from an observational study evaluating the effectiveness of right heart catheterization on critically ill patients. Overall, the DR-CIT and g-CIT showed good performance and outperformed the CT algorithms.

To estimate the conditional probability of treatment and the conditional expectation of the outcome in CIT algorithms, users needs to select among using 1) the whole dataset, 2) the data in the parent node being considered for splitting and 3) the data in each of the child nodes for

each possible split. Using subsets of data for estimation is more robust to model misspecification, but leads to increased variability (due to the smaller sample size) and is more computationally intensive. The results presented in the main paper use the data in the parent node being considered for splitting for estimation, which offers a reasonable balance between model misspecification and variability.

Ensemble-based methods that average multiple trees, such as bagging or random forest, usually improve prediction accuracy over single trees. Single trees have the advantage of partitioning the covariate space into identifiable subsets with differential treatment effects, and therefore construct an easily interpretable stratification rule. In contrast, ensemble methods that average multiple trees result in black box prediction models that do not provide interpretable treatment effect stratification. Future research should consider using CITs as building blocks for ensemble methods.

Another area for future research is confidence interval construction for subgroup-specific treatment effects after subgroup identification. Here, we used data-splitting which is both inefficient (as only part of the data is used for tree building and treatment effect estimation) and potentially sensitive to how the data is split. Alternative methods that improve efficiency and stability would be of practical value.

Chapter 4

Longitudinal Identification of Subgroup Algorithm: Finding Subgroups with Heterogeneous Treatment Effects in Longitudinal Data

4.1 Introduction

Dolutegravir (DTG) is an orally administered antiretroviral drug that in combination with other antiretroviral drugs has been approved for treating patients with human immunodeficiency virus infection (HIV) [Organization et al., 2018]. DTG-containing antiretroviral therapies (ART) have many advantages such as promising treatment efficacy and low cost [Phillips et al., 2019, Group, 2019, Venter et al., 2019]. However, several randomized trials [Venter et al., 2019, 2020, Group, 2019, Calmy et al., 2020, Sax et al., 2020] and observational studies [Bourgi et al., 2020b,a, Esber et al., 2022, Surial et al., 2021] have found larger average weight gain among patients receiving DTG-containing ART regimens. Weight gains can have negative consequences including increased risk of cardiovascular and metabolic diseases as well as other comorbidities among patients living

with HIV [Herrin et al., 2016, Achhra et al., 2016, Kim et al., 2012]. Some studies have found that weight gain differs based on gender and race among HIV patients and they looked at subgroup-specific weight gain for DTG-containing ART regimens and other ART regimens [Bourgi et al., 2020a, Calmy et al., 2020, Venter et al., 2020, Sax et al., 2020]. Nevertheless, these analyses did not formally test the heterogeneous effect of the DTG-containing ART regimens on weight modified by gender or race; additionally, they focused on subgroups that were pre-specified by the investigators and did not attempt data-driven identification of subgroups with differential effect of DTG on weight gain.

Data-driven subgroup identification requires larger sample sizes than estimation of population average treatment effects. Unlike randomized controlled trials, which are usually powered to detect average treatment effects in the overall population, electronic health records (EHRs) often contain rich longitudinal data collected on a large number of patients in a "real world" setting. This makes EHRs well suited to more fine grained treatment effect analysis such as identifying subgroups with heterogeneous treatment effects. However, treatment effect estimation using electronic health records comes with several challenges including: i) participants in the EHRs have repeated measures over time, and these measures are not temporally aligned; ii) treatments are not randomized and can vary over time leading to potential time-varying confounding; iii) there is potential for informative dropout (i.e., participants that dropout of the EHRs systematically differ from those who don't dropout).

Unlike other blackbox machine learning algorithms that produce completely personalized predictions, tree-based methods stratify the covariate space into interpretable subgroups making them ideal for subgroup identification. Previous work on tree-based subgroup identification methods has mostly focused on cross-sectional randomized [Steingrimsson and Yang, 2019, Foster et al., 2011, Su et al., 2009, Seibold et al., 2016] or non-randomized data [Yang et al., 2021, Athey and Imbens, 2016]. Previous tree-based methods for treatment effect estimation with longitudinal data [Wei et al., 2020, Su et al., 2011] base splitting decisions on linear mixed effects models or generalised estimating equations and those require modeling assumptions. None can handle both non-randomized time-varying treatments and informative dropout.

In this paper, we develop the first tree-based algorithm for identifying subgroups with heterogeneous treatment effects using longitudinal data with non-randomized time-varying treatment

sequences and potentially informative dropout. The algorithm we develop, referred to as the Longitudinal Identification of Subgroup Algorithm (LISA), extends the generalized interaction tree algorithm [Yang et al., 2021] and uses longitudinal targeted maximum likelihood estimation (TMLE) [van der Laan and Gruber, 2012, Van der Laan et al., 2011, Van Der Laan and Rubin, 2006, Petersen et al., 2014, Stitelman et al., 2012, Van der Laan and Rose, 2018] for subgroup identification. More specifically, LISA recursively splits the covariate space into disjoint subgroups where splitting decisions are based on maximizing the heterogeneity of subgroup-specific TMLE treatment effect estimators. TMLE estimators of treatment effects have the advantage of being doubly robust [Robins et al., 1994, Bang and Robins, 2005], efficient [Van der Laan and Rose, 2018, Stitelman et al., 2012], and are guaranteed to take values in the support of the outcome. We apply the algorithm to identify subgroups with differential effect of DTG-containing ART regimens on weight using electronic health records on HIV patients in Western Kenya.

In Section 4.2, we introduce the data-structure and the target parameter. In Section 4.3, we show how the generalized interaction tree algorithm can be combined with the longitudinal TMLE estimator for subgroup identification. In Section 4.4, we apply the methods to EHRs on HIV patients in Western Kenya to identify subgroups where DTG has differential effects on weight.

4.2 Data structure and target parameter

For $k \in \{0, \ldots, K\}$, let $\boldsymbol{L}_k$ be a vector of time-dependent covariates (that includes measures of the outcome Y_k) measured at time k, A_k be the treatment indicator at time k, and C_k be an indicator whether a person drops out at time k, and Y_{K+1} be the outcome at time $K+1$. The baseline covariates $\boldsymbol{L}_0$ can include covariates that do not vary over time and baseline values of time-varying covariates. We assume the following time ordering of the data

$$\boldsymbol{O} = (\boldsymbol{L}_0, A_0, C_0, \boldsymbol{L}_1, A_1, C_1, \ldots, \boldsymbol{L}_K, A_K, C_K, Y_{K+1}).$$

By the time ordering assumption, each covariate has no causal effect on those measured before. For a vector $\boldsymbol{X} = (X_0, \ldots, X_K)$, define $\overline{\boldsymbol{X}}_k = (X_0, \ldots, X_k)$, $\overrightarrow{\boldsymbol{X}}_k = (X_{k+1}, \ldots, X_K)$, and $\boldsymbol{X}$ without subscript indicates a vector from time 0 to K. We assume that if an observation drops out at

time k, all quantities measured after time k are not observed. That is, $C_k = 1$ implies that $\overrightarrow{\boldsymbol{C}}_k$, $\overrightarrow{\boldsymbol{L}}_k$, $\overrightarrow{\boldsymbol{A}}_k$, and Y_{K+1} are not observed. Let $\overline{\boldsymbol{Y}}_{K+1,i}^{\boldsymbol{a},\boldsymbol{c}=\boldsymbol{0}}$ be the vector of potential outcomes [Rubin, 1974, Robins and Greenland, 2000] if participant i follows treatment sequence $\boldsymbol{a}$ and participant i is not censored during the study (i.e., artificially setting $\boldsymbol{C}_i = \boldsymbol{0}$). Similarly, let $\overline{\boldsymbol{L}}_{K,i}^{\boldsymbol{a},\boldsymbol{c}=\boldsymbol{0}}$ be the potential covariate trajectory if, possibly contrary to fact, participant i follows treatment sequence $\boldsymbol{a}$ and is not censored.

Let $\boldsymbol{L}_0$ take values in $\boldsymbol{\mathcal{L}}_0$; any set $w \subseteq \boldsymbol{\mathcal{L}}_0$ defines a subgroup consisting of observations with $\boldsymbol{L}_0 \in w$. Define $n(w) = \sum_{i=1}^{n} I(\boldsymbol{L}_{0,i} \in w)$ as the number of observations in subgroup w. Define the treatment effect for a subgroup w as $\mu_{TE}(w) = \mathrm{E}[Y_{K+1}^{\boldsymbol{1},\boldsymbol{c}=\boldsymbol{0}} | \boldsymbol{L}_0 \in w] - \mathrm{E}[Y_{K+1}^{\boldsymbol{0},\boldsymbol{c}=\boldsymbol{0}} | \boldsymbol{L}_0 \in w]$. That is, $\mu_{TE}(w)$ is the subgroup-specific average treatment effect contrasting the treatment sequences $\boldsymbol{1}$ and $\boldsymbol{0}$. Our objective is to find $\widehat{M}$ mutually exclusive subgroups of $\boldsymbol{\mathcal{L}}_0$ that maximize the between subgroup treatment effect difference. The methods we develop can be used to compare any treatment sequences. Without loss of generality, we focus on contrasts between always treated (i.e., $\boldsymbol{A} = \boldsymbol{1}$) and never treated (i.e., $\boldsymbol{A} = \boldsymbol{0}$).

4.3 Longitudinal subgroup identification algorithm

4.3.1 Generalized interaction tree algorithm

In this section we shortly review the generalized interaction tree algorithm that has been used to identify subgroups with heterogeneous treatment effects for cross-sectional randomized trial data [Steingrimsson and Yang, 2019, Su et al., 2009] and cross-sectional observational data [Yang et al., 2021].

The generalized interaction tree algorithm focuses on maximizing between subgroup specific treatment effects based on some estimators of subgroup-specific treatment effects ($\widehat{\mu}_{TE}(w)$). In Section 4.3.2 we describe three subgroup specific treatment effect estimators appropriate for the data structures described in Section 4.2 including a TMLE estimator. For a given subgroup specific treatment effect estimator $\widehat{\mu}_{TE}(w)$, define the splitting criterion for splitting a subgroup w into two

disjoint and exhaustive subgroups l and r as

$$\left(\frac{\widehat{\mu}_{TE}(l) - \widehat{\mu}_{TE}(r)}{\sqrt{\widehat{\text{Var}}\left[\widehat{\mu}_{TE}(l) - \widehat{\mu}_{TE}(r)\right]}} \right)^2 . \tag{4.1}$$

The splitting criterion (4.1) measures a standardized difference between the treatment effects in the two subgroups l and r. Selecting the subgroups l and r by maximizing the splitting criterion (4.1) results in a split that tries to identify the pair of subgroups that have the largest difference in treatment effects (i.e., maximizing between subgroup treatment effect heterogeneity).

The generalized interaction tree algorithm consists of three steps: initial tree building, pruning and final tree selection. Algorithm describes the initial tree building process where the generalized interaction tree algorithm recursively partitions the covariate space into subgroups using splitting criterion (4.1) until a pre-determined stopping criterion is met (e.g., there are too few observations in each subgroup to justify further splitting).

Algorithm Initial Tree Building Process for Longitudinal Subgroup Identification Algorithm

1: Define the root node as consisting of all observations. Set the root node as the node under consideration.
2: Define $\boldsymbol{L}_0^{(j)}$ as the j-th component of the baseline covariate vector. In the node under consideration, identify all permissible $(\boldsymbol{L}_0^{(j)}, c)$ pairs that split the covariate space into two groups defined by $l = \{\boldsymbol{L}_0^{(j)} < c\}$ and $r = \{\boldsymbol{L}_0^{(j)} \geq c\}$.
3: Run through all possible splits in Step 2 and split the node under consideration into two new subgroups based on the split that maximizes the splitting criterion (4.1).
4: If a pre-determined stopping criterion is met, stop the tree building process; otherwise, on every node that is not already split and has not met the stopping criterion, repeat Steps 2-4.

The generalized interaction tree algorithm starts by splitting the data into an initial tree building and a validation dataset. Algorithm is run on the initial tree building dataset and the result is a large tree $\widehat{\psi}_{\max}$, referred to as the initial tree. The initial tree usually substantially overfits to the data and to reduce overfitting a pruning algorithm uses the initial tree building dataset to create a set of subtrees of $\widehat{\psi}_{\max}$ and the final tree (i.e., the final model) is selected from that sequence. The pruning algorithm is described in detail in Appendix C.1. In short, the pruning algorithm is based

on the split complexity defined as

$$G^{\lambda}(\psi) = \sum_{i \in S_{\psi}} G_i(\psi) - \lambda |S_{\psi}|, \tag{4.2}$$

where for a tree ψ the set of internal nodes is S_{ψ}, $G_i(\psi)$ is the value of the splitting criterion for internal node i, $|S_{\psi}|$ is the number of internal nodes, and λ is a positive penalization parameter. The first term in equation (4.2) is a measure of the treatment effect heterogeneity of the tree and each additional split in the tree is guaranteed to increase (or at least not decrease) the first term. To offset that, the second term penalizes the size of the tree (a measure of the complexity of the model). For a fixed λ, a subtree of the initial tree is said to be an optimal subtree if it maximizes $G^{\lambda}(\psi)$. Varying λ from 0 to ∞ results in a sequence of subtrees of $\widehat{\psi}_{Max}$, $\widehat{\psi}_1, \ldots \widehat{\psi}_{\widehat{D}}$, that all are optimal for some interval of λ.

In the original classification and regression tree algorithm [Breiman et al., 1984] for outcome prediction, the final tree from the sequence of trees created by the pruning algorithm is selected by optimizing a cross-validated objective function. But as the true treatment effect is unknown for all participants (we only observe at most one of the potential outcomes $Y_{K+1,i}^{\mathbf{1},\boldsymbol{c}=\mathbf{0}}$ and $Y_{K+1,i}^{\mathbf{0},\boldsymbol{c}=\mathbf{0}}$), there is no observed treatment effect for each participant that the treatment effect predictions can be cross-validated against. Hence, cross-validation cannot be applied directly to select the final tree in the generalized interaction tree algorithm. To overcome that, the generalized interaction tree algorithm selects the final tree by estimating a validation set split complexity. For a given tree $\widehat{\psi}_d$, $d \in \{1, \ldots, \widehat{D}\}$ and a fixed λ, we calculate the validation set split complexity by recalculating the split complexity in expression (4.2) using the validation data. The final tree is selected as the tree that maximizes the validation split complexity. This relies on specifying a value of λ and one option is to use a quantile of the χ_1^2 distribution (the asymptotic distribution of the splitting statistics (4.1) when there is no treatment effect heterogeneity).

By implementing initial tree building, pruning, and final tree selection, the generalized interaction tree algorithm stratifies the covariate space $\mathcal{L}_0$ into mutually exclusive and exhaustive subgroups $\widehat{w}_1, \ldots, \widehat{w}_{\widehat{M}}$. For each of the subgroups, we calculate a subgroup specific estimator $\widehat{\mu}_{TE}(w)$ (terminal node estimator).

4.3.2 Subgroup-specific treatment effect estimation

Implementation of the generalized interaction tree algorithm relies on estimating the subgroup specific treatment effect $\mu_{TE}(w) = \mathrm{E}[Y_{K+1}^{\mathbf{1},\boldsymbol{c}=\mathbf{0}}|\boldsymbol{L}_0 \in w] - \mathrm{E}[Y_{K+1}^{\mathbf{0},\boldsymbol{c}=\mathbf{0}}|\boldsymbol{L}_0 \in w]$. In this section, we describe how TMLE can be used for to estimate $\mathrm{E}[Y_{K+1}^{\boldsymbol{a},\boldsymbol{c}=\mathbf{0}}|\boldsymbol{L}_0 \in w]$ for a treatment sequence $\boldsymbol{a}$. The TMLE estimator makes the following identifiability assumptions $\mu_{TE}(w)$ [Robins, 1986, Pearl and Robins, 1995, van der Laan and Gruber, 2012]:

- Consistency: For time $k = 0, \ldots, K$ and for every individual receiving treatment $\overline{\boldsymbol{A}}_k = \overline{\boldsymbol{a}}_k$ that has not dropped out of the study, their observed outcome at time $k+1$ (Y_{k+1}) is equal to their potential outcome $Y_{k+1}^{\overline{\boldsymbol{a}}_k,\overline{\boldsymbol{c}}_k=\mathbf{0}}$ and their observed covariate trajectory is equal to their potential covariate trajectory.

- Sequential exchengeability: At time $k, k = 0, \ldots, K$, the potential outcomes $(Y_{k+1}^{\overline{\boldsymbol{a}}_k,\overline{\boldsymbol{c}}_k}, \ldots, Y_{K+1}^{\boldsymbol{a},\boldsymbol{c}})$ are independent of treatment A_k and censoring status C_k conditional on past treatment and covariate trajectories.

- Positivity: We assume that for $k = 0$ and all $\bar{\boldsymbol{l}}_0$ such that the density $f(\overline{\boldsymbol{L}}_0 = \bar{\boldsymbol{l}}_0) > 0$, $\Pr(A_0 = a_0|\overline{\boldsymbol{L}}_0 = \bar{\boldsymbol{l}}_0) > 0$, and for $k = 1, \ldots, K$ and all $\bar{\boldsymbol{l}}_k$ such that the density $f(\overline{\boldsymbol{L}}_k = \bar{\boldsymbol{l}}_k, \overline{\boldsymbol{A}}_{k-1} = \overline{\boldsymbol{a}}_{k-1}, \overline{\boldsymbol{C}}_{k-1} = \mathbf{0}) > 0$,

$$\Pr(A_k = a_k|\overline{\boldsymbol{L}}_k = \bar{\boldsymbol{l}}_k, \overline{\boldsymbol{A}}_{k-1} = \overline{\boldsymbol{a}}_{k-1}, \overline{\boldsymbol{C}}_{k-1} = \mathbf{0}) > 0.$$

  That is, conditional on past treatment and covariate history there is a positive probability of continuing to follow the treatment sequence of interest.

Before describing the longitudinal TMLE estimator we start by describing inverse probability weighting (IPW) and g-computation estimators. Denote the probability of following the treatment sequence of interest at time $k, k = 1 \ldots, K$, conditional on observed treatment and covariate history and not having dropped out from the study as

$$g_{A,k}(\overline{\boldsymbol{L}}_k, \overline{\boldsymbol{A}}_{k-1}; \overline{\boldsymbol{C}}_{k-1} = \mathbf{0}) = \Pr(A_k = a_k|\overline{\boldsymbol{L}}_k, \overline{\boldsymbol{A}}_{k-1}, \overline{\boldsymbol{C}}_{k-1} = \mathbf{0})$$

and at time $k = 0$, $g_{A,0}(\boldsymbol{L}_0) = \Pr(A_0 = a_0|\boldsymbol{L}_0)$. For simplicity, denote $g_{A,k}(\overline{\boldsymbol{L}}_k, \overline{\boldsymbol{A}}_{k-1}; \overline{\boldsymbol{C}}_{k-1} = \mathbf{0})$

or $g_{A,0}(\boldsymbol{L}_0)$ as $g_{A,k}$. Define $g_A = \prod_{k=0}^{K} g_{A,k}$. Similarly, denote the probability of not dropping out from the study at time $k, k = 1, \ldots, K$, conditional on observed treatment and covariate history as

$$g_{C,k}(\overline{\boldsymbol{L}}_k, \overline{\boldsymbol{A}}_k; \overline{\boldsymbol{C}}_{k-1} = \mathbf{0}) = \Pr(C_k = 0 | \overline{\boldsymbol{L}}_k, \overline{\boldsymbol{A}}_k = \overline{\boldsymbol{a}}_k, \overline{\boldsymbol{C}}_{k-1} = \mathbf{0})$$

and at time $k = 0$, $g_{C,0}(\boldsymbol{L}_0, A_0) = \Pr(C_0 = 0 | \boldsymbol{L}_0, A_0)$. For simplicity, denote $g_{C,k}(\overline{\boldsymbol{L}}_k, \overline{\boldsymbol{A}}_k; \overline{\boldsymbol{C}}_{k-1} = \mathbf{0})$ or $g_{C,0}(\boldsymbol{L}_0)$ as $g_{C,k}$. Define $g_C = \prod_{k=0}^{K} g_{C,k}$. For any parameter g, let $\widehat{g}$ be an estimator for g. The IPW estimator [Horvitz and Thompson, 1952, Robins and Rotnitzky, 1992] is given by

$$\widehat{\mathrm{E}}[Y_{K+1,i}^{\boldsymbol{a},\boldsymbol{c}=\mathbf{0}} | \boldsymbol{L}_{0,i} \in w] = \frac{\sum_{\boldsymbol{L}_{0,i} \in w} \frac{I(\boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{C}_i = \mathbf{0}) Y_{K+1,i}}{\widehat{g}_{A,i} \widehat{g}_{C,i}}}{\sum_{\boldsymbol{L}_{0,i} \in w} \frac{I(\boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{C}_i = \mathbf{0})}{\widehat{g}_{A,i} \widehat{g}_{C,i}}}$$

The g-computation estimator is based on writing the potential outcome mean as a series of iterated conditional expectations of the observed outcome [Robins, 1986]

$$\mathrm{E}[Y_{K+1}^{\boldsymbol{a},\boldsymbol{c}=\mathbf{0}} | \boldsymbol{L}_0 \in w] = \mathrm{E}\left[\mathrm{E}[\ldots \mathrm{E}\left[Y_{K+1} | \overline{\boldsymbol{L}}_K, \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{C} = \mathbf{0}\right] \ldots | \boldsymbol{L}_0, A_0 = a_0, C_0 = 0\right] | \boldsymbol{L}_\mathbf{0} \in w\right]. \tag{4.3}$$

The g-computation estimator is calculated by sequential estimation of conditional expectations using the formula above, where all the estimators for the conditional expectation are restricted to using data from the subgroup w. In the first step of the g-computation estimator, regress Y_{K+1} on the observed treatment and covariate history among the participant that did not dropout to estimate the innermost expectation ($\mathrm{E}\left[Y_{K+1} | \overline{\boldsymbol{L}}_K, \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{C} = \mathbf{0}\right]$). Use the regression model fit in the first step to create predictions for each observation with $\overline{\boldsymbol{C}}_{K-1} = \mathbf{0}$ setting their treatment sequence to $\boldsymbol{A} = \boldsymbol{a}$. Denote the predictions by Q_{K+1}. In the second step, use Q_{K+1} as a pseudo-outcome and regress Q_{K+1} on $\overline{\boldsymbol{L}}_{K-1}, \overline{\boldsymbol{A}}_{K-1}$ among those with $\overline{\boldsymbol{C}}_{K-1} = \mathbf{0}$. Repeat this iterative process until the regression is only conditional on the baseline covariates and treatment at visit 0. The g-computation estimator is then calculated by averaging over the predictions from the last regression (i.e., averaging over the empirical distribution of the observations with $\boldsymbol{L}_0 \in w$ regardless of treatment status).

The longitudinal TMLE estimator updates the conditional expectation of the outcome in each iterative regression model above using the treatment sequence information, targeted to reduce bias

[van der Laan and Gruber, 2012, Kreif et al., 2017]. Specifically, when estimating the innermost expectation of expression (4.3) ($\mathrm{E}\left[Y_{K+1}|\overline{\boldsymbol{L}}_K, \boldsymbol{A}=\boldsymbol{a}, \boldsymbol{C}=\boldsymbol{0}\right]$), as for the g-computation estimator, regress Y_{K+1} on the observed treatment and covariate history among the participant that did not dropout of the study and create the predictions Q_{K+1}. In the "targeting" step we fit an intercept only model using Y_{K+1} as the outcome with Q_{K+1} as an offset and with weights $\frac{I(\boldsymbol{A}=\boldsymbol{a},\boldsymbol{C}=\boldsymbol{0})}{\widehat{g}_{A,0:K}\widehat{g}_{C,0:K}}$. We denote the resulting predictions by Q^*_{K+1} and this targeting step is performed for each iterative regression. TMLE, in contrast to the inverse probability weighting and g-computation estimators, is doubly robust [van der Laan and Gruber, 2012]. That is, it is consistent for the potential outcome mean when either i) all models for the treatment assignment and the dropout probability are correctly specified or ii) all models for the conditional expectation of the outcome are correctly specified. And it is efficient [van der Laan and Gruber, 2012] when all the models are correctly specified. The TMLE is a substitution estimator, and is therefore, under mild assumptions, guaranteed to fall within the support of Y. Although we focus on TMLE in the remainder of the manuscript, other doubly robust estimators such as those given by Bang and Robins [2005] could be used.

4.3.3 Longitudinal subgroup identification algorithm

The Longitudinal identification of subgroups algorithm (LISA) is defined by using the longitudinal TMLE estimator as the treatment effect estimator in the generalized interaction tree algorithm. LISA splits the covariate space into disjoint and exhaustive subgroups based on maximizing between subgroup treatment effect heterogeneity and within each subgroup a treatment effect estimator is calculated (terminal node estimator). In Web Appendix C.2 we proof the following theorem that shows that when the number of terminal nodes grows at a rate of $o_P\left(\frac{n}{\log(n)}\right)$, the estimator is L_2 consistent.

The subgroups are selected by searching over multiple (covariate, split-point) combinations and selecting the combination that maximizes between subgroup heterogeneity. As a result, the observed treatment effect heterogeneity is likely overestimated resulting in biased within subgroup treatment effect estimators. Data splitting can be used to get unbiased within subgroup treatment effect estimators [Fithian et al., 2014, Yang et al., 2021]. Data splitting splits the data into two disjoint and exhaustive parts. One part is used for fitting the LISA algorithm resulting in a list of subgroups and the other part of the data is used for subgroup-specific treatment effect estimation

and confidence interval construction. This ensures that the data used for identification of the subgroups (tree building) and the data used for subgroup-specific treatment effect estimation and confidence interval construction are independent, resulting in unbiased treatment effect estimators and asymptotically valid confidence intervals.

4.4 Identification of subgroups with differential weight gain when on DTG-containing ART regimens

DTG belongs to the class of integrase strand transfer inhibitor drugs and has in combination with other antiretroviral drugs shown promising results in clinical trials [Llibre et al., 2018]. A concern with DTG-containing ART regimens is the emergence of substantial weight gain as a potential side effect. A recent phase III trial conducted in South Africa [Venter et al., 2019] compared two DTG-containing ART regimens to standard of care over a 96 week follow-up period. Substantially higher weight gains were seen in the two arms receiving DTG-containing regimens compared to the standard of care arm. Of particular concern is that weight gain was associated with increased truncal fat, a known risk factor for cardiovascular outcomes [Lake, 2017].

Not all patients are expected to be at the same risk for increased weight gain associated with DTG-containing regimens. In the South African trial previously described, weight gains were more severe for females and patients with lower CD4 counts and higher viral loads. Another recent phase III trial conducted in Cameroon compared a DTG-containing regimen to a low-dose efavirenz-based regimen [Group, 2019] and reached similar conclusions (higher weight gain and increased obesity for the DTG-containing regimen and weight gain was more severe among women). Identifying subgroups of patients for whom DTG is more likely to cause substantial weight gain is important as it allows treatment recommendations or monitoring plans to be tailored to a patient's risk profile.

The Academic Model Providing Access to Healthcare (AMPATH) partnership is a consortium of research institutions focusing on prevention and treatment of HIV in Kenya [Tierney et al., 2007]. AMPATH administers the AMPATH's open-source electronic medical record system (AMRS), which is a large-scale electronic health record database that has information on clinical encounters, lab measurements, and demographic characteristics. We implemented the LISA algorithm on data from AMRS to identify subgroups with differential effect of DTG-containing ART

regimens on weight.

In the analysis we included HIV patients who are at least 18 years old, are initiating or on ART and have data on or after July 1, 2016 in AMRS. A total of $88,367$ HIV patients met these inclusion criteria. For each patient, we defined day 0 as July 1, 2016 if the patient was enrolled in AMPATH before that time and as the enrollment date if the patient was enrolled after July 1, 2016. To structure the data, we used time-windows of length 200 days and captured data at time k within days $[dk, d(k+1))$, $k \in \{0, \ldots, K+1\}$ (e.g., we used data from days 0-199 to create $\boldsymbol{L}_0, A_0, C_0$ and days 200-399 to create $\boldsymbol{L}_1, A_1, C_1$). We focused on the first five time-windows (1,000 days) as the data on patients that were always on DTG-containing regimens became sparse over time, leading to instability in the estimation. We censored participants at time k if there was no clinical visit in the time windows after time k. We excluded 1,475 women that were pregnant at time 0 and censored patients before they were first recorded as pregnant. Last, we excluded one patient who had ART start date recorded before date of birth and $2,446$ participants that had no outcome or treatment information available. Hence, the analysis set consisted of 84,445 participants who had a total of 1,178,016 unique visits.

The outcome of interest was weight in kilograms at time four (i.e., days 800-999) and treatment assignment at each time was defined as 1 if the ART regimen included DTG and 0 if not. In the analysis we included systolic blood pressure, diastolic blood pressure, weight, height, whether the patient had active tuberculosis, being treated for tuberculosis, married/living with partner, viral load, and whether covered by the National Health Insurance Fund as time-dependent covariates. The vector of baseline covariates $\boldsymbol{L}_0$ included the value of time-varying covariates at time 0, gender, age when starting ART, age at time 0, time on ART at time 0 and whether enrolled on or after July 1, 2016. For the analysis, all continuous variables were standardized. Further details on how the data were pre-processed, how we handled missing visits, missing data, and multiple clinic visits within a time-window are provided in Web Appendix C.3.

When implementing the LISA algorithm, we applied data splitting where we randomly selected 60% of the dataset (50,666 participants) for tree building using the LISA algorithm and 40% (33,779 participants) for subgroup specific treatment effect estimation and confidence interval construction using the non-parametric bootstrap with 1,000 bootstrap samples. In the tree building dataset, we randomly sampled 80% (40,533 participants) for initial tree building and the remaining 20%

(10,133 participants) as the validation set used for final tree selection.

Implementation of the LISA algorithm requires fitting models for the probability of treatment assignment, the probability of dropping out of the study, and the regression of the pseudo-outcome at each visit. For all models we used generalized linear regression models with linear main effects of all past covariate and treatment variables.

Table 4.1 shows the estimated average weight since time one had all participants been always and never on DTG-containing regimens. Table 4.1 also shows the average weight gain when always being on DTG-containing regimens compared to never being on DTG-containing regimens. The confidence intervals are estimated using the non-parametric bootstrap with 1,000 bootstrap samples. The results show that being always on a DTG-containing regimen leads to increased weight gain compared to never being on a DTG-containing regimen and the estimated weight gain increases from 1.08 at time 1 (i.e., days 200-399) to 1.41 at 800 days (i.e., days 800-999).

Time	Uncensored	Always DTG		Never DTG		
k	n	n	$\widehat{E}\left[Y_k^{\mathbf{1},\boldsymbol{c}=0}\right]$	n	$\widehat{E}\left[Y_k^{\mathbf{0},\boldsymbol{c}=0}\right]$	DTG Effect
1	71,431	5,700	63.26 (63.04, 63.48)	65,731	62.18 (62.09, 62.27)	1.08 (0.86, 1.29)
2	64,353	3,083	63.80 (63.54, 64.06)	59,550	62.60 (62.51, 62.69)	1.20 (0.95, 1.45)
3	58,646	1,350	64.10 (63.78, 64.41)	53,164	62.83 (62.73, 62.92)	1.27 (0.96, 1.58)
4	54,005	517	64.58 (64.17, 65.00)	44,473	63.17 (63.07, 63.28)	1.41 (1.00, 1.83)

Table 4.1: The table shows results from analysis of the whole analysis set. The tables shows the total number of patients with data available at each time k and the number of patients that were always or never on DTG-containing regimen up to and including that time. The column labeled $\widehat{E}\left[Y_k^{\mathbf{1},\boldsymbol{c}=0}\right]$ shows the estimated weight using targeted maximum likelihood estimation (TMLE) if always on DTG-containing regimens and the column labeled $\widehat{E}\left[Y_k^{\mathbf{0},\boldsymbol{c}=0}\right]$ shows the estimated weight if never on DTG-containing regimens. The DTG Effect column shows the estimated causal effect of always vs never being on a DTG-containing regimen on weight gain since time 0.

Figure 4.1 shows the final tree structure identified by the LISA algorithm with subgroup specific treatment effect estimators and associated 95% confidence intervals. The final tree first splits on gender and then makes another split depending on whether the females are younger or older than 43.2 years old at time 0. The results suggest that the effects of DTG-containing ART regimens are larger for females than males (2.27 vs 0.87kg) although the 95% bootstrap confidence intervals are slightly overlapping ($[0.02, 1.66]$ for males and $[1.40, 3.14]$ for females). Results from randomized trials and other data-sources support the finding that additional weight gains due to being on DTG-

containing regimens are larger among females than males [Bourgi et al., 2020a, Calmy et al., 2020, Venter et al., 2020, Sax et al., 2020]. The result adds to the current research that being female is associated with more weight gain among people living with HIV [Venter et al., 2019, Group, 2019, Calmy et al., 2020, Venter et al., 2020]. Among females, the DTG effect is 2.07 kilograms among those that are younger than 42.8 years old and 2.68 kilograms among those that are older than or equal to 42.8 years old, but the corresponding bootstrap confidence intervals are highly overlapping ($[0.86, 3.29]$ for younger females and $[1.40, 3.90]$ for older females). In Web Appendix C.3 we present additional results on the data analysis.

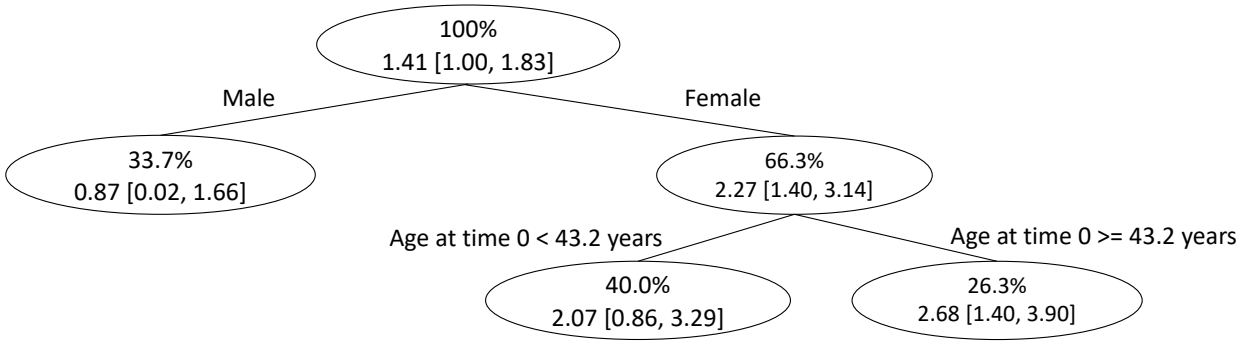

Figure 4.1: Final tree structure when the LISA algorithm is applied to data from electronic health records on HIV patients in Kenya to identify subgroups with heterogeneous effects of DTG-based regimens on weight gain. Within each subgroup, the first row is the percentage of the dataset that falls into the node and the second row is the estimator of the additional weight gain comparing being always vs never on DTG-containing regimens and associated 95% confidence interval.

4.5 Discussion

We developed the longitudinal identification of subgroups algorithm (LISA) for subgroup identification with longitudinal data. The algorithm combines the generalized interaction tree algorithm with longitudinal targeted maximum likelihood estimation. We used the algorithm to identify subgroups with differential weight gain when using DTG-based ART regimens using data from AMPATH. Our results identified gender as the primary source of heterogeneity in weight gain with females gaining more weight than males. Our analysis focused on comparing DTG-containing ART regimens to other regimens and did not distinguish among different combinations of ART regimens that contain DTG. Previous studies have evaluated different DTG-containing ART regimens [Ven-

ter et al., 2019] and the results show that the weight gains differ depending on what drugs DTG is combined with. Further investigation into specific DTG-containing regimens will be helpful to gain a more fine-grained understanding into effect of DTG on weight gain.

The LISA algorithm defines subgroups based on baseline covariates. Extending the algorithm to dynamically update the subgroups using time-varying data when new information becomes available is of interest [Sun et al., 2020]. Furthermore, the LISA algorithm results in a single tree structure that creates an interpretable treatment effect stratification rule by partitioning the covariate space into identifiable subgroups with differential treatment effects. Ensemble methods, such as random forest or bagging, average multiple single trees with the aim of improving prediction accuracy. But the resulting black-box model no longer has the structure of a single tree and unlike the single tree it does not identify interpretable subgroups. It would be of interest to develop ensemble based methods for prediction with EHR data using the LISA trees as the building block.

Appendix A

Appendix for Chapter 2

A.1 Derivations for the variance of the shrunken estimates

Write the variance as

$$\begin{aligned}
\text{Var}(\hat{\delta}_{i^*j^*} - \delta_{i^*j^*}) =& \text{Var}\left[(\hat{\delta} + \hat{\gamma}_{\delta i^*j^*}) - (\delta + \gamma_{\delta i^*j^*})\right] \\
=& \text{Var}(\hat{\delta}) + \text{Var}(\hat{\gamma}_{\delta i^*j^*} - \gamma_{\delta i^*j^*}) + 2\text{cov}(\hat{\delta}, \hat{\gamma}_{\delta i^*j^*} - \gamma_{\delta i^*j^*}) \\
=& \boldsymbol{C_\theta}\text{Var}(\hat{\boldsymbol{\theta}})\boldsymbol{C_\theta}^T + \boldsymbol{C_b}\text{Var}(\hat{\boldsymbol{b}}_{i^*j^*} - \boldsymbol{b}_{i^*j^*})\boldsymbol{C_b}^T + 2\boldsymbol{C_\theta}\text{cov}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{b}}_{i^*j^*})\boldsymbol{C_b}^T \\
& - 2\boldsymbol{C_\theta}\text{cov}(\hat{\boldsymbol{\theta}}, \boldsymbol{b}_{i^*j^*})\boldsymbol{C_b}^T
\end{aligned} \tag{A.1}$$

where $\boldsymbol{C_\theta}$ are summarized in Table 2.1; $\boldsymbol{C_b}$ is the contrast matrix that pulls off the appropriate element from the random effects vector or the variance matrix for the random effects and equals 1 for the model with fixed intercepts and random slopes and equals $(0, 1)$ for the model with random intercepts and random slopes.

Using $\hat{\boldsymbol{\theta}} = \left[\sum_{i=1}^{I}\sum_{j=1}^{J_i}\mathbf{X}_{ij}^T\boldsymbol{\Sigma}_{ij}^{-1}\mathbf{X}_{ij}\right]^{-1}\left[\sum_{i=1}^{I}\sum_{j=1}^{J_i}\mathbf{X}_{ij}^T\boldsymbol{\Sigma}_{ij}^{-1}\boldsymbol{Y}_{ij}\right]$ and

$\hat{\boldsymbol{b}}_{i^*j^*} = \mathbf{D}\mathbf{Z}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1}(\boldsymbol{Y}_{i^*j^*} - \mathbf{X}_{i^*j^*}\hat{\boldsymbol{\theta}})$ [Laird and Ware, 1982],

$$
\begin{aligned}
\operatorname{cov}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{b}}_{i^*j^*}) =& \operatorname{cov}\left\{\hat{\boldsymbol{\theta}}, \mathbf{D}\mathbf{Z}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \boldsymbol{Y}_{i^*j^*}\right\} - \operatorname{cov}\left\{\hat{\boldsymbol{\theta}}, \mathbf{D}\mathbf{Z}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \mathbf{X}_{i^*j^*}\hat{\boldsymbol{\theta}}\right\} \\
=& \operatorname{cov}\left\{ \left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} \mathbf{X}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{X}_{ij}\right]^{-1} \mathbf{X}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \boldsymbol{Y}_{i^*j^*}, \mathbf{D}\mathbf{Z}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \boldsymbol{Y}_{i^*j^*} \right\} \\
& - \operatorname{Var}(\hat{\boldsymbol{\theta}}) \mathbf{X}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \mathbf{Z}_{i^*j^*} \mathbf{D} \\
& \qquad \text{(participants are independent of each other)} \\
=& 0
\end{aligned} \tag{A.2}
$$

$$
\begin{aligned}
\operatorname{cov}(\hat{\boldsymbol{\theta}}, \boldsymbol{b}_{i^*j^*}) =& \operatorname{cov}\left\{ \left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} \mathbf{X}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{X}_{ij}\right]^{-1} \mathbf{X}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \boldsymbol{Y}_{i^*j^*}, \boldsymbol{b}_{i^*j^*} \right\} \\
=& \operatorname{cov}\left\{ \left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} \mathbf{X}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{X}_{ij}\right]^{-1} \mathbf{X}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \mathbf{Z}_{i^*j^*} \boldsymbol{b}_{i^*j^*}, \boldsymbol{b}_{i^*j^*} \right\} \\
& \qquad (\boldsymbol{b}_{i^*j^*} \perp\!\!\!\perp \boldsymbol{\epsilon}_{i^*j^*}) \\
=& \left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} \mathbf{X}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{X}_{ij}\right]^{-1} \mathbf{X}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \mathbf{Z}_{i^*j^*} \mathbf{D}.
\end{aligned} \tag{A.3}
$$

Plug $\operatorname{Var}(\hat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^{I}\sum_{j=1}^{J_i} \mathbf{X}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{X}_{ij}\right)^{-1}$, $\operatorname{Var}(\hat{\boldsymbol{b}}_{i^*j^*} - \boldsymbol{b}_{i^*j^*}) = \mathbf{D} - \mathbf{D}\mathbf{Z}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \mathbf{Z}_{i^*j^*} \mathbf{D}$ $+\mathbf{D}\mathbf{Z}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \mathbf{X}_{i^*j^*} \left(\sum_{i=1}^{I}\sum_{j=1}^{J_i} \mathbf{X}_{ij}^T \boldsymbol{\Sigma}_{ij}^{-1} \mathbf{X}_{ij}\right)^{-1} \mathbf{X}_{i^*j^*}^T \boldsymbol{\Sigma}_{i^*j^*}^{-1} \mathbf{Z}_{i^*j^*} \mathbf{D}$ [Laird and Ware, 1982], (A.2) and (A.3) into (A.1) gives $\operatorname{Var}(\hat{\delta}_{i^*j^*} - \delta_{i^*j^*})$.

A.2 Additional comparisons of designs

We include additional illustrations in this section. Unless specified, the parameter choices were the same as those described in Section 2.7.

A.2.1 Results for models with fixed and random intercepts

Figure A.1 shows how the average required number of measurements across a series of n-of-1 trials ($IJKL$) changes as a function of the number of measurements per participant (KL, left) or the

number of participants (IJ, right) for optimized designs when the four possible models in Table 2.1 are used for estimating the population average treatment effect.

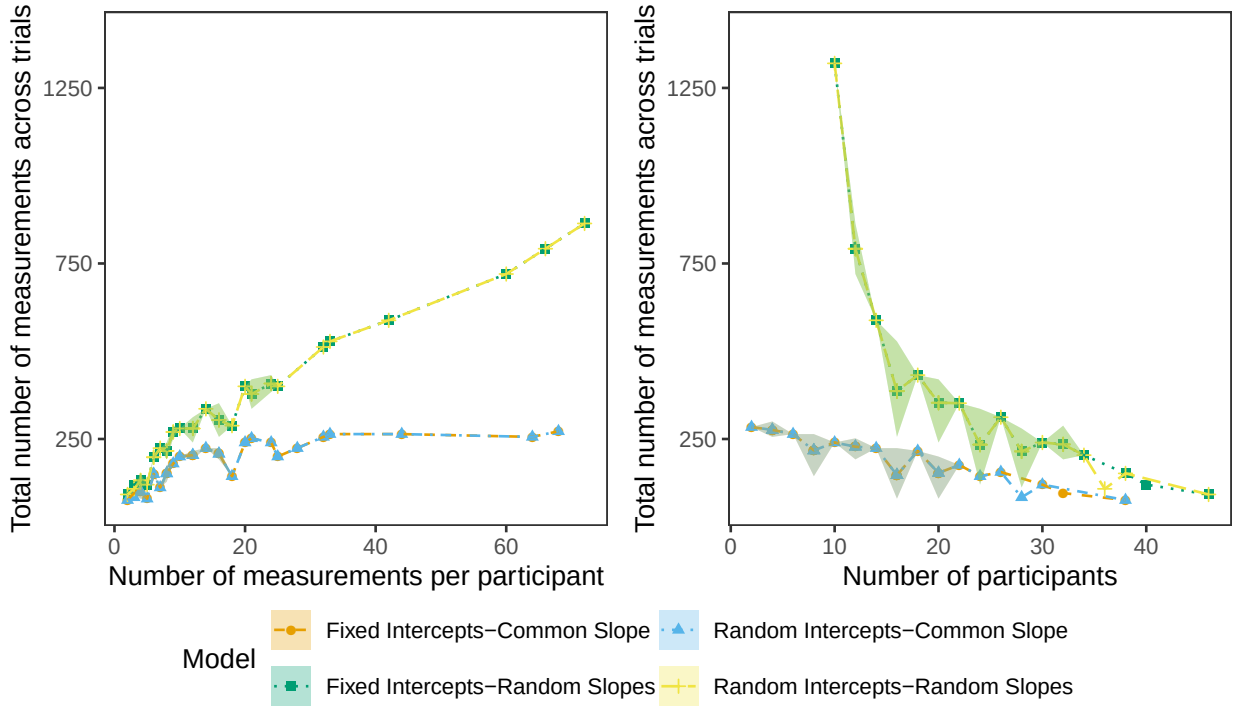

Figure A.1: Average required number of measurements across a series of n-of-1 trials versus number of measurements per participant (KL, left) and number of participants (IJ, right) for optimized designs when the four possible models in Table 2.1 are used for estimating the population average treatment effect.

We find in Figure A.1 that the optimized designs for estimating the population average treatment effect are similar if we choose to model the slopes in the same way regardless of how we choose to model the intercepts.

Therefore in Section 2.7 and the following sections, we will present results for models with fixed intercepts and by the form of slopes.

A.2.2 Results with alternative type I and II errors

Figure A.2 and A.3 show how the average required number of measurements across a series of n-of-1 trials ($IJKL$) changes as a function of type I and II errors, respectively, for optimized designs when we use a common slope (left) and the average of random slopes (right) to estimate the population average treatment effect fixing the number of measurements per participant (KL, top) and the number of participants (IJ, bottom). Other parameter choices were the same as those described

in Section 2.7.

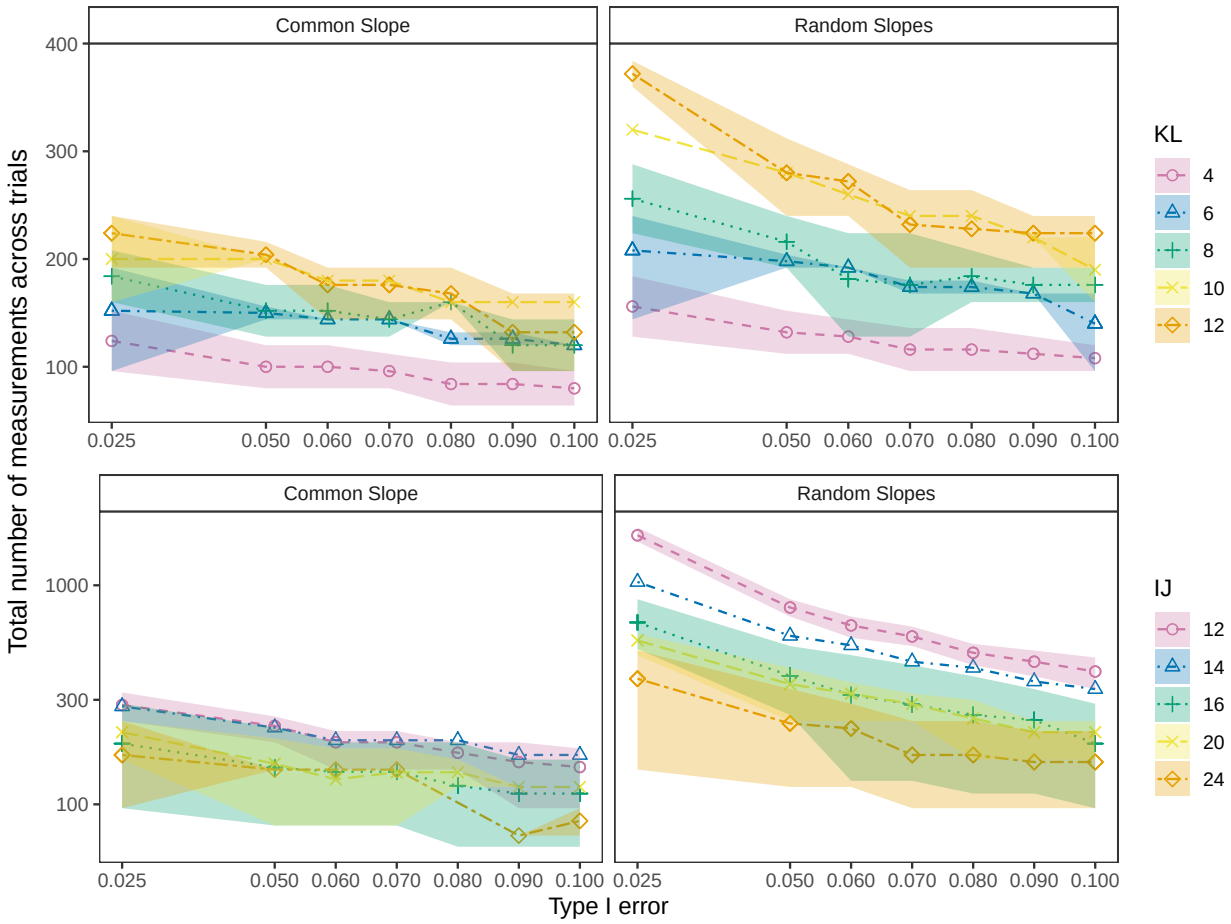

Figure A.2: Average required number of measurements across a series of n-of-1 trials versus type I error rates for optimized designs when we use a common slope (left) and the average of random slopes (right) to estimate the population average treatment effect fixing the number of measurements per participant (KL, top) and the number of participants (IJ, bottom).

As expected, the results show that the average required number of measurements across trials increases with smaller type I and II errors. The increasing rate is higher 1) if we want to reduce smaller type I and II errors, 2) if the number of measurements per participant is larger and 3) if the number of participants in trials is smaller.

A.2.3 Results with alternative parameterizations for residual error variance matrix

Figure A.4 shows how the average required number of measurements across a series of n-of-1 trials ($IJKL$) changes as a function of the number of measurements per participant (KL, top) and the

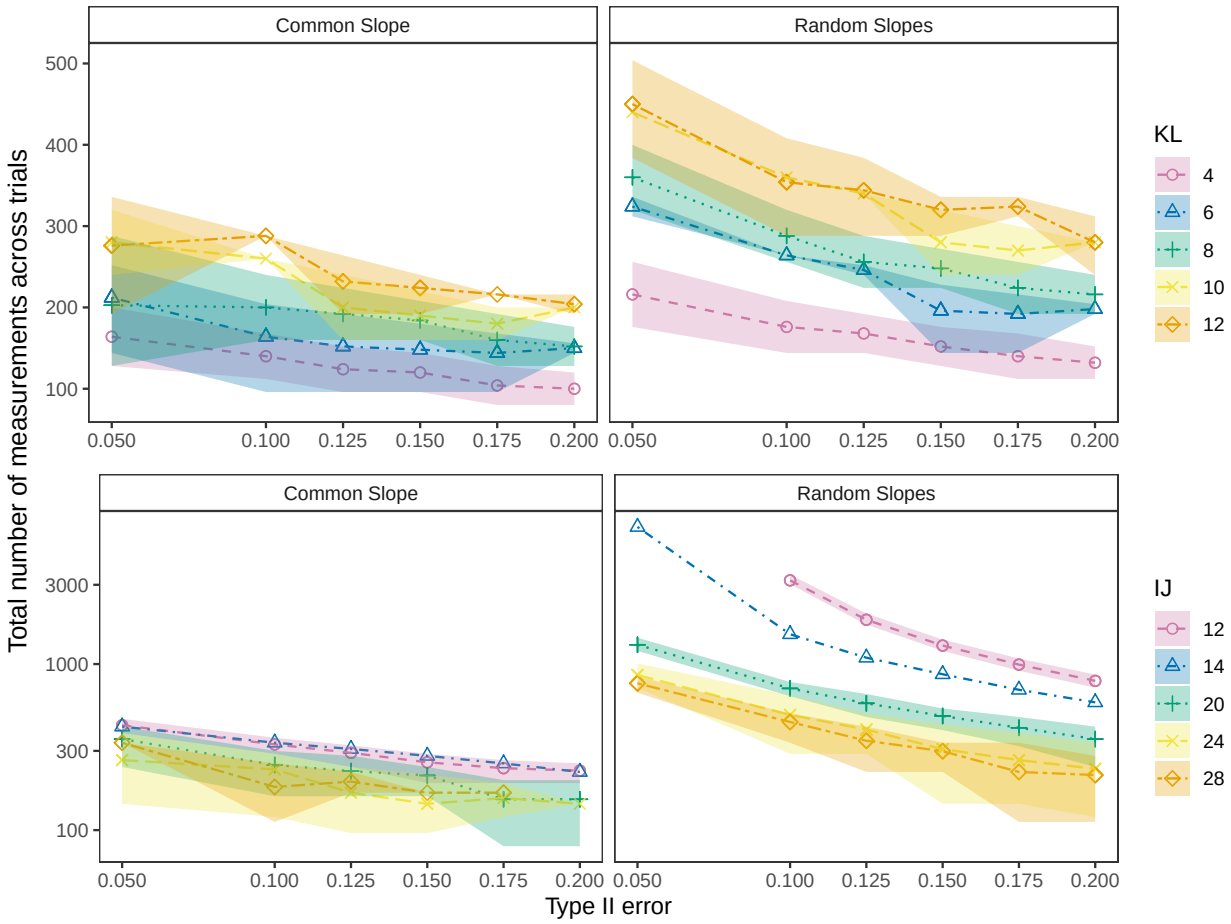

Figure A.3: Average required number of measurements across a series of n-of-1 trials versus type II error rates for optimized designs when we use a common slope (left) and the average of random slopes (right) to estimate the population average treatment effect fixing the number of measurements per participant (KL, top) and the number of participants (IJ, bottom).

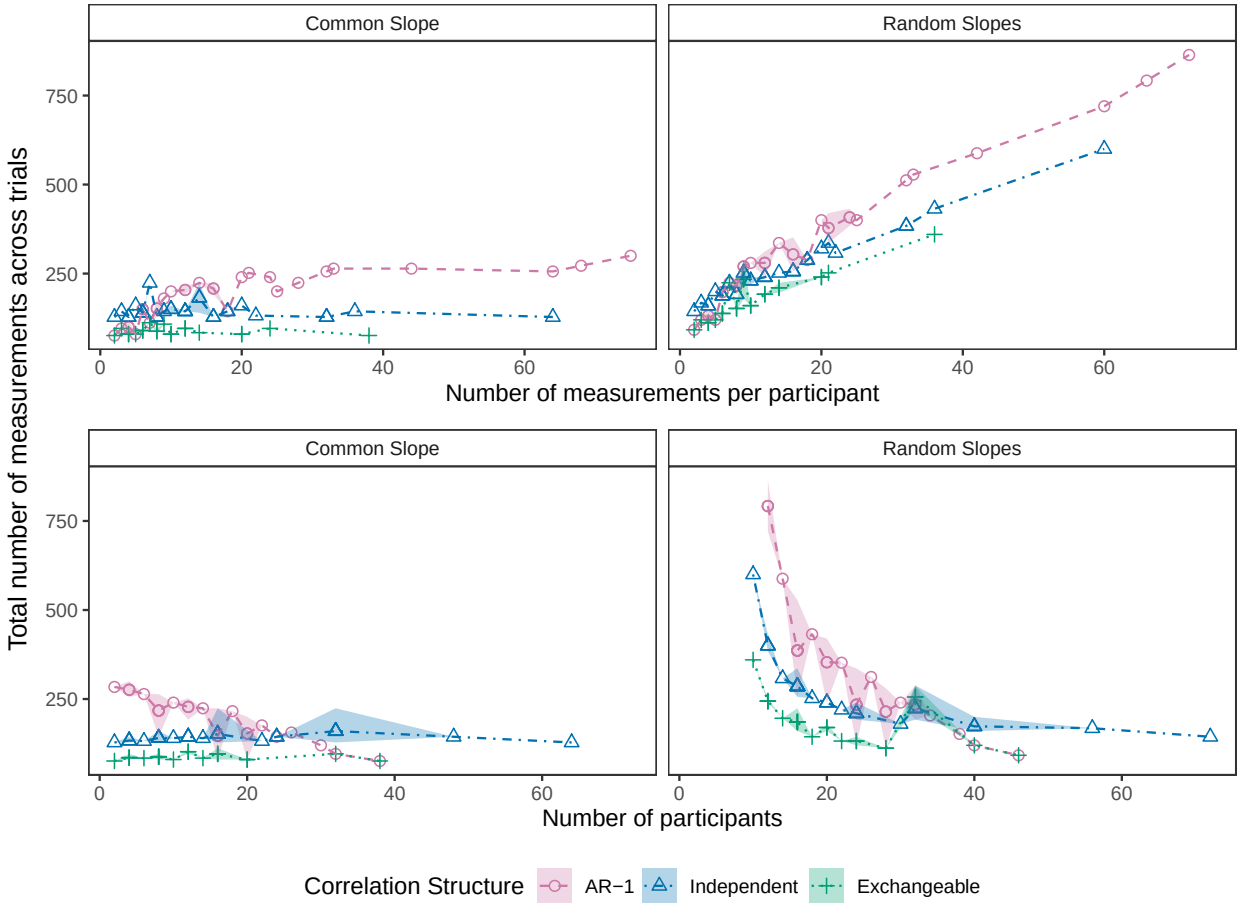

Figure A.4: Average required number of measurements across a series of n-of-1 trials versus number of measurements per participant (KL, top) and the number of participants (IJ, bottom) for optimized designs assuming independent, exchangeable and first order autoregressive correlation structures among the repeated measurements on a single participant when we use a common slope (left) and the average of random slopes (right) to estimate the population average treatment effect.

number of participants (IJ, bottom) for optimized designs assuming independent, exchangeable and first order autoregressive correlation structures among the repeated measurements on a single participant when we use a common slope (left) and the average of random slopes (right) to estimate the population average treatment effect. Figure A.5 and A.6 show how the average required number of measurements across trials ($IJKL$) changes as a function of the residual error variance and the residual correlation coefficient, respectively, fixing the number of measurements per participant (KL, top) and the number of participants (IJ, bottom) for optimized designs. Other parameter choices were the same as those described in Section 2.7.

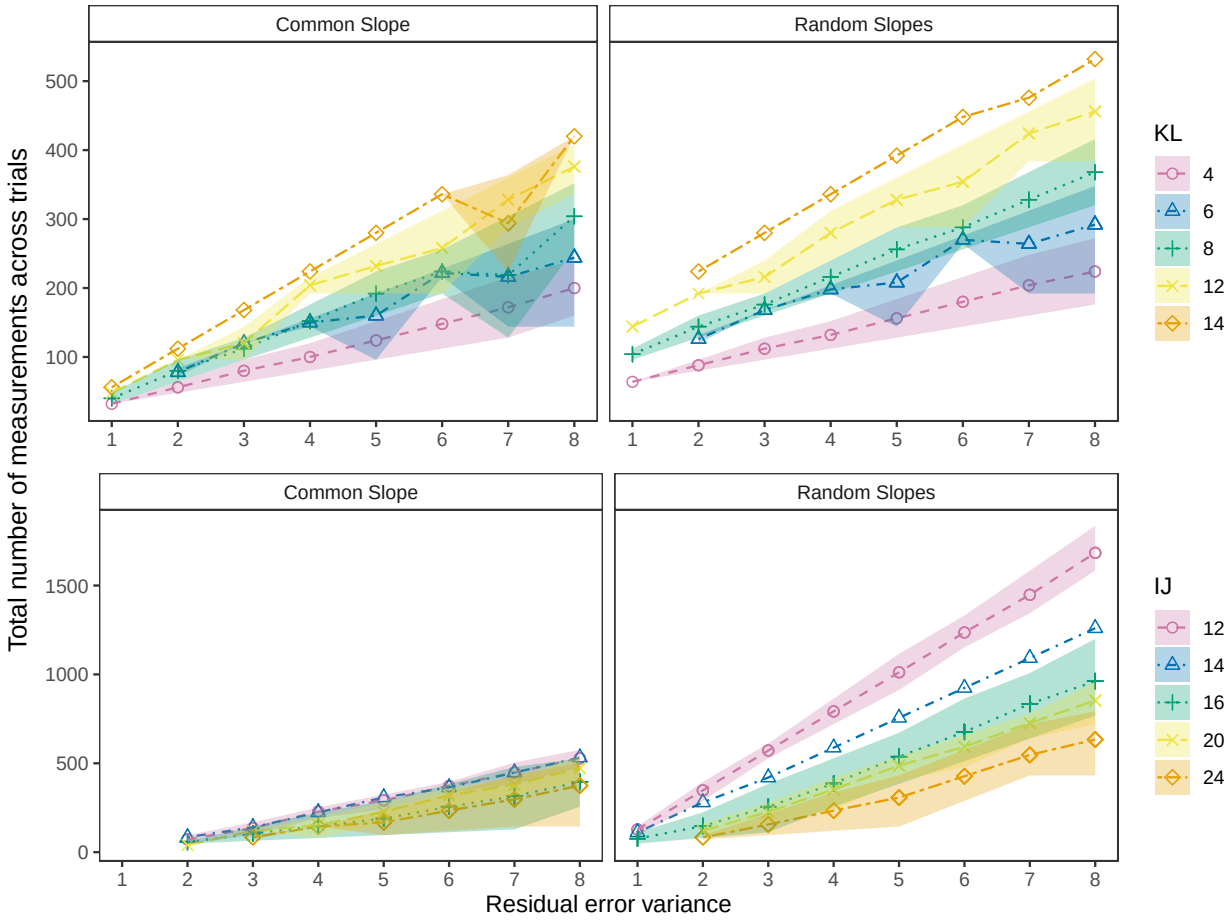

Figure A.5: Average required number of measurements across a series of n-of-1 trials versus residual error variance for optimized designs when we use a common slope (left) and the average of random slopes (right) to estimate the population average treatment effect fixing the number of measurements per participant (KL, top) and the number of participants (IJ, bottom).

Trials with exchangeable correlation structure require the fewest measurements across trials, followed by independent correlation structure. Under first order autoregressive correlation struc-

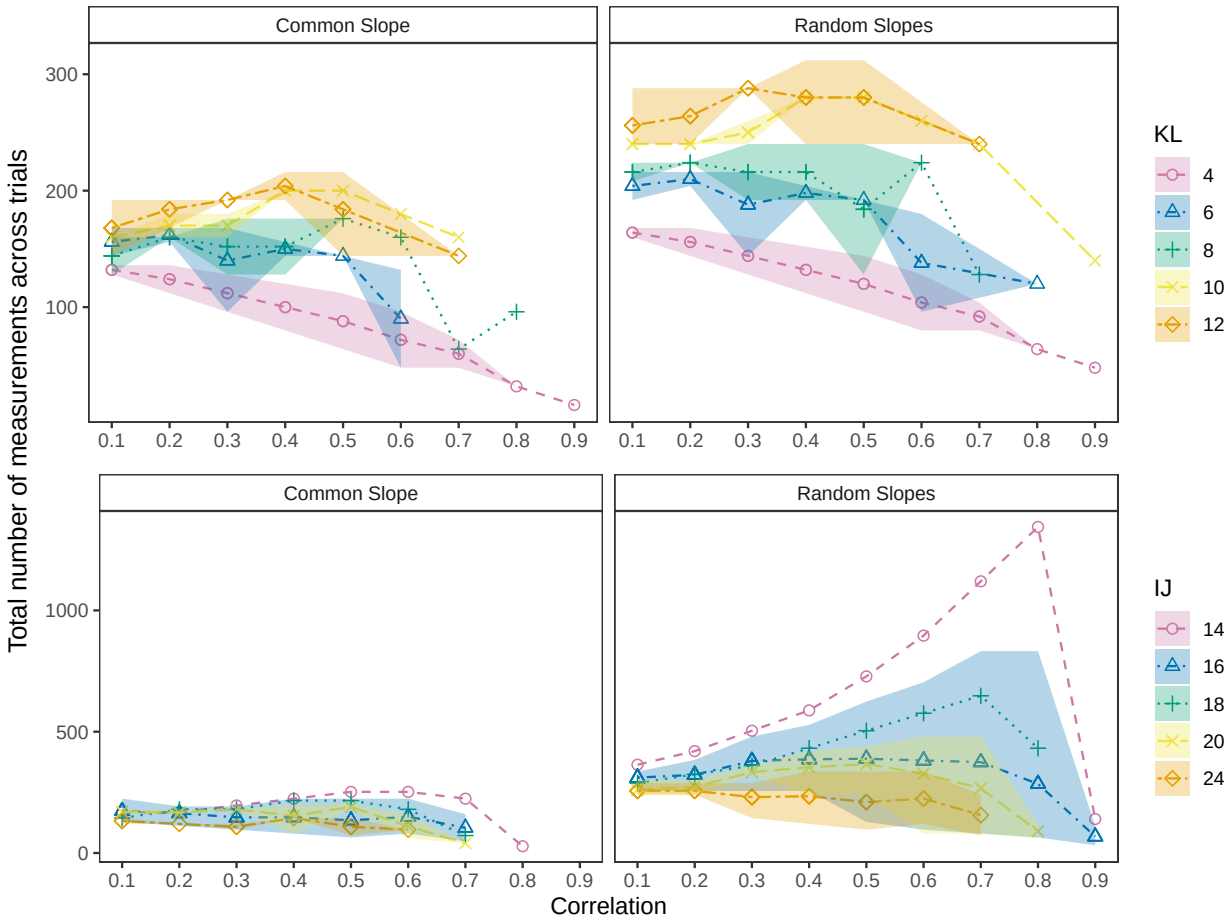

Figure A.6: Average required number of measurements across a series of n-of-1 trials versus residual correlation coefficient for optimized designs when we use a common slope (left) and average of random slopes (right) to estimate the population average treatment effect fixing the number of measurements per participant (KL, top) and the number of participants (IJ, bottom).

ture, the required number of measurements across trials 1) is similar to that under exchangeable correlation structure when the number of measurements per participant is small and the correlation coefficient is large and 2) becomes larger than that under independent correlation structure when the number of measurements per participant is large and the correlation coefficient is small, and 3) increases linearly with homogeneous variance and the increasing rate is larger when the number of measurements per participant is larger and when the number of participants in trials is smaller.

A.2.4 Results with alternative parameterizations for random-effect variance matrix

Figure A.7, A.8 and A.9 show how the average required number of measurements across a series of n-of-1 trials changes as a function of the variance of random intercepts, correlation between random intercepts and slopes and the variance of random slopes, respectively, when applicable in the model for optimized designs.

The results show that variance of random intercepts and correlation between random intercepts and slopes do not affect the optimized designs that satisfy the power requirement. Holding the number of measurements per participant the same, the average required number of measurements across trials increases linearly with the variance of random slopes; the increasing rate is larger when the number of measurements per participant is larger. Holding the number of participants the same, the average required number of measurements across trials increases more at larger values of variance of random slopes.

A.2.5 Results with alternative minimal clinically important treatment effects

Figure A.10 shows how the average required number of measurements across a series of n-of-1 trials ($IJKL$) changes as a function of minimal clinically important treatment effect when we use a common slope (left) and the average of random slopes (right) to estimate the population average treatment effect fixing the number of measurements per participant (KL, top) and the number of participants (IJ, bottom) for optimized designs.

The results show that the required number of measurements across trials increases with smaller minimal clinically important treatment effect. The increasing rate is higher 1) if we want to reduce

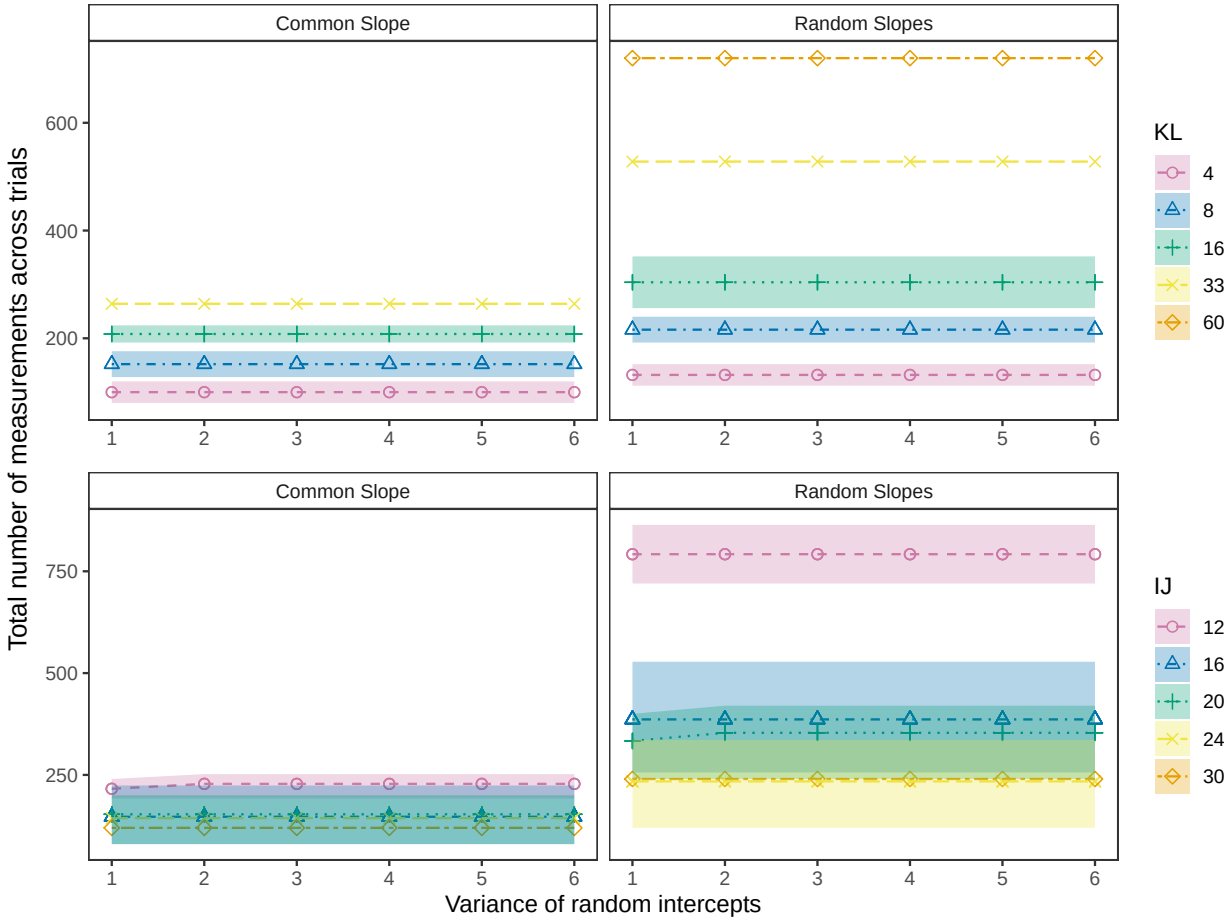

Figure A.7: Average required number of measurements across a series of n-of-1 trials versus variance of random intercepts when we use random intercepts-common slope model (left) and random intercepts-random slopes model (right) to estimate the population average treatment effect fixing the number of measurements per participant (KL, top) and the number of participants (IJ, bottom) for optimized designs.

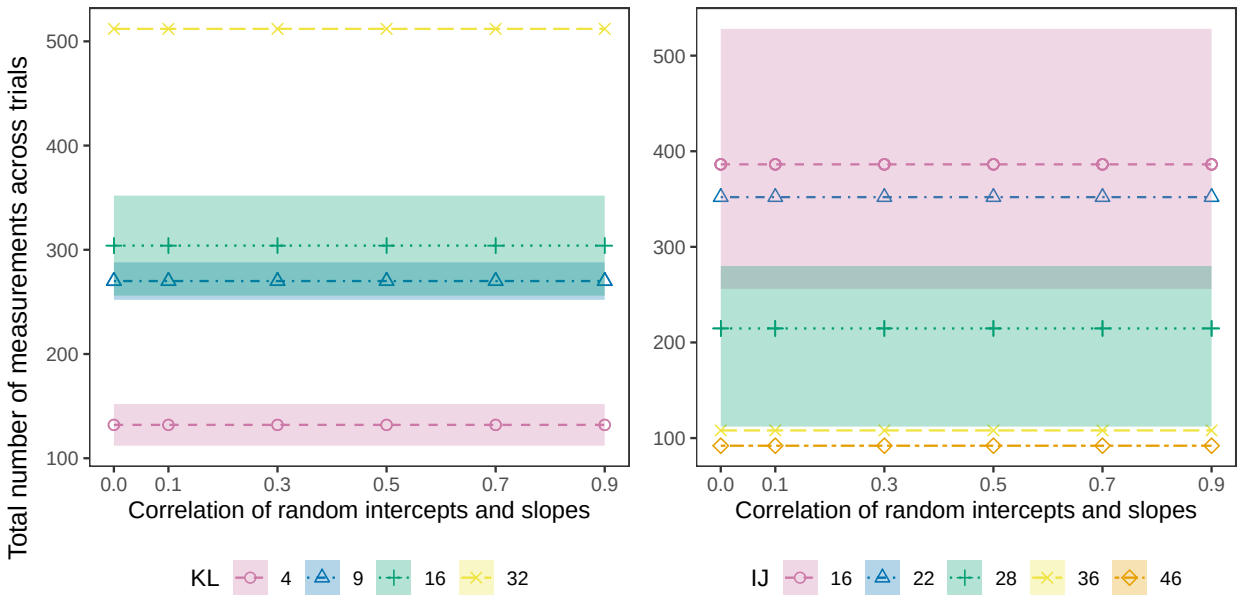

Figure A.8: Average required number of measurements across a series of n-of-1 trials versus correlation between random intercepts and random slopes when we use random intercepts-random slopes model to estimate the population average treatment effect fixing the number of measurements per participant (KL, left) and the number of participants (IJ, right) for optimized designs.

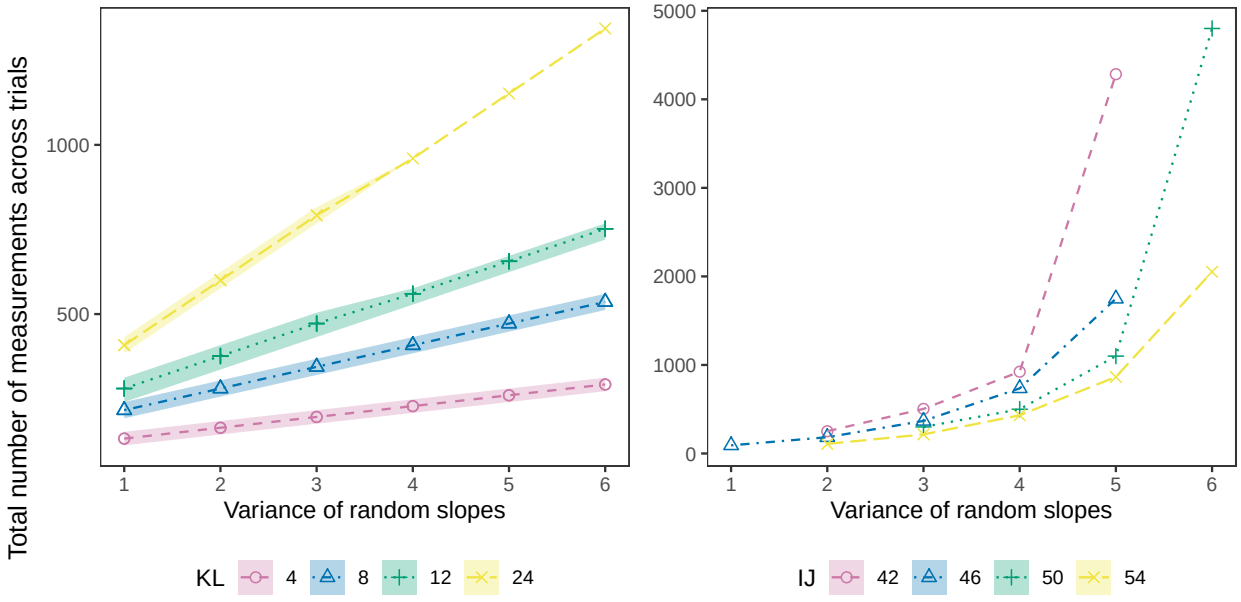

Figure A.9: Average required number of measurements across a series of n-of-1 trials versus variance of random slopes when we use the average of random slopes to estimate the population average treatment effect fixing the number of measurements per participant (KL, left) and the number of participants (IJ, right) for optimized designs.

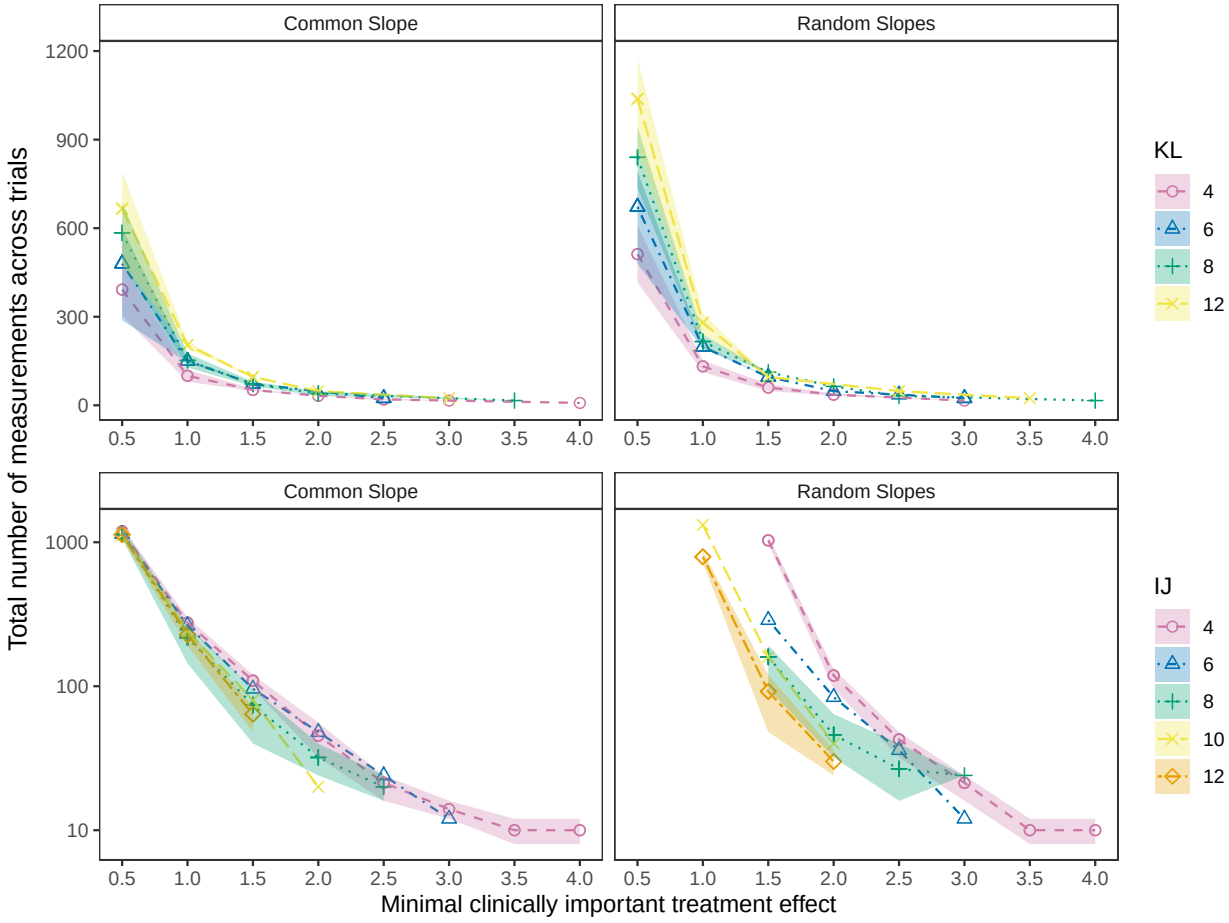

Figure A.10: Average required number of measurements across a series of n-of-1 trials versus minimal clinically important treatment effect when we use a common slope (left) and the average of random slopes (right) to estimate the population average treatment effect fixing the number of measurements per participant (KL, top) and the number of participants (IJ, bottom) for optimized designs.

a smaller minimal clinically important treatment effect, 2) if the number of measurements per participant is larger and 3) if the number of participants in trials is smaller.

Appendix B

Appendix for Chapter 3

B.1 Properties of the Subgroup-Specific Treatment Effect Estimators

B.1.1 Proof of identifiability of subgroup-specific treatment effect estimators

Proof of Theorem 3.3.1: We start by showing the identifiability result (3.3)

$$\begin{aligned}
\mathrm{E}(Y^a|\boldsymbol{X} \in w) &= \mathrm{E}\left\{\mathrm{E}\left(Y^a|\boldsymbol{X}\right)\middle|\boldsymbol{X} \in w\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left(Y^a|\boldsymbol{X}, A=a\right)\middle|\boldsymbol{X} \in w\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left(Y|\boldsymbol{X}, A=a\right)\middle|\boldsymbol{X} \in w\right\}.
\end{aligned}$$

For result (3.4), we have

$$\begin{aligned}
\mathrm{E}(Y^a|\boldsymbol{X} \in w) &= \mathrm{E}\left\{\mathrm{E}\left(Y|\boldsymbol{X}, A=a\right)\middle|\boldsymbol{X} \in w\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left[\frac{I(A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}Y|\boldsymbol{X}\right]\middle|\boldsymbol{X} \in w\right\} \\
&= \mathrm{E}\left[\frac{I(\boldsymbol{X} \in w)}{\mathrm{P}(\boldsymbol{X} \in w)}\mathrm{E}\left\{\frac{I(A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}Y|\boldsymbol{X}\right\}\right] \\
&= \frac{1}{\mathrm{P}(\boldsymbol{X} \in w)}\mathrm{E}\left[\mathrm{E}\left\{\frac{I(\boldsymbol{X} \in w, A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}Y|\boldsymbol{X}\right\}\right] \\
&= \frac{1}{\mathrm{P}(\boldsymbol{X} \in w)}\mathrm{E}\left\{\frac{I(\boldsymbol{X} \in w, A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}Y\right\}.
\end{aligned}$$

□

B.1.2 Proof of Theorem 3.3.2

Proof of Theorem 3.3.2: We derive the asymptotic variance when the conditional probability of treatment is estimated using a correctly specified logistic regression model using the data falling in the union of the subgroups $p = (l \cup r)$. For simplicity of notation, we denote $e = e_1(\boldsymbol{X};\boldsymbol{\beta}) = \{1 + \exp(-\boldsymbol{X}\boldsymbol{\beta})\}^{-1}$ and $1 - e = e_0(\boldsymbol{X};\boldsymbol{\beta})$, $\boldsymbol{e_\beta} = \partial e/\partial\boldsymbol{\beta}$, $\gamma_l = \mathrm{P}(\boldsymbol{X} \in l|\boldsymbol{X} \in p)$. Define $\boldsymbol{\theta} = (T(l,r), \boldsymbol{\beta}^T, \gamma_l)^T$ as the vector of unknown parameters with the true value denoted as $\boldsymbol{\theta}_0$. The joint set of estimating equations used to estimate $\boldsymbol{\theta}_0$ using the data falling in the union of the subgroups p is given by

$$\begin{cases} \frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\psi_1(Y_i, A_i, \boldsymbol{X}_i, T(l,r), \boldsymbol{\beta}, \gamma_l) = \frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\left[\frac{I(\boldsymbol{X}_i\in l)}{\gamma_l}\left\{\frac{A_iY_i}{e_i} - \frac{(1-A_i)Y_i}{1-e_i}\right\}\right. \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. -\frac{I(\boldsymbol{X}_i\in r)}{1-\gamma_l}\left\{\frac{A_iY_i}{e_i} - \frac{(1-A_i)Y_i}{1-e_i}\right\} - T(l,r)\right] = 0 \\ \frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\boldsymbol{\psi}_2(Y_i, A_i, \boldsymbol{X}_i, T(l,r), \boldsymbol{\beta}, \gamma_l) = \frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\frac{(A_i-e_i)\boldsymbol{e_{\beta i}}}{e_i(1-e_i)} = 0 \\ \frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\psi_3(Y_i, A_i, \boldsymbol{X}_i, T(l,r), \boldsymbol{\beta}, \gamma_l) = \frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\{I(\boldsymbol{X}_i \in l) - \gamma_l\} = 0, \end{cases}$$

Define $\widehat{\boldsymbol{\theta}}_0$ as the solution to the three estimating equations listed above and let $\boldsymbol{O} = (Y, A, \boldsymbol{X})$ and $\boldsymbol{\psi} = (\psi_1, \boldsymbol{\psi}_2, \psi_3)$. Here, $\boldsymbol{\psi}_2$ is the estimating equation corresponding to the maximum likelihood estimation from the logistic regression model. Straightforward calculations give

$$\mathbf{A}(\boldsymbol{\theta}_0) = \mathrm{E}\left\{-\frac{\partial}{\partial\theta}\boldsymbol{\psi}(\boldsymbol{O}, \boldsymbol{\theta}_0)\right\} = \begin{pmatrix} 1 & \boldsymbol{H}^T_{\boldsymbol{\beta},l} - \boldsymbol{H}^T_{\boldsymbol{\beta},r} & \frac{1}{\gamma_l}\{\mu_1(l) - \mu_0(l)\} + \frac{1}{\gamma_r}\{\mu_1(r) - \mu_0(r)\} \\ \mathbf{0} & \mathbf{E}_{\boldsymbol{\beta\beta}} & \mathbf{0} \\ 0 & \mathbf{0} & 1 \end{pmatrix}$$

$$\Rightarrow [\mathbf{A}(\boldsymbol{\theta}_0)]^{-1} = \begin{pmatrix} 1 & -(\boldsymbol{H}_{\boldsymbol{\beta},l} - \boldsymbol{H}_{\boldsymbol{\beta},r})^T\mathbf{E}^{-1}_{\boldsymbol{\beta\beta}} & -\frac{1}{\gamma_l}\{\mu_1(l) - \mu_0(l)\} - \frac{1}{\gamma_r}\{\mu_1(r) - \mu_0(r)\} \\ \mathbf{0} & \mathbf{E}^{-1}_{\boldsymbol{\beta\beta}} & \mathbf{0} \\ 0 & \mathbf{0} & 1 \end{pmatrix},$$

where as in main manuscript $\boldsymbol{H}_{\boldsymbol{\beta},w} = \mathrm{E}\left[\left\{\frac{AY}{e} + \frac{(1-A)Y}{1-e}\right\}\boldsymbol{e}_{\boldsymbol{\beta}}\middle|\boldsymbol{X} \in w\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, for $w \in \{l, r\}$, and $\mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \mathrm{E}\left\{\frac{\boldsymbol{e}_{\boldsymbol{\beta}}\boldsymbol{e}_{\boldsymbol{\beta}}^T}{e(1-e)}\middle|\boldsymbol{X} \in p\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$. Also,

$$\mathbf{B}(\boldsymbol{\theta}_0) = \mathrm{E}\left\{\boldsymbol{\psi}(\boldsymbol{O}, \boldsymbol{\theta}_0)\boldsymbol{\psi}(\boldsymbol{O}, \boldsymbol{\theta}_0)^T\right\} = \begin{pmatrix} \{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,1} & (\boldsymbol{H}_{\boldsymbol{\beta},l} - \boldsymbol{H}_{\boldsymbol{\beta},r})^T & \{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,3} \\ \boldsymbol{H}_{\boldsymbol{\beta},l} - \boldsymbol{H}_{\boldsymbol{\beta},r} & \mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{0} \\ \{\mathbf{B}(\boldsymbol{\theta}_0)\}_{3,1} & \mathbf{0} & \gamma_l(1-\gamma_l) \end{pmatrix},$$

where $\{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,1} = \frac{1}{\gamma_l}\,\mathrm{E}\left[\frac{(AY)^2}{e} + \frac{\{(1-A)Y\}^2}{1-e}\middle|\boldsymbol{X} \in l\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \frac{1}{\gamma_r}\,\mathrm{E}\left[\frac{(AY)^2}{e} + \frac{\{(1-A)Y\}^2}{1-e}\middle|\boldsymbol{X} \in r\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} - T(l,r)^2$ and $\{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,3} = \{\mathbf{B}(\boldsymbol{\theta}_0)\}_{3,1} = \gamma_r\{\mu_1(l) - \mu_0(l)\} + \gamma_l\{\mu_1(r) - \mu_0(r)\}$.

Results in Stefanski and Boos [2002] imply that the asymptotic variance of $\widehat{\boldsymbol{\theta}}_0$ is given by

$$\mathbf{V}(\boldsymbol{\theta}_0) = \{\mathbf{A}(\boldsymbol{\theta}_0)\}^{-1}\,\mathbf{B}(\boldsymbol{\theta}_0)\left\{\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\right\}^T = \begin{pmatrix} \{\mathbf{V}(\boldsymbol{\theta}_0)\}_{1,1} & \mathbf{0}^T & 0 \\ \mathbf{0} & \mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} & \mathbf{0} \\ 0 & \mathbf{0} & \gamma_l(1-\gamma_l) \end{pmatrix}.$$

where $\{\mathbf{V}(\boldsymbol{\theta}_0)\}_{1,1} = \{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,1} - (\boldsymbol{H}_{\boldsymbol{\beta},l} - \boldsymbol{H}_{\boldsymbol{\beta},r})^T\mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{H}_{\boldsymbol{\beta},l} - \boldsymbol{H}_{\boldsymbol{\beta},r}) - \frac{1}{\gamma_l\gamma_r}\left[\gamma_r\{\mu_1(l) - \mu_0(l)\}\right.$ $\left.+\gamma_l\{\mu_1(r) - \mu_0(r)\}\right]^2$. $\{\mathbf{V}(\boldsymbol{\theta}_0)\}_{1,1}$ gives the asymptotic variance of $\widehat{T}_{\mathrm{IPW}}(l,r)$. □

Theorem B.1.1 *Assume that the the conditional probability of treatment is estimated using a correctly specified logistic regression model using the data falling in the union of the subgroups $p = (l \cup r)$. The asymptotic variance of $\widehat{T}_{IPW}(l,r)$ when the union of the subgroups $p = (l \cup r)$ is split into l and r can be consistently estimated by*
$\widehat{Var}\left\{\widehat{T}_{IPW}(l,r)\right\} = \frac{1}{n(p)}\left[\frac{1}{n(p)}\sum_{i=1}^n I(\boldsymbol{X}_i \in p)\widehat{I}_i^2 - \frac{1}{\widehat{\gamma}_l\widehat{\gamma}_r}\left[\widehat{\gamma}_r\{\widehat{\mu}_1(l) - \widehat{\mu}_0(l)\} + \widehat{\gamma}_l\{\widehat{\mu}_1(r) - \widehat{\mu}_0(r)\}\right]^2\right]$,
where $\widehat{I}_i = \left\{\frac{I(\boldsymbol{X}_i \in l)A_iY_i}{\widehat{\gamma}_l\cdot\widehat{e}_i} - \frac{I(\boldsymbol{X}_i \in l)(1-A_i)Y_i}{\widehat{\gamma}_l\cdot(1-\widehat{e}_i)}\right\} - \left\{\frac{I(\boldsymbol{X}_i \in r)A_iY_i}{\widehat{\gamma}_r\cdot\widehat{e}_i} - \frac{I(\boldsymbol{X}_i \in r)(1-A_i)Y_i}{\widehat{\gamma}_r\cdot(1-\widehat{e}_i)}\right\} - \widehat{T}_{IPW}(l,r) - (A_i - \widehat{e}_i)\left(\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},l} - \widehat{\boldsymbol{H}}_{\boldsymbol{\beta},r}\right)^T\widehat{\mathbf{E}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\boldsymbol{X}_i$; $\widehat{\gamma}_w = \frac{n(w)}{n(p)}$, $\widehat{e}_i = e_1(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}_{LR})$, $\widehat{\mathbf{E}}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{1}{n(p)}\sum_{i=1}^n I(\boldsymbol{X}_i \in p)\widehat{e}_i(1 - \widehat{e}_i)\boldsymbol{X}_i\boldsymbol{X}_i^T$, $\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},w} = \frac{1}{n(p)}\sum_{i=1}^n\left\{\frac{A_iY_i(1-\widehat{e}_i)}{\widehat{e}_i} + \frac{(1-A_i)Y_i\widehat{e}_i}{(1-\widehat{e}_i)}\right\}\frac{\boldsymbol{X}_iI(\boldsymbol{X}_i \in w)}{\widehat{\gamma}_w}$ *for* $w \in \{l, r\}$.

Proof of Theorem B.1.1: By the consistency of $\widehat{\mu}_a(w), a \in \{0,1\}, w \in \{l, r\}$, $\widehat{\gamma}_l$, and $\widehat{\gamma}_r$ we have $\frac{1}{\widehat{\gamma}_l\widehat{\gamma}_r}\left[\widehat{\gamma}_r\{\widehat{\mu}_1(l) - \widehat{\mu}_0(l)\} + \widehat{\gamma}_l\{\widehat{\mu}_1(r) - \widehat{\mu}_0(r)\}\right]^2 \to \frac{1}{\gamma_l\gamma_r}\left[\gamma_r\{\mu_1(l) - \mu_0(l)\} + \gamma_l\{\mu_1(r) - \mu_0(r)\}\right]^2$ in probability. Also, the empirical mean of the squares of each of the first 3 terms of $\widehat{I}_i$ converges in probability to $\{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,1}$.

The square of the last term in $\widehat{I}_i$ and cross product terms when squaring $\widehat{I}_i$ converge to

$$
\begin{aligned}
&\mathrm{E}\left\{(A-\widehat{e})^2\left(\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},l}-\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},r}\right)^T\widehat{\mathbf{E}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\boldsymbol{X}\boldsymbol{X}^T\widehat{\mathbf{E}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\left(\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},l}-\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},r}\right)\right\}\\
&-2\,\mathrm{E}\left[(A-\widehat{e})\left(\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},l}-\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},r}\right)^T\widehat{\mathbf{E}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\boldsymbol{X}\left[\left\{\frac{I(\boldsymbol{X}\in l)AY}{\frac{n(l)}{n(p)}\cdot\widehat{e}}-\frac{I(\boldsymbol{X}\in l)(1-A)Y}{\frac{n(l)}{n(p)}\cdot(1-\widehat{e})}\right\}\right.\\
&\left.-\left\{\frac{I(\boldsymbol{X}\in r)AY}{\frac{n(r)}{n(p)}\cdot\widehat{e}}-\frac{I(\boldsymbol{X}\in r)(1-A)Y}{\frac{n(r)}{n(p)}\cdot(1-\widehat{e})}\right\}-\widehat{T}_{\mathrm{IPW}}(l,r)\right]\\
&=(\boldsymbol{H}_{\boldsymbol{\beta},l}-\boldsymbol{H}_{\boldsymbol{\beta},r})^T\mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\,\mathrm{E}\left\{(A-\widehat{e})^2\boldsymbol{X}\boldsymbol{X}^T\right\}\mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{H}_{\boldsymbol{\beta},l}-\boldsymbol{H}_{\boldsymbol{\beta},r})-2\,(\boldsymbol{H}_{\boldsymbol{\beta},l}-\boldsymbol{H}_{\boldsymbol{\beta},r})^T\mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}\\
&\cdot\mathrm{E}\left[\frac{\boldsymbol{X}I(\boldsymbol{X}\in l)}{n(l)/n(p)}\left\{\frac{AY(1-\widehat{e})}{\widehat{e}}+\frac{(1-A)Y\widehat{e}}{1-\widehat{e}}\right\}-\frac{\boldsymbol{X}I(\boldsymbol{X}\in r)}{n(r)/n(p)}\left\{\frac{AY(1-\widehat{e})}{\widehat{e}}+\frac{(1-A)Y\widehat{e}}{1-\widehat{e}}\right\}\right.\\
&\left.-(A-\widehat{e})\boldsymbol{X}\widehat{T}_{\mathrm{IPW}}(l,r)\right]\\
&=-(\boldsymbol{H}_{\boldsymbol{\beta},l}-\boldsymbol{H}_{\boldsymbol{\beta},r})^T\mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{H}_{\boldsymbol{\beta},l}-\boldsymbol{H}_{\boldsymbol{\beta},r}).
\end{aligned}
$$

Therefore, $\lim_{n\to\infty}\frac{1}{n(p)}\sum_{i=1}^{n}I(\boldsymbol{X}_i\in p)\widehat{I}_i^2=\{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,1}-(\boldsymbol{H}_{\boldsymbol{\beta},l}-\boldsymbol{H}_{\boldsymbol{\beta},r})^T\mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\boldsymbol{H}_{\boldsymbol{\beta},l}-\boldsymbol{H}_{\boldsymbol{\beta},r})$, which completes the proof. □

B.1.3 Deriving asymptotic variance of $\widehat{T}_{\mathrm{IPW}}(l,r)$ and a consistent variance estimator when the conditional probability of treatment is estimated separately in the two subgroups

Implementation of the CIT algorithms requires choosing what part of the data is used to estimate the conditional probability of treatment (e.g., fit a single model in the union of the subgroups $l\cup r$ or fit a separate model in the subgroups l and r). Estimating the conditional probability of treatment using all the data in $l\cup r$ results in a fitted model using a larger sample size compared to fitting a separate model in each of the subgroups, but using a separate model for each of the subgroups is more flexible as it allows the relationship between the treatment and the covariates to differ between the two subgroups. Another alternative that we explore in the simulations is to build a single model before the start of the tree building process and use that model to estimate the conditional probability of treatment in all the tree building process.

In the following two theorems we derive the asymptotic variance and provide a consistent variance estimator for the case when the conditional probability of treatment for the inverse probability

weighting estimator is estimated using a separate logistic regression model in each subgroup.

Theorem B.1.2 *Assume that the conditional probability of treatment is estimated using two separate correctly specified logistic regression models in two subgroups l and r. The asymptotic variance of $\widehat{T}_{IPW}(l,r)$ when $p=(l\cup r)$ is split into two subgroups l and r is given by* $\frac{1}{\mathrm{P}(\boldsymbol{X}\in l|\boldsymbol{X}\in p)}$ $\mathrm{E}\left[\frac{(YA)^2}{e_1(\boldsymbol{X};\boldsymbol{\beta}_{LR,l})}+\frac{\{Y(1-A)\}^2}{e_0(\boldsymbol{X};\boldsymbol{\beta}_{LR,l})}\middle|\boldsymbol{X}\in l\right]+\frac{1}{\mathrm{P}(\boldsymbol{X}\in r|\boldsymbol{X}\in p)}\,\mathrm{E}\left[\frac{(YA)^2}{e_1(\boldsymbol{X};\boldsymbol{\beta}_{LR,r})}+\frac{\{Y(1-A)\}^2}{e_0(\boldsymbol{X};\boldsymbol{\beta}_{LR,r})}\middle|\boldsymbol{X}\in r\right]-T(l,r)^2$ $-\boldsymbol{H}^T_{\boldsymbol{\beta}_{LR,l}}\mathbf{E}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta},l}\boldsymbol{H}_{\boldsymbol{\beta}_{LR,l}}-\boldsymbol{H}^T_{\boldsymbol{\beta}_{LR,r}}\mathbf{E}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta},r}\boldsymbol{H}_{\boldsymbol{\beta}_{LR,r}}-\frac{[\mathrm{P}(\boldsymbol{X}\in r|\boldsymbol{X}\in p)\{\mu_1(l)-\mu_0(l)\}+\mathrm{P}(\boldsymbol{X}\in l|\boldsymbol{X}\in p)\{\mu_1(r)-\mu_0(r)\}]^2}{\mathrm{P}(\boldsymbol{X}\in l|\boldsymbol{X}\in p)\,\mathrm{P}(\boldsymbol{X}\in r|\boldsymbol{X}\in p)}$. *Here, for subgroup $w\in\{l,r\}$, $\boldsymbol{\beta}_{LR,w}$ is the limit of the maximum likelihood logistic regression estimator,* $\boldsymbol{H}_{\boldsymbol{\beta}_{LR,w}}=\mathrm{E}\left[\left\{\frac{AY}{e_1(\boldsymbol{X};\boldsymbol{\beta}_{LR,w})}+\frac{(1-A)Y}{e_0(\boldsymbol{X};\boldsymbol{\beta}_{LR,w})}\right\}\frac{\partial}{\partial\boldsymbol{\beta}}e_1(\boldsymbol{X};\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{LR,w}}\middle|\boldsymbol{X}\in w\right]$ *and* $\mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta},w}=$

$$\mathrm{E}\left[\left.\frac{I(\boldsymbol{X}\in w)\frac{\partial}{\partial\boldsymbol{\beta}}e_1(\boldsymbol{X};\boldsymbol{\beta})\Big|_{\beta=\beta_{LR,w}}\left\{\frac{\partial}{\partial\boldsymbol{\beta}}e_1(\boldsymbol{X};\boldsymbol{\beta})\Big|_{\beta=\beta_{LR,w}}\right\}^T}{e_1(\boldsymbol{X};\boldsymbol{\beta}_{LR,w})e_0(\boldsymbol{X};\boldsymbol{\beta}_{LR,w})}\right|\boldsymbol{X}\in w\right].$$

Proof of Theorem B.1.2: For simplicity of notation define $e_{i,w}=e_1(\boldsymbol{X}_i;\boldsymbol{\beta}_{\mathrm{LR},w})=\{1+\exp(-\boldsymbol{X}_i\boldsymbol{\beta}_{LR,w})\}^{-1}$ and $1-e_{i,w}=e_0(\boldsymbol{X}_i;\boldsymbol{\beta}_{\mathrm{LR},w})$, and define $\boldsymbol{e}_{\boldsymbol{\beta}_{i,w}}=\partial e_{i,w}/\partial\boldsymbol{\beta}_{LR,w}$ for $w\in\{l,r\}$. The joint set of estimating equations using the data in the union of the subgroups p for $\boldsymbol{\theta}=(T(l,r),\boldsymbol{\beta}^T_{LR,l},\boldsymbol{\beta}^T_{LR,r},\gamma_l)^T$ is given by

$$\begin{cases}\frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\psi_1(Y_i,A_i,\boldsymbol{X}_i,\boldsymbol{\theta})=\frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\left[\frac{I(\boldsymbol{X}_i\in l)}{\gamma_l}\left\{\frac{A_iY_i}{e_{i,l}}-\frac{(1-A_i)Y_i}{1-e_{i,l}}\right\}-\frac{I(\boldsymbol{X}_i\in r)}{1-\gamma_l}\left\{\frac{A_iY_i}{e_{i,r}}\right.\right.\\ \qquad\qquad\left.\left.-\frac{(1-A_i)Y_i}{1-e_{i,r}}\right\}-T(l,r)\right]=0\\ \frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\boldsymbol{\psi}_2(Y_i,A_i,\boldsymbol{X}_i,\boldsymbol{\theta})=\frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\frac{I(\boldsymbol{X}_i\in l)(A_i-e_{i,l})\boldsymbol{e}_{\boldsymbol{\beta}i,l}}{e_{i,l}(1-e_{i,l})}=0\\ \frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\boldsymbol{\psi}_3(Y_i,A_i,\boldsymbol{X}_i,\boldsymbol{\theta})=\frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\frac{I(\boldsymbol{X}_i\in r)(A_i-e_{i,r})\boldsymbol{e}_{\boldsymbol{\beta}i,r}}{e_{i,r}(1-e_{i,r})}=0\\ \frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\psi_4(Y_i,A_i,\boldsymbol{X}_i,\boldsymbol{\theta})=\frac{1}{n(p)}\sum_{i:\boldsymbol{X}_i\in p}\{I(\boldsymbol{X}_i\in l)-\gamma_l\}=0,\end{cases}$$

Let $\boldsymbol{\theta}_0$ be the true value of the set of parameters $\boldsymbol{\theta}$, $\boldsymbol{O}=(Y,A,\boldsymbol{X})$ and define $\boldsymbol{\psi}=(\psi_1,\boldsymbol{\psi}_2,\boldsymbol{\psi}_3,\psi_4)$.

Analogous calculations to those in the proof of Theorem 3.3.2 show that

$$\begin{aligned}
\mathbf{A}(\boldsymbol{\theta}_0) &= \mathrm{E}\left\{-\frac{\partial}{\partial\theta}\boldsymbol{\psi}(\boldsymbol{O},\boldsymbol{\theta}_0)\right\} \\
&= \begin{pmatrix} 1 & \boldsymbol{H}^T_{\boldsymbol{\beta}_{LR,l}} & -\boldsymbol{H}^T_{\boldsymbol{\beta}_{LR,r}} & \frac{1}{\gamma_l}\{\mu_1(l)-\mu_0(l)\}+\frac{1}{\gamma_r}\{\mu_1(r)-\mu_0(r)\} \\ \mathbf{0} & \mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta},l} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta},r} & \mathbf{0} \\ 0 & \mathbf{0} & \mathbf{0} & 1 \end{pmatrix} \\
\Rightarrow \mathbf{A}^{-1}(\boldsymbol{\theta}_0) &= \begin{pmatrix} 1 & -\boldsymbol{H}^T_{\boldsymbol{\beta}_{LR,l}}\mathbf{E}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta},l} & \boldsymbol{H}^T_{\boldsymbol{\beta}_{LR,r}}\mathbf{E}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta},r} & -\frac{1}{\gamma_l}\{\mu_1(l)-\mu_0(l)\}-\frac{1}{\gamma_r}\{\mu_1(r)-\mu_0(r)\} \\ \mathbf{0} & \mathbf{E}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta},l} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{E}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta},r} & \mathbf{0} \\ 0 & \mathbf{0} & \mathbf{0} & 1 \end{pmatrix}.
\end{aligned}$$

We have

$$\mathbf{B}(\boldsymbol{\theta}_0) = \mathrm{E}\left\{\boldsymbol{\psi}(\boldsymbol{O},\boldsymbol{\theta}_0)\boldsymbol{\psi}(\boldsymbol{O},\boldsymbol{\theta}_0)^T\right\} = \begin{pmatrix} \{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,1} & \boldsymbol{H}^T_{\boldsymbol{\beta}_{LR,l}} & -\boldsymbol{H}^T_{\boldsymbol{\beta}_{LR,r}} & \{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,4} \\ \boldsymbol{H}_{\boldsymbol{\beta}_{LR,l}} & \mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta},l} & \mathbf{0} & \mathbf{0} \\ -\boldsymbol{H}_{\boldsymbol{\beta}_{LR,r}} & \mathbf{0} & \mathbf{E}_{\boldsymbol{\beta}\boldsymbol{\beta},r} & \mathbf{0} \\ \{\mathbf{B}(\boldsymbol{\theta}_0)\}_{4,1} & \mathbf{0} & \mathbf{0} & \gamma_l(1-\gamma_l) \end{pmatrix},$$

where $\{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,1} = \mathrm{E}\left\{\psi_1^2(Y_i,A_i,\boldsymbol{X}_i,\boldsymbol{\theta})\right\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ and $\{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,4} = \{\mathbf{B}(\boldsymbol{\theta}_0)\}_{4,1} = \gamma_r\{\mu_1(l)-\mu_0(l)\} + \gamma_l\{\mu_1(r)-\mu_0(r)\}$. Results in Stefanski and Boos [2002] imply that the asymptotic variance of $\widehat{\boldsymbol{\theta}}_0$ is given by $\mathbf{V}(\boldsymbol{\theta}_0) = \{\mathbf{A}(\boldsymbol{\theta}_0)\}^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\left[\{\mathbf{A}(\boldsymbol{\theta}_0)\}^{-1}\right]^T$. It follows that the asymptotic variance of $\widehat{T}_{\mathrm{IPW}}(l,r)$ is $\{\mathbf{B}(\boldsymbol{\theta}_0)\}_{1,1} - \boldsymbol{H}^T_{\boldsymbol{\beta},l}\mathbf{E}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta},l}\boldsymbol{H}_{\boldsymbol{\beta},l} - \boldsymbol{H}^T_{\boldsymbol{\beta},r}\mathbf{E}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta},r}\boldsymbol{H}_{\boldsymbol{\beta},r} - \frac{1}{\gamma_l\gamma_r}\left[\gamma_r\{\mu_1(l)-\mu_0(l)\}+\gamma_l\{\mu_1(r)-\mu_0(r)\}\right]^2$. □

From the above theorem we see that the variance estimator can be decomposed into four parts: one corresponding to the variance when the true conditional probability of treatment and subgroup probabilities are used to calculate $\widehat{T}_{\mathrm{IPW}}(l,r)$ and three terms corresponding to the variance reduction associated with estimating each of the two models for the conditional probability of treatment and the subgroup probability.

Theorem B.1.3 *Assume that the conditional probability of treatment is estimated using two sep-*

arate correctly specified logistic regression models in the two subgroups l and r. The asymptotic variance of $\widehat{T}_{IPW}(l,r)$ when the union of the subgroups $p=(l\cup r)$ is split into two subgroups l and r can be consistently estimated by

$\widehat{Var}\left[\widehat{T}_{IPW}(l,r)\right]=\frac{1}{n(p)}\left[\frac{1}{n(p)}\sum_{i=1}^{n}I(\boldsymbol{X}_i\in p)\widehat{I}_i^2-\frac{1}{\widehat{\gamma}_l\widehat{\gamma}_r}\left[\widehat{\gamma}_r\{\widehat{\mu}_1(l)-\widehat{\mu}_0(l)\}+\widehat{\gamma}_l\{\widehat{\mu}_1(r)-\widehat{\mu}_0(r)\}\right]^2\right]$,
where $\widehat{I}_i=\left\{\frac{I(\boldsymbol{X}_i\in l)A_iY_i}{\widehat{\gamma}_l\cdot\widehat{e}_{i,l}}-\frac{I(\boldsymbol{X}_i\in l)(1-A_i)Y_i}{\widehat{\gamma}_l\cdot(1-\widehat{e}_{i,l})}\right\}-\left\{\frac{I(\boldsymbol{X}_i\in r)A_iY_i}{\widehat{\gamma}_r\cdot\widehat{e}_{i,r}}-\frac{I(\boldsymbol{X}_i\in r)(1-A_i)Y_i}{\widehat{\gamma}_r\cdot(1-\widehat{e}_{i,l})}\right\}-\widehat{T}_{IPW}(l,r)-I(\boldsymbol{X}_i\in l)(A_i-\widehat{e}_{i,l})\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},l}^T\widehat{\mathbf{E}}_{\boldsymbol{\beta}\boldsymbol{\beta},l}^{-1}\boldsymbol{X}_i-I(\boldsymbol{X}_i\in r)(A_i-\widehat{e}_{i,r})\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},r}^T\widehat{\mathbf{E}}_{\boldsymbol{\beta}\boldsymbol{\beta},r}^{-1}\boldsymbol{X}_i$; $\widehat{\gamma}_w=\frac{n(w)}{n(p)}$, $\widehat{e}_{i,w}=e_1(\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}}_{LR,w})$,
$\widehat{\mathbf{E}}_{\boldsymbol{\beta}\boldsymbol{\beta},w}=\frac{1}{n(w)}\sum_{i=1}^{n}I(\boldsymbol{X}_i\in w)\widehat{e}_{i,w}(1-\widehat{e}_{i,w})\boldsymbol{X}_i\boldsymbol{X}_i^T$ *and*
$\widehat{\boldsymbol{H}}_{\boldsymbol{\beta},w}=\frac{1}{n(p)}\sum_{i=1}^{n}\left\{\frac{A_iY_i(1-\widehat{e}_{i,w})}{\widehat{e}_{i,w}}+\frac{(1-A_i)Y_i\widehat{e}_{i,w}}{(1-\widehat{e}_{i,w})}\right\}\frac{\boldsymbol{X}_iI(\boldsymbol{X}_i\in w)}{\widehat{\gamma}_w}$ *for* $w\in\{l,r\}$.

The proof is omitted as it is similar to the proof of Theorem B.1.1.

B.1.4 The first order influence function of $\mu_a(w)$ under non-parametric model

Let $\{p_t:t\in[0,1)\}$ be a regular parametric submodel with $t=0$ being the true "data law". Let $\boldsymbol{O}=(\boldsymbol{X},A,Y)$ be the observed data, and $l(\cdot)$ be the score function. Calculating the pathwise derivative of the target parameter using identification result (3.3) and evaluating at $t=0$ gives

$$\begin{aligned}
&\frac{\partial}{\partial t}\,\mathrm{E}_{p_t}\{\mathrm{E}_{p_t}(Y|\boldsymbol{X},A=a)|\boldsymbol{X}\in w\}\Big|_{t=0}\\
=&\frac{\partial}{\partial t}\,\mathrm{E}_{p_t}\{\mathrm{E}(Y|\boldsymbol{X},A=a)|\boldsymbol{X}\in w\}\Big|_{t=0}+\mathrm{E}\left\{\frac{\partial}{\partial t}\,\mathrm{E}_{p_t}(Y|\boldsymbol{X},A=a)\Big|_{t=0}|\boldsymbol{X}\in w\right\}\\
=&\underbrace{\mathrm{E}\left\{\,\mathrm{E}(Y|\boldsymbol{X},A=a)l(\boldsymbol{O}|\boldsymbol{X}\in w)|\boldsymbol{X}\in w\right\}}_{T_1}+\underbrace{\mathrm{E}\left[\,\mathrm{E}\{Yl(Y|\boldsymbol{X},A=a)|\boldsymbol{X},A=a\}|\boldsymbol{X}\in w\right]}_{T_2}.
\end{aligned}$$

Since $\mathrm{E}\{\mu_a(w)l(\boldsymbol{O}|\boldsymbol{X}\in w)|\boldsymbol{X}\in w\}=\mu_a(w)\,\mathrm{E}\{l(\boldsymbol{O}|\boldsymbol{X}\in w)|\boldsymbol{X}\in w\}=0$, we have

$$\begin{aligned}
T_1=&\,\mathrm{E}\left[\{\mathrm{E}(Y|\boldsymbol{X},A=a)-\mu_a(w)\}\,l(\boldsymbol{O}|\boldsymbol{X}\in w)|\boldsymbol{X}\in w\right]\\
=&\frac{1}{\mathrm{P}(\boldsymbol{X}\in w)}\,\mathrm{E}\left[I(\boldsymbol{X}\in w)\,\{\mathrm{E}(Y|\boldsymbol{X},A=a)-\mu_a(w)\}\,l(\boldsymbol{O}|\boldsymbol{X}\in w)\right]\\
=&\frac{1}{\mathrm{P}(\boldsymbol{X}\in w)}\,\mathrm{E}\left[I(\boldsymbol{X}\in w)\,\{\mathrm{E}(Y|\boldsymbol{X},A=a)-\mu_a(w)\}\,\frac{\partial}{\partial t}\log\mathrm{P}_t(\boldsymbol{O}|\boldsymbol{X}\in w)\right]\\
=&\frac{1}{\mathrm{P}(\boldsymbol{X}\in w)}\,\mathrm{E}\left[I(\boldsymbol{X}\in w)\,\{\mathrm{E}(Y|\boldsymbol{X},A=a)-\mu_a(w)\}\,l(\boldsymbol{O})\right]\\
&-\underbrace{\frac{1}{\mathrm{P}(\boldsymbol{X}\in w)}\,\mathrm{E}\left[I(\boldsymbol{X}\in w)\,\{\mathrm{E}(Y|\boldsymbol{X},A=a)-\mu_a(w)\}\,\frac{\partial}{\partial t}\log\mathrm{P}_t(\boldsymbol{X}\in w)\right]}_{T_3}
\end{aligned}$$

and

$$\begin{aligned}T_3 =& \frac{\partial}{\partial t}\log \mathrm{P}_t(\boldsymbol{X}\in w)\,\mathrm{E}\left\{\mathrm{E}(Y|\boldsymbol{X},A=a)-\mu_a(w)|\boldsymbol{X}\in w\right\}\\ =& \frac{\partial}{\partial t}\log \mathrm{P}_t(\boldsymbol{X}\in w)\left[\mathrm{E}\left\{\mathrm{E}(Y|\boldsymbol{X},A=a)|\boldsymbol{X}\in w\right\}-\mu_a(w)\right]=0.\end{aligned}$$

Using that $\mathrm{E}\left[\mathrm{E}\{\mathrm{E}(Y|\boldsymbol{X},A=a)l(Y|\boldsymbol{X},A=a)|\boldsymbol{X},A=a\}|\boldsymbol{X}\in w\right]=$
$\mathrm{E}\left[\mathrm{E}(Y|\boldsymbol{X},A=a)\,\mathrm{E}\{l(Y|\boldsymbol{X},A=a)|\boldsymbol{X},A=a\}|\boldsymbol{X}\in w\right]=0$, we get

$$\begin{aligned}T_2 =& \mathrm{E}\left[\mathrm{E}\left[\{Y-\mathrm{E}(Y|\boldsymbol{X},A=a)\}\,l(Y|\boldsymbol{X},A=a)|\boldsymbol{X},A=a\right]|\,\boldsymbol{X}\in w\right]\\ =& \mathrm{E}\left[\mathrm{E}\left[\frac{I(A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}\{Y-\mathrm{E}(Y|\boldsymbol{X},A=a)\}\,l(Y|\boldsymbol{X},A=a)\middle|\,\boldsymbol{X}\right]\middle|\,\boldsymbol{X}\in w\right]\\ =& \mathrm{E}\left[\mathrm{E}\left[\frac{I(A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}\{Y-\mathrm{E}(Y|\boldsymbol{X},A=a)\}\,l(Y|\boldsymbol{X},A)\middle|\,\boldsymbol{X}\right]\middle|\,\boldsymbol{X}\in w\right]\\ =& \frac{1}{\mathrm{P}(\boldsymbol{X}\in w)}\mathrm{E}\left[\frac{I(\boldsymbol{X}\in w)I(A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}\{Y-\mathrm{E}(Y|\boldsymbol{X},A=a)\}\,l(Y|\boldsymbol{X},A)\right]\\ =& \frac{1}{\mathrm{P}(\boldsymbol{X}\in w)}\mathrm{E}\left[\frac{I(\boldsymbol{X}\in w)I(A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}\{Y-\mathrm{E}(Y|\boldsymbol{X},A=a)\}\,l(\boldsymbol{O})\right]\\ &-\underbrace{\frac{1}{\mathrm{P}(\boldsymbol{X}\in w)}\mathrm{E}\left[\frac{I(\boldsymbol{X}\in w)I(A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}\{Y-\mathrm{E}(Y|\boldsymbol{X},A=a)\}\,l(\boldsymbol{X},A)\right]}_{T_4}\end{aligned}$$

Now $T_4 = \mathrm{E}\left[\mathrm{E}\left[\frac{I(A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}\{Y-\mathrm{E}(Y|\boldsymbol{X},A=a)\}\,l(\boldsymbol{X},A=a)\middle|\,\boldsymbol{X}\right]\middle|\,\boldsymbol{X}\in w\right]=0$. Combining the previous results gives $\frac{\partial}{\partial t}\mathrm{E}_{p_t}[\mathrm{E}_{p_t}[Y|\boldsymbol{X},A=a]|\boldsymbol{X}\in w]\big|_{t=0}=$
$\mathrm{E}\left[\frac{1}{\mathrm{P}(\boldsymbol{X}\in w)}\left[I(\boldsymbol{X}\in w)\{\mathrm{E}(Y|\boldsymbol{X},A=a)-\mu_a(w)\}+\frac{I(\boldsymbol{X}\in w)I(A=a)}{\mathrm{P}(A=a|\boldsymbol{X})}\{Y-\mathrm{E}(Y|\boldsymbol{X},A=a)\}\right]l(\boldsymbol{O})\right]$,
which completes the derivation of the influence function.

B.1.5 Consistency and asymptotic properties of the doubly robust estimator

Proof of Theorem 3.3.3:

Unless otherwise stated, all convergence results in this section refer to convergence in probability.

1. *Consistency:* By the law of large numbers, $\widehat{\mu}_{\mathrm{DR},a}(w)\to\mathrm{E}\left[\frac{I(\boldsymbol{X}\in w)}{\mathrm{P}(\boldsymbol{X}\in w)}\left[g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*)+\right.\right.$
$\left.\left.\frac{I(A=a)}{e_a(\boldsymbol{X};\boldsymbol{\beta}^*)}\{Y-g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*)\}\right]\right]$. We now study the asymptotic limit of the doubly robust estimator by considering the two different cases in Assumption A.4.

- When $g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a) \to \mathrm{E}(Y|\boldsymbol{X}, A=a)$ holds, using the law of total expectation,

$$\begin{aligned}\widehat{\mu}_{\mathrm{DR},a}(w) \to & \mathrm{E}\left[\mathrm{E}(Y|\boldsymbol{X}, A=a) + \frac{I(A=a)}{e_a(\boldsymbol{X};\boldsymbol{\beta}^*)}\left\{Y - \mathrm{E}(Y|\boldsymbol{X}, A=a)\right\}\middle| \boldsymbol{X} \in w\right] \\ = & \mathrm{E}\left\{\mathrm{E}(Y|\boldsymbol{X}, A=a)|\boldsymbol{X} \in w\right\} \\ = & \mu_a(w)\end{aligned}$$

- When $e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}}) \to P(A=a|\boldsymbol{X})$ holds, using the law of total expectation gives

$$\begin{aligned}\widehat{\mu}_{\mathrm{DR},a}(w) \to & \mathrm{E}\left[g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*) + \frac{I(A=a)}{P(A=a|\boldsymbol{X})}\left\{Y - g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*)\right\}\middle| \boldsymbol{X} \in w\right] \\ = & \mathrm{E}\left[g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*) + \frac{\mathrm{E}\left\{I(A=a)Y|\boldsymbol{X}\right\} - g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*)P(A=a|\boldsymbol{X})}{P(A=a|\boldsymbol{X})}\middle| \boldsymbol{X} \in w\right] \\ = & \mathrm{E}\left\{\mathrm{E}\left(Y|\boldsymbol{X}, A=a\right)|\,\boldsymbol{X} \in w\right\} \\ = & \mu_a(w)\end{aligned}$$

This completes the consistency proof of $\widehat{\mu}_{\mathrm{DR},a}(w)$.

2. *Rate of convergence:*

Decompose

$$\begin{aligned}&\sqrt{n}\left\{\widehat{\mu}_{\mathrm{DR,a}}(w) - \mu_a(w)\right\} \\ =& \underbrace{\left[\mathbb{G}_n\left\{H\left(e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}}), g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a), A, Y, \widehat{\gamma}_w\right)\right\} - \mathbb{G}_n\left\{H(e_a(\boldsymbol{X};\boldsymbol{\beta}^*), g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*), A, Y, \gamma_w)\right\}\right]}_{T_1} \\ &+ \underbrace{\mathbb{G}_n\left\{H(e_a(\boldsymbol{X};\boldsymbol{\beta}^*), g_a(\boldsymbol{X};\boldsymbol{\eta}_a^*), A, Y, \gamma_w)\right\}}_{T_2} \\ &+ \underbrace{\sqrt{n}\left[\mathrm{E}\left\{H(e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}}), g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a), A, Y, \widehat{\gamma}_w)\right\} - \mu_a(w)\right]}_{T_3}\end{aligned}$$

By assumptions A.1 and A.2, $T_1 = o_P(1)$. By the central limit theorem and assumption A.3,

T_2 is asymptotically normal and $\sqrt{n}$ consistent. Using that $\sqrt{n}\left(\frac{1}{\widehat{\gamma}_w} - \frac{1}{\gamma_w}\right) = O_P(1)$ we have

$$\begin{aligned}\frac{1}{\sqrt{n}}T_3 =& \mathrm{E}\left[\frac{I(\boldsymbol{X} \in w)}{\widehat{\gamma}_w}\left[g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a) + \frac{I(A=a)}{e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}})}\left\{Y - g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a)\right\}\right]\right] - \mu_a(w) \\ =& \mathrm{E}\left[\frac{I(\boldsymbol{X} \in w)}{\gamma_w}\left[g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a) + \frac{I(A=a)}{e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}})}\left\{Y - g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a)\right\}\right]\right] - \mu_a(w) + O_P\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Using the law of total expectation, the Cauchy Schwarz inequality, and the identifiability result (3.3) we have

$$\begin{aligned}&\mathrm{E}\left[\frac{I(\boldsymbol{X} \in w)}{\gamma_w}\left[g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a) + \frac{I(A=a)}{e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}})}\left\{Y - g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a)\right\}\right]\right] - \mu_a(w) \\ &\leq O_P\left\{||e_a(\boldsymbol{X};\widehat{\boldsymbol{\beta}}) - \mathrm{P}(A=a|\boldsymbol{X})||_2 \times ||g_a(\boldsymbol{X};\widehat{\boldsymbol{\eta}}_a) - \mathrm{E}(Y|\boldsymbol{X},A=a)||_2\right\}.\end{aligned}$$

The rate of convergence follows by combining the above results.

B.1.6 Consistency of the Causal Interaction Tree algorithms

In this section, we show that the CIT-based treatment effect estimator is L_2 consistent both when the models for the conditional probability of treatment assignment and/or the conditional expectation of the outcome are fit using the whole dataset and when the models are fit separately in each terminal node.

Let $\mathcal{X}$ be the sample space of $\boldsymbol{X}$; assume that the outcome is bounded, $|Y| \leq B_1 < \infty$; define $\mathcal{O} = \mathcal{X} \times \{0,1\} \times [-B_1, B_1]$ as the sample space of the observed data $\boldsymbol{O} = (\boldsymbol{X}, A, Y)$. Borrowing notation from Nobel et al. [1996] and Sun et al. [2019], let $\widehat{\psi}_n$ be an empirical partition rule (tree) with $\widehat{K}_n$ partitions (nodes). Let π be any partitioning of $\mathcal{X}$ and define Π as the range of π and Π_n as the range [Nobel et al., 1996] of $\widehat{\psi}_n$, respectively. For each element $\pi = \{u_k\} \in \Pi$, define an associated partition $\widetilde{\pi} = \{u_k \times \{0,1\} \times [-B_1, B_1]\}$ on $\mathcal{O}$ and let $\widetilde{\Pi}$ be the range of $\widetilde{\pi}$. Analogously, for each element $\pi_n = \{w_k\} \in \Pi_n$, define an associated partition $\widetilde{\pi}_n = \{w_k \times \{0,1\} \times [-B_1, B_1]\}$ on $\mathcal{O}$ and let $\widetilde{\Pi}_n$ be the range of $\widetilde{\pi}_n$. Finally, let $w_\pi(\boldsymbol{x})$ be the unique partition (terminal node) in partition rule π that contains the covariate vector $\boldsymbol{x}$.

As the proof of the IPW-CIT and g-CIT algorithms are similar to that of the DR-CIT algorithm,

we focus on the DR-CIT algorithm. Define the transformed outcome

$$
\begin{aligned}
&Y^*_{\mathrm{DR}}\{\boldsymbol{O};\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta}_1,\boldsymbol{\eta}_0)\}\\
&=\left[g_1(\boldsymbol{X};\boldsymbol{\eta}_1)+\frac{I(A=1)\{Y-g_1(\boldsymbol{X};\boldsymbol{\eta}_1)\}}{e_1(\boldsymbol{X};\boldsymbol{\beta})}\right]-\left[g_0(\boldsymbol{X};\boldsymbol{\eta}_0)+\frac{I(A=0)\{Y-g_0(\boldsymbol{X};\boldsymbol{\eta}_0)\}}{e_0(\boldsymbol{X};\boldsymbol{\beta})}\right],
\end{aligned}
$$

where the treatment and outcome models are indexed by $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta}_1,\boldsymbol{\eta}_0)$. The transferred outcome $Y^*_{\mathrm{DR}}\{\boldsymbol{O};\widehat{\boldsymbol{\theta}}=(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\eta}}_1,\widehat{\boldsymbol{\eta}}_0)\}$ uses models for the conditional probability of treatment assignment and the conditional expectation of the outcome that are estimated using the whole dataset. Define a second transformed outcome as

$$
\begin{aligned}
&Y^*_{w_{\pi_n}(\boldsymbol{X}),\mathrm{DR}}\{\boldsymbol{O};\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{X})}=(\boldsymbol{\beta},\boldsymbol{\eta}_1,\boldsymbol{\eta}_0)\}\\
&=\left[g_{1,w_{\pi_n}(\boldsymbol{X})}(\boldsymbol{X};\boldsymbol{\eta}_1)+\frac{I(A=1)\{Y-g_{1,w_{\pi_n}(\boldsymbol{X})}(\boldsymbol{X};\boldsymbol{\eta}_1)\}}{e_{1,w_{\pi_n}(\boldsymbol{X})}(\boldsymbol{X};\boldsymbol{\beta})}\right]\\
&\quad-\left[g_{0,w_{\pi_n}(\boldsymbol{X})}(\boldsymbol{X};\boldsymbol{\eta}_0)+\frac{I(A=0)\{Y-g_{0,w_{\pi_n}(\boldsymbol{X})}(\boldsymbol{X};\boldsymbol{\eta}_0)\}}{e_{0,w_{\pi_n}(\boldsymbol{X})}(\boldsymbol{X};\boldsymbol{\beta})}\right],
\end{aligned}
$$

where the treatment and outcome models are indexed by $\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{X})}=(\boldsymbol{\beta},\boldsymbol{\eta}_1,\boldsymbol{\eta}_0)$. The transferred outcome $Y^*_{w_{\pi_n}(\boldsymbol{X}),\mathrm{DR}}\{\boldsymbol{O};\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{X})}=(\boldsymbol{\beta},\boldsymbol{\eta}_1,\boldsymbol{\eta}_0)\}$ uses models for the conditional probability of treatment assignment and the conditional expectation of the outcome that are fit using terminal node data. For simplicity of notation, we use $Y^*_{\mathrm{DR}}(\boldsymbol{\theta})$ and $Y^*_{w_{\pi_n}(\boldsymbol{X}),\mathrm{DR}}\left(\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{X})}\right)$ to denote $Y^*_{\mathrm{DR}}\{\boldsymbol{O};\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\eta}_1,\boldsymbol{\eta}_0)\}$ and $Y^*_{w_{\pi_n}(\boldsymbol{X}),\mathrm{DR}}\{\boldsymbol{O};\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{X})}=(\boldsymbol{\beta},\boldsymbol{\eta}_1,\boldsymbol{\eta}_0)\}$, respectively.

Using this notation the doubly robust CIT-based treatment effect estimators for $r(\boldsymbol{x})=\mathrm{E}(Y|A=1,\boldsymbol{X}=\boldsymbol{x})-\mathrm{E}(Y|A=0,\boldsymbol{X}=\boldsymbol{x})$ can be written as

$$
\widehat{r}_{1n,\mathrm{DR}}(\boldsymbol{x})=\frac{\sum_{i=1}^n I\{\boldsymbol{X}_i\in w_{\widehat{\psi}_n}(\boldsymbol{x})\}Y^*_{i,\mathrm{DR}}(\widehat{\boldsymbol{\theta}})}{\sum_{i=1}^n I\{\boldsymbol{X}_i\in w_{\widehat{\psi}_n}(\boldsymbol{x})\}},
$$

when the models for the conditional probability of treatment assignment and the conditional expectation of the outcome are estimated using the whole dataset and as

$$
\widehat{r}_{2n,\mathrm{DR}}(\boldsymbol{x})=\frac{\sum_{i=1}^n I\{\boldsymbol{X}_i\in w_{\widehat{\psi}_n}(\boldsymbol{x})\}Y^*_{i,w_{\widehat{\psi}_n}(\boldsymbol{x}),\mathrm{DR}}\left(\widehat{\boldsymbol{\theta}}_{w_{\widehat{\psi}_n}(\boldsymbol{x})}\right)}{\sum_{i=1}^n I\{\boldsymbol{X}_i\in w_{\widehat{\psi}_n}(\boldsymbol{x})\}},
$$

when the models are using terminal node data.

To prove the consistency of $\widehat{r}_{1n,\mathrm{DR}}(\boldsymbol{x})$ we make the following assumptions:

B.1 There exists a $B_1 > 0$ such that $|Y| \leq B_1 < \infty$.

B.2 The number of terminal nodes satisfies $\widehat{K}_n \to \infty$ and $\widehat{K}_n = o\left\{\frac{n}{\log(n)}\right\}$.

B.3 The derivative $\frac{\partial Y^*_{\mathrm{DR}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ exists and is uniformly bounded.

B.4 For any $\boldsymbol{x} \in \mathcal{X}$ and $a \in \{0,1\}$, at least one of $g_a(\boldsymbol{x};\boldsymbol{\eta}^*_a) = \mathrm{E}(Y|A=a,\boldsymbol{x})$ or $e_a(\boldsymbol{x};\boldsymbol{\beta}^*) = \mathrm{P}(A=a|\boldsymbol{x})$ holds.

Note that assumption B.4 is the same as the doubly robust assumption A.4. The following theorem shows that the DR-CIT algorithm is L_2 consistent.

Theorem B.1.4 *Under assumptions B.1-B.4 and if the models for the conditional probability of treatment assignment and the conditional expectation of the outcome are estimated using the whole dataset, the doubly robust Causal Interaction Tree algorithm is L_2 consistent.*

Proof of Theorem B.1.4: Define

$$\widetilde{r}_n(\boldsymbol{x}) = \frac{\sum_{i=1}^n I\{\boldsymbol{X}_i \in w_{\widehat{\psi}_n}(\boldsymbol{x})\} r(\boldsymbol{X}_i)}{\sum_{i=1}^n I\{\boldsymbol{X}_i \in w_{\widehat{\psi}_n}(\boldsymbol{x})\}}.$$

It is easy to show that

$$\int_{\mathcal{X}} |r(\boldsymbol{x}) - \widehat{r}_{1n,\mathrm{DR}}(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) \leq 2\int_{\mathcal{X}} |\widehat{r}_{1n,\mathrm{DR}}(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) + 2\int_{\mathcal{X}} |r(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}),$$

where $\mathrm{P}(\boldsymbol{x})$ is the distribution function for $\boldsymbol{x}$. We will show that both terms on the right hand side of the above equation are $o_P(1)$.

By assumption B.1 and the positivity assumption for the conditional probability of treatment assignment, there exists a $B_2 > 0$ such that $\max\left\{\left|Y^*_{\mathrm{DR}}(\widehat{\boldsymbol{\theta}})\right|, \left|Y^*_{\mathrm{DR}}(\boldsymbol{\theta}^*)\right|\right\} \leq B_2 < \infty$ for $\boldsymbol{O} \in \mathcal{O}$, where $\boldsymbol{\theta}^*$ is the limit of $\widehat{\boldsymbol{\theta}}$. Define $B_1^* = \max(B_1, B_2)$. Assumption B.1 and equation (18) in Nobel

et al. [1996] imply

$$
\begin{aligned}
&\int_{\mathcal{X}} |\widehat{r}_{1n,\mathrm{DR}}(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) \\
\leq & 2B_1^* \sum_{w \in \widehat{\psi}_n} \gamma_w \left| \frac{1}{n(w)} \sum_{i=1}^n I\{\boldsymbol{X}_i \in w\} \{Y^*_{i,\mathrm{DR}}(\widehat{\boldsymbol{\theta}}) - r(\boldsymbol{X}_i)\} \right| \\
\leq & 2B_1^* \sum_{w \in \widehat{\psi}_n} \widehat{\gamma}_w \left| \frac{1}{n(w)} \sum_{i=1}^n I\{\boldsymbol{X}_i \in w\} \{Y^*_{i,\mathrm{DR}}(\widehat{\boldsymbol{\theta}}) - r(\boldsymbol{X}_i)\} \right| \\
& + 2B_1^* \sum_{w \in \widehat{\psi}_n} \left| \frac{1}{n(w)} \sum_{i=1}^n I\{\boldsymbol{X}_i \in w\} \{Y^*_{i,\mathrm{DR}}(\widehat{\boldsymbol{\theta}}) - r(\boldsymbol{X}_i)\} \right| |\widehat{\gamma}_w - \gamma_w| \\
\leq & 2B_1^* \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} \widehat{\gamma}_w \left| \frac{1}{n(w)} \sum_{i=1}^n I\{\boldsymbol{X}_i \in w\} \{Y^*_{i,\mathrm{DR}}(\widehat{\boldsymbol{\theta}}) - r(\boldsymbol{X}_i)\} \right| \\
& + 4B_1^{*2} \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} |\widehat{\gamma}_w - \gamma_w| \,.
\end{aligned}
$$

Under assumption B.2, applying Proposition 2, Lemma 3, and Theorem 7 in Nobel et al. [1996] gives $\sup_{\pi \in \Pi_n} \sum_{w \in \pi} |\widehat{\gamma}_w - \gamma_w| = o_P(1)$. It follows that

$$
\int_{\mathcal{X}} |\widehat{r}_{1n,\mathrm{DR}}(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) \leq 2B_1^* \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} \left| \frac{1}{n} \sum_{i=1}^n I\{\boldsymbol{X}_i \in w\} \{Y^*_{i,\mathrm{DR}}(\widehat{\boldsymbol{\theta}}) - r(\boldsymbol{X}_i)\} \right| + o_P(1).
$$

Using assumption B.3 and Taylor series arguments,

$$
\begin{aligned}
&\int_{\mathcal{X}} |\widehat{r}_{1n,\mathrm{DR}}(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) \\
\leq & 2B_1^* \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} \left| \frac{1}{n} \sum_{i=1}^n I(\boldsymbol{X}_i \in w) \left\{ Y^*_{i,\mathrm{DR}}(\boldsymbol{\theta}^*) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \left. \frac{\partial Y^*_{i,\mathrm{DR}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} - r(\boldsymbol{X}_i) \right\} \right| + o_P(1) \\
= & 2B_1^* \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} \left| \frac{1}{n} \sum_{i=1}^n I(\boldsymbol{X}_i \in w) \{Y^*_{i,\mathrm{DR}}(\boldsymbol{\theta}^*) - r(\boldsymbol{X}_i)\} \right| + o_P(1),
\end{aligned}
$$

where the last equality follows because $\widehat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}^*$.

Define $h\{\boldsymbol{o} = (\boldsymbol{x}, a, y); \boldsymbol{\theta}^*\} = y^*_{\mathrm{DR}}(\boldsymbol{\theta}^*) - r(\boldsymbol{x})$ on $\mathcal{O}$. By assumption B.1 and the positivity assumption for the conditional probability of treatment, $h(\boldsymbol{o}; \boldsymbol{\theta}^*)$ is bounded. By assumption B.4, $\int_{u \times \{0,1\} \times [-B_1, B_1]} h(\boldsymbol{o}; \boldsymbol{\theta}^*) \mathrm{d}\,\mathrm{P}(\boldsymbol{o}) = 0$ for every measurable subset u of $\mathcal{X}$, where $\mathrm{P}(\boldsymbol{o})$ is the joint distribution function for $\boldsymbol{o}$. By assumption B.2, Proposition 2, Lemma 3 and Equation (13) in

Nobel et al. [1996],

$$\begin{aligned}
&\int_{\mathcal{X}} |\widehat{r}_{1n,\mathrm{DR}}(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) \\
\leq & 2B_1^* \sup_{\widetilde{\pi}_n \in \widetilde{\Pi}_n} \sum_{\widetilde{w} \in \widetilde{\pi}_n} \left| \mathbb{P}_n \left\{ I(\boldsymbol{O} \in \widetilde{w}) h(\boldsymbol{O}; \boldsymbol{\theta}^*) \right\} - \int I(\boldsymbol{o} \in \widetilde{w}) h(\boldsymbol{o}; \boldsymbol{\theta}^*) \mathrm{d}\,\mathrm{P}(\boldsymbol{o}) \right| + o_P(1) = o_P(1).
\end{aligned}$$

By exactly the same arguments in the Proof of Theorem 1 in Nobel et al. [1996], $\int_{\mathcal{X}} |r(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) = o_P(1)$. It then follows that $\int_{\mathcal{X}} |r(\boldsymbol{x}) - \widehat{r}_{1n,\mathrm{DR}}(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) = o_P(1)$.

□

To show the consistency of $\widehat{r}_{2n,\mathrm{DR}}(\boldsymbol{x})$, we replace conditions B.3 and B.4 by

B.3* The derivative $\left. \frac{\partial Y^*_{w_{\pi_n}(\boldsymbol{X}),\mathrm{DR}}(\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{X})})}{\partial \boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{X})}} \right|_{\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{X})} = \boldsymbol{\theta}^*_{w_{\pi_n}(\boldsymbol{X})}}$ exists for $\pi_n \in \Pi_n$ and is uniformly bounded.

B.4* For $a \in \{0,1\}$ and any $\widetilde{\pi} = \{\widetilde{u}_k\} \in \widetilde{\Pi}$ and $\boldsymbol{o} \in \widetilde{u}_k$, at least one of $g_{a,\widetilde{u}_k}(\boldsymbol{x}; \boldsymbol{\eta}^*_{a,\widetilde{u}_k}) = \mathrm{E}(Y|A = a, \boldsymbol{x})$ or $e_{a,\widetilde{u}_k}(\boldsymbol{x}; \boldsymbol{\beta}^*) = \mathrm{P}(A = a|\boldsymbol{x})$ holds.

Theorem B.1.5 *Under assumptions B.1, B.2, B.3*, and B.4* and if the models for the conditional probability of treatment assignment and the conditional expectation of the outcome are fit separately in each terminal node, the doubly robust Causal Interaction Tree algorithm is L_2 consistent.*

Proof of Theorem B.1.5:

It follows from the proof of Theorem B.1.4 and the inequality

$$\int_{\mathcal{X}} |r(\boldsymbol{x}) - \widehat{r}_{2n,\mathrm{DR}}(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) \leq 2 \int_{\mathcal{X}} |\widehat{r}_{2n,\mathrm{DR}}(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) + 2 \int_{\mathcal{X}} |r(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}),$$

that it is sufficient to show that

$$\int_{\mathcal{X}} |\widehat{r}_{2n,\mathrm{DR}}(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) = o_P(1).$$

By assumption B.1 and the positivity assumption for the conditional probability of treatment assignment, there exists a $B_3 > 0$ such that $\max \left\{ \left| Y^*_{w_{\pi_n}(\boldsymbol{X}),\mathrm{DR}} \left(\widehat{\boldsymbol{\theta}}_{w_{\pi_n}(\boldsymbol{X})} \right) \right|, \left| Y^*_{w_{\pi_n}(\boldsymbol{X}),\mathrm{DR}} \left(\boldsymbol{\theta}^*_{w_{\pi_n}(\boldsymbol{X})} \right) \right| \right\} \leq B_3 < \infty$ for $\boldsymbol{O} \in \mathcal{O}$ and $\pi_n \in \Pi_n$, where $\boldsymbol{\theta}^*_{w_{\pi_n}(\boldsymbol{X})}$ is the limit of $\widehat{\boldsymbol{\theta}}_{w_{\pi_n}(\boldsymbol{X})}$. Define $B_2^* =$

$\max(B_1, B_3)$. Applying assumption B.3* and similar arguments as in the proof of Theorem B.1.4 gives

$$\int_{\mathcal{X}} |\widehat{r}_{2n,\mathrm{DR}}(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) \leq 2B_2^* \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} \left| \frac{1}{n} \sum_{i=1}^{n} I\{\boldsymbol{X}_i \in w\} \{Y^*_{i,w,\mathrm{DR}}(\boldsymbol{\theta}^*_w) - r(\boldsymbol{X}_i)\} \right| + o_P(1).$$

Define a class of functions

$$\mathcal{H} = \left\{ h\left(\boldsymbol{o} = (\boldsymbol{x}, a, y); \widetilde{\pi} = \{\widetilde{u}_k\}, \{\boldsymbol{\theta}^*_{\widetilde{u}_k}\}\right) = \sum_{\boldsymbol{o} \in \widetilde{u}_k} I\{\boldsymbol{o} \in \widetilde{u}_k\} \{y^*_{\widetilde{u}_k,\mathrm{DR}}(\boldsymbol{\theta}^*_{\widetilde{u}_k}) - r(\boldsymbol{x})\} : \widetilde{\pi} \in \widetilde{\Pi} \right\}.$$

Under assumption B.1 and the positivity assumption for the conditional probability of treatment, $h\left(\boldsymbol{o}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}_k}\}\right)$ is bounded. Under assumption B.4*, $\int_{u \times \{0,1\} \times [-B^*, B^*]} h\left(\boldsymbol{o}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}_k}\}\right) \mathrm{d}\,\mathrm{P}(\boldsymbol{o}) = 0$ for every measurable subset u of $\mathcal{X}$ and $h\left(\boldsymbol{o}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}_k}\}\right) \in \mathcal{H}$. Therefore, by assumption B.2, Proposition 2, Lemma 3 and Equation (13) in Nobel et al. [1996],

$$\begin{aligned}
&\int_{\mathcal{X}} |\widehat{r}_{2n,\mathrm{DR}}(\boldsymbol{x}) - \widetilde{r}_n(\boldsymbol{x})|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{x}) \\
\leq & 2B_2^* \sup_{\widetilde{\pi}_n \in \widetilde{\Pi}_n} \sum_{\widetilde{w} \in \widetilde{\pi}_n} \left| \mathbb{P}_n\left\{ I(\boldsymbol{O} \in \widetilde{w}) h(\boldsymbol{O}; \widetilde{\pi}_n, \{\boldsymbol{\theta}^*_{\widetilde{w}}\}) \right\} - \int I(\boldsymbol{o} \in \widetilde{w}) h(\boldsymbol{o}; \widetilde{\pi}_n, \{\boldsymbol{\theta}^*_{\widetilde{w}}\}) \mathrm{d}\,\mathrm{P}(\boldsymbol{o}) \right| + o_P(1) \\
\leq & 2B_2^* \sup_{h_{\widetilde{\pi}} \in \mathcal{H}} \sup_{\widetilde{\pi}_n \in \widetilde{\Pi}_n} \sum_{\widetilde{w} \in \widetilde{\pi}_n} \left| \mathbb{P}_n\left\{ I(\boldsymbol{O} \in \widetilde{w}) h(\boldsymbol{O}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}}\}) \right\} - \int I(\boldsymbol{o} \in \widetilde{w}) h(\boldsymbol{o}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}}\}) \mathrm{d}\,\mathrm{P}(\boldsymbol{o}) \right| + o_P(1) \\
= & o_P(1).
\end{aligned}$$

□

B.2 Alternative Final Tree Selection Method

The alternative final tree selection method is adapted from the tree selection method proposed in Steingrimsson and Yang [2019]. We will now briefly describe the method, but refer to Steingrimsson and Yang [2019] for further details. In the Classification and Regression Tree algorithm for outcome prediction [Breiman et al., 1984], the final tree is selected based on minimizing cross-validation error. As the treatment effect is not observed on any participant, cross-validation cannot be directly used for final tree selection using treatment effect prediction error.

To overcome this difficulty, we will use the random forest algorithm [Breiman, 2001] as a surrogate for the true treatment effect estimate and select the tree that gives the prediction closest to the random forest predictions. We start by splitting the data into a training and a validation set, fitting an inverse probability weighted random forest model to the training data, and calculating the treatment effect predictions on the validation set. We refer to the predictions as the random forest treatment effect validation set predictions.

In the final tree selection step, for a given split into a training and a validation set and a candidate tree $\widehat{\psi}_m$, $m \in \{1, \ldots, \widehat{M}\}$, re-estimate the terminal node estimators of $\widehat{\psi}_m$ using only the training data falling in each terminal node. Use the re-estimated terminal node estimators to predict the treatment effect for all participants in the validation set and refer to the predictions as the CIT validation set predictions. Calculate the cross-validation error for tree $\widehat{\psi}_m$ corresponding to this particular split into validation and training set as the average L_2 distance between the CIT validation set predictions and the random forest treatment effect validation set predictions. The final tree is selected as the tree that results in the smallest cross-validation error averaged over all splits into validation and training sets.

B.3 Additional Simulation Results

We use 1000 simulations to evaluate all the algorithms in this section, except in Web Appendix B.3.3, where the running time for the algorithms compared in Section 3.4 is evaluated using 10,000 simulations.

B.3.1 Selecting a splitting rule and cross-validation method for Causal Tree algorithms

Following, Athey and Imbens [2016], for the "Original CT" algorithm, we set both the splitting rule (`split.Rule`) and the cross-validation method (`"cv.option"`) to `"CT"`. We also set `split.Honest` and `cv.Honest` to `TRUE` for honest splitting and cross-validation.

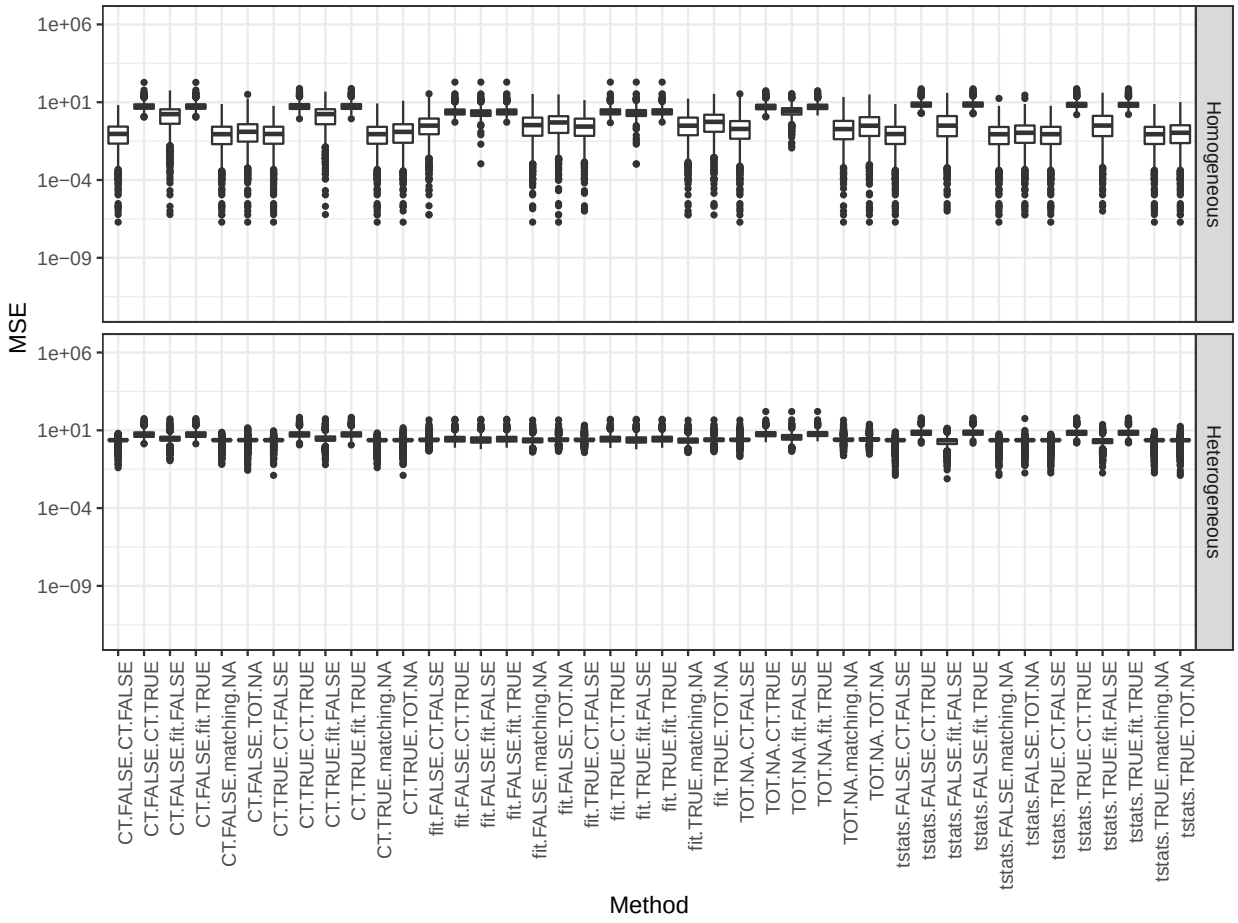

Figure B.1: Mean squared error (MSE) for different choices of splitting rules and cross-validation methods for the Causal Tree algorithms for the homogeneous (top) and the heterogeneous (bottom) settings described in Section 3.4.1. Lower values indicate better performance. Splitting rule (`split.Rule`) can be chosen from transformed outcome trees (`TOT`), causal tree (`CT`), fit-based trees (`fit`) and squared t-statistic trees (`tstats`). Adaptive (`split.Honest = FALSE`) and honest (`split.Honest = TRUE`) versions are available except for `TOT`. Cross-validation method (`cv.option`) can be chosen from `TOT`, matching (`matching`), `CT` and `fit`. Adaptive (`cv.Honest = FALSE`) and honest (`cv.honest = TRUE`) versions are available for `CT` and `fit`. The naming pattern for the choices are "(`split.Rule`).(`split.Honest`).(`cv.option`).(`cv.Honest`)". If there is no honest version of the chosen splitting rule or cross-validation method, `split.Honest` or `cv.Honest` is `NA`.

Table B.1: Average MSE (lower is better), proportion of correct trees (higher is better), average number of noise variables used for splitting (lower is better), pairwise prediction similarity (higher is better), and proportion of trees making the first split correctly (only for heterogeneous setting, higher is better) for the ten combinations of splitting rules and cross-validation methods included in `causalTree` that have the highest rank in terms of average MSE. Columns 2, 3, 4, 5, and 6 show simulation results when the treatment effect is homogeneous, and columns 7, 8, 9, 10, 11, and 12 show simulation results when the treatment effect is heterogeneous. Splitting rule (`split.Rule`) can be chosen from transformed outcome trees (`TOT`), causal tree (`CT`), fit-based trees (`fit`) and squared t-statistic trees (`tstats`). Adaptive (`split.Honest = FALSE`) and honest (`split.Honest = TRUE`) versions are available except for `TOT`. Cross-validation method (`cv.option`) can be chosen from `TOT`, matching (`matching`), `CT` and `fit`. Adaptive (`cv.Honest = FALSE`) and honest (`cv.honest = TRUE`) versions are available for `CT` and `fit`. The naming pattern for the choices are "(`split.Rule`) (`split.Honest`) (`cv.option`) (`cv.Honest`)". If there is no honest version of the chosen splitting rule or cross-validation method, `split.Honest` or `cv.Honest` is `NA`.

	Homogeneous Effect					Heterogeneous Effect					
Parameter Setting	MSE	Correct Trees	Number Noise	PPS	Time	MSE	Correct Trees	Number Noise	PPS	Time	Correct First Split
tstats TRUE matching NA	0.23	0.98	0.04	1.00	0.02	2.14	0.15	0.08	0.58	0.02	0.63
tstats TRUE CT FALSE	0.23	0.98	0.05	0.99	0.05	2.24	0.11	0.08	0.56	0.05	0.63
tstats FALSE matching NA	0.23	0.99	0.03	1.00	0.02	2.17	0.13	0.10	0.57	0.02	0.62
tstats FALSE CT FALSE	0.23	0.98	0.06	0.99	0.05	2.24	0.10	0.07	0.56	0.05	0.62
tstats TRUE TOT NA	0.35	0.89	0.50	0.95	0.02	2.20	0.12	0.47	0.59	0.02	0.63
CT TRUE matching NA	0.24	0.97	0.04	1.00	0.02	2.29	0.08	0.12	0.55	0.02	0.35
CT TRUE CT FALSE	0.24	0.96	0.06	0.99	0.06	2.35	0.06	0.16	0.54	0.06	0.35
tstats FALSE TOT NA	0.36	0.90	0.46	0.95	0.02	2.27	0.10	0.42	0.57	0.02	0.62
CT FALSE CT FALSE	0.24	0.96	0.07	0.99	0.06	2.34	0.06	0.12	0.54	0.06	0.32
CT FALSE matching NA	0.24	0.96	0.06	1.00	0.02	2.34	0.06	0.13	0.54	0.02	0.32

In addition, we also compared the performance of the CIT algorithms against a CT algorithm with the combination of tuning parameters that has the highest rank on average in terms of minimizing MSE across the two simulation settings. We refer to this parameter combination as "Optimized CT" in simulation results.

The splitting rule (`split.Rule`) can be chosen from transformed outcome trees (`TOT`), causal trees (`CT`), fit-based trees (`fit`) and squared t-statistic trees (`tstats`). Both adaptive and honest versions of the estimators are available for the splitting rules except for `TOT`. This leads to 7 possible choices of splitting rules. The cross-validation method (`cv.option`) can be chosen from transformed outcome (`TOT`), matching (`matching`), `CT` and `fit`. Adaptive and honest versions of the criterion are available for `CT` and `fit`. This leads to 6 possible choices of cross-validation methods.

Figure B.1 shows boxplots of MSE using 1000 simulations for both the homogeneous and the heterogeneous simulation settings described in Section 3.4.1. All combinations of splitting rules and cross-validation methods are included in the boxplots (a total of $7 \times 6 = 42$ combinations). The model for estimating the conditional probability of treatment for the `weights` parameter includes the main effects of all covariates. Table B.1 shows the average MSE, proportion of correct trees, average number of noise variables, pairwise prediction similarity, the average running time for both simulation settings, and the proportion of trees making a correct first split in the heterogeneous setting for the ten combinations of splitting rules and cross-validation methods that have the highest rank on average in terms of minimizing MSE.

The results in Figure B.1 and Table B.1 show that setting `split.Rule` equal to `"tstats"` with the honest version (`split.Honest = TRUE`) and `cv.option` to `"matching"` performs the best and is therefore used for Optimized CT in the main manuscript.

B.3.2 Simulations comparing the performance of regular and honest Causal Trees

Figure B.2 presents the simulation results for the homogeneous and the heterogeneous settings described in Section 3.4.1 when regular (`causalTree`) and honest (`honest.causalTree`) CT algorithms are fitted using the same tuning parameter selections as for Original CT and Optimized CT described in Section 3.4.

Figure B.2 shows that regular CTs give similar or lower MSE than their honest peers under all

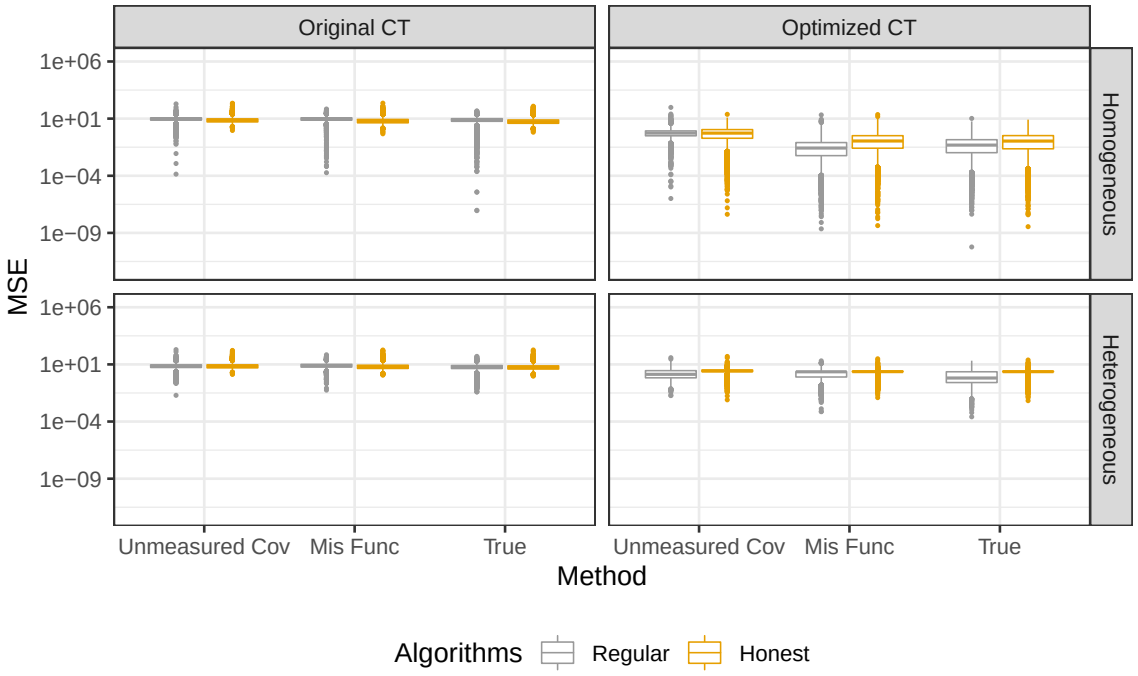

Figure B.2: Mean squared error (MSE) for regular (`causalTree`) and honest (`honest.causalTree`) Causal Trees for the homogeneous (top) and the heterogeneous (bottom) settings described in Section 3.4.1. Lower values indicate better performance. Original CT refers to the original Causal Tree algorithms by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters.

model specifications for the conditional probability of treatment. Therefore, results from regular CTs are used in the comparisons presented in the main simulations in Section 3.4.4.

B.3.3 Comparisons of running time

Table B.2 lists the average time in seconds it takes to implement the methods compared in the simulations in Section 3.4. The results show that the DR-CITs run on average 8-30 times faster than IPW-CITs and g-CITs. The reason is that the variance estimators for the inverse probability weighting and g-formula estimators require calculating the inverse of the quadratic form of the design matrix and thus we need to ensure the design matrix is full rank every time we calculate the splitting statistic. On the other hand, the variance estimator for the doubly robust estimator does not involve that extra step, which substantially reduces the computational complexity.

Table B.2: The average time in seconds it takes to implement the methods (lower is better) corresponding to algorithms in Figure 3.1 and Table 3.1. Original CT refers to the original Causal Tree algorithms by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Section 3.4.3, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

		Homogeneous	Heterogeneous
Algorithm	Model	Time	Time
Original CT	Unmeasured Cov	0.37	0.35
	Mis Func	0.36	0.34
	True	0.38	0.34
Optimized CT	Unmeasured Cov	0.10	0.09
	Mis Func	0.10	0.08
	True	0.10	0.07
IPW-CIT	Unmeasured Cov	77.74	69.41
	Mis Func	102.72	96.18
	True	80.12	73.41
g-CIT	Unmeasured Cov	159.11	156.09
	Mis Func	191.90	188.74
	True	140.09	74.39
DR-CIT	Both Unmeasured Cov	9.95	8.64
	Both Mis Func	12.55	10.47
	True Treat Mis Func Out	12.09	10.15
	True Out Mis Func Treat	13.14	11.35
	Both True	12.85	11.12

Table B.3: Proportion of correct trees (higher is better), average number of noise variables used for splitting (lower is better), pairwise prediction similarity (higher is better), the average time in seconds it takes to implement the methods (lower is better), and proportion of trees making the first split correctly (only for heterogeneous setting, higher is better) corresponding to algorithms in Figure B.3 when the alternative final tree selection method described in B.2 is used. Columns 3, 4, 5, and 6 show simulation results when the treatment effect is homogeneous, and columns 7, 8, 9, 10, and 11 show simulation results when the treatment effect is heterogeneous. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Section 3.4.3, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

	Homogeneous Effect				Heterogeneous Effect				
Algorithm	Correct Trees	Number Noise	PPS	Time	Correct Trees	Number Noise	PPS	Time	Correct First Split
IPW-CIT									
Unmeasured Cov	0.96	0.05	0.98	144.09	0.49	0.27	0.74	127.76	0.76
Mis Func	0.96	0.05	0.98	187.63	0.15	0.25	0.59	174.52	0.21
True	0.98	0.04	0.99	148.98	0.32	0.18	0.65	135.04	0.54
g-CIT									
Unmeasured Cov	0.84	0.44	0.90	261.45	0.43	0.70	0.78	255.23	0.97
Mis Func	0.69	0.95	0.81	308.33	0.40	1.18	0.78	307.52	0.97
True	0.23	6.94	0.43	227.53	0.99	0.00	1.00	115.01	1.00
DR-CIT									
Both Unmeasured Cov	0.98	0.03	0.99	43.83	0.87	0.08	0.93	41.62	0.98
Both Mis Func	0.89	0.21	0.97	50.40	0.83	0.21	0.94	46.82	0.99
True Treat Mis Func Out	0.97	0.06	0.99	48.53	0.92	0.10	0.96	45.71	1.00
True Out Mis Func Treat	0.84	0.28	0.94	57.51	0.75	0.47	0.96	55.65	1.00
Both True	0.84	0.27	0.94	56.12	0.76	0.43	0.97	53.87	1.00

B.3.4 Simulations when the alternative final tree selection method described in B.2 is used

Figure B.3 shows boxplots of MSE and Table B.3 shows the proportion of correct trees, average number of noise variables, pairwise prediction similarity, the average running time for both simulation settings, and the proportion of trees making a correct first split in the heterogeneous setting when the final tree selection method described in B.2 is used in connection with the CIT algorithms.

Figure B.3 shows that the relative performance of the CIT algorithms in terms of MSE is similar to what is seen in Figure 3.1 except that in the homogeneous setting, the MSE of the inverse probability weighting trees using misspecified functional form of covariates in the model for

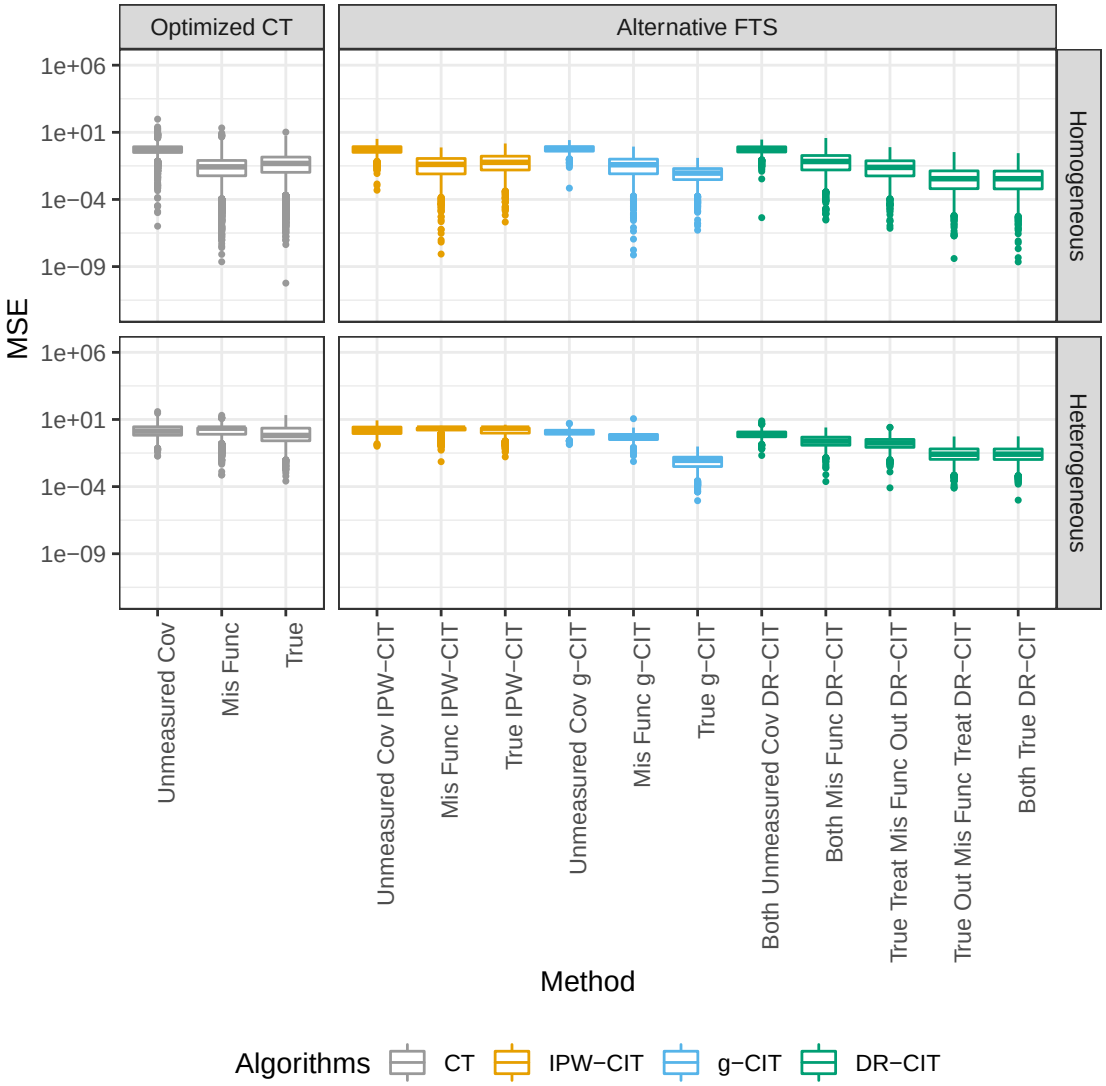

Figure B.3: Mean squared error (MSE) for the different tree building algorithms for the homogeneous (top) and the heterogeneous (bottom) settings described in Section 3.4.1 when the alternative final tree selection method described in B.2 is used. Lower values indicate better performance. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Alternative FTS" refers to the Causal Interaction Tree algorithms using the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Section 3.4.3, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

the conditional probability of treatment is slightly smaller than that using the correctly specified model.

The results in Figure B.3 and Table B.3 show that when the final tree selection method developed in Steingrimsson and Yang [2019] is used, the CIT algorithms outperform the CT algorithms. However, when comparing the different CIT algorithms the results do not always match what is expected based on the properties of the estimators for $\mu_a(w)$ used in the tree building process. For example, the g-CITs with a correctly specified model for the conditional expectation of the outcome overfit (build larger trees than the true tree) in the homogeneous setting and perform worse on some evaluation measures than the g-CITs with misspecified models. Similar trend is seen for the DR-CITs in both settings.

B.3.5 Simulations when the models are fitted on the whole dataset or separately within each child node

Figure B.4 shows boxplots of MSE and Table B.4 shows the proportion of correct trees, average number of noise variables, pairwise prediction similarity, the average running time for both simulation settings, and the proportion of trees making a correct first split in the heterogeneous setting when the models for the conditional probability of treatment and/or the models for the conditional expectation of the outcome are fitted prior to the tree building process using the whole dataset. Figure B.5 and Table B.5 show the results when the models are fitted separately in each potential child node.

The performance of the CIT algorithms when the models are fitted on the whole dataset presented in Figure B.4 and Table B.4 is similar to when the models are fitted using the data in the node that is being considered for splitting (Figure 3.1, Table 3.1, Figure B.3 and Table B.3).

When the models are fitted separately within each potential child node, the performance of the CIT algorithms is similar to when the models are fitted using the data falling in the node being considered for splitting or using the whole dataset except that 1) there are more outliers in the mean squared error and 2) the performance of DR-CITs with correctly specified model for the conditional expectation of the outcome using the final tree selection method described in the main manuscript becomes worse. These indicate that the CIT algorithms are more likely to overfit when the models are fitted separately within each potential child node. A potential explanation is that

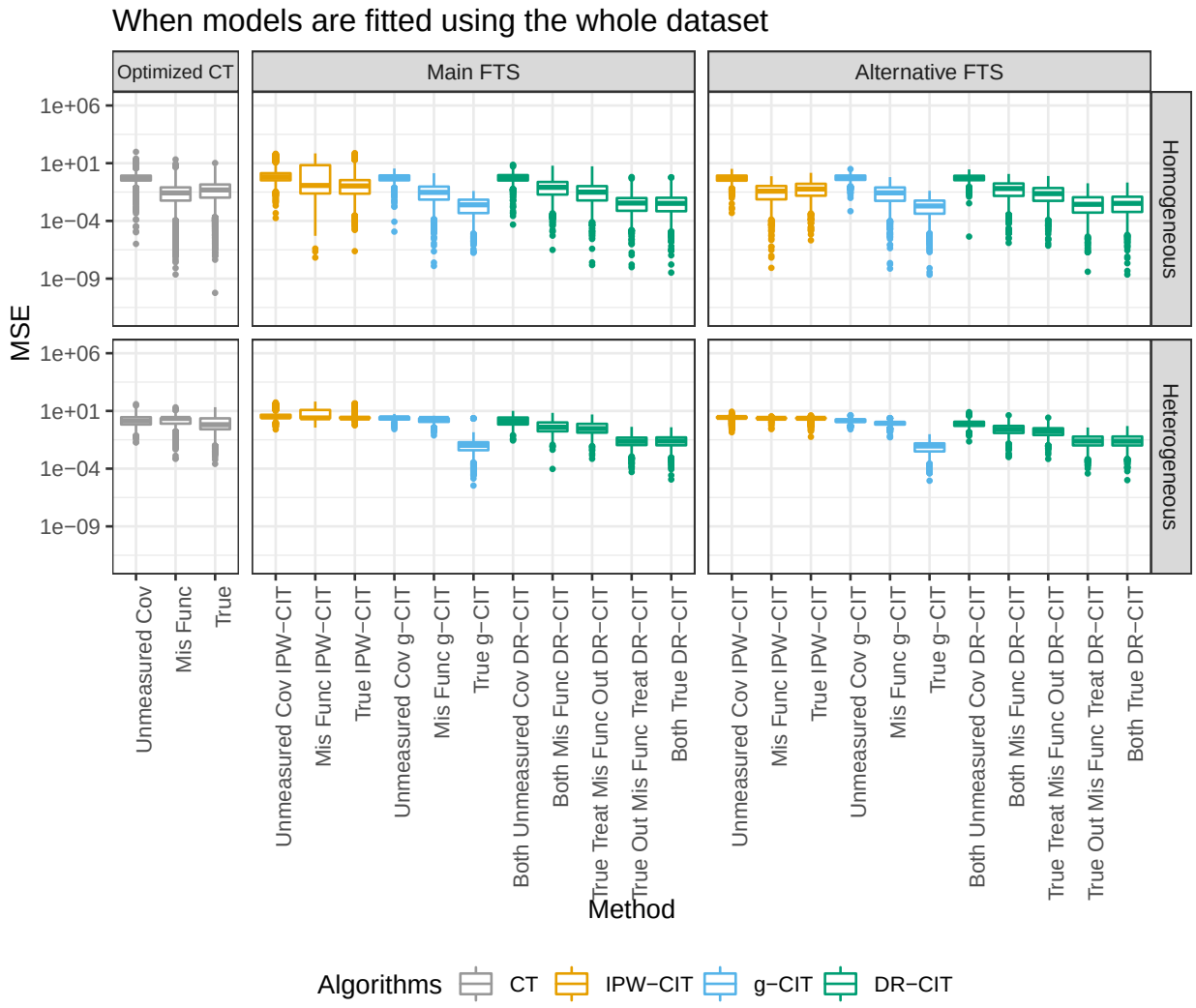

Figure B.4: Mean squared error (MSE) for the different tree building algorithms for the homogeneous (top) and the heterogeneous (bottom) settings described in Section 3.4.1 when the models for the conditional probability of treatment and/or the models for the conditional expectation of the outcome are fitted prior to the tree building process using the whole dataset. Lower values indicate better performance. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Section 3.4.3, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

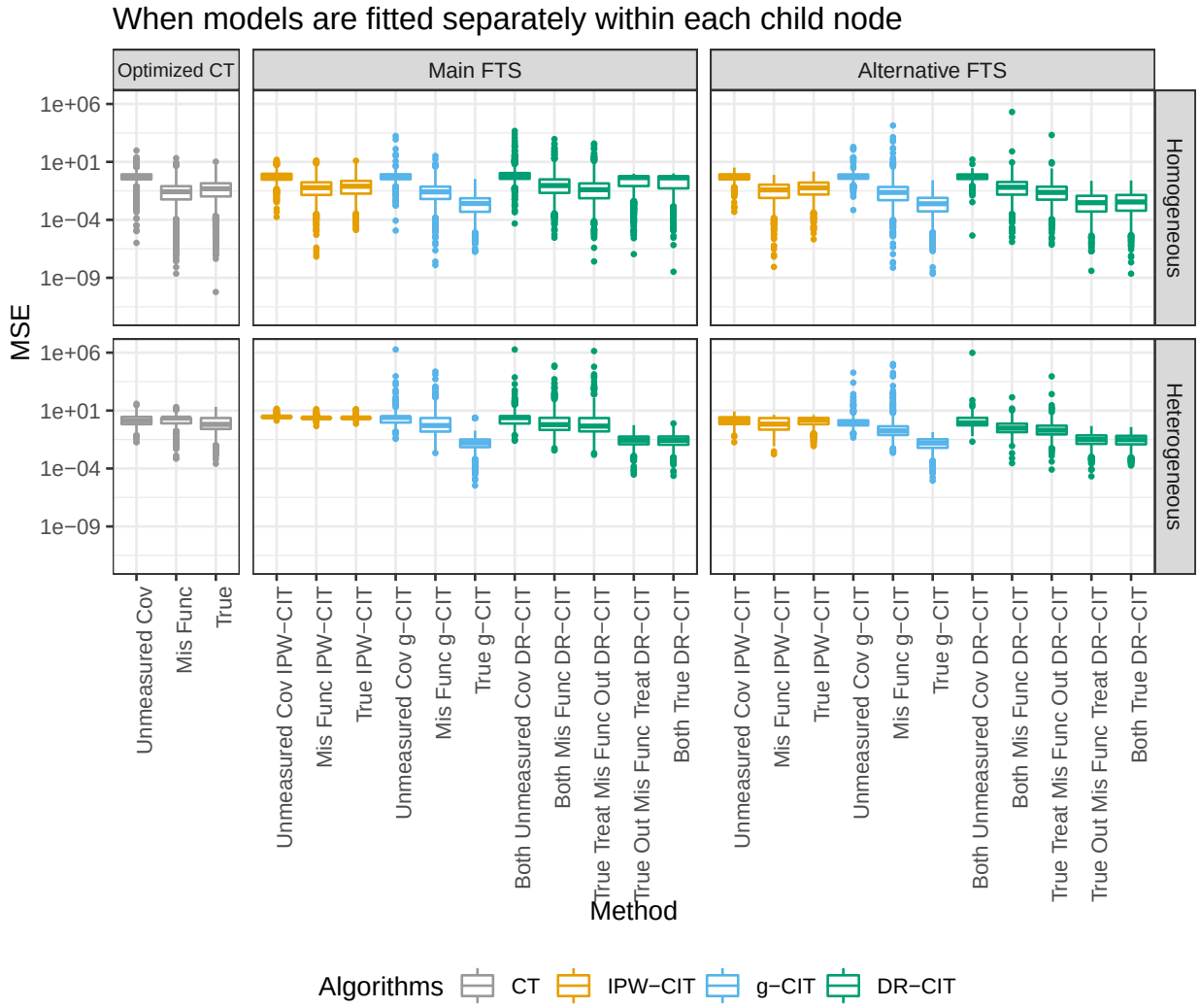

Figure B.5: Mean squared error (MSE) for the different tree building algorithms for the homogeneous (top) and the heterogeneous (bottom) settings described in Section 3.4.1 when the models for the conditional probability of treatment and/or the models for the conditional expectation of the outcome are fitted within each potential child node separately. Lower values indicate better performance. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Section 3.4.3, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

Table B.4: Proportion of correct trees (higher is better), average number of noise variables used for splitting (lower is better), pairwise prediction similarity (higher is better), the average time in seconds it takes to implement the methods (lower is better), and proportion of trees making the first split correctly (only for heterogeneous setting, higher is better) corresponding to algorithms in Figure B.4 when the models for the conditional probability of treatment and/or the models for the conditional expectation of the outcome are fitted prior to the tree building process using the whole dataset. Columns 3, 4, 5, and 6 show simulation results when the treatment effect is homogeneous, and columns 7, 8, 9, 10, and 11 show simulation results when the treatment effect is heterogeneous. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Section 3.4.3, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

	Homogeneous Effect				Heterogeneous Effect				
Algorithm	Correct Trees	Number Noise	PPS	Time	Correct Trees	Number Noise	PPS	Time	Correct First Split
Main FTS									
IPW-CIT									
Unmeasured Cov	0.82	1.77	0.86	84.00	0.04	1.71	0.55	77.29	0.71
Mis Func	0.69	2.37	0.80	122.26	0.01	2.53	0.53	114.14	0.22
True	0.85	1.32	0.90	87.06	0.03	1.32	0.53	81.18	0.50
g-CIT									
Unmeasured Cov	0.92	0.10	0.96	161.39	0.32	0.07	0.61	158.13	0.96
Mis Func	0.93	0.10	0.96	197.96	0.33	0.04	0.62	194.56	0.96
True	1.00	0.00	1.00	151.42	0.99	0.00	0.99	82.40	1.00
DR-CIT									
Both Unmeasured Cov	0.92	0.16	0.97	10.31	0.58	0.17	0.79	8.49	0.97
Both Mis Func	0.92	0.21	0.97	13.34	0.73	0.19	0.87	10.57	0.99
True Treat Mis Func Out	0.94	0.15	0.98	13.14	0.77	0.19	0.88	10.33	1.00
True Out Mis Func Treat	0.95	0.17	0.98	12.98	0.94	0.12	0.99	10.29	1.00
Both True	0.94	0.18	0.98	13.39	0.94	0.10	0.99	10.45	1.00
Alternative FTS									
IPW-CIT									
Unmeasured Cov	1.00	0.00	1.00	126.07	0.18	0.00	0.57	114.65	0.76
Mis Func	1.00	0.00	1.00	181.78	0.01	0.00	0.51	171.12	0.21
True	1.00	0.00	1.00	130.03	0.07	0.00	0.53	120.42	0.54
g-CIT									
Unmeasured Cov	0.80	0.73	0.88	236.16	0.59	0.82	0.78	227.09	0.97
Mis Func	0.91	0.25	0.95	286.19	0.59	0.57	0.79	277.76	0.97
True	0.27	6.51	0.54	216.11	0.98	0.00	1.00	113.42	1.00
DR-CIT									
Both Unmeasured Cov	1.00	0.00	1.00	21.78	0.92	0.02	0.94	18.77	0.98
Both Mis Func	1.00	0.00	1.00	27.55	0.93	0.02	0.94	22.82	0.99
True Treat Mis Func Out	1.00	0.00	1.00	26.92	0.96	0.02	0.97	22.57	1.00
True Out Mis Func Treat	0.87	0.20	0.95	26.92	0.77	0.39	0.97	22.52	1.00
Both True	0.85	0.25	0.94	26.29	0.78	0.36	0.97	22.51	1.00

Table B.5: Proportion of correct trees (higher is better), average number of noise variables used for splitting (lower is better), pairwise prediction similarity (higher is better), the average time in seconds it takes to implement the methods (lower is better), and proportion of trees making the first split correctly (only for heterogeneous setting, higher is better) corresponding to algorithms in Figure B.5 when the models for the conditional probability of treatment and/or the models for the conditional expectation of the outcome are fitted within each potential child node separately. Lower values indicate better performance. Columns 3, 4, 5, and 6 show simulation results when the treatment effect is homogeneous, and columns 7, 8, 9, 10, and 11 show simulation results when the treatment effect is heterogeneous. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Section 3.4.3, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

	Homogeneous Effect				Heterogeneous Effect				
Algorithm	Correct Trees	Number Noise	PPS	Time	Correct Trees	Number Noise	PPS	Time	Correct First Split
Main FTS									
IPW-CIT									
Unmeasured Cov	0.97	0.17	0.99	369.57	0.00	0.12	0.50	290.36	0.66
Mis Func	0.93	0.17	0.98	498.16	0.01	0.10	0.51	393.91	0.69
True	0.95	0.14	0.99	433.71	0.01	0.06	0.50	335.55	0.64
g-CIT									
Unmeasured Cov	0.99	0.02	1.00	452.43	0.36	0.09	0.69	407.65	0.91
Mis Func	0.99	0.02	1.00	650.68	0.54	0.07	0.78	551.47	0.92
True	0.99	0.06	0.99	757.01	0.99	0.00	0.99	313.05	1.00
DR-CIT									
Both Unmeasured Cov	0.83	0.67	0.94	699.52	0.40	0.73	0.77	482.48	0.79
Both Mis Func	0.87	0.58	0.95	1022.13	0.57	0.79	0.86	639.57	0.88
True Treat Mis Func Out	0.88	0.45	0.96	945.72	0.58	0.59	0.86	597.50	0.88
True Out Mis Func Treat	0.21	10.43	0.36	927.01	0.93	0.16	0.98	607.27	1.00
Both True	0.23	9.88	0.38	856.06	0.93	0.19	0.98	578.88	1.00
Alternative FTS									
IPW-CIT									
Unmeasured Cov	0.99	0.01	1.00	554.95	0.52	0.14	0.79	436.94	0.70
Mis Func	0.99	0.01	1.00	757.16	0.51	0.13	0.80	598.35	0.73
True	1.00	0.01	1.00	661.16	0.29	0.20	0.72	498.39	0.67
g-CIT									
Unmeasured Cov	0.98	0.02	0.99	671.50	0.79	0.05	0.89	611.97	0.92
Mis Func	0.97	0.03	1.00	979.15	0.84	0.08	0.93	841.21	0.94
True	0.93	0.26	0.96	1144.92	0.77	0.20	0.98	426.33	1.00
DR-CIT									
Both Unmeasured Cov	1.00	0.00	1.00	1124.23	0.71	0.07	0.86	755.15	0.84
Both Mis Func	0.98	0.04	1.00	1662.19	0.78	0.10	0.90	1020.50	0.91
True Treat Mis Func Out	0.99	0.01	1.00	1543.16	0.84	0.06	0.92	947.59	0.92
True Out Mis Func Treat	0.86	0.26	0.97	1457.26	0.73	0.51	0.96	926.53	1.00
Both True	0.83	0.30	0.96	1326.09	0.74	0.41	0.97	873.73	1.00

instability is induced into the modeling procedure when only a small subset of the data is used to fit the models.

B.3.6 Simulations when the models are estimated using random forests

Non-parametric methods can be used to estimate the models for the conditional probability of treatment and/or the models for the conditional expectation of the outcome. However, many flexible data-driven algorithms violate the Donsker condition detailed in Assumption A.1 in Section 3.3.4, which can lead to bias in the treatment effect estimation [Chernozhukov et al., 2018].

In this section, we evaluate the performance of DR-CITs that use the random forest algorithm to fit the models for the conditional probability of treatment and the models for the conditional expectation of the outcome needed for the implementation of $\widehat{\mu}_{\text{DR},a}$. We estimate the variance using the influence function approach The models are fitted prior to the tree building process using the whole dataset.

We implemented two versions of DR-CITs. In the first version, the models were fitted using random forests on all the covariates in their original form. The second version mimics the scenario where there is an unmeasured common cause of the outcome and treatment assignment and we excluded $X^{(2)}$ from the tree building process.

Figure B.6 shows boxplots of MSE and Table B.6 shows the proportion of correct trees, average number of noise variables, pairwise prediction similarity, the average running time for both simulation settings, and the proportion of trees making a correct first split in the heterogeneous setting for DR-CITs when the models for the conditional probability of treatment and the models for the conditional expectation of the outcome are estimated using random forests.

In general, the performance of DR-CITs when the models are estimated by random forests is similar to the results presented in Web Appendix B.3.5 where the models are estimated by generalized linear models using the whole dataset and the final tree selection method described in the main manuscript is used.

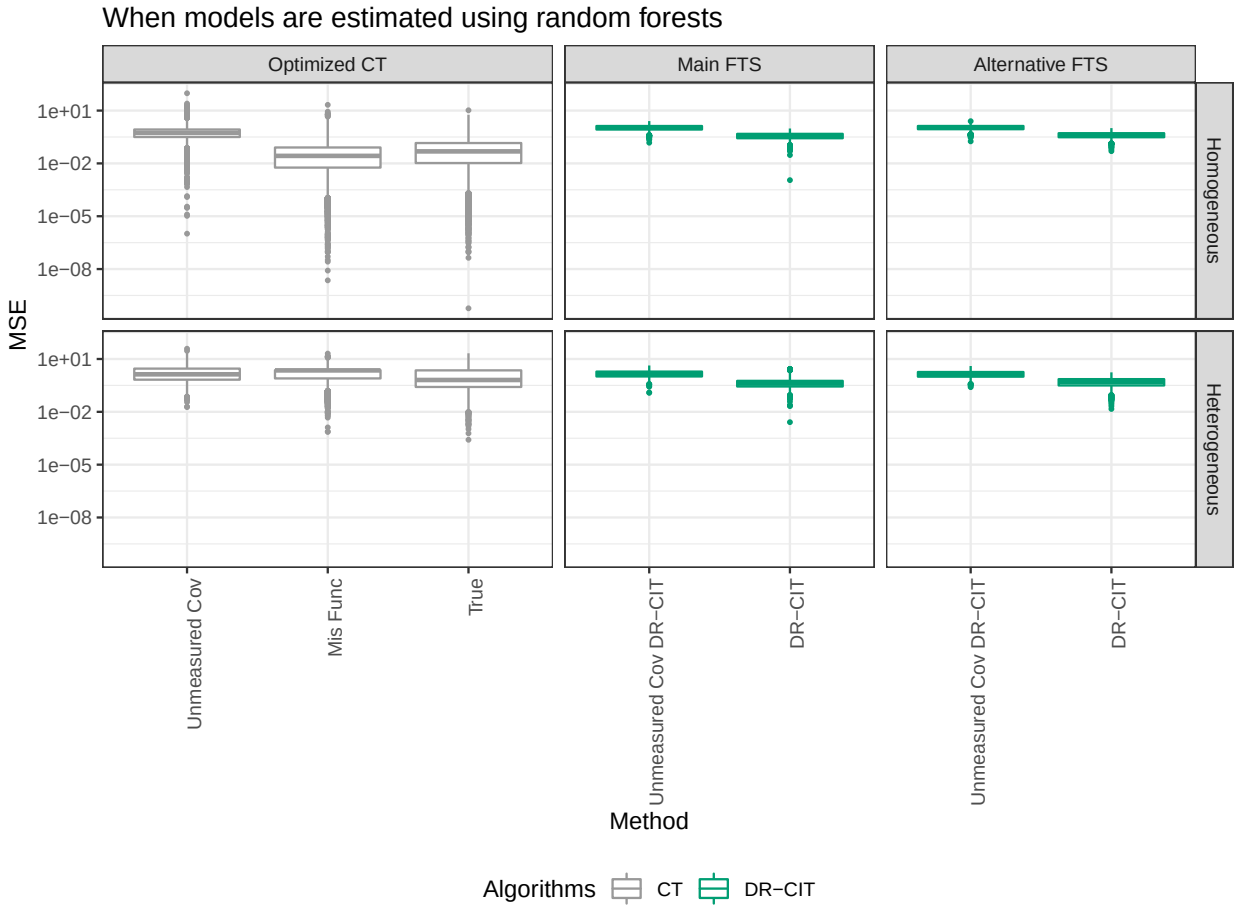

Figure B.6: Mean squared error (MSE) for Optimized CTs and DR-CITs for the homogeneous (top) and the heterogeneous (bottom) settings described in Section 3.4.1 when the models for the conditional probability of treatment and the models for the conditional expectation of the outcome are estimated using random forests. Lower values indicate better performance. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. "Unmeasured Cov DR-CIT" refers to having an unmeasured common cause of outcome and treatment assignment as described in Web Appendix B.3.6. "DR-CIT" refers to including all covariates when fitting random forests as described in Web Appendix B.3.6.

Table B.6: Proportion of correct trees (higher is better), average number of noise variables used for splitting (lower is better), pairwise prediction similarity (higher is better), the average time in seconds it takes to implement the methods (lower is better), and proportion of trees making the first split correctly (only for heterogeneous setting, higher is better) for DR-CITs when the models for the conditional probability of treatment and the models for the conditional expectation of the outcome are estimated using random forests. Columns 2, 3, 4, and 5 show simulation results when the treatment effect is homogeneous, and columns 6, 7, 8, 9, and 10 show simulation results when the treatment effect is heterogeneous. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. "Unmeasured Cov DR-CIT" refers to having an unmeasured common cause of outcome and treatment assignment as described in Web Appendix B.3.6. "DR-CIT" refers to including all covariates when fitting random forests as described in Web Appendix B.3.6.

	Homogeneous Effect				Heterogeneous Effect				
Algorithm	Correct Trees	Number Noise	PPS	Time	Correct Trees	Number Noise	PPS	Time	Correct First Split
Main FTS									
Unmeasured Cov DR-CIT	0.90	0.17	0.95	16.87	0.74	0.13	0.89	15.82	1.00
DR-CIT	0.91	0.17	0.97	22.98	0.80	0.16	0.95	20.85	1.00
Alternative FTS									
Unmeasured Cov DR-CIT	0.53	0.65	0.78	48.08	0.40	1.22	0.89	46.70	1.00
DR-CIT	0.20	1.95	0.63	56.41	0.17	3.04	0.85	54.67	1.00

B.3.7 Simulations for a binary outcome with continuous and categorical covariates

In this section, we present simulation results where the outcome is binary and the covariate vector includes both continuous and categorical variables. We furthermore use simulations to evaluate coverage of confidence intervals that use data-splitting.

B.3.7.1 Simulation Results

The covariate vector included six variables. The first three components were generated from a 3-dimensional mean zero multivariate normal distribution, where $\forall j, k \in \{1, 2, 3\}$, $\text{cov}(X^{(j)}, X^{(k)}) = 0.3$ when $j \neq k$, and $\text{Var}(X^{(j)}) = 1$. For the other three components, $j \in \{4, 5, 6\}$, $X^{(j)}$ was generated from a discrete uniform distribution with j levels and for convenience the levels were labelled with the first j captalized letters. For example, $X^{(4)}$ took on the value of "A", "B", "C", or "D". The treatment indicator A was simulated from a Bernoulli(p) distribution, where $p = \text{expit}\left\{0.3X^{(2)} - 0.3X^{(3)} + 0.3I(X^{(6)} \in \{\text{"B", "C"}\})\right\}$. As in the simulations in the main manuscript, the outcome was simulated from two settings, one with a homogeneous treatment

effect and one with a heterogeneous treatment effect.

- In the homogeneous treatment effect setting, the outcome Y was simulated from a Bernoulli distribution with $\mathrm{P}(Y = 1|\boldsymbol{X}, A) = 0.15 + 0.1A + \text{expit}\left(0.2X^{(2)}\right) - 0.4I(X^{(4)} \in \{\text{“B”, “D”}\})$. For this setting, the treatment effect is the same for all covariate values and the correct tree consists only of the root node.

- In the heterogeneous treatment effect setting, the outcome Y was simulated from a Bernoulli distribution with $\mathrm{P}(Y = 1|\boldsymbol{X}, A) = 0.1 + 0.1A + \text{expit}\left(0.2X^{(2)}\right) - 0.4AI(X^{(4)} \in \{\text{“B”, “D”}\})$. For this setting, the treatment effect differs depending on whether $X^{(4)} \in \{\text{“B”, “D”}\}$ or not. The correct tree splits the dataset into $X^{(4)} \in \{\text{“B”, “D”}\}$ and $X^{(4)} \in \{\text{“A”, “C”}\}$ groups.

As in the main manuscript, we implemented three versions of each tree-based algorithm. For the correct model specification, the correct logistic regression model for estimating the conditional probability of treatment in $\widehat{\mu}_{\text{IPW},a}(w)$, $\widehat{\mu}_{\text{DR},a}(w)$ and the `weights` in CT algorithms includes main effects of $X^{(2)}, X^{(3)}$ and $I(X^{(6)} \in \{\text{“B”, “C”}\})$. The correct logistic regression model for estimating the conditional expectation of the outcome in $\widehat{\mu}_{\text{g},a}(w)$, $\widehat{\mu}_{\text{DR},a}(w)$ includes $A, X^{(2)}$ and $I(X^{(4)} \in \{\text{“B”, “D”}\})$ for the homogeneous setting and an interaction between A and $I(X^{(4)} \in \{\text{“B”, “D”}\})$ instead of $I(X^{(4)} \in \{\text{“B”, “D”}\})$ for the heterogeneous setting. The current implementation of the CT algorithms can only be implemented treating categorical covariates as ordinal.

For the version with misspecified functional forms of covariates, the model for the conditional probability of treatment includes exponentiated form of all continuous covariates and dummy coding of all categorical variables; the model for the conditional expectation of the outcome includes main effects of treatment and all covariates and all two way treatment-covariate interactions in the original form of all continuous variables and all dummy coded categorical variables.

For the version that had an unmeasured common cause of the outcome and treatment assignment, we exclude $X^{(2)}$ from the dataset. The model for the conditional probability of treatment includes main effects of all the remaining covariates; the model for the conditional expectation of the outcome includes main effect of treatment and all the remaining covariates and all the remaining two-way treatment-covariate interactions.

Optimized CT in this simulation setting is when we set `split.Rule` to `"tstats"` without honest splitting (`split.Honest = FALSE`) and `cv.option` to `"fit"` without honest cross-validation

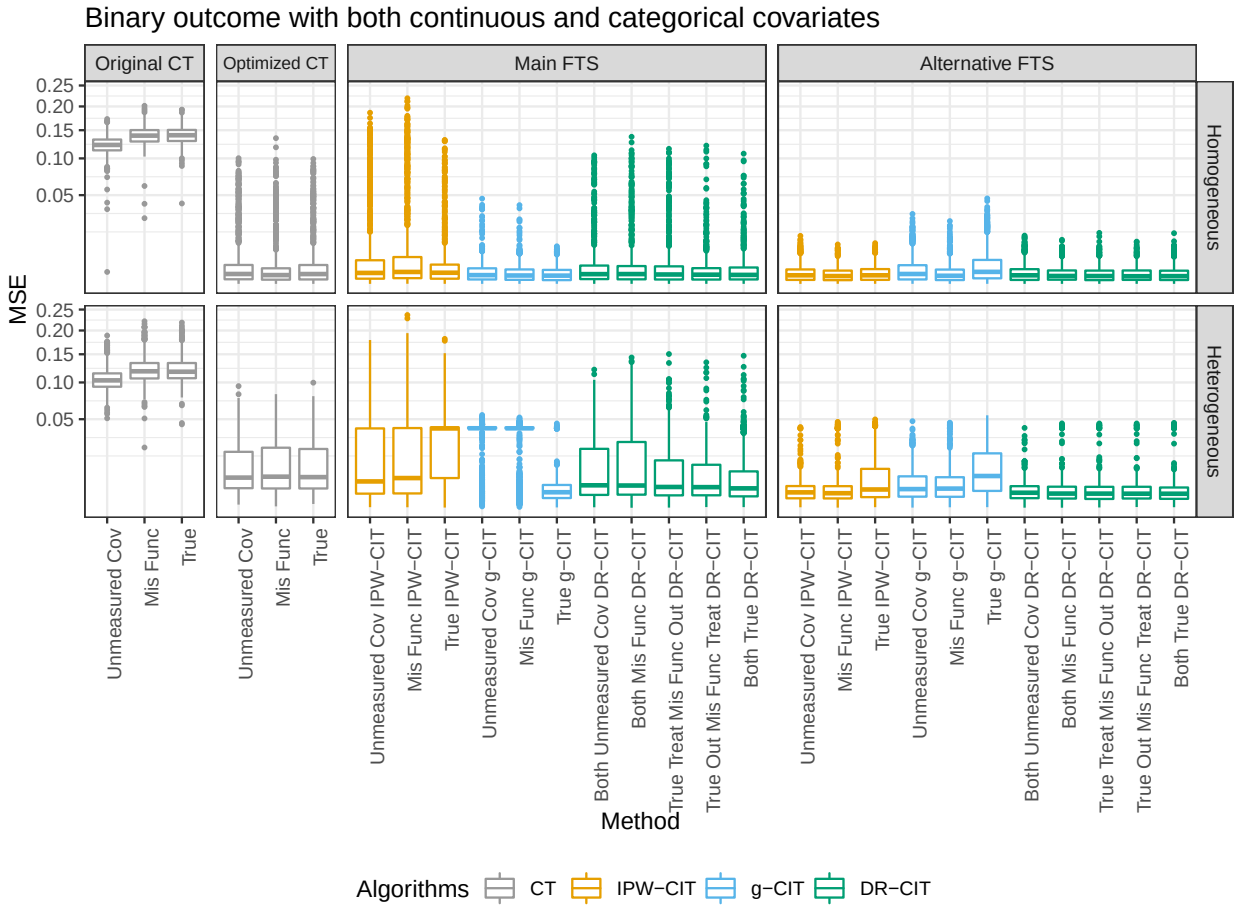

Figure B.7: Mean squared error (MSE) for the different tree building algorithms for the homogeneous (top) and the heterogeneous (bottom) settings described in Web Appendix B.3.7 when the outcome is binary and the covariate vector includes both continuous and categorical variables. Lower values indicate better performance. Original CT refers to the original Causal Tree algorithm by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

Table B.7: Proportion of correct trees (higher is better), average number of noise variables used for splitting (lower is better), pairwise prediction similarity (higher is better), the average time in seconds it takes to implement the methods (lower is better), and proportion of trees making the first split correctly (only for heterogeneous setting, higher is better) corresponding to algorithms in Figure B.7 when the outcome is binary and the covariate vector includes both continuous and categorical variables. Columns 3, 4, 5, and 6 show simulation results when the treatment effect is homogeneous, and columns 7, 8, 9, 10, and 11 show simulation results when the treatment effect is heterogeneous. Original CT refers to the original Causal Tree algorithms by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Web Appendix B.3.7, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

	Homogeneous Effect				Heterogeneous Effect				
Algorithm	Correct Trees	Number Noise	PPS	Time	Correct Trees	Number Noise	PPS	Time	Correct First Split
CT									
Original CT									
Unmeasured Cov	0.00	29.27	0.04	0.31	0.00	25.96	0.53	0.31	0.00
Mis Func	0.00	29.64	0.04	0.32	0.00	26.34	0.53	0.31	0.00
True	0.00	29.93	0.04	0.32	0.00	26.57	0.53	0.30	0.00
Optimized CT									
Unmeasured Cov	0.85	1.30	0.90	0.30	0.00	0.61	0.72	0.32	0.00
Mis Func	0.89	0.78	0.92	0.30	0.00	0.59	0.71	0.31	0.00
True	0.87	0.79	0.92	0.29	0.00	0.56	0.71	0.30	0.00
Main FTS									
IPW-CIT									
Unmeasured Cov	0.81	1.89	0.87	61.06	0.61	1.89	0.86	51.11	0.98
Mis Func	0.78	2.12	0.85	99.38	0.55	2.10	0.84	81.10	0.95
True	0.90	0.81	0.94	46.60	0.26	0.83	0.67	40.93	0.72
g-CIT									
Unmeasured Cov	0.98	0.22	0.99	130.35	0.18	0.18	0.60	117.64	0.90
Mis Func	0.99	0.12	0.99	208.01	0.19	0.11	0.60	183.34	0.91
True	1.00	0.00	1.00	98.86	0.97	0.00	0.99	99.62	1.00
DR-CIT									
Both Unmeasured Cov	0.90	0.64	0.94	12.68	0.69	0.61	0.89	12.05	0.98
Both Mis Func	0.90	0.71	0.94	15.85	0.68	0.90	0.89	14.87	0.97
True Treat Mis Func Out	0.90	0.70	0.94	15.16	0.72	0.60	0.91	14.21	0.98
True Out Mis Func Treat	0.94	0.25	0.97	13.03	0.74	0.30	0.90	12.20	0.97
Both True	0.94	0.25	0.97	12.55	0.76	0.38	0.92	11.81	0.98
Alternative FTS									
IPW-CIT									
Unmeasured Cov	0.97	0.03	0.99	102.54	0.96	0.05	0.99	87.18	0.99
Mis Func	0.98	0.02	0.99	160.59	0.93	0.04	0.98	132.16	0.99
True	1.00	0.00	1.00	79.71	0.70	0.18	0.90	71.38	0.82
g-CIT									
Unmeasured Cov	0.84	0.70	0.92	198.37	0.68	1.32	0.94	179.86	0.96
Mis Func	0.94	0.36	0.97	311.28	0.62	1.81	0.93	275.66	0.96
True	0.70	3.47	0.79	158.99	0.24	8.99	0.80	165.97	1.00
DR-CIT									
Both Unmeasured Cov	0.97	0.03	0.99	53.79	0.98	0.01	1.00	54.11	0.99
Both Mis Func	0.99	0.01	1.00	62.14	0.97	0.01	0.99	61.10	0.99
True Treat Mis Func Out	0.99	0.01	1.00	59.60	0.97	0.01	0.99	59.25	1.00
True Out Mis Func Treat	0.99	0.01	1.00	51.44	0.98	0.01	0.99	52.22	1.00
Both True	0.99	0.01	1.00	47.55	0.98	0.01	0.99	48.78	1.00

(`cv.Honest = FALSE`). Additionally, regular CTs (`causalTree`) give lower MSE than their honest peers (`honest.causalTree`) from simulations, so we present the results from regular CTs. Other implementation choices for the CT and CIT algorithms were as described in Section 3.4.3.

Figure B.7 shows boxplots of MSE and Table B.7 shows the proportion of correct trees, average number of noise variables, pairwise prediction similarity, the average running time for both simulation settings, and the proportion of trees making a correct first split in the heterogeneous setting (the dataset is split into $X^{(4)} \in \{\text{“B”, “D”}\}$ and $X^{(4)} \in \{\text{“A”, “C”}\}$ groups) when the outcome is binary and the covariate vector includes both continuous and categorical variables.

In general, the results follow the trends seen in the main simulations presented in Section 3.4.4. Additionally, in the heterogeneous setting, although CT algorithms split on fewer noise variables than some of the CIT algorithms, both the proportion of correct trees and of trees making a correct first split are 0. This is because the current CT algorithm implementation give a default order to the levels in categorical variables when building trees, so the trees cannot split at points that do not preserve the order. For example, in the heterogeneous setting, the CT algorithm implementation can only split on $X^{(4)}$ based on whether or not $X^{(4)} \in \{\text{“D”}\}$, $X^{(4)} \in \{\text{“C”, “D”}\}$, or $X^{(4)} \in \{\text{“B”, “C”, “D”}\}$ but cannot split based on whether or not $X^{(4)} \in \{\text{“B”, “D”}\}$. Therefore, the trees usually first split based on whether or not $X^{(4)} \in \{\text{“D”}\}$ and produce trees with 4 terminal nodes with each level of $X^{(4)}$ in one node in our simulation setting. When the data generation process of the heterogeneous setting is modified so that the treatment effect differs depending on whether $X^{(4)} \in \{\text{“C”, “D”}\}$, CT algorithms have the ability to identify the correct split point.

B.3.7.2 Evaluating data-splitting based confidence interval coverage

We use simulations to evaluate the coverage of confidence intervals that use data-splitting. The data is simulated as described in Web Appendix B.3.7. Each simulated dataset was split into two equal sized parts, the CIT algorithms were fit using the first part of the data and the terminal node treatment effect estimators and bootstrap based 95% confidence intervals were calculated using the second part of the data. Table B.8 shows the coverage probabilities for the trees that have the correct tree structure as for these trees the true treatment effect can easily be calculated. The results show coverage probabilities close to 95% for all settings with slightly lower than the expected 95% coverage when the model for the conditional probability of treatment and/or the

Table B.8: Coverage probabilities of 95% confidence intervals for the subgroup-specific treatment effect estimators when the correct tree structure is identified by the Causal Interaction Trees corresponding to algorithms in Figure B.7 and Table B.7 when the outcome is binary and the covariate vector includes both continuous and categorical variables. Column 3 shows the coverage probabilities when the treatment effect is homogeneous, and column 4 and 5 show the coverage probabilities for the two terminal nodes when the treatment effect is heterogeneous. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Web Appendix B.3.7, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

		Homogeneous Effect	Heterogeneous Effect	
Algorithm	Model	Root Node	$X^{(4)} \in \{$"A", "C"$\}$	$X^{(4)} \in \{$"B", "D"$\}$
Main FTS				
IPW-CIT	Unmeasured Cov	0.93	0.95	0.94
	Mis Func	0.95	0.95	0.96
	True	0.95	0.94	0.97
g-CIT	Unmeasured Cov	0.93	0.93	0.91
	Mis Func	0.95	0.93	0.92
	True	0.95	0.95	0.94
DR-CIT	Both Unmeasured Cov	0.92	0.94	0.94
	Both Mis Func	0.95	0.95	0.95
	True Treat Mis Func Out	0.95	0.95	0.95
	True Out Mis Func Treat	0.95	0.95	0.95
	Both True	0.95	0.96	0.95
Alternative FTS				
IPW-CIT	Unmeasured Cov	0.93	0.94	0.94
	Mis Func	0.95	0.96	0.95
	True	0.95	0.95	0.95
g-CIT	Unmeasured Cov	0.93	0.94	0.93
	Mis Func	0.95	0.96	0.95
	True	0.95	0.95	0.95
DR-CIT	Both Unmeasured Cov	0.93	0.95	0.94
	Both Mis Func	0.95	0.95	0.95
	True Treat Mis Func Out	0.95	0.95	0.95
	True Out Mis Func Treat	0.95	0.95	0.95
	Both True	0.95	0.96	0.95

model for the conditional expectation of the outcome are mis-specified or there is an unmeasured common cause of outcome and treatment assignment.

B.3.8 Simulations with lower signal-to-noise ratio and modified correlation structure

In this section, we present simulations with a lower signal-to-noise ratio and a a different covariate correlation structure compared to the simulation settings used in the main paper.

The covariate vector was simulated from a 6-dimensional mean zero multivariate normal distribution where $\text{cov}(X^{(j)}, X^{(k)}) = 0.6^{k-j}$ for $1 \leq j \leq k \leq 6$. The treatment indicator A was simulated from a Bernoulli(p) distribution, where $p = \text{expit}(0.6X^{(1)} - 0.6X^{(2)} + 0.6X^{(3)})$. As in the simulations in the main manuscript, the outcome was simulated using two settings, one with a homogeneous treatment effect and one with a heterogeneous treatment effect.

- In the homogeneous setting, $Y = 2+0.5A+2I(X^{(1)} < 0)+\exp(X^{(2)})+I(X^{(4)} > 0)+(X^{(5)})^3+\epsilon$, where $\epsilon \sim N(0, 1)$. For this setting, the treatment effect is the same for all covariate values and the correct tree consists only of the root node.
- In the heterogeneous setting, $Y = 2 + 0.5A + 2I(X^{(1)} < 0) + \exp(X^{(2)}) + AI(X^{(4)} > 0) + (X^{(5)})^3 + \epsilon$, where $\epsilon \sim N(0, 1)$. For this setting, the treatment effect differs depending on whether $X^{(4)} > 0$ or not and the correct tree splits on $X^{(4)}$ at 0.

For both simulation settings, the algorithms were fit on a training set generated by drawing 1000 independent samples from the joint distribution of $(\boldsymbol{X}, A, Y)$ and evaluated on a test set of size 1000. We implemented the tree-based algorithms using the same correct model specification and the two versions of model misspecification as described in Section 3.4.3.

Optimized CT in this simulation setting is when we set `split.Rule` to `"tstats"` with honest splitting (`split.Honest = TRUE`) and `cv.option` to `"CT"` without honest cross-validation (`cv.Honest = FALSE`). Additionally, for Original CTs, honest CTs (`honest.causalTree`) give lower MSE than their peers (`causalTree`) from simulations, and for Optimized CTs, regular CTs give lower MSE. Therefore, we present the results from honest CTs for Original CT and those from regular CTs for Optimized CT. Other implementation choices for the CT and CIT algorithms were as described in Section 3.4.3.

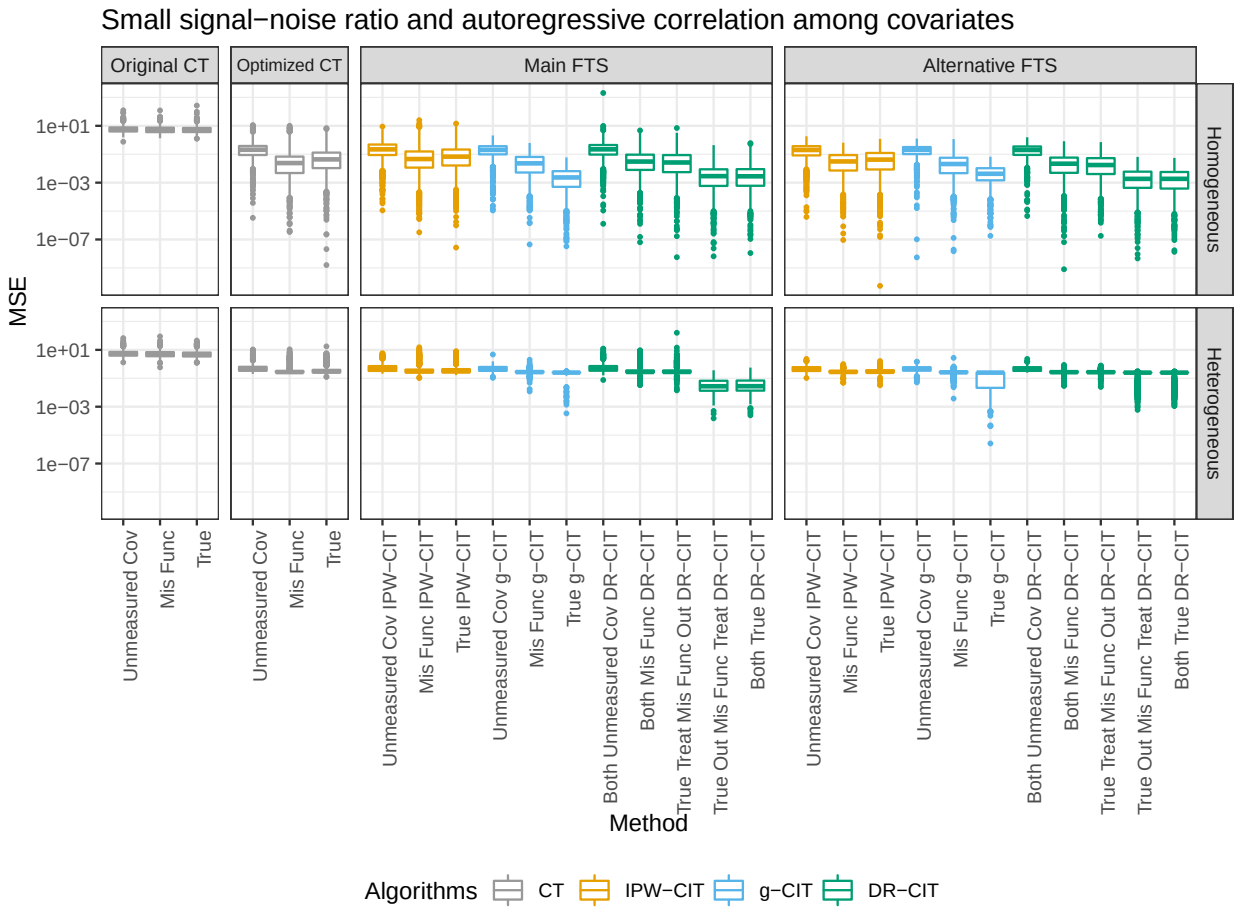

Figure B.8: Mean squared error (MSE) for the different tree building algorithms for the homogeneous (top) and the heterogeneous (bottom) settings described in Web Appendix B.3.8 when the signal-noise ratio is one and there is an autoregressive correlation structure among covariates. Lower values indicate better performance. Original CT refers to the original Causal Tree algorithm by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

Table B.9: Proportion of correct trees (higher is better), average number of noise variables used for splitting (lower is better), pairwise prediction similarity (higher is better), the average time in seconds it takes to implement the methods (lower is better), and proportion of trees making the first split correctly (only for heterogeneous setting, higher is better) corresponding to algorithms in Figure B.8 when the signal-noise ratio is one and there is an autoregressive correlation structure among covariates. Columns 3, 4, 5, and 6 show simulation results when the treatment effect is homogeneous, and columns 7, 8, 9, 10, and 11 show simulation results when the treatment effect is heterogeneous. Original CT refers to the original Causal Tree algorithms by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Web Appendix B.3.8, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

	Homogeneous Effect				Heterogeneous Effect				
Algorithm	Correct Trees	Number Noise	PPS	Time	Correct Trees	Number Noise	PPS	Time	Correct First Split
CT									
Original CT									
Unmeasured Cov	0.00	41.72	0.03	0.09	0.00	33.27	0.52	0.08	0.22
Mis Func	0.00	41.81	0.03	0.08	0.00	34.91	0.51	0.08	0.18
True	0.00	41.74	0.03	0.08	0.00	34.60	0.51	0.08	0.21
Optimized CT									
Unmeasured Cov	0.95	0.15	0.98	0.30	0.01	0.10	0.51	0.30	0.28
Mis Func	0.96	0.10	0.98	0.31	0.00	0.12	0.50	0.29	0.16
True	0.97	0.09	0.98	0.31	0.03	0.11	0.52	0.29	0.31
Main FTS									
IPW-CIT									
Unmeasured Cov	0.90	0.65	0.94	87.68	0.01	0.39	0.51	86.61	0.24
Mis Func	0.86	0.77	0.93	116.79	0.01	0.73	0.51	112.58	0.12
True	0.91	0.52	0.95	90.15	0.01	0.47	0.51	89.60	0.20
g-CIT									
Unmeasured Cov	0.95	0.15	0.97	169.76	0.03	0.20	0.52	172.17	0.40
Mis Func	0.96	0.10	0.97	206.63	0.03	0.15	0.52	209.39	0.38
True	1.00	0.00	1.00	163.36	0.02	0.00	0.51	80.65	1.00
DR-CIT									
Both Unmeasured Cov	0.90	0.66	0.94	13.18	0.02	0.54	0.51	12.89	0.35
Both Mis Func	0.92	0.56	0.95	16.81	0.06	0.47	0.54	15.46	0.49
True Treat Mis Func Out	0.92	0.49	0.96	16.23	0.06	0.50	0.54	15.04	0.51
True Out Mis Func Treat	0.95	0.14	0.98	17.60	0.81	0.15	0.90	15.18	1.00
Both True	0.94	0.17	0.98	17.52	0.80	0.19	0.90	15.17	1.00
Alternative FTS									
IPW-CIT									
Unmeasured Cov	0.99	0.01	1.00	137.50	0.01	0.01	0.50	136.68	0.26
Mis Func	0.99	0.01	1.00	179.44	0.01	0.02	0.50	173.93	0.11
True	0.98	0.04	0.99	140.73	0.02	0.03	0.51	139.39	0.22
g-CIT									
Unmeasured Cov	0.94	0.12	0.97	261.95	0.04	0.19	0.52	264.26	0.46
Mis Func	0.82	0.56	0.89	313.73	0.03	0.54	0.53	316.83	0.43
True	0.38	2.88	0.62	240.99	0.27	0.00	0.63	111.39	1.00
DR-CIT									
Both Unmeasured Cov	1.00	0.00	1.00	40.20	0.00	0.00	0.50	39.68	0.39
Both Mis Func	0.97	0.04	0.99	46.15	0.01	0.01	0.51	44.77	0.50
True Treat Mis Func Out	0.98	0.04	0.99	45.25	0.01	0.02	0.51	43.78	0.55
True Out Mis Func Treat	1.00	0.00	1.00	52.87	0.20	0.03	0.60	51.30	1.00
Both True	1.00	0.00	1.00	51.35	0.20	0.02	0.60	49.93	1.00

Figure B.8 shows boxplots of MSE and Table B.9 shows the proportion of correct trees, average number of noise variables, pairwise prediction similarity, the average running time for both simulation settings, and the proportion of trees making a correct first split for the heterogeneous setting.

The relative performance of the CIT and the CT algorithms follow the trends seen in Section 3.4.4 and Web Appendix B.3.4. Additionally, when compared to the performance on the dataset with a larger signal-noise ratio and exchangeable correlation structure among covariates (Table 3.1 and Table B.3), the relative performance of the CIT algorithms are slightly better in the homogeneous setting and becomes worse in the heterogeneous setting. g-CITs and DR-CITs perform better than both versions of the CT algorithms.

B.3.9 Simulations when the true subgroups do not follow a tree structure

In this section, we present simulation results when the true subgroups do not have a tree structure.

The covariate vector was simulated from a 6-dimensional mean zero multivariate normal distribution, where $\text{cov}(X^{(j)}, X^{(k)}) = 0.3$ when $j \neq k$, and $\text{Var}(X^{(j)}) = 1$. The treatment indicator A was simulated from a Bernoulli(p) distribution, where $p = \text{expit}\left(0.6X^{(1)} - 0.6X^{(2)} + 0.6X^{(3)}\right)$. The outcome was simulated from a homogeneous setting and a heterogeneous setting:

- In the homogeneous setting, $Y = 1+A+I\left(X^{(1)} < 0\right)+\exp\left(X^{(2)}\right)+X^{(4)}-\left(X^{(5)}\right)^3-\left(X^{(6)}\right)^2+\epsilon$, where $\epsilon \sim N(0,1)$. For this setting, the treatment effect is the same for all covariate values and the correct tree consists only of the root node.

- In the heterogeneous setting, $Y = 1 + A + I\left(X^{(1)} < 0\right) + \exp\left(X^{(2)}\right) + AX^{(4)} - \left(X^{(5)}\right)^3 - A\left(X^{(6)}\right)^2 + \epsilon$, where $\epsilon \sim \text{N}(0,1)$. For this setting, the treatment effect differs depending on the values of $X^{(4)}$ and $X^{(6)}$ and we expect the trees to make splits on these two covariates but there is no "correct" split point.

Of the evaluation measures listed in Section 3.4.2, we will not use the proportion of correct trees, pairwise prediction similarity, and proportion of correct first splits for the heterogeneous setting as the rely on the correct model being a tree. Instead we define two new evaluation measures:

- Size of trees: The average number of terminal nodes in the trees.

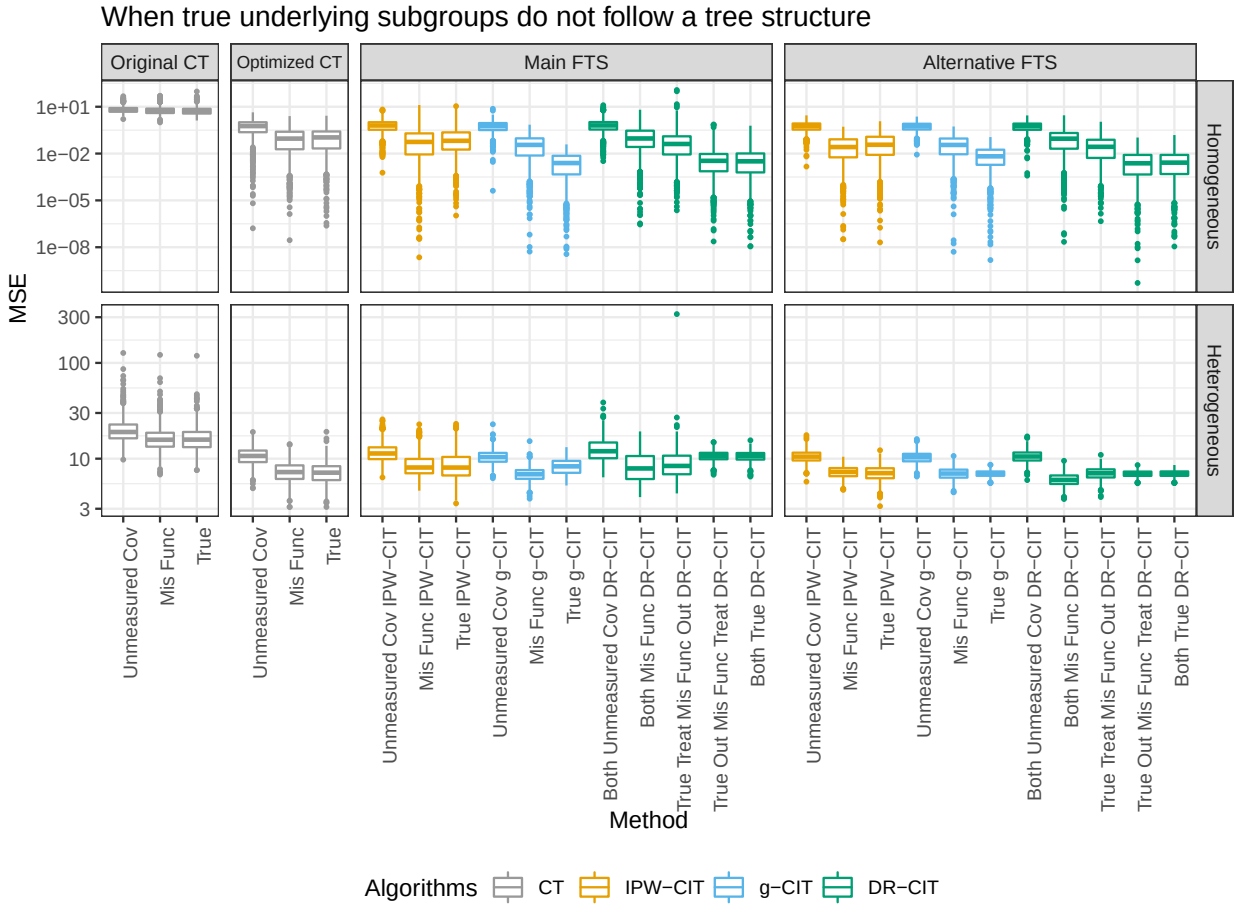

Figure B.9: Mean squared error (MSE) for the different tree building algorithms for the homogeneous (top) and the heterogeneous (bottom) settings described in Web Appendix B.3.9 when the true underlying subgroups do not follow a tree structure. Lower values indicate better performance. Original CT refers to the original Causal Tree algorithm by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

Table B.10: Proportion of correct trees (only for homogeneous setting, higher is better), average size of the trees, average number of noise variables used for splitting (lower is better), the average time in seconds it takes to implement the methods (lower is better), and average proportion of splits made on signal variables in the first 3 and 4 levels of the trees (only for heterogeneous setting, higher is better) corresponding to algorithms in Figure B.9 when the true underlying subgroups do not follow a tree structure. Columns 3, 4, 5, and 6 show simulation results when the treatment effect is homogeneous, and columns 7, 8, 9, 10 and 11 show simulation results when the treatment effect is heterogeneous. Original CT refers to the original Causal Tree algorithms by Athey and Imbens [2016]. Optimized CT refers to the Causal Tree algorithm with optimized tuning parameters. "Main FTS" refers to the Causal Interaction Tree algorithms described in the main manuscript. "Alternative FTS" refers to the final tree selection method described in B.2. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively. As described in Web Appendix B.3.9, "Unmeasured Cov" refers to having an unmeasured common cause of outcome and treatment assignment; "Mis Func" refers to using misspecified functional forms of covariates in the model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome; "True" refers to using a correctly specified model for the conditional probability of treatment and/or the model for the conditional expectation of the outcome. For DR-CITs, "Treat" and "Out" stand for the model for the conditional probability of treatment and the model for the conditional expectation of the outcome, respectively.

	Homogeneous Effect				Heterogeneous Effect				
								Correct Splits	
Algorithm	Correct Trees	Size Trees	Number Noise	Time	Size Trees	Number Noise	Time	Level 3	Level 4
CT									
Original CT									
Unmeasured Cov	0.00	42.12	41.12	0.08	42.02	22.85	0.08	0.54	0.51
Mis Func	0.00	41.57	40.57	0.08	41.48	24.88	0.08	0.49	0.48
True	0.00	41.45	40.45	0.08	41.70	24.81	0.08	0.49	0.47
Optimized CT									
Unmeasured Cov	1.00	1.00	0.00	0.06	1.09	0.04	0.06	0.02	0.02
Mis Func	0.99	1.03	0.03	0.06	1.11	0.04	0.06	0.03	0.03
True	1.00	1.01	0.01	0.06	1.06	0.02	0.06	0.03	0.03
Main FTS									
IPW-CIT									
Unmeasured Cov	0.90	1.55	0.56	89.62	2.36	0.46	94.14	0.30	0.30
Mis Func	0.81	1.90	0.90	120.34	2.62	0.80	124.97	0.28	0.27
True	0.89	1.64	0.64	93.43	2.18	0.43	97.74	0.35	0.35
g-CIT									
Unmeasured Cov	0.92	1.16	0.16	172.58	1.49	0.08	172.73	0.25	0.26
Mis Func	0.90	1.17	0.17	205.07	1.52	0.12	205.50	0.25	0.25
True	1.00	1.00	0.00	162.98	2.22	0.00	215.95	0.68	0.68
DR-CIT									
Both Unmeasured Cov	0.88	1.61	0.61	13.16	3.00	0.48	13.78	0.47	0.46
Both Mis Func	0.91	1.54	0.54	16.59	3.48	0.52	16.45	0.61	0.60
True Treat Mis Func Out	0.90	1.53	0.53	16.02	3.31	0.51	15.76	0.56	0.55
True Out Mis Func Treat	0.95	1.18	0.18	17.16	7.91	0.31	18.16	1.00	1.00
Both True	0.94	1.15	0.15	16.99	7.81	0.31	17.93	1.00	1.00
Alternative FTS									
IPW-CIT									
Unmeasured Cov	0.99	1.01	0.01	140.68	1.02	0.00	147.41	0.01	0.01
Mis Func	0.95	1.06	0.06	186.26	1.01	0.01	193.54	0.00	0.00
True	0.99	1.01	0.01	145.49	1.00	0.00	152.32	0.00	0.00
g-CIT									
Unmeasured Cov	0.93	1.11	0.11	261.74	1.39	0.08	264.58	0.22	0.22
Mis Func	0.69	1.73	0.73	308.49	1.02	0.01	309.55	0.00	0.00
True	0.40	4.27	3.27	242.95	1.00	0.00	323.30	0.00	0.00
DR-CIT									
Both Unmeasured Cov	1.00	1.00	0.00	40.45	1.02	0.00	42.31	0.02	0.02
Both Mis Func	0.92	1.18	0.18	46.33	1.00	0.00	48.57	0.00	0.00
True Treat Mis Func Out	0.97	1.04	0.04	45.32	1.00	0.00	46.88	0.00	0.00
True Out Mis Func Treat	0.96	1.05	0.05	52.11	1.00	0.00	58.08	0.00	0.00
Both True	0.96	1.06	0.06	50.11	1.00	0.00	55.93	0.00	0.00

- Proportion of splits made on signal variables in the first k levels of the trees (only applicable to the heterogeneous setting): The average proportion of splits that are made on signal variables (i.e., $\{X^{(4)}, X^{(6)}\}$) in the first k levels of the trees. For the simulations we use both $k = 3$ and $k = 4$.

As in the main manuscript, we implemented three versions of each tree-based algorithm. For the correct model specification, the correct linear regression model for the conditional expectation of the outcome in $\widehat{\mu}_{\mathrm{g},a}(w)$ and $\widehat{\mu}_{\mathrm{DR},a}(w)$ includes A, $I\left(X^{(1)} < 0\right)$, $\exp\left(X^{(2)}\right)$, $X^{(4)}$, $\left(X^{(5)}\right)^3$ and $\left(X^{(6)}\right)^2$ for the homogeneous setting and interactions between A and $X^{(4)}$, A and $\left(X^{(6)}\right)^2$ instead of the corresponding main effect terms for the heterogeneous setting. The correct logistic regression model for estimating the conditional probability of treatment in $\widehat{\mu}_{\mathrm{IPW},a}(w)$ and $\widehat{\mu}_{\mathrm{DR},a}(w)$ and the other two versions of the tree-based algorithms were implemented as described in Section 3.4.3.

Optimized CT in this simulation setting is when we set `split.Rule` to `"tstats"` with honest splitting (`split.Honest = TRUE`) and `cv.option` to `"CT"` without honest cross-validation (`cv.Honest = FALSE`). Additionally, honest CTs (`honest.causalTree`) give lower MSE than their peers (`causalTree`), so we present the results from honest CTs. Other implementation choices for the CT and CIT algorithms were as described in Section 3.4.3.

Figure B.9 shows boxplots of MSE and Table B.10 shows the proportion of correct trees for the homogeneous setting, the average size of the trees, average number of noise variables, average running time for both simulation settings, and average proportion of splits made on signal variables in the first 3 and 4 levels of the trees for the heterogeneous setting when the true underlying subgroups do not follow a tree structure.

The performance of the CIT algorithms using the final tree selection method described in the main manuscript follows the trends seen in Section 3.4.4. The performance of CIT algorithms using the final tree selection method described in B.2 is similar to that seen in Web Appendix B.3.4 for the homogeneous setting, and becomes worse for the heterogeneous setting.

The CIT algorithms using the final tree selection method described in the main manuscript are able to identify the signal variables that cause treatment effect heterogeneity in the heterogeneous setting and produce root-only trees in the homogeneous setting. The CIT algorithms also outperform the CT algorithms where the Original CTs overfit the data and the Optimized CTs fails to

identify any signal in the heterogeneous setting.

B.4 Additional Information on Analysis of SUPPORT Dataset

In this section we provide additional analysis of the SUPPORT data.

B.4.1 Covariates included in analysis

The analysis of the SUPPORT dataset included the following covariates: Age, sex, race, years of education, income, type of medical insurance, primary disease category, secondary disease category, Duke Activity Status Index (DASI), do-not-resuscitate (DNR) status on day 1, cancer status, SUPPORT model estimate of the probability of surviving 2 months, APACHE score, Glasgow Coma Score, weight, temperature, mean blood pressure, respiratory rate, heart rate, PaO_2/FIO_2 ratio, $PaCO_2$, pH, white blood cell count, hematocrit, sodium, potassium, creatinine, bilirubin, albumin, 10 categories of admission diagnosis, and 12 categories of comorbidities illness. Additionally, following Hirano and Imbens [2001], we a) do not use activities of daily living scale and urin output in the analysis due to large amount of missing data (75% and 53% missingness, respectively); and, b) create an additional indicator variable denoting if a participant's weight is recorded as 0 or not (there are 515 individuals with weight equal to 0).

B.4.2 Stability of the Causal Interaction Tree algorithms

We used subsampling to evaluate how stable the trees produced by the CIT algorithms are to variation in the data used to build them. We create 1000 random samples by sampling 2868 observations (half of the data) without replacement from the SUPPORT data. For each dataset, we apply the CIT algorithms and record the size of the tree. Figure B.10 presents the distribution of the size of trees given by the three CIT algorithms. The results show that the IPW-CIT, g-CIT, and DR-CIT algorithms produce root-only trees around 82.8%, 99.7%, and 87.9% of the times respectively.

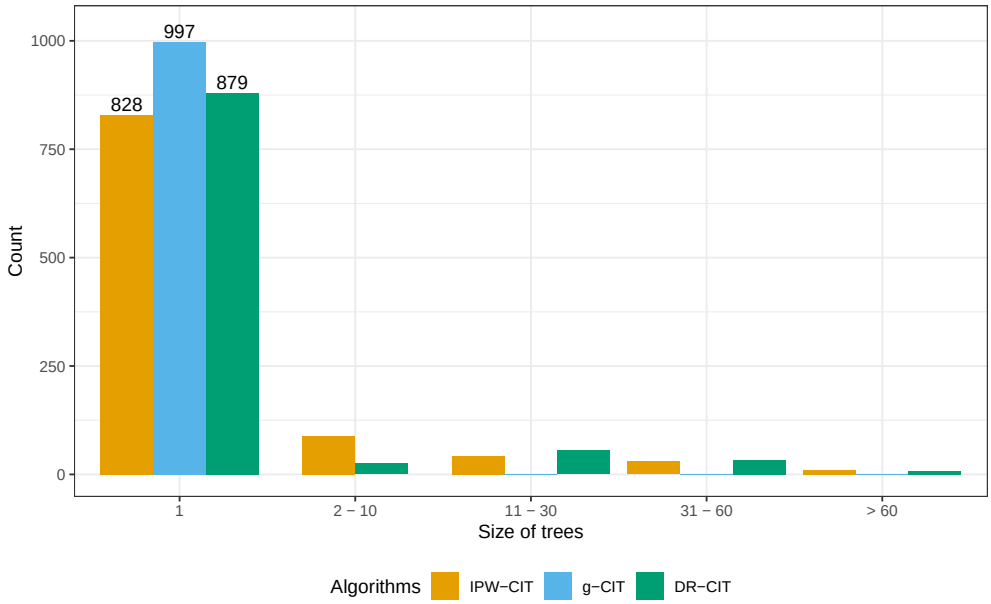

Figure B.10: Distribution of the size of trees produced by the Causal Interaction Tree algorithms across 1000 subsamples of the SUPPORT data. The subsamples were created by randomly sampling without replacement half of the observations from the SUPPORT data. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively.

B.4.3 Data-splitting confidence intervals for the SUPPORT data analysis

Bootstrap intervals that are calculated using the same data as was used for tree building do not produce valid confidence intervals. Data-splitting can be used to create asymptotically valid confidence intervals [Fithian et al., 2014] and we implemented data-splitting for the CIT algorithms. For data-splitting, half of the data was used for fitting the CIT algorithms and the other half was used for treatment effect estimation and confidence interval construction. We randomly split the SUPPORT data 1000 times. For the DR-CITs that only produce a root node, the average treatment effect is 0.067 and the average lower and upper limits of the 95% bootstrap confidence intervals are 0.027 and 0.108 respectively. Figure B.11 shows histograms of the average treatment effects and lower and upper limits of the 95% confidence intervals across different data splits from the three CIT algorithms.

The large tree $\psi_{\max}$ built by DR-CITs first splits on if the SUPPORT model estimate of the probability of surviving 2 months at study entry is greater than or equal to 0.85 or not. The average subgroup treatment effect estimator for those with a survival probability greater than 0.85

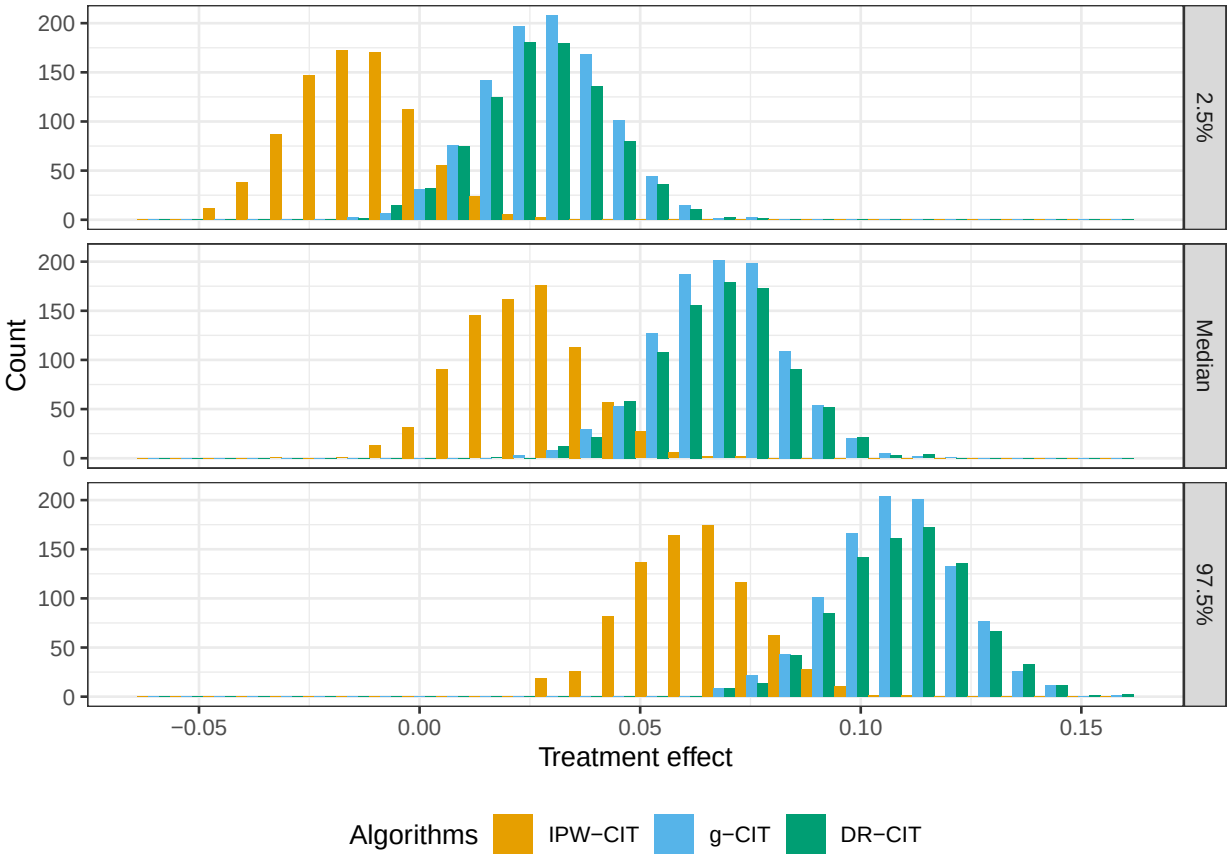

Figure B.11: Histograms of the average treatment effects and lower and upper limits of the 95% confidence intervals estimated by the Causal Interaction Tree algorithms across different subsamples of the SUPPORT data. The subsamples were created by randomly sampling without replacement half of the observations from the SUPPORT data. "Median" refers to the average treatment effect; "2.5%" and "97.5%" refers to the lower and upper limits respectively. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively.

was -0.007 and the average lower and upper limits of the 95% bootstrap confidence intervals were -0.090 and 0.081 respectively; for those with the survival probability lower than 0.85, the estimated average treatment effect was 0.066 with the average lower and upper limits being 0.022 and 0.108. The treatment effect difference between the two subgroups was estimated to be 0.076 with the average lower and upper limits being -0.014 and 0.169, respectively. Figure B.12 shows histograms of the average treatment effects and lower and upper limits of the 95% confidence intervals in the two subgroups.

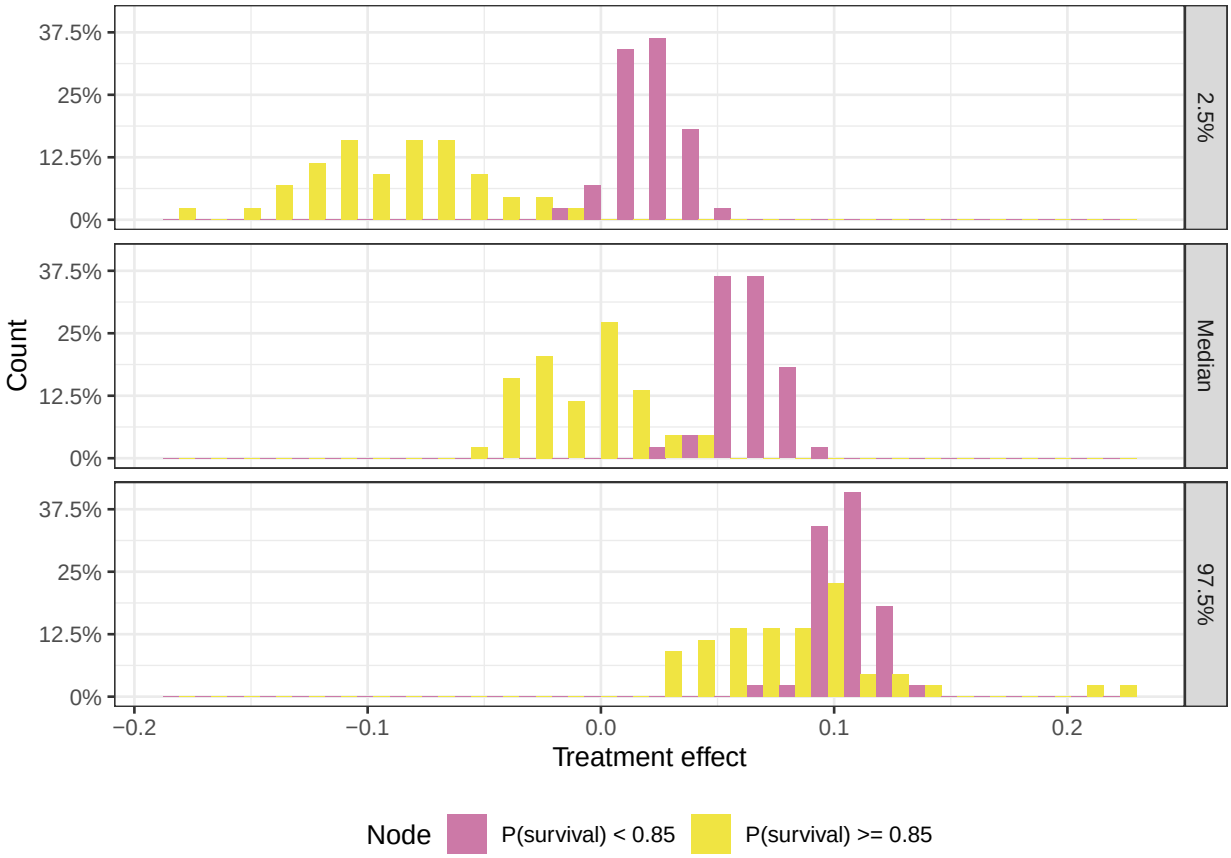

Figure B.12: Histograms of the average treatment effects and lower and upper limits of the 95% confidence intervals in the two subgroups produced by the Doubly Robust Causal Interaction Tree algorithms where the large tree $\psi_{\max}$ first splits on if the SUPPORT model estimate of the probability of surviving 2 months at study entry is greater than or equal to 0.85 across different subsamples of the SUPPORT data. The subsamples were created by randomly sampling without replacement half of the observations from the SUPPORT data. "Median" refers to the average treatment effect; "2.5%" and "97.5%" refers to the lower and upper limits respectively. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively.

B.4.4 Simulations based on SUPPORT study

In this section we use simulations from the SUPPORT data to evaluate the performance of the CIT algorithms in the presence of treatment effect heterogeneity. The analysis includes the covariates that are most highly associated with either the outcome or the treatment indicator. We used step-wise regression to select the five covariates that are most associated with the outcome. The selected covariates were primary disease category, secondary disease category, do-not-resuscitate (DNR) status on day 1, SUPPORT model estimate of the probability of surviving 2 months, and bilirubin and we denote the set of these covariates as $\boldsymbol{X}^{(Y)}$. Similarly, the five covariates that have the strongest association with the treatment indicator using step-wise regression are primary disease category, secondary disease category, mean blood pressure, respiratory rate, and PaO_2/FIO_2 ratio and we denote the set of these covariates as $\boldsymbol{X}^{(A)}$. The covariates included in the analysis are the covariates selected by at least one of the step-wise regression and the variable age is also included. We denote the set of covariates that are included in the analysis as $\boldsymbol{X}$.

Define $\widehat{\boldsymbol{\beta}}^{(A)}$ as the estimated coefficients from a logistic regression model regressing the treatment indicator A on $\boldsymbol{X}^{(A)}$ and let $\widehat{\boldsymbol{\beta}}^{(Y)}$ and $\widehat{\beta}_A$ be the estimated coefficients associated with $\boldsymbol{X}^{(Y)}$ and A respectively in the linear regression model regressing the outcome Y on $\boldsymbol{X}^{(Y)}$ and the treatment indicator A. Table B.11 presents the coefficients $\widehat{\boldsymbol{\beta}}^{(A)}$ and $\widehat{\boldsymbol{\beta}}^{(Y)}$ that are used for the simulations. The treatment effect $\widehat{\beta}_A$ is estimated to be 0.07.

We simulated the treatment indicator $A^{(\text{sim})}$ from a Bernoulli($\widehat{p}$) distribution with $\widehat{p} = \text{expit}\left\{(\boldsymbol{X}^{(A)})^T\widehat{\boldsymbol{\beta}}^{(A)}\right\}$. The outcome was simulated using two different simulation settings, one with a homogeneous treatment effect and one with a heterogeneous treatment effect.

- In the homogeneous treatment effect setting, the outcome $Y^{(\text{sim})}$ is simulated from a Bernoulli distribution with $P\left(Y^{(\text{sim})} = 1|\boldsymbol{X}, A^{(\text{sim})}\right) = f\left(\widehat{p}_{\text{unscaled}}, 0.05, 0.95\right)$, with $\widehat{p}_{\text{unscaled}} = (\boldsymbol{X}^{(Y)})^T\widehat{\boldsymbol{\beta}}^{(Y)} + A^{(\text{sim})}\widehat{\beta}_A$ and $f(p^*, p_{\min}, p_{\max}) = \frac{\{p^* - \min(p^*)\}(p_{\max} - p_{\min})}{\max(p^*) - \min(p^*)} + p_{\min}$. The function $f(p^*, p_{\min}, p_{\max})$ scales the range of the simulated probabilities to $[p_{\min}, p_{\max}]$. For this setting, the treatment effect is $\frac{\widehat{\beta}_A(0.95-0.05)}{\max(\widehat{p}_{\text{unscaled}}) - \min(\widehat{p}_{\text{unscaled}})}$ for all covariate values and the correct tree consists only of the root node.

- In the heterogeneous treatment effect setting, the outcome $Y^{(\text{sim})}$ is simulated from a Bernoulli distribution with $P\left(Y^{(\text{sim})} = 1|\boldsymbol{X}, A^{(\text{sim})}\right) = f\left(\widehat{p}_{\text{unscaled}}, 0.05, 0.55\right) + 0.4A^{(\text{sim})}I\big($primary dis-

Table B.11: Estimated coefficients $\widehat{\boldsymbol{\beta}}^{(A)}$ and $\widehat{\boldsymbol{\beta}}^{(Y)}$ that are used to simulate the treatment indicator $A^{(\text{sim})}$ and the outcome $Y^{(\text{sim})}$ respectively. Standard errors are presented in the parentheses. ARF, CHF, COPD, MOSF refers to acute respiratory failure, congestive heart failure, chronic obstructive pulmonary disease and multiorgan system failure, respectively.

Covariates	$\widehat{\boldsymbol{\beta}}^{(A)}$ (s.e.)	$\widehat{\boldsymbol{\beta}}^{(Y)}$ (s.e.)
Primary disease category		
ARF (reference level)	-	-
CHF	0.98 (0.12)	-0.02 (0.02)
Cirrhosis	-0.43 (0.18)	0.08 (0.03)
Colon cancer	-0.73 (1.09)	-0.33 (0.16)
Nontraumatic coma	-0.49 (0.14)	0.20 (0.02)
COPD	-1.26 (0.15)	-0.02 (0.02)
Lung cancer	-0.90 (0.49)	0.05 (0.07)
MOSF with malignancy	0.02 (0.12)	0.06 (0.03)
MOSF with sepsis	1.11 (0.08)	0.02 (0.02)
Secondary disease category		
Cirrhosis (reference level)	-	-
Colon cancer	1.12 (1.46)	-0.23 (0.31)
Nontraumatic coma	-0.57 (0.46)	-0.24 (0.08)
Lung cancer	-0.91 (0.86)	-0.05 (0.13)
MOSF with malignancy	-0.24 (0.40)	-0.27 (0.08)
MOSF with sepsis	0.87 (0.38)	-0.35 (0.07)
Missing secondary disease category	0.14 (0.37)	-0.37 (0.07)
Do-not-resuscitate (DNR) status on day 1	-	0.21 (0.02)
SUPPORT model estimate of the probability of surviving 2 months	-	-0.64 (0.04)
Mean blood pressure	-0.01 (0.00)	-
Respiratory rate	-0.02 (0.00)	-
PaO_2/FIO_2 ratio	0.00 (0.00)	-
Bilirubin	-	0.01 (0.00)

ease category $\in$ {acute respiratory failure (ARF), multiorgan system failure (MOSF) with malignancy or sepsis}). For this setting, the treatment effect is $\frac{\widehat{\beta}_A(0.55-0.05)}{\max(\widehat{p}_{\text{unscaled}})-\min(\widehat{p}_{\text{unscaled}})}+0.4$ for participants with primary disease category falling into acute respiratory failure or multiorgan system failure with malignancy or sepsis and the treatment effect is $\frac{\widehat{\beta}_A(0.55-0.05)}{\max(\widehat{p}_{\text{unscaled}})-\min(\widehat{p}_{\text{unscaled}})}$ for the remaining participants. The correct tree splits the dataset on primary disease category depending on whether participants have primary disease as either acute respiratory failure or multiorgan system failure with either malignancy or sepsis.

For both simulation settings, we simulated 1000 dataset and used subsampling to evaluate the performance of the CIT algorithms (to mimic the procedure used for the actual SUPPORT analysis). The subsamples are created by randomly sampling half of the observations (2868 observations) without replacement. We fit the CIT algorithms on the first part of the data using the implementation choices as described in Section 3.5.

Table B.12: Proportion of correct trees (higher is better), average number of noise variables used for splitting (lower is better), pairwise prediction similarity (higher is better), proportion of trees making the correct first split (only for heterogeneous setting, higher is better) and proportion of trees making the first split on the correct variable (only for heterogeneous setting, higher is better) for the Causal Interaction Trees for the homogeneous and the heterogeneous settings described in Web Appendix B.4.4.

Setting	Algorithm	Correct Trees	Number Noise	PPS	Correct First Split	Correct First Variable
Homogeneous	IPW-CIT	0.847	2.595	0.917	-	-
	g-CIT	0.919	0.240	0.994	-	-
	DR-CIT	0.808	3.796	0.869	-	-
Heterogeneous	IPW-CIT	0.331	2.751	0.878	0.494	0.946
	g-CIT	0	0.379	0.598	0.004	0.776
	DR-CIT	0.144	5.216	0.807	0.309	0.941

Figure B.13 shows boxplots of MSE and distribution of the size of trees across the 1000 datasets and Table B.12 shows the proportion of correct trees, average number of noise variables, pairwise prediction similarity for both simulation settings, proportion of trees making the correct first split and proportion of trees making the first split on the correct variable in the heterogeneous setting for the CIT algorithms. The results demonstrate the differential performance of the CIT algorithms in the homogeneous and the heterogeneous treatment effect setting. In the homogeneous setting, the CIT algorithms produce the correct root-only trees more than 80% of the times. In the het-

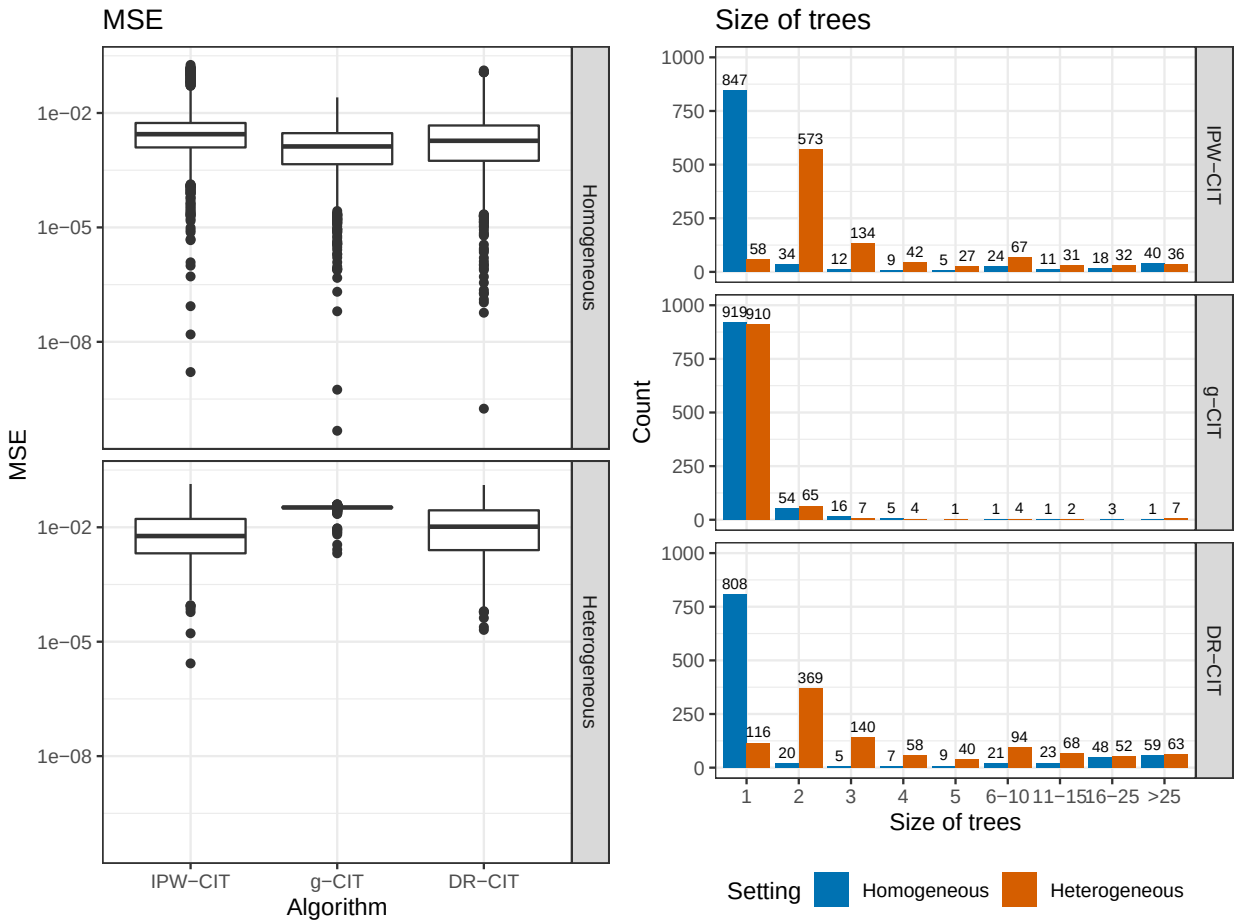

Figure B.13: Mean squared error (MSE, left) and distribution of the size of trees (right) for the Causal Interaction Tree algorithms for the homogeneous and the heterogeneous settings described in Web Appendix B.4.4. IPW-CIT, g-CIT and DR-CIT refer to the inverse probability weighting, g-formula, and doubly robust Causal Interaction Tree algorithms, respectively.

erogeneous setting, the IPW-CIT, g-CIT and DR-CIT split first on the correct variable 95%, 78% and 94% of the times, respectively. The variable that defines the correct tree (primary disease) is nominal with 9 levels with only 7 participants falling in the colon cancer category making it hard to identify exactly the correct split on this variable.

Appendix C

Appendix for Chapter 4

C.1 Pruning and final tree selection for the LISA algorithm

Now we give a precise definition of the pruning and final tree selection for the LISA algorithm. The initial tree building process results in a large tree that usually substantially overfits the data. The goal of the pruning step is to create a sequence of subtrees of the initial tree that are candidates for being the final model selected. This reduces the computational complexity compared to looking at all possible subtrees.

The pruning algorithm we use is an extension of weakest link pruning Breiman et al. [1984] to the setting of causal treatment effect estimation Su et al. [2008], Yang et al. [2021]. The algorithm builds a sequence of subtrees of $\widehat{\psi}_{\max}$ by sequentially cutting the "weakest link" defined as the branch of the tree with the smallest split complexity.

For a non-terminal node k define the branch that consists of the node k and all its descendants as ψ_k^*. The split complexity of the subtree ψ_k^* is zero at

$$\lambda^* = \frac{\sum_{i \in S_{\psi_k^*}} G_i(\psi_k^*)}{|S_{\psi_k^*}|}.$$

If $\lambda > \lambda*$ removing the branch ψ_k^* is preferred if the objective is to maximize split complexity and if $\lambda < \lambda*$ keeping keeping branch ψ_k^* is preferred if the objective is to maximize split complexity. Weakest link pruning is given by the following algorithm

1. Initialize the algorithm by setting $\widehat{\psi}_0 = \widehat{\psi}_{\max}$ and set $d = 0$.

2. Set $d = d + 1$.

3. Define a function

$$g(h) = \begin{cases} \frac{\sum_{i \in S_{\psi_h^*}} G_i(\psi_h^*)}{S_{\psi_h^*}|} & \text{if } h \in S_{\psi_h^*} \\ \infty & \text{otherwise.} \end{cases}$$

The weakest branch of the tree is the branch rooted in node $h = \min_{h' \in |S_{\widehat{\psi}^*_{m-1}}|} g(h')$. Define the subtree $\widehat{\psi}_d$ as the subtree of $\widehat{\psi}_{d-1}$ with all descendants of node h removed.

4. Repeat steps 2 and 3 until the tree $\widehat{\psi}_d$ consists only of the root node.

The weakest link pruning algorithm creates a finite sequence of trees $\widehat{\psi}_0, \ldots, \widehat{\psi}_{\widehat{D}}$. The final tree selection step selects a single tree from that algorithm.

C.2 Consistency of the longitudinal subgroup identification algorithms

In this section, we show that the longitudinal subgroup identification algorithms give estimators that are L_2 consistent.

Recall that $\boldsymbol{\mathcal{L}}_0$ is the sample space of $\boldsymbol{L}_0$ and further let $\boldsymbol{\mathcal{L}}_k$ be the sample space of $\boldsymbol{L}_k$, $k = 1, \ldots, K$; assume that the outcome of interest at visit $K+1$ is bounded, $|Y_{K+1}| \leq B_1 < \infty$; define the sample space of the observed data as $\mathcal{O} = \boldsymbol{\mathcal{L}}_0 \times \{0,1\} \times \{0,1\} \times \boldsymbol{\mathcal{L}}_1 \times \cdots \times [-B_1, B_1]$. Borrowing notation from Nobel et al. [1996], let π be any partition of $\boldsymbol{\mathcal{L}}_0$ and define Π as the range of π; let $\hat{\psi}_n$ be an empirical rule (tree) with $\hat{K}_n$ partitions (nodes) and define Π_n as the range of $\hat{\psi}_n$. For each element $\pi = \{u_k\} \in \Pi$, define an associated partition $\widetilde{\pi} = \{u_k \times \{0,1\} \times \cdots \times [-B_1, B_1]\}$ on $\mathcal{O}$ and let $\widetilde{\Pi}$ be the range of $\widetilde{\pi}$. Analogously, for each element $\pi_n = \{w_k\} \in \Pi_n$, define an associated partition $\widetilde{\pi}_n = \{w_k \times \{0,1\} \times \cdots \times [-B_1, B_1]\}$ on $\mathcal{O}$ and let $\widetilde{\Pi}_n$ be the range of $\widetilde{\pi}_n$. Finally, let $w_\pi(\boldsymbol{l}_0)$ be the unique partition (terminal node) in partition rule π that contains the covariate vector $\boldsymbol{l}_0$.

Define a transformed outcome $Y^*_{w_{\pi_n}(\boldsymbol{L}_0)}\{\boldsymbol{O}; \boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{L}_0)}\}$ where $\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{L}_0)}$ is a parameter vector that includes the parameters indexing the probability of treatment, probability of censoring and conditional expectation of outcome models estimated within terminal node $w_{\pi_n}(\boldsymbol{L}_0)$; for simplicity,

we denote the transformed outcome as $Y^*_{w_{\pi_n}(\boldsymbol{L}_0)}\left(\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{L_0})}\right)$. Similar to the proof in Yang et al. [2021], we write our estimator for $r(\boldsymbol{l}_0) = \mathrm{E}[Y^{\mathbf{1},\boldsymbol{c}=\mathbf{0}}_{K+1}|\boldsymbol{L}_0 = \boldsymbol{l}_0] - \mathrm{E}[Y^{\mathbf{0},\boldsymbol{c}=\mathbf{0}}_{K+1}|\boldsymbol{L}_0 = \boldsymbol{l}_0]$ as

$$\widehat{r}_n(\boldsymbol{l}_0) = \frac{\sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w_{\widehat{\psi}_n}(\boldsymbol{l_0})\} Y^*_{i,w_{\widehat{\psi}_n}(\boldsymbol{l_0})}\left(\widehat{\boldsymbol{\theta}}_{w_{\widehat{\psi}_n}(\boldsymbol{l_0})}\right)}{\sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w_{\widehat{\psi}_n}(\boldsymbol{l_0})\}},$$

The following theorem shows that the longitudinal subgroup identification algorithms are L_2 consistent.

Theorem C.2.1 *Assuming:*

B.1 Y_{K+1} *is bounded such that* $|Y_{K+1}| \leq B_1 < \infty$.

B.2 The number of terminal nodes $\widehat{K}_n \to \infty$ *and* $\widehat{K}_n = o\left\{\frac{n}{\log(n)}\right\}$.

B.3 The derivative $\left.\frac{\partial Y^*_{w_{\pi_n}(\boldsymbol{L}_0)}(\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{L}_0)})}{\partial \boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{L}_0)}}\right|_{\boldsymbol{\theta}_{w_{\pi_n}(\boldsymbol{L}_0)}=\boldsymbol{\theta}^*_{w_{\pi_n}(\boldsymbol{L}_0)}}$ *exists for* $\pi_n \in \Pi_n$ *and is uniformly bounded, where* $\boldsymbol{\theta}^*_{w_{\pi_n}(\boldsymbol{L}_0)}$ *is the limit of* $\widehat{\boldsymbol{\theta}}_{w_{\pi_n}(\boldsymbol{L}_0)}$.

B.4 For given $\boldsymbol{a}$ *and* $\boldsymbol{c}$, *any* $\widetilde{\pi} = \{\widetilde{u}_k\} \in \widetilde{\Pi}$ *and* $\boldsymbol{o} \in \widetilde{u}_k$, *either the iterated conditional expectation of the observed outcome or the treatment sequence mechanism and the censoring mechanism,* $g_{A,0:K}g_{C,0:K}$, *is correctly specified.*

the longitudinal subgroup identification algorithms are L_2 *consistent.*

Proof of Theorem C.2.1: Define

$$\widetilde{r}_n(\boldsymbol{l}_0) = \frac{\sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w_{\widehat{\psi}_n}(\boldsymbol{l}_0)\} r(\boldsymbol{L}_{0,i})}{\sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w_{\widehat{\psi}_n}(\boldsymbol{l}_0)\}}.$$

We can show that

$$\int_{\boldsymbol{\mathcal{L}}_0} |r(\boldsymbol{l}_0) - \widehat{r}_n(\boldsymbol{l}_0)|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{l}_0) \leq 2\int_{\boldsymbol{\mathcal{L}}_0} |\widehat{r}_n(\boldsymbol{l}_0) - \widetilde{r}_n(\boldsymbol{l}_0)|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{l}_0) + 2\int_{\boldsymbol{\mathcal{L}}_0} |r(\boldsymbol{l}_0) - \widetilde{r}_n(\boldsymbol{l}_0)|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{l}_0),$$

where $\mathrm{P}(\boldsymbol{l}_0)$ is the distribution function for $\boldsymbol{l}_0$. We will show that both terms on the right hand side of the above equation are $o_P(1)$ which will lead to the consistency of $\widehat{r}_n(\boldsymbol{l}_0)$.

Assumption B.1 and the positivity assumption for the treatment sequence of interest give that there exists a $B_2 > 0$ such that $\max\left\{\left|Y^*_{w_{\pi_n}(\boldsymbol{L}_0)}\left(\widehat{\boldsymbol{\theta}}_{w_{\pi_n}(\boldsymbol{L}_0)}\right)\right|, \left|Y^*_{w_{\pi_n}(\boldsymbol{L}_0)}\left(\boldsymbol{\theta}^*_{w_{\pi_n}(\boldsymbol{L}_0)}\right)\right|\right\}$

$\leq B_2 < \infty$ for $\boldsymbol{O} \in \mathcal{O}$ and $\pi_n \in \Pi_n$, where $\boldsymbol{\theta}^*_{w_{\pi_n}(\boldsymbol{L}_0)}$ is the limit of $\widehat{\boldsymbol{\theta}}_{w_{\pi_n}(\boldsymbol{L}_0)}$. Define $B^* = \max(B_1, B_2)$, $\gamma_w = \mathrm{P}(\boldsymbol{L}_0 \in w)$ and $n(w) = \sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w\}$. Assumption B.1 and equation (18) in Nobel et al. [1996] imply

$$
\begin{aligned}
&\int_{\boldsymbol{\mathcal{L}}_0} |\widehat{r}_n(\boldsymbol{l}_0) - \widetilde{r}_n(\boldsymbol{l}_0)|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{l}_0) \\
\leq & 2B^* \sum_{w \in \widehat{\psi}_n} \gamma_w \left| \frac{1}{n(w)} \sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w\}\{Y^*_{w,i}(\widehat{\boldsymbol{\theta}}_w) - r(\boldsymbol{L}_{0,i})\} \right| \\
\leq & 2B^* \sum_{w \in \widehat{\psi}_n} \widehat{\gamma}_w \left| \frac{1}{n(w)} \sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w\}\{Y^*_{w,i}(\widehat{\boldsymbol{\theta}}_w) - r(\boldsymbol{L}_{0,i})\} \right| \\
& + 2B^* \sum_{w \in \widehat{\psi}_n} \left| \frac{1}{n(w)} \sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w\}\{Y^*_{w,i}(\widehat{\boldsymbol{\theta}}_w) - r(\boldsymbol{L}_{0,i})\} \right| |\widehat{\gamma}_w - \gamma_w| \\
\leq & 2B^* \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} \widehat{\gamma}_w \left| \frac{1}{n(w)} \sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w\}\{Y^*_{w,i}(\widehat{\boldsymbol{\theta}}_w) - r(\boldsymbol{L}_{0,i})\} \right| \\
& + 4B^{*2} \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} |\widehat{\gamma}_w - \gamma_w| \,.
\end{aligned}
$$

Under assumption B.2, applying Proposition 2, Lemma 3, and Theorem 7 in Nobel et al. [1996] gives

$\sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} |\widehat{\gamma}_w - \gamma_w| = o_P(1)$. It follows that

$$
\int_{\boldsymbol{\mathcal{L}}_0} |\widehat{r}_n(\boldsymbol{l}_0) - \widetilde{r}_n(\boldsymbol{l}_0)|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{l}_0) \leq 2B_1^* \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} \left| \frac{1}{n} \sum_{i=1}^n I\{\boldsymbol{L}_{0,i} \in w\}\{Y^*_{w,i}(\widehat{\boldsymbol{\theta}}_w) - r(\boldsymbol{L}_{0,i})\} \right| + o_P(1).
$$

Using assumption B.3 and Taylor series arguments,

$$
\begin{aligned}
&\int_{\boldsymbol{\mathcal{L}}_0} |\widehat{r}_n(\boldsymbol{l}_0) - \widetilde{r}_n(\boldsymbol{l}_0)|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{l}_0) \\
\leq & 2B^* \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} \left| \frac{1}{n} \sum_{i=1}^n I(\boldsymbol{L}_{0,i} \in w) \left\{ Y^*_{w,i}(\boldsymbol{\theta}^*_w) + (\widehat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}^*_w) \left. \frac{\partial Y^*_{w,i}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*_w} - r(\boldsymbol{L}_{0,i}) \right\} \right| + o_P(1) \\
= & 2B^* \sup_{\pi_n \in \Pi_n} \sum_{w \in \pi_n} \left| \frac{1}{n} \sum_{i=1}^n I(\boldsymbol{L}_{0,i} \in w)\{Y^*_{w,i}(\boldsymbol{\theta}^*_w) - r(\boldsymbol{L}_{0,i})\} \right| + o_P(1),
\end{aligned}
$$

where the last equality follows because $\widehat{\boldsymbol{\theta}}_w$ converges in probability to $\boldsymbol{\theta}^*_w$.

Define a class of functions

$$\mathcal{H} = \left\{ h\left(\boldsymbol{o}; \widetilde{\pi} = \{\widetilde{u}_k\}, \{\boldsymbol{\theta}^*_{\widetilde{u}_k}\}\right) = \sum_{\boldsymbol{o} \in \widetilde{u}_k} I\{\boldsymbol{o} \in \widetilde{u}_k\}\{y^*_{\widetilde{u}_k}(\boldsymbol{\theta}^*_{\widetilde{u}_k}) - r(\boldsymbol{l}_0)\} : \widetilde{\pi} \in \widetilde{\Pi} \right\}.$$

Under assumption B.1 and the positivity assumption for the conditional probability of treatment, $h\left(\boldsymbol{o}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}_k}\}\right)$ is bounded. Under assumption B.4*, $\int_{u \times \{0,1\} \times \cdots \times [-B_1, B_1]} h\left(\boldsymbol{o}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}_k}\}\right) \mathrm{d}\,\mathrm{P}(\boldsymbol{o}) = 0$ for every measurable subset u of $\boldsymbol{\mathcal{L}}_0$ and $h\left(\boldsymbol{o}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}_k}\}\right) \in \mathcal{H}$. Therefore, by assumption B.2, Proposition 2, Lemma 3 and Equation (13) in Nobel et al. [1996],

$$\begin{aligned}
&\int_{\boldsymbol{\mathcal{L}}_0} |\widehat{r}_n(\boldsymbol{l}_0) - \widetilde{r}_n(\boldsymbol{l}_0)|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{l}_0) \\
\leq & 2B^* \sup_{\widetilde{\pi}_n \in \widetilde{\Pi}_n} \sum_{\widetilde{w} \in \widetilde{\pi}_n} \left| \mathbb{P}_n\{I(\boldsymbol{O} \in \widetilde{w}) h(\boldsymbol{O}; \widetilde{\pi}_n, \{\boldsymbol{\theta}^*_{\widetilde{w}}\})\} - \int I(\boldsymbol{o} \in \widetilde{w}) h(\boldsymbol{o}; \widetilde{\pi}_n, \{\boldsymbol{\theta}^*_{\widetilde{w}}\}) \mathrm{d}\,\mathrm{P}(\boldsymbol{o}) \right| + o_P(1) \\
\leq & 2B^* \sup_{h_{\widetilde{\pi}} \in \mathcal{H}} \sup_{\widetilde{\pi}_n \in \widetilde{\Pi}_n} \sum_{\widetilde{w} \in \widetilde{\pi}_n} \left| \mathbb{P}_n\{I(\boldsymbol{O} \in \widetilde{w}) h(\boldsymbol{O}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}}\})\} - \int I(\boldsymbol{o} \in \widetilde{w}) h(\boldsymbol{o}; \widetilde{\pi}, \{\boldsymbol{\theta}^*_{\widetilde{u}}\}) \mathrm{d}\,\mathrm{P}(\boldsymbol{o}) \right| + o_P(1) \\
= & o_P(1).
\end{aligned}$$

By exactly the same arguments in the Proof of Theorem 1 in Nobel et al. [1996], $\int_{\mathcal{X}} |r(\boldsymbol{l}_0) - \widetilde{r}_n(\boldsymbol{l}_0)|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{l}_0) = o_P(1)$. It then follows that $\int_{\boldsymbol{\mathcal{L}}_0} |r(\boldsymbol{l}_0) - \widehat{r}_n(\boldsymbol{l}_0)|^2 \,\mathrm{d}\,\mathrm{P}(\boldsymbol{l}_0) = o_P(1)$.

C.3 Additional information on analysis of AMPATH data

To capture treatment in the window $k, k \in \{0, \ldots, K\}$, we use data at visits with more than or equal to 3 drugs because antiretroviral therapy consists of a combination of three or more antiretroviral drugs Organization et al. [2016] and at visits if they are on DTG to ensure that we use all DTG information; treatment data at visits where there are fewer than 3 therapies and DTG is not among one of the therapies are disregarded. After disregarding treatment at those visits, we let treatment at visit$k$ equal 1 when the patient is on DTG at the last visit in the window, equal 0 when not on DTG at the last visit and be missing if there is no remaining treatment data in the window. We then collect the last observed outcome and time-varying covariates before or at the visit where treatment is captured if there is no switch of DTG status in the window and before or at where the last switch happens if there is switch of DTG status. If treatment is missing in the window,

we collect the last observed outcome and time-varying covariates in the window. In window $K+1$, we collect the last observed outcome measurement as Y_{K+1}. Outcome or time-varying covariates will be missing in the window if there are no observed measurements satisfying the forementioned condition.

For missingness in outcome and treatment at visit0, we re-define day 0 until both outcome and treatment are observed at visit0; after visit0 we carry over from the previous window. If there is no visit at visit0, we re-define day 0 as the date of patient's first visit after July 1, 2016. If viral load is still missing at visit0 after re-defining day 0, we further re-define day 0 as the date of the visit with the first viral load measurement. If treatment is still missing at visit0 after re-defining day 0, we move the date of visit with the first observed treatment to the last day of visit0 (i.e., day $d-1$ for a window length of d days); therefore, day 0 is re-defined as $(d-1)$ days before date of visit with the first observed treatment. We exclude participants who still have missing values in outcome and treatment at visit0 with forementioned processing. That is, participants 1) with no observed weight measurements, 2) with no observed treatment or 3) with no observed treatment after the first observed weight measurement.

For missingness in viral load, we create an ordinal variable where missingness is the lowest category. The new variable is ordered as 1) missing, 2) $[0, 10^3)$, 3) $[10^3, 10^4)$, and 4) $[10^4, \infty)$. We drop the covariates with more than 10% missing values in uncensored participants. For the remaining covariates, we carry over from the previous window if there are observed measurements to carry over; otherwise for continuous covariates, we impute the missing values using the average of the observed values at the same time, and for binary covariates, we impute the missing values by 0. Table C.1 and C.2 present the number and proportion of missing values in covariates that are dropped and in time-varying covariates that are used, respectively. Table C.3 present the summary for each covariate used in the analysis before and after imputation, if not dropped.

Table C.4 presents the number of subgroups having differential effects of DTG-containing ART regimens on weight gain identified by LISA algorithms in 1,000 repetitions. Table C.5 presents the number of times the final trees make splits on the variables in $\boldsymbol{L}_0$ in the repetitions.

Covariate Uncensored patients	visit0 (%) $n = 84,445$	visit1 (%) $n = 71,431$	visit2 (%) $n = 64,353$	visit3 (%) $n = 58,646$
Employed outside home	31,807 (37.7%)	-	-	-
Electricity in home	31,669 (37.5%)	-	-	-
Water piped in home	31,928 (37.8%)	-	-	-
CD4	77,760 (92.1%)	66,453 (93.0%)	60,252 (93.6%)	55,306 (94.3%)
Blood glucose	84,407 (> 99.9%)	71,376 (99.9%)	64,300 (99.9%)	58,582 (99.9%)
Fasting serum glucose	84,438 (> 99.9%)	71,420 (> 99.9%)	64,338 (> 99.9%)	58,628 (> 99.9%)

Table C.1: Number and proportion of missing values in covariates that are dropped in AMPATH dataset.

Covariate Uncensored patients	visit0 $n = 84,445$	visit1 $n = 71,431$	visit2 $n = 64,353$	visit3 $n = 58,646$
Systolic blood pressure	1,269 (1.5%)	179 (0.3%)	66 (0.1%)	35 (0.1%)
Diastolic blood pressure	1,276 (1.5%)	181 (0.3%)	68 (0.1%)	36 (0.1%)
Height	4,276 (5.1%)	1,659 (2.3%)	862 (1.3%)	405 (0.7%)
TB treatment	56 (0.1%)	11 (0%)	4 (0%)	3 (0%)
Married/living with partner	3,019 (3.6%)	484 (0.7%)	174 (0.3%)	71 (0.1%)
Covered by NHIF	6,362 (7.5%)	2,305 (3.2%)	1,039 (1.6%)	496 (0.8%)

Table C.2: Number and proportion of missing values in time-varying covariates in AMPATH dataset. TB refers to tuberculosis; NHIF refers to National Health Insurance Fund.

	$\boldsymbol{L}_0$ ($n = 84,445$)		$\boldsymbol{L}_1$ ($n = 71,431$)	
Covariates	Before	After	Before	After
Male, n(%)	28,440 (33.7%)		-	
Age when starting ART, years	38.1 ± 10.4		-	
Age at visit 0, years	42.1 ± 11.0		-	
Time on ART at time 0, years	4.0 ± 3.9		-	
Whether enrolled on or after July 1, 2016, n(%)	24,688 (29.2%)		-	
Weight, kg	62.2 ± 12.5		-	
Viral load, n(%)				
Missing	48,309 (57.2%)		10,043 (14.1%)	
$[0, 10^3)$	30,025 (35.6%)		54,954 (76.9%)	
$[10^3, 10^4)$	3,166 (3.7%)		3,459 (4.8%)	
$[10^4, \infty)$	2,945 (3.5%)		2,975 (4.2%)	
Systolic blood pressure, mmHg	118.9 ± 18.7	118.9 ± 18.6	119.8 ± 18.9	119.8 ± 18.9
Diastolic blood pressure, mmHg	73.9 ± 12.1	73.9 ± 12.0	73.9 ± 12.3	73.9 ± 12.2
Height, cm	165.7 ± 9.2	165.7 ± 9.0	165.6 ± 9.5	165.6 ± 9.3
Active TB, n(%)	2,821 (3.3%)		2,479 (3.5%)	
TB treatment, n(%)	1,936 (2.3%)	1,936 (2.3%)	596 (0.8%)	596 (0.8%)
Married/living with partner, n(%)	47,759 (58.7%)	47,759 (56.6%)	41,065 (57.9%)	41,065 (57.5%)
Covered by NHIF, n(%)	16,485 (21.1%)	16,485 (19.5%)	16,104 (23.3%)	16,104 (22.5%)
	$\boldsymbol{L}_2$ ($n = 64,353$)		$\boldsymbol{L}_3$ ($n = 58,646$)	
	Before	After	Before	After
Viral load, n(%)				
Missing	4,622 (7.2%)		2,368 (4.0%)	
$[0, 10^3)$	54,117 (84.1%)		51,749 (88.2%)	
$[10^3, 10^4)$	3,269 (5.1%)		2,797 (4.8%)	
$[10^4, \infty)$	2,345 (3.6%)		1,732 (3.0%)	
Systolic blood pressure, mmHg	120.6 ± 18.9	120.6 ± 18.9	121.2 ± 18.8	121.2 ± 18.8
Diastolic blood pressure, mmHg	74.6 ± 12.1	74.6 ± 12.1	75.1 ± 12.1	75.1 ± 12.1
Height, cm	165.7 ± 9.3	165.7 ± 9.2	165.8 ± 9.0	165.8 ± 9.0
Active TB, n(%)	2,349 (3.7%)		2,167 (3.7%)	
TB treatment, n(%)	369 (0.6%)	369 (0.6%)	277 (0.5%)	277 (0.5%)
Married/living with partner, n(%)	37,061 (57.7%)	37,061 (57.6%)	33,164 (56.6%)	33,164 (56.5%)
Covered by NHIF, n(%)	15,970 (25.2%)	15,970 (24.8%)	15,779 (27.1%)	15,779 (26.9%)

Table C.3: Summary of covariates used in AMPATH dataset before and after imputation. For binary or ordinal covariates, we present the number and proportion of participants falling in the listed category; for continuous covariates, we present the mean and standard deviation. ART refers to antiretroviral therapy; TB refers to tuberculosis; NHIF refers to National Health Insurance Fund.

Subgroups	1	2	3	4
n	245	206	322	227

Table C.4: Number of subgroups with differential effect of DTG-containing ART regimens on weight identified by LISA algorithm.

Covariate	First split	Other splits	Total
Male	482	61	543
Age when starting ART	66	90	156
Age at time 0	64	264	328
Time on ART at time 0	13	2	15
Whether enrolled on or after July 1, 2016	11	0	11
Weight	14	138	152
Viral load	3	46	49
Systolic blood pressure	12	58	70
Diastolic blood pressure	9	34	43
Height	69	56	125
Active TB	0	0	0
TB treatment	0	0	0
Married/living with partner	12	25	37
Covered by NHIF	0	2	2

Table C.5: Number of times the final trees make splits on the variables in $\boldsymbol{L}_0$. ART refers to antiretroviral therapy; TB refers to tuberculosis; NHIF refers to National Health Insurance Fund.

Bibliography

A. Achhra, A. Mocroft, P. Reiss, C. Sabin, L. Ryom, S. De Wit, C. Smith, A. d'Arminio Monforte, A. Phillips, R. Weber, et al. Short-term weight gain after antiretroviral therapy initiation and subsequent risk of cardiovascular disease and diabetes: the d: A: D study. *HIV medicine*, 17(4): 255–268, 2016.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

C. Barr, M. Marois, I. Sim, C. H. Schmid, B. Wilsey, D. Ward, N. Duan, R. D. Hays, J. Selsky, J. Servadio, et al. The preempt study-evaluating smartphone-assisted n-of-1 trials in patients with chronic pain: study protocol for a randomized controlled trial. *Trials*, 16(1):1–11, 2015.

D. Benkeser and M. Van Der Laan. The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE, 2016.

K. Bourgi, C. A. Jenkins, P. F. Rebeiro, B. E. Shepherd, F. Palella, R. D. Moore, K. N. Althoff, J. Gill, C. S. Rabkin, S. J. Gange, et al. Weight gain among treatment-naïve persons with hiv starting integrase inhibitors compared to non-nucleoside reverse transcriptase inhibitors or protease inhibitors in a large observational cohort in the united states and canada. *Journal of the International AIDS Society*, 23(4):e25484, 2020a.

K. Bourgi, P. F. Rebeiro, M. Turner, J. L. Castilho, T. Hulgan, S. P. Raffanti, J. R. Koethe,

and T. R. Sterling. Greater weight gain in treatment-naive persons starting dolutegravir-based antiretroviral therapy. *Clinical Infectious Diseases*, 70(7):1267–1274, 2020b.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and regression trees. wadsworth int. *Group*, 37(15):237–251, 1984.

S. T. Brookes, E. Whitely, M. Egger, G. D. Smith, P. A. Mulheran, and T. J. Peters. Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test. *Journal of clinical epidemiology*, 57(3):229–236, 2004.

A. Calmy, T. T. Sanchez, C. Kouanfack, M. Mpoudi-Etame, S. Leroy, S. Perrineau, M. L. Wandji, D. T. Tata, P. O. Bassega, T. A. Bwenda, et al. Dolutegravir-based and low-dose efavirenz-based regimen for the initial treatment of hiv-1 infection (namsal): week 96 results from a two-group, multicentre, randomised, open label, phase 3 non-inferiority trial in cameroon. *The lancet HIV*, 7(10):e677–e687, 2020.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, et al. Double/debiased machine learning for treatment and structural parameters, 2018.

A. F. Connors, T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, D. Wagner, et al. The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, 276 (11):889–897, 1996.

I. J. Dahabreh, R. Hayward, and D. M. Kent. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International journal of epidemiology*, 45(6):2184–2193, 2016.

I. J. Dahabreh, T. A. Trikalinos, D. M. Kent, and C. H. Schmid. Heterogeneity of treatment effects. In *Methods in comparative effectiveness research*, pages 247–292. Chapman and Hall/CRC, 2017.

N. Duan, R. L. Kravitz, and C. H. Schmid. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of clinical epidemiology*, 66(8):S21–S28, 2013.

A. L. Esber, D. Chang, M. Iroezindu, E. Bahemana, H. Kibuuka, J. Owuoth, V. Singoei, J. Maswai, N. F. Dear, T. A. Crowell, et al. Weight gain during the dolutegravir transition in the african cohort study. *Journal of the International AIDS Society*, 25(4):e25899, 2022.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

S. Greenfield, R. Kravitz, N. Duan, and S. H. Kaplan. Heterogeneity of treatment effects: implications for guidelines, payment, and quality assessment. *The American journal of medicine*, 120 (4):S3–S9, 2007.

N. A. . S. Group. Dolutegravir-based or low-dose efavirenz–based regimen for the treatment of hiv-1. *New England Journal of Medicine*, 381(9):816–826, 2019.

M. Herrin, J. P. Tate, K. M. Akgün, A. A. Butt, K. Crothers, M. S. Freiberg, C. L. Gibert, D. A. Leaf, D. Rimland, M. C. Rodriguez-Barradas, et al. Weight gain and incident diabetes among hiv infected-veterans initiating antiretroviral therapy compared to uninfected individuals. *Journal of acquired immune deficiency syndromes (1999)*, 73(2):228, 2016.

K. Hirano and G. W. Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278, 2001.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

D. Jackson, M. Law, T. Stijnen, W. Viechtbauer, and I. R. White. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in medicine*, 37(7):1059–1085, 2018.

T. Johannessen and D. Fosstvedt. Statistical power in single subject trials. *Family practice*, 8(4): 384–387, 1991.

J. Kang, X. Su, L. Liu, and M. L. Daviglus. Causal inference of interaction effects with inverse propensity weighting, g-computation and tree-based standardization. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(5):323–336, 2014.

D. J. Kim, A. O. Westfall, E. Chamot, A. L. Willig, M. J. Mugavero, C. Ritchie, G. A. Burkholder, H. M. Crane, J. L. Raper, M. S. Saag, et al. Multimorbidity patterns in hiv-infected patients: the role of obesity in chronic disease clustering. *Journal of acquired immune deficiency syndromes (1999)*, 61(5):600, 2012.

R. L. Kravitz, N. Duan, and J. Braslow. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82(4):661–687, 2004.

R. L. Kravitz, A. Aguilera, E. J. Chen, Y. K. Choi, E. Hekler, C. Karr, K. K. Kim, S. Phatak, S. Sarkar, S. M. Schueller, et al. Feasibility, acceptability, and influence of mhealth-supported n-of-1 trials for enhanced cognitive and emotional well-being in us volunteers. *Frontiers in public health*, 8:260, 2020.

N. Kreif, L. Tran, R. Grieve, B. De Stavola, R. C. Tasker, and M. Petersen. Estimating the comparative effectiveness of feeding interventions in the pediatric intensive care unit: a demonstration of longitudinal targeted maximum likelihood estimation. *American journal of epidemiology*, 186 (12):1370–1379, 2017.

S. W. Lagakos et al. The challenge of subgroup analyses-reporting without distorting. *New England Journal of Medicine*, 354(16):1667, 2006.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

J. E. Lake. The fat of the matter: obesity and visceral adiposity in treated hiv infection. *Current HIV/AIDS Reports*, 14(6):211–219, 2017.

A. Legha, R. D. Riley, J. Ensor, K. I. Snell, T. P. Morris, and D. L. Burke. Individual participant data meta-analysis of continuous outcomes: a comparison of approaches for specifying and estimating one-stage models. *Statistics in medicine*, 37(29):4404–4420, 2018.

I. Lipkovich, A. Dmitrienko, and R. B D'Agostino Sr. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1):136–196, 2017.

J. M. Llibre, C.-C. Hung, C. Brinson, F. Castelli, P.-M. Girard, L. P. Kahl, E. A. Blair, K. Angelis, B. Wynne, K. Vandermeulen, et al. Efficacy, safety, and tolerability of dolutegravir-rilpivirine for the maintenance of virological suppression in adults with hiv-1: phase 3, randomised, non-inferiority sword-1 and sword-2 studies. *The Lancet*, 391(10123):839–849, 2018.

J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

J. Nikles and G. Mitchell. *The essential guide to N-of-1 trials in health*. Springer, 2015.

A. Nobel et al. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105, 1996.

L. Olsen, D. Aisner, J. M. McGinnis, et al. *The learning healthcare system: workshop summary*. Natl Academy Pr, 2007.

W. H. Organization et al. *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. World Health Organization, 2016.

W. H. Organization et al. Updated recommendations on first-line and second-line antiretroviral regimens and post-exposure prophylaxis and recommendations on early infant diagnosis of hiv: interim guidelines: supplement to the 2016 consolidated guidelines on the use of antiretroviral drugs for treating and preventing hiv infection. Technical report, World Health Organization, 2018.

J. Pearl and J. M. Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *UAI*, volume 95, pages 444–453. Citeseer, 1995.

M. Petersen, J. Schwab, S. Gruber, N. Blaser, M. Schomaker, and M. van der Laan. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of causal inference*, 2(2):147–185, 2014.

A. N. Phillips, F. Venter, D. Havlir, A. Pozniak, D. Kuritzkes, A. Wensing, J. D. Lundgren, A. De Luca, D. Pillay, J. Mellors, et al. Risks and benefits of dolutegravir-based antiretroviral drug regimens in sub-saharan africa: a modelling study. *The Lancet HIV*, 6(2):e116–e127, 2019.

S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

R. D. Riley, A. Legha, D. Jackson, T. P. Morris, J. Ensor, K. I. Snell, I. R. White, and D. L. Burke. One-stage individual participant data meta-analysis models for continuous and binary outcomes: comparison of treatment coding options and estimation methods. *Statistics in medicine*, 39(19): 2536–2555, 2020.

S. E. Robertson, A. Leith, C. H. Schmid, and I. J. Dahabreh. Assessing heterogeneity of treatment effects in observational studies. *American Journal of Epidemiology (in press)*, 2020.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

J. M. Robins and S. Greenland. Causal inference without counterfactuals: comment. *Journal of the American Statistical Association*, 95(450):431–435, 2000.

J. M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pages 297–331. Springer, 1992.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

K. J. Rothman, S. Greenland, and A. M. Walker. Concepts of interaction. *American journal of epidemiology*, 112(4):467–470, 1980.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

P. E. Sax, K. M. Erlandson, J. E. Lake, G. A. Mccomsey, C. Orkin, S. Esser, T. T. Brown, J. K. Rockstroh, X. Wei, C. C. Carter, et al. Weight gain following initiation of antiretroviral therapy: risk factors in randomized comparative clinical trials. *Clinical Infectious Diseases*, 71(6):1379–1389, 2020.

H. Seibold, A. Zeileis, and T. Hothorn. Model-based recursive partitioning for subgroup analyses. *The international journal of biostatistics*, 12(1):45–63, 2016.

S. Senn. Sample size considerations for n-of-1 trials. *Statistical methods in medical research*, 28(2): 372–383, 2019.

S. S. Senn. *Cross-over trials in clinical research*, volume 5. John Wiley & Sons, 2002.

L. A. Stefanski and D. D. Boos. The calculus of m-estimation. *The American Statistician*, 56(1): 29–38, 2002.

J. A. Steingrimsson and J. Yang. Subgroup identification using covariate-adjusted interaction trees. *Statistics in medicine*, 2019.

J. A. Steingrimsson, L. Diao, A. M. Molinaro, and R. L. Strawderman. Doubly robust survival trees. *Statistics in medicine*, 35(20):3595–3612, 2016.

O. M. Stitelman, V. De Gruttola, and M. J. van der Laan. A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *The international journal of biostatistics*, 8(1), 2012.

X. Su, T. Zhou, X. Yan, J. Fan, and S. Yang. Interaction trees with censored survival data. *The international journal of biostatistics*, 4(1), 2008.

X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

X. Su, K. Meneses, P. McNees, and W. O. Johnson. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):457–474, 2011.

X. Su, J. Kang, J. Fan, R. A. Levine, and X. Yan. Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research*, 13(Oct):2955–2994, 2012.

Y. Sun, S. H. Chiou, and M.-C. Wang. Roc-guided survival trees and ensembles. *Biometrics*, 2019.

Y. Sun, S. H. Chiou, C. O. Wu, M. McGarry, and C.-Y. Huang. Dynamic risk prediction using survival tree ensembles with application to cystic fibrosis. *arXiv preprint arXiv:2011.07175*, 2020.

B. Surial, C. Mugglin, A. Calmy, M. Cavassini, H. F. Günthard, M. Stöckle, E. Bernasconi, P. Schmid, P. E. Tarr, H. Furrer, et al. Weight and metabolic changes after switching from tenofovir disoproxil fumarate to tenofovir alafenamide in people living with hiv: a cohort study. *Annals of internal medicine*, 174(6):758–767, 2021.

W. M. Tierney, J. K. Rotich, T. J. Hannan, A. M. Siika, P. G. Biondich, B. W. Mamlin, W. M. Nyandiko, S. Kimaiyo, K. Wools-Kaloustian, J. E. Sidle, et al. The ampath medical record system: creating, implementing, and sustaining an electronic medical record system to support hiv/aids care in western kenya. *Studies in health technology and informatics*, 129(1):372, 2007.

M. J. van der Laan and S. Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics*, 8(1), 2012.

M. J. Van der Laan and S. Rose. *Targeted learning in data science*. Springer, 2018.

M. J. Van Der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

M. J. Van der Laan, S. Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

T. J. VanderWeele. On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871, 2009.

W. D. Venter, M. Moorhouse, S. Sokhela, L. Fairlie, N. Mashabane, M. Masenya, C. Serenata, G. Akpomiemie, A. Qavi, N. Chandiwana, et al. Dolutegravir plus two different prodrugs of tenofovir to treat hiv. *New England Journal of Medicine*, 381(9):803–815, 2019.

W. D. Venter, S. Sokhela, B. Simmons, M. Moorhouse, L. Fairlie, N. Mashabane, C. Serenata, G. Akpomiemie, M. Masenya, A. Qavi, et al. Dolutegravir with emtricitabine and tenofovir alafenamide or tenofovir disoproxil fumarate versus efavirenz, emtricitabine, and tenofovir disoproxil fumarate for initial treatment of hiv-1 infection (advance): week 96 results from a randomised, phase 3, non-inferiority trial. *The Lancet HIV*, 7(10):e666–e676, 2020.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

R. Wang, S. W. Lagakos, J. H. Ware, D. J. Hunter, and J. M. Drazen. Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194, 2007.

Y. Wei, L. Liu, X. Su, L. Zhao, and H. Jiang. Precision medicine: Subgroup identification in longitudinal trajectories. *Statistical Methods in Medical Research*, page 0962280220904114, 2020.

I. R. White, R. M. Turner, A. Karahalios, and G. Salanti. A comparison of arm-based and contrast-based models for network meta-analysis. *Statistics in medicine*, 38(27):5197–5213, 2019.

J. Yang, I. J. Dahabreh, and J. A. Steingrimsson. Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. *Biometrics*, 2021.

S. L. Zeger. A regression model for time series of counts. *Biometrika*, 75(4):621–629, 1988.

D. R. Zucker, R. Ruthazer, and C. H. Schmid. Individual (n-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *Journal of clinical epidemiology*, 63(12):1312–1323, 2010.