Received: 31 August 2021 | Revised: 2 February 2023 | Accepted: 3 February 2023

DOI: 10.1002/sim.9686

RESEARCH ARTICLE

Statistics in Medicine WILEY

Ranking of average treatment effects with generalized random forests for time-to-event outcomes

Helene C. W. Rytgaard[1] | **Claus T. Ekstrøm**[1] | **Lars V. Kessing**[2] | **Thomas A. Gerds**[1]

[1]Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark

[2]Copenhagen Affective Disorder research Center (CADIC), Psychiatric Center Copenhagen, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark

Correspondence
Helene C. W. Rytgaard, Section of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, 1014 Copenhagen, Denmark.
Email: hely@sund.ku.dk

Abstract

Linkage between drug claims data and clinical outcome allows a data-driven experimental approach to drug repurposing. We develop an estimation procedure based on generalized random forests for estimation of time-point specific average treatment effects in a time-to-event setting with competing risks. To handle right-censoring, we propose a two-step procedure for estimation, applying inverse probability weighting to construct time-point specific weighted outcomes as input for the generalized random forest. The generalized random forests adaptively handle covariate effects on the treatment assignment by applying a splitting rule that targets a causal parameter. Using simulated data we demonstrate that the method is effective for a causal search through a list of treatments to be ranked according to the magnitude of their effect on clinical outcome. We illustrate the method using the Danish national health registries where it is of interest to discover drugs with an unexpected protective effect against relapse of severe depression.

KEYWORDS

average treatment effect, competing risks, random forests, time-to-event

1 | INTRODUCTION

Drug repurposing is an important low-cost method for drug discovery, which can be based on a data-driven experimental approach.[1] We are motivated by drug repurposing studies of electronic health records where drug claims data are linked with clinical outcomes and the aim is to rank a potentially very long list of treatment variables according to their effect on a time-to-event outcome. As always in observational studies, confounder control is important, but having to correctly specify regression models for bulk analysis is a near impossible task. We propose to let the data "do the modeling" by using generalized random forests[2] which we embed into a time-to-event setting with competing risks. In order to rank the list of treatments we estimate effects on the crude and the net probabilities, respectively, which we define in a counterfactual framework.[3,4] Under a set of structural and distributional assumptions the parameters are linked to the observed data. These considerations are closely related to the work of Young et al.[5]

A random forest[6] is a popular data-driven algorithm that can be used for variable importance analysis.[7,8] The common variable importance measures are based on prediction performance,[6,7] or on the tree building process of the forests.[9,10] Our approach is different in that we consider the causal treatment effect as a variable importance measure. Similar approaches have also been considered in the context of high-dimensional biomarker discovery, see, for example, References 11-13. The generalized random forest (GRF)[2,14] is an extension of Breiman's random forests that has been developed for

nonparametric inference on heterogeneous treatment effects in settings with real-valued and uncensored outcomes of interest. A recent extension called causal survival forests estimates conditional average treatment effects on the expected counterfactual survival time.[15] Our approach is different as we formulate causal parameters in terms of average differences of event probabilities at pre-specified time horizons of interest, allowing us to report a time-point specific measure of the effect of a particular treatment.

We illustrate our methods using a study on drug claims data and development of psychiatric disorders in Danish national registries where the goal is to discover if drugs that are already in clinical use may have a protective effect against depression. Psychiatric disorders is a field where the pharmaceutical industry has substantially withdrawn from developing new drugs; thus, in the absence of new randomized clinical trials, and to supplement the expensive and time-consuming generation of data from clinical trials, a systematic search through all drug purchases in the registry data is a cost-efficient way to identify new treatments as well as to discover adverse side-effects. Specific findings can subsequently be further investigated and motivate new randomized trials. For proof of concept and illustration, we include all Danish citizens who have a first time diagnosis with depression registered. We follow these patients until depression relapse (event of interest), death without relapse (competing risk), or right-censoring, whatever comes first. We implement GRFs for the right censored time-to-event outcomes by using inverse probability weighting in a two step approach.

The article is organized as follows. In Section 2, we introduce the setting and notation for survival and competing risks data. In Section 3, we define our target parameters in terms of counterfactual outcomes, and discuss the distributional assumptions under which we can identify the parameters from the observed data. In Section 4, we review the generalized random forest methodology and present our weighting approach for making the methodology applicable to time-to-event data. In Section 5, we study the performance of our method using simulated data. In Section 6, we analyze Danish registry data. We close with a discussion in Section 7.

2 | SETTING AND NOTATION

We consider a time-to-event setting where subjects are observed from study entry to the occurrence of an event of interest or a competing event. If no event of any kind is observed within the subject-specific follow-up time, the subject is right-censored. Specifically, we consider a competing risks situation with $J \geq 2$ mutually exclusive types of events. For sake of presentation, we assume throughout that $J = 2$. We denote by T the uncensored event time, by $\Delta \in \{1, 2\}$ the event type and by C the censoring time, such that the observed data are $\tilde{T} = \min(T, C)$ and $\tilde{\Delta} = \mathbb{1}\{T \leq C\}\Delta$. Moreover, $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ is a vector of baseline covariates and $\boldsymbol{A} = (A_1, \dots, A_K) \in \{0, 1\}^K$ is a vector of $K \in \mathbb{N}$ binary treatment variables with $A_k = 1$ indicating treatment and $A_k = 0$ no treatment, $k = 1, \dots, K$. The data consist of $n \in \mathbb{N}$ independent samples, $\left\{(\boldsymbol{X}_1, \boldsymbol{A}_1, \tilde{T}_1, \tilde{\Delta}_1), \dots, (\boldsymbol{X}_n, \boldsymbol{A}_n, \tilde{T}_n, \tilde{\Delta}_n)\right\}$. We are interested in estimating the effect of each treatment variable A_k on the probability of events of type $j = 1$. We refer to the other type of events ($j = 2$) as competing events, or competing risks. We define our target parameter in terms of counterfactuals, using a notation with superscripts to define interventions. In particular, for $a = 0, 1$, we define T^a as the uncensored counterfactual event time and Δ^a as the corresponding event indicator that would result from setting treatment A_k to a. Further, for $j = 1, 2$, we use $T^{j,a}$ to denote the uncensored counterfactual event time of type j that would result if treatment A_k had been set to a in a hypothetical world where cause j is the only cause and the competing risks do not exist. Note that we distinguish between the counterfactual event time variable T^a with a single superscript and the counterfactual event time variable $T^{j,a}$ with double superscript. Note also that when studying the treatment A_k, the other treatments can enter the vector of baseline covariates. We thus introduce the notation

$$\boldsymbol{X}^{(\boldsymbol{k})} = (\boldsymbol{X}, A_1, \dots, A_{k-1}, A_{k+1}, \dots, A_K),$$

for $k = 1, \dots, K$.

3 | RANKING TREATMENT EFFECTS IN PRESENCE OF COMPETING RISKS

Suppose we have a list of treatments, $A_1, A_2, \dots, A_K$, $K \in \mathbb{N}$, that we would like to rank according to their effect on a time-to-event outcome. For the purpose of ranking in the presence of competing risks, we distinguish between crude and net probabilities.[16] The crude probability (ie, cumulative incidence) describes the probability of the outcome event in the

presence of the other risks. The net probability describes the probability of the event in a hypothetical world where the competing risks do not exist. The problem with crude probabilities is that they reflect a mixture of effects on the hazard rate of the event of interest and effects on the hazard rate of the competing risks. Net effects allow us to study the effect of a particular drug in a way that is independent of the effect that this drug may have on the hazard rate of the competing events. The problem with net probabilities is that they require strong identifiability assumptions (Section 3.3). We argue that, for the purpose of drug discovery, it may be desirable to search for drugs that have net effects — even though bias may occur due to unmeasured confounders.

3.1 | Effects on crude probabilities

Recall that the random variables T^0 and T^1 denote the uncensored counterfactual event times that would result if treatment had been set to $A_k = 0$ or $A_k = 1$, respectively. The average treatment effect (ATE) of A_k on the crude risk of events of type 1 before a fixed time horizon $t_0 > 0$ is defined as follows

$$\overline{\theta}_{\text{crude}} = P(T^1 \leq t_0, \Delta^1 = 1) - P(T^0 \leq t_0, \Delta^0 = 1). \tag{1}$$

The quantities $P(T^a \leq t_0, \Delta^a = 1)$, $a = 0, 1$, in (1), referred to as the crude probabilities, are the cumulative incidence functions[17] of the event of interest for a hypothetical treated and a hypothetical untreated population, respectively. These crude probabilities also depend on the hazard rate of the competing event, since, at any time, the event of interest can only occur for subjects who have survived all risks so far. A treatment which reduces the hazard rate of the competing risk increases the event-free survival probability and thereby indirectly increases the crude risk of the event of interest, and vice versa. Particularly, a treatment effect reflected in a non-zero value of $\overline{\theta}_{\text{crude}}$ will occur also if there is only an indirect effect via the hazard rate of the competing event.

3.2 | Effects on net probabilities

Recall that the counterfactual random variables $T^{1,0}$ and $T^{1,1}$ are the (uncensored) counterfactual event times that would have been observed in a hypothetical world in which cause $j = 1$ is the only cause and where treatment had been set to $A_k = 0$ and $A_k = 1$, respectively. Particularly, $T^{1,0}$ and $T^{1,1}$ are latent times that are not always observed in the real world due to cause $j = 2$ events and due to right-censoring. The average treatment effect of A_k on the net risk of events of type 1 is defined as follows

$$\overline{\theta}_{\text{net}} = P(T^{1,1} \leq t_0) - P(T^{1,0} \leq t_0). \tag{2}$$

We emphasize that, opposed to the crude risks $P(T^a \leq t_0, \Delta^a = 1)$, $a = 0, 1$, in Equation (1), the net risks $P(T^{1,a} \leq t_0)$, $a = 0, 1$, are not affected by the (indirect) effect that a treatment may have on the hazard rate of the competing risk. They are interpreted as net probabilities for the event of interest in a hypothetical world where the competing event cannot happen. A treatment effect reflected in a non-zero value $\overline{\theta}_{\text{net}}$ will only occur if the studied treatment has a direct effect on the event of interest.

3.3 | Identifiability of treatment effects on crude and net probabilities

The average treatment effects on crude probabilities $\overline{\theta}_{\text{crude}}$ and net probabilities $\overline{\theta}_{\text{net}}$ are defined in terms of counterfactual random variables, and are identified from the observed data only under causal assumptions.[18] We review these assumptions in the supplementary material (Appendix A) separately for $\overline{\theta}_{\text{crude}}$ and $\overline{\theta}_{\text{net}}$. Here we briefly discuss the assumption of *no unmeasured confounding* to contrast the choice between the two parameters. For identifiability of the crude effects, this is an assumption of conditional independence between the counterfactuals and the treatment and censoring mechanisms, as follows, $(T^a, \Delta^a) \perp\!\!\!\perp A_k \mid \boldsymbol{X}^{(k)}$, for $a = 0, 1$, and $(T, \Delta) \perp\!\!\!\perp C \mid A_k, \boldsymbol{X}^{(k)}$. To move from crude to net effects, one needs additionally that $T^{1,A_k} \perp\!\!\!\perp T^{2,A_k} \mid A_k, \boldsymbol{X}^{(k)}$. As previously mentioned, we stress that this is a very strong assumption: Whether A_k and $\boldsymbol{X}$ together include all factors that we believe to be predictive of both event types depends very much on the nature of the competing events and how rich the measured set of covariates is.

3.4 | Ranking of treatments

To obtain a ranking of the list of treatments, we apply the two-step approach described in Section 4 which controls for confounding by using generalized random forests and deals with right censored data by using inverse probability weighting. The approach yields estimates $\hat{\bar{\theta}}_{\text{crude},k}$ and $\hat{\bar{\theta}}_{\text{net},k}$ for the treatment effects on the crude and the net probability scale, respectively, for all drugs A_k, $k = 1, \dots, K$. We can then order the treatments according to their crude and according to their net effects.

4 | GENERALIZED RANDOM FORESTS WITH INVERSE PROBABILITY WEIGHTED OUTCOMES

Generalized random forests (GRFs)[2] are a recent generalization of the original random forest algorithm,[6] a machine learning tool that adaptively searches the covariate space by recursive sample splitting. Generally, a forest consists of $B \in \mathbb{N}$ randomized trees, where the bth tree of the forest is grown by recursively splitting the covariate space according to some split criterion. GRFs provide a data-adaptive approach to estimation of conditional treatment effects for uncensored data, particularly, for a generic outcome variable $Y \in \mathbb{R}$, $\theta(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X^{(k)}} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X^{(k)}} = \boldsymbol{x}]$. A key part of the generalized random forest algorithm is the splitting rule that targets specifically the estimation of the quantity $\theta(\boldsymbol{x})$ of interest; particularly, each tree applies a splitting rule that adaptively makes binary partitions of the covariate space such as to maximize heterogeneity in $\theta(\boldsymbol{x})$. By averaging over neighborhoods defined by the trees, the forest produces a neighborhood function that is used as a kernel for estimation of $\theta(\boldsymbol{x})$. In the supplementary material (Appendix C) we describe the local gradient-based criterion for making splits and the kernel-based estimator for average treatment effects for uncensored data as proposed by Reference 2.

The problem in our setting is that we do not observe the actual outcomes of interest. For the parameter $\bar{\theta}_{\text{crude}}$, for example, we do not observe $Y := \mathbb{1}\{T \le t_0, \Delta = 1\}$ due to right-censoring. In this section we assume that we are given an estimator $\hat{G}$ for the conditional distribution function $G(t \mid \boldsymbol{A}, \boldsymbol{X}) = P(C > t \mid \boldsymbol{A}, \boldsymbol{X})$. Based on $\hat{G}$, we define the inverse probability weighted outcome for estimation of crude effects:

$$\tilde{Y} := \frac{\mathbb{1}\{\tilde{T} \le t_0, \tilde{\Delta} = 1\}}{\hat{G}(\tilde{T} - \mid \boldsymbol{A}, \boldsymbol{X})}. \tag{3}$$

For this outcome, we show in Section 4.1 below that

$$\theta_{\text{crude}} = \mathbb{E}\left[\mathbb{E}[\tilde{Y} \mid \boldsymbol{X^{(k)}} = \boldsymbol{x}, A_k = 1] - \mathbb{E}[\tilde{Y} \mid \boldsymbol{X^{(k)}} = \boldsymbol{x}, A_k = 0]\right].$$

The idea is that we can apply GRFs directly to our weighted outcome $\tilde{Y}$. This provides an estimator $\hat{\theta}_{\text{crude}}(\boldsymbol{x})$ for the conditional effect $\theta_{\text{crude}}(\boldsymbol{x}) = P(T^1 \le t_0, \Delta^1 = 1 \mid \boldsymbol{X^{(k)}} = \boldsymbol{x}) - P(T^0 \le t_0, \Delta^0 = 1 \mid \boldsymbol{X^{(k)}} = \boldsymbol{x})$ and thereby an estimator for the corresponding average effect $\hat{\bar{\theta}}_{\text{crude}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{\text{crude}}(\boldsymbol{X^{(k)}}_i)$. This leads to the following two-step approach:

Step 1. The conditional distribution function G is estimated based on the full dataset and is used to construct the weighted outcome $\tilde{Y}$ as defined by Equation (3).

Step 2. A generalized random forest is applied with $\tilde{Y}$ as outcome, yielding estimates $\hat{\theta}_{\text{crude}}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}$, and the ATE is then estimated simply by averaging.

A similar two-step approach is utilized to estimate the effect on net probabilities. We note that this requires, in addition to an estimator for the conditional distribution G, an estimator $\hat{G}_2$ for the conditional distribution $G_2(t \mid \boldsymbol{A}, \boldsymbol{X}) = P(T^{2,a} > t \mid \boldsymbol{A}, \boldsymbol{X})$. We construct the inverse probability weighted outcome for estimation of net effects:

$$\tilde{Y}' := \frac{\mathbb{1}\{\tilde{T} \le t_0, \tilde{\Delta} = 1\}}{\hat{G}(\tilde{T} - \mid \boldsymbol{A}, \boldsymbol{X})\hat{G}_2(\tilde{T} - \mid \boldsymbol{A}, \boldsymbol{X})}. \tag{4}$$

Thus, to estimate the effects on the net probability, we need to model the survival functions of both the latent time to a competing risk event and the censoring time.

4.1 | Identifiability by inverse probability weighting

The causal assumptions (see Section 3.3) allow us to link the distribution of the counterfactual variables to the observed data distribution. Since,

$$\mathbb{E}\left[\tilde{Y} \mid \boldsymbol{X}, \boldsymbol{A}\right] = \mathbb{E}\left[\tilde{Y} \mid \boldsymbol{X}^{(\boldsymbol{k})}, A_k\right] = \mathbb{E}\left[\frac{\mathbb{1}\{\tilde{T} \leq t_0, \tilde{\Delta} = 1\}}{G(\tilde{T}- \mid \boldsymbol{X}^{(\boldsymbol{k})}, A_k)} \;\middle|\; \boldsymbol{X}^{(\boldsymbol{k})}, A_k\right] = \mathbb{E}\left[\mathbb{1}\{T \leq t_0, \Delta = 1\} \mid \boldsymbol{X}^{(\boldsymbol{k})}, A_k\right]$$

it follows that,

$$\begin{aligned}\overline{\theta}_{\text{crude}}(\boldsymbol{x}) &= \mathbb{E}\left[P(T^1 \leq t_0, \Delta^1 = 1 \mid \boldsymbol{X}^{(\boldsymbol{k})} = \boldsymbol{x}) - P(T^0 \leq t_0, \Delta^0 = 1 \mid \boldsymbol{X}^{(\boldsymbol{k})} = \boldsymbol{x})\right] \\ &= \mathbb{E}\left[P(T \leq t_0, \Delta = 1 \mid \boldsymbol{X}^{(\boldsymbol{k})} = \boldsymbol{x}, A_k = 1) - P(T \leq t_0, \Delta = 1 \mid \boldsymbol{X}^{(\boldsymbol{k})} = \boldsymbol{x}, A_k = 0)\right] \\ &= \mathbb{E}\left[\mathbb{E}[\tilde{Y} \mid \boldsymbol{X}^{(\boldsymbol{k})} = \boldsymbol{x}, A_k = 1] - \mathbb{E}[\tilde{Y} \mid \boldsymbol{X}^{(\boldsymbol{k})} = \boldsymbol{x}, A_k = 0]\right].\end{aligned}$$

Similarly, we identify $\theta_{\text{net}}(\boldsymbol{x})$. More details can be found in the supplementary material (Appendix B).

4.2 | Estimation of inverse probability weights

To implement our two-step approach, we need consistent estimators for the nuisance parameters G and G_2 on $[0, t_0]$. Even though inverse probability weighted estimators have been around for many years, there is no genuine solution to balance the bias-variance tradeoff (see also Section 7). In the following, we discuss three different strategies to estimate the nuisance parameters based on, respectively, marginal Kaplan-Meier estimators, stratified Kaplan-Meier estimators and random survival forests. For different choices of (data-adaptive) kernels K these strategies occur as special cases of the following estimator of the conditional survival function of the censoring time distribution:

$$\hat{G}(t \mid \boldsymbol{a}, \boldsymbol{x}) = \prod_{t_k \leq t}\left(1 - \frac{\sum_{i=1}^n \mathbb{1}\{T_i = t_k, \Delta_i = 0\} K(\boldsymbol{a}, \boldsymbol{A}_i, \boldsymbol{x}, \boldsymbol{X}_i)}{\sum_{i=1}^n \left(\mathbb{1}\{T_i \geq t_k\} - \mathbb{1}\{T_i = t_k, \Delta_i > 0\}\right) K(\boldsymbol{a}, \boldsymbol{A}_i, \boldsymbol{x}, \boldsymbol{X}_i)}\right).$$

Ties in the event times are handled with the usual convention that the event of interest happens before competing events and censoring events. Similarly, we estimate G_2 with

$$\hat{G}_2(t \mid \boldsymbol{a}, \boldsymbol{x}) = \prod_{t_k \leq t}\left(1 - \frac{\sum_{i=1}^n \mathbb{1}\{T_i = t_k, \Delta_i = 2\} K_2(\boldsymbol{a}, \boldsymbol{A}_i, \boldsymbol{x}, \boldsymbol{X}_i)}{\sum_{i=1}^n \left(\mathbb{1}\{T_i \geq t_k\} - \mathbb{1}\{T_i = t_k, \Delta_i \neq 2\}\right) K_2(\boldsymbol{a}, \boldsymbol{A}_i, \boldsymbol{x}, \boldsymbol{X}_i)}\right).$$

In some applications, we may be willing to strengthen the identifiability assumptions of Section 3.3. For example, it may be reasonable to assume that the right-censored times are independent of the event time outcome and the covariates, that is, $(T, \Delta, \boldsymbol{A}, \boldsymbol{X}) \perp\!\!\!\perp C$. When this assumption holds, consistent weights for the first step of our algorithm for crude effects can be obtained with the reverse Kaplan-Meier estimator, that is, the Kaplan-Meier for the marginal censoring survival distribution, represented by $K(\boldsymbol{a}, \boldsymbol{A}, \boldsymbol{x}, \boldsymbol{X}) = 1$ and $K_2(\boldsymbol{a}, \boldsymbol{A}, \boldsymbol{x}, \boldsymbol{X}) = 1$. However, in most applications covariates and treatments will have effects on the distribution G_2 and the independence assumption would rarely be plausible for estimating net effects. The following less stringent, but still quite strong, assumptions could then be used: There exists a subset of categorical covariates $\boldsymbol{Z} \subset \{\boldsymbol{A}, \boldsymbol{X}\}$ such that $(T, \Delta) \perp\!\!\!\perp C \mid \boldsymbol{Z}$ and $T^{1,A_k} \perp\!\!\!\perp T^{2,A_k} \mid \boldsymbol{Z}$. Under these assumptions we can estimate the censoring survival distribution function G, conditional on $\boldsymbol{Z}$, with the stratified Kaplan-Meier estimator which corresponds to setting $K(\boldsymbol{a}, \boldsymbol{A}, \boldsymbol{x}, \boldsymbol{X}) = \mathbb{1}\{\boldsymbol{z} = \boldsymbol{Z}\}$ and $K_2(\boldsymbol{a}, \boldsymbol{A}, \boldsymbol{x}, \boldsymbol{X}) = \mathbb{1}\{\boldsymbol{z} = \boldsymbol{Z}\}$.

In general settings, one may need a different more flexible approach. We address the more general case by using a random forest approach and test this choice in our empirical studies. We estimate each of the distributions $G(t|\boldsymbol{X}, \boldsymbol{A})$ and $G_2(t|\boldsymbol{X}, \boldsymbol{A})$ based on a separate random survival forest.[19] Specifically, we draw B bootstrap subsamples without replacement where we set the subsample size to 63.2% of the sample size of the full data. In each bootstrap subsample we grow a survival tree for which we use a specific splitting rule that is based on maximally selected rank statistics.[20] The values of

the further hyper parameters of the random survival forest are discussed in Section 5. For any fixed value $(\boldsymbol{a}, \boldsymbol{x})$, we denote the terminal node of the bth survival tree that contains the value by $\mathcal{T}_b(\boldsymbol{a}, \boldsymbol{x})$. The kernel corresponding to the random survival forest estimator of G (and similar for G_2) is given by

$$K(\boldsymbol{a}, \boldsymbol{A}, \boldsymbol{x}, \boldsymbol{X}) = \sum_{b=1}^{B} \mathbb{1}\{\{\boldsymbol{A}, \boldsymbol{X}\} \in T_b(\boldsymbol{a}, \boldsymbol{x})\}.$$

Note that even though we sample without replacement, a single subject from the full data set still can contribute multiple times to the nominator and denominator of the Nelson-Aalen estimator. This happens when subjects from the full data share the terminal node of multiple survival trees with each other. Note also that the tuning of the hyperparameters of the random survival forest for the first step of our two-step algorithm may require a simulation study which synthesizes the setting of the data application characterized by, for example, the number of treatments, the number of covariates, the sample size, the prevalence of the treatments.

4.3 | Confidence intervals

A delta method argument using the standard errors $\hat{\sigma}_n(\boldsymbol{x})$ for the conditional estimates, as provided by Theorem 5 and Section 6 in Athey et al (2019),[2] yields asymptotic normality of the forest estimators $\hat{\overline{\theta}}_{\text{net},k}$, $\hat{\overline{\theta}}_{\text{crude},k}$ for the average treatment effects. The standard errors obtained in this way do not reflect the fact that the censoring distribution was estimated in Step 1 of our procedure. Although the asymptotic standard errors also contain a contribution from the uncertainty of the weights constructed in Step 1 of our procedure, these contributions are in our experience often very small in real data applications. In our simulations and illustrative data analysis, we only show confidence intervals which ignore the statistical uncertainty due to Step 1. Despite these shortcomings, we note that in our simulation studies (Section 5) the estimated standard errors agree nicely with the empirical standard deviations of estimates across simulation repetitions. It may be possible to obtain standard errors that include the uncertainty from estimating the weights by adapting a bootstrap procedure,[21] or the jackknife-after-bootstrap.[22]

5 | EMPIRICAL STUDIES

To evaluate the performance of our proposed methodology, and as a proof of concept, we test our algorithm on simulated data. Our simulations further compare our method with some existing approaches and illustrate the difference between ranking according to treatment effects on the crude and net probability scales. We here explain the design of the simulations and then show selected simulation results. Further details are provided in Appendix E and reproducible results are available on github, see Section 8. Overall we report results based on $M = 2000$ simulation repetitions to decrease the Monte Carlo error. However, for the sample size $n = 5000$ the Monte Carlo error was much smaller and we kept $M = 1000$ to reduce the computational burden of our study.

5.1 | Design

In all settings we simulate ten binary treatment variables with different prevalences, five binary covariates variables and two continuous covariates. The distribution of the first treatment variable can depend on the other variables according to a logistic regression model. Latent time variables are simulated with Cox-Weibull distributions that allow effects of the treatment variables and covariates via regression in the shape parameter.[23] The observed event times (and the corresponding event status variable) are then constructed as the minimum of the latent event times and the censoring time. The dependence between all variables is controlled by a setting specific list of structural equations using odds ratios and hazard ratios. The true values of the parameters $\overline{\theta}_{\text{net}}$ and $\overline{\theta}_{\text{crude}}$ are obtained by drawing large datasets in the hypothetical world where treatment variables are set to 1 and 0, respectively, for all subjects. Specific values of for example treatment prevalences and functional forms of data-generating mechanisms can be found in Appendix E as well as in the github repository.

5.2 | Estimation methods

In all simulations (except when marked otherwise) we apply our two-step algorithm where in the first step we obtain inverse probability weights based on a random survival forest with 100 survival trees for sample sizes smaller than 5000 and 50 survival trees for sample size equal to 5000. As our splitting rule we use maximally selected rank statistics[20] and we set the number of variables tried at each split to 17 such that all variables are tried at each split.

We compare our algorithm with G-formula estimators that use a Fine-Gray regression model $\hat{F}_1^{\text{FGR}}$[24] and a the combination of two cause-specific Cox regression models $\hat{F}_1^{CSC}$,[25,26] respectively, to estimate the conditional absolute risk of cause one, $F_1(t|\boldsymbol{a},\boldsymbol{x})$. The G-formula estimator[18] of the crude effects of treatment A_k is then given for method $m = (\text{FGR}, CSC)$ by setting the treatment value first to one and then to zero:

$$\hat{\theta}^m_{\text{crude}} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{F}_1^m(t|A_{1,i}\ \ldots\ ,A_{k-1,i},1,A_{k+1,i},\ \ldots\ ,A_{K,i},\boldsymbol{X_i}) - \hat{F}_1^m(t|A_{1,i},\ \ldots\ ,A_{k-1,i},0,A_{k+1,i},\ \ldots\ ,A_{K,i},\boldsymbol{X_i})\}.$$

In the specification of the cause-specific hazards models and the Fine and Gray regression model in our simulation studies we included additive main effects of all covariates and treatments into the linear predictors. In the case where the data generating event time distributions involved a non-linear dependence on a continuous covariate, particularly a squared effect of X_6, these regression models are misspecified.

5.3 | Simulation results

Figure 1 shows crude effect estimation performance in terms of coverage (based on standard error estimates provided by the forest) and bias across different sample sizes, varying amount of censoring by the time-horizon of interest and

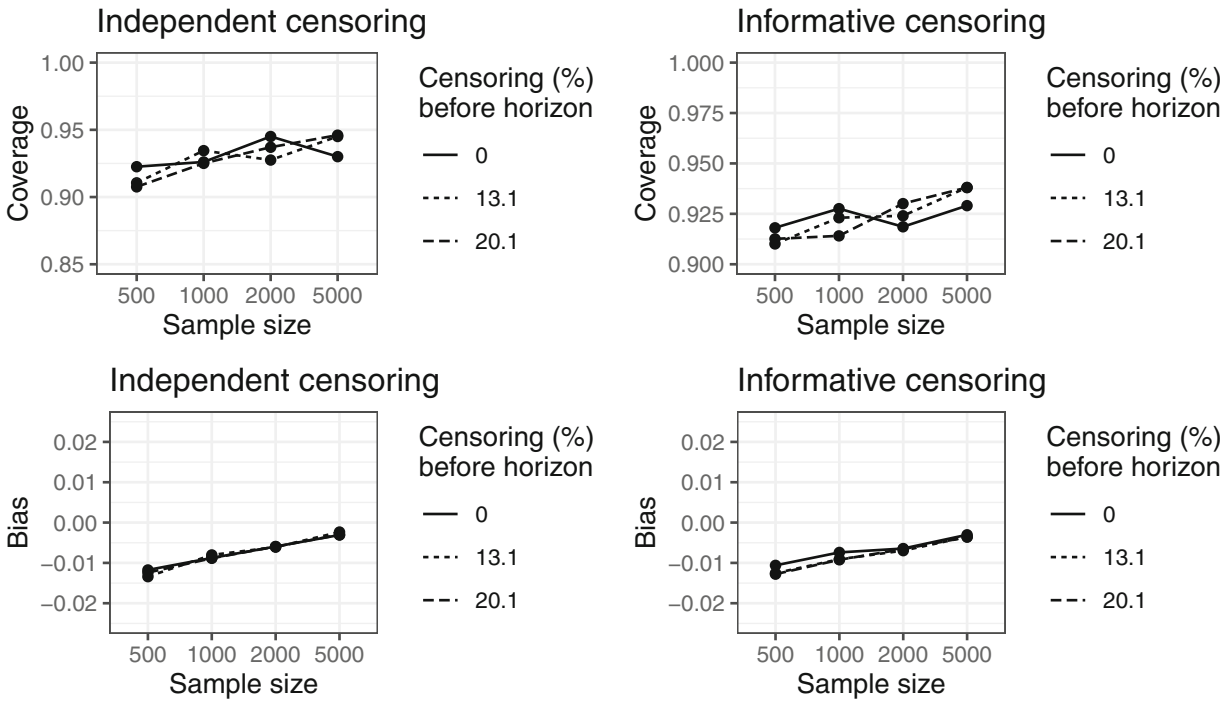

FIGURE 1 Shown are the results from the simulation study for estimation of crude effects ($\bar{\theta}_{\text{crude},A_1}$) of the treatment variable A_1 across simulation repetitions. The figure shows performance in terms of coverage (upper panel) and bias (lower panel), varying the sample size (x-axis), the (in)dependence of censoring on covariates (independent in the left panel, dependent in the right panel) and amount of censoring by the time-horizon

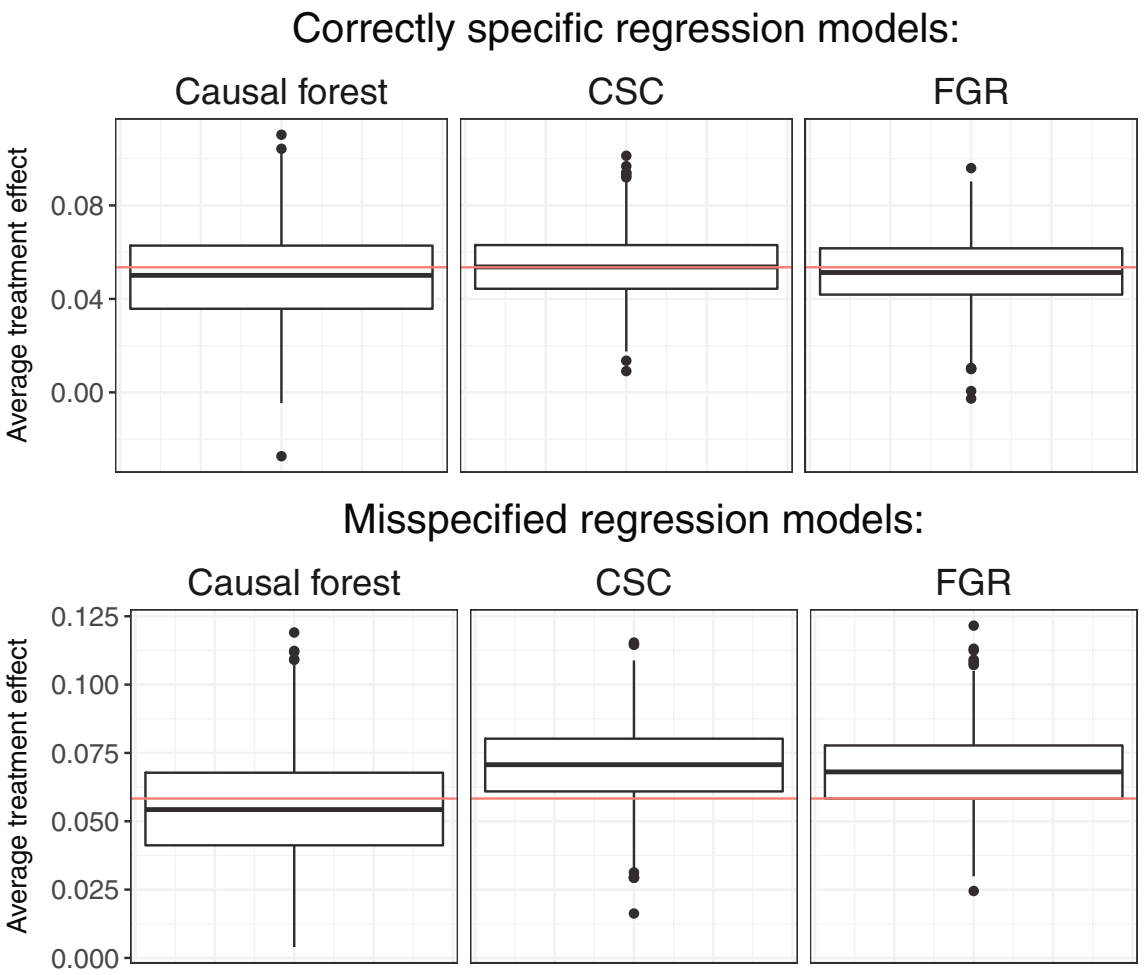

FIGURE 2 Shown are the results from the simulation study for estimation of crude effects ($\bar{\theta}_{\text{crude},A_1}$) of the treatment variable A_1 based on the causal forest approach and the two parametric methods summarized in Section 5.2. The figure shows the distribution of estimates obtained across simulation repetitions when the parametric methods are correctly specified (upper panel), and when they are misspecified (lower panel). We note that the causal forest approach provides unbiased estimation for both settings, but that the regression-based methods are more efficient — since they rely on much smaller statistical models — in the correctly specified case

for covariate dependent and independent censoring. For increasing sample sizes, bias is seen to diminish and coverage to approach 95%. We remark that a certain seed effect was present, giving rise to the non-monotone lines in the upper right panel; also, positivity issues induced by low treatment prevalence and high amount of right-censoring complicates the estimation problem drastically and likely explains the lower coverage seen in this graph. For estimation of the net effects (see Appendix E and github), on the other hand, coverage was lying nicely around 95%. Figure 2 shows the results from applying our causal forest approach to the two comparison methods summarized in Section 5.2 and demonstrates the bias inflicted by the misspecification corresponding to specifying only a linear main effect of X_6.

To assess ranking effectiveness, we calculate the frequency (across simulation repetitions) that a treatment variable is ranked most important among the ten treatment variables according to estimated crude and net effects. Figure 3 shows this frequency for the variables A_1, A_2, A_3 for varying sample sizes and for varying effects of treatment variable A_2 on the latent cause two event time. The figure shows that treatment variable A_1 (which has a moderate effect on the latent time to cause one) is ranked most important with increasing probability for increasing sample size. The figure also shows that when ranking is based on crude effects, and when the variable A_2 has a large effect on the latent time to cause two, A_2 is frequently detected as important. This does not happen when ranking is based on net effects. The variable A_3 has no effect on the latent times of both causes and is not detected as important.

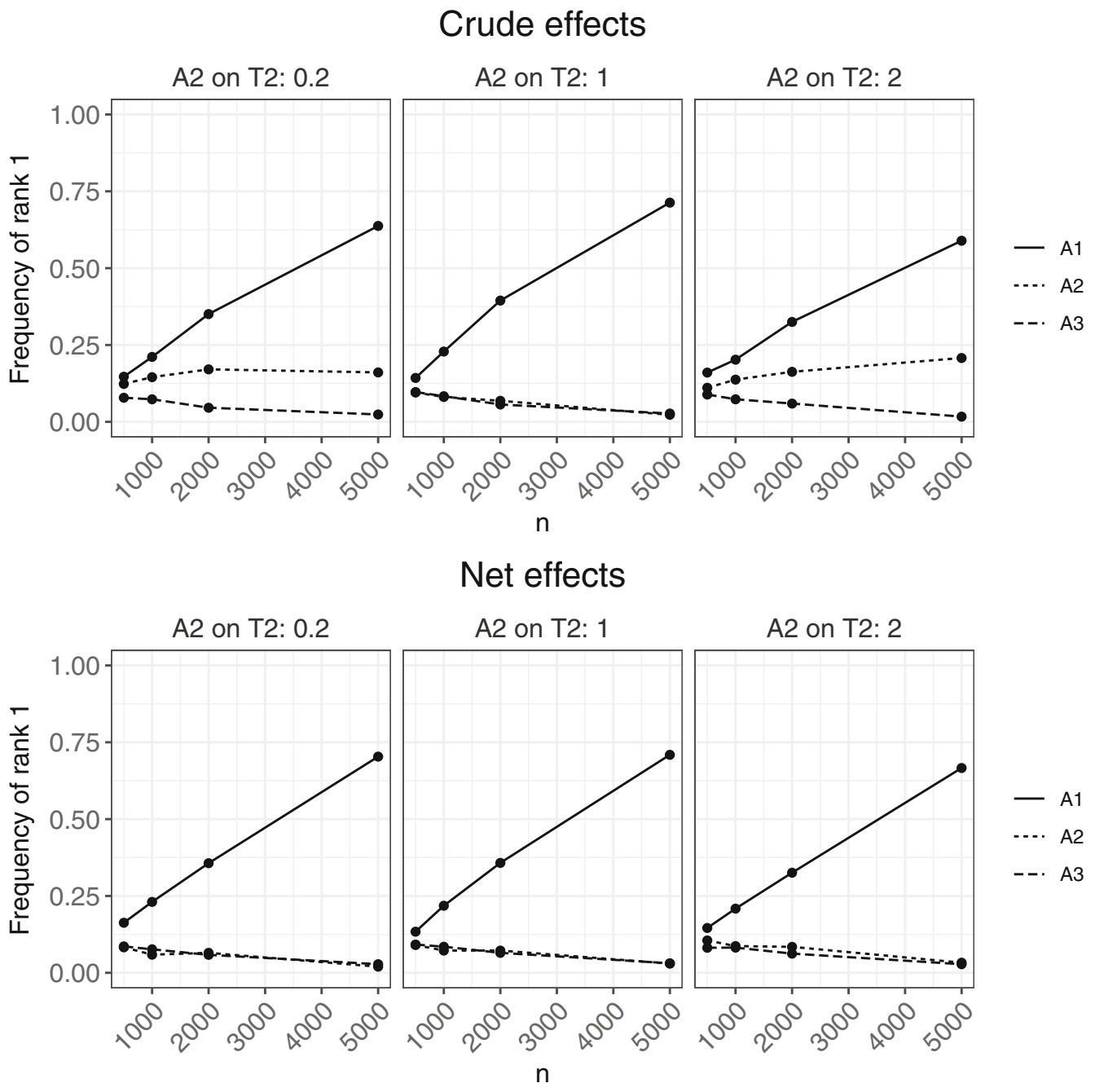

FIGURE 3 Results of the simulation studies. Shown are the fraction of times that the three treatment variables A_1, A_2, A_3 was ranked most important across simulation repetitions. Rankings according to crude effects are shown in the three upper panels which vary the effect (hazard ratio) of the treatment variable A_2 on the latent time to cause two. Similarly, the three lower panels show rankings according to net effects. The hazard ratio of A_1 on the latent time to cause one has a value of 1.25 in all 6 panels

6 | REGISTRY STUDY

We apply our method to our motivating example in which it is of interest to study whether the use of any particular drug decreases the risk of relapse of depression resulting in psychiatric hospitalization. We here report estimates of effects on the net probabilities as well as those on the crude probabilities. Our aim is to discover new active substances; for this purpose, net probabilities will allow us to rank drugs according to their direct effect on depression, isolating this effect from what effect that drug may have on competing events.

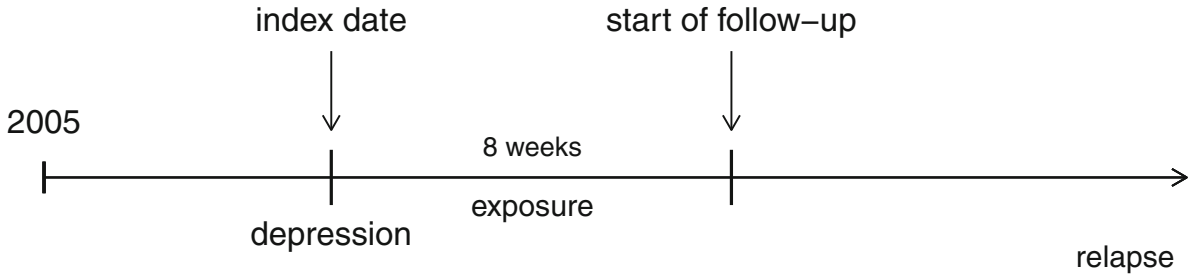

FIGURE 4 Illustration of our study design. The date of first contact with depression is defined as the index date. Patients with a psychiatric hospitalization in the 8 weeks window following the index date are excluded. ATC drug codes are grouped after their first three digits to define binary exposure variables with the value 1 if there was at least one prescribed purchase within the ATC group in the eight weeks window. Information on comorbidity is collected during a ten year period before the index date and included as covariates in the analysis, along with sex and age at the index date

The data we work with are obtained by linking Danish population-based registers that contain data on all prescribed medical purchases at pharmacies since 1995 and data on all patients treated at hospitals since 1977. A total of 78 700 patients were included who all had a first-time admission with depression after 2005. Figure 4 illustrates our design. The date of first contact with depression was defined as the index date. Patients with a psychiatric hospitalization in the eight weeks window following the index date were excluded. The first three digits of ATC codes, a general system to group medical drugs according to the organs and systems on which they work, and how they work, were used to define treatment variables. In this way each treatment variable reflects an anatomic main group (first digit of ATC) and a therapeutic main group (next two digits of ATC). In our analysis, we defined treatment as a binary variable indicating if there was at least one prescribed purchase within the ATC group in the eight weeks window. Information on comorbidity was collected during a ten year period before the index date and included as covariates in the analysis, along with sex and age at the index date. Subjects were followed for 5 years from the end of the exposure window until depression relapse ($\Delta = 1$), competing event ($\Delta = 2$), or loss to follow-up ($\Delta = 0$). Competing risks were death or a diagnosis with shizophrenia or bipolar depression. Summary statistics on comorbidities, exposure and number of events can be found in the supplementary material (Appendix D).

To estimate the treatment effect of each considered drug group A_k on the net and crude probabilities, $\overline{\theta}_{\text{net},A_k}$ and $\overline{\theta}_{\text{crude},A_k}$, the inverse probability weights were adjusted for all treatment variables, sex, age, and comorbidity history. We applied random survival forests as described in Section 4.2 using a splitting rule based on maximally selected rank statistics to estimate the weights G and G_2, respectively. We included all treatments and all covariates (sex, age and comorbidity history) to estimate the weights based on 50 survival trees and setting the number of variables tried at each split to $mtry = 20$. In the causal forest we used $B = 50$ trees and included all variables (treatments, sex, age group and comorbidities). For comparison we include in Appendix F rankings based on the cause-specific Cox regression G-formula approach.

6.1 | Results

Figure 5 shows the causal forest estimates of the effect on net probabilities, $\overline{\theta}_{\text{net}}$, and of the effect on crude probabilities, $\overline{\theta}_{\text{crude}}$, for each drug group. The size of the estimates allows us to rank the treatment groups according to their effect on relapse with depression. As we saw in the simulation study, there can be a substantial difference between ranking based on $\overline{\theta}_{\text{net}}$ and $\overline{\theta}_{\text{crude}}$. Here we see in Figure 5, as well, that the estimates of the two parameters lead to slightly differing conclusions. Consider, for example the drug group 'C08' (calcium channel blockers), 'A03' (drugs for functional gastrointestinal disorders), 'H02' (corticosteroids for systemic use) that are ranked higher in terms of net probabilities than in terms of crude probabilities, or the drug groups 'C07' (beta blocking agents), 'D07' (corticosteroids, dermatological preparations) and 'G03' (sex hormones and modulators of the genital system) that are ranked higher in terms of crude probabilities than in terms of net probabilities. Recall that net effects, if we believe in the assumptions required to go from a crude to a net interpretation (Section 3.3), allow us to rank drugs according to their direct effect on the depression relapse without interference from indirect effects on the competing events. Thus, we can avoid pitfalls

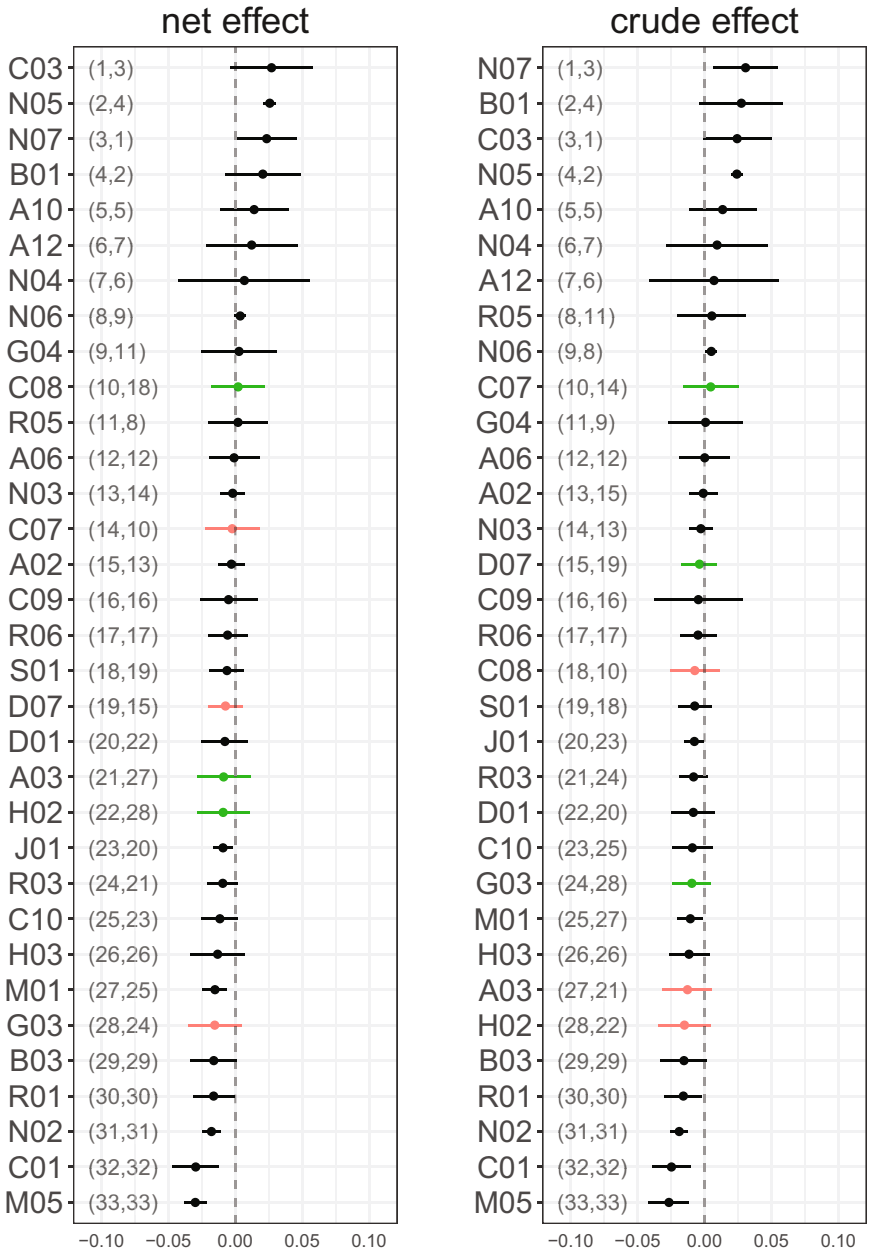

FIGURE 5 Results from the data analysis. Shown are for each ATC group (marked on the *y*-axis) the estimates of $\overline{\theta}_{\text{net}}$ (left) and of $\overline{\theta}_{\text{crude}}$ (right). The numbers written in parenthesis indicate the rank of that ATC group by the current method followed by the other method. The colors highlight ATC groups ranked differently by the two methods

that we see in our simulation study, like reporting a large treatment effect simply if that treatment increases the rate of a competing event.

7 | DISCUSSION

In this paper we have considered average treatment effect estimation for the purpose of ranking treatments according to their effect on a specific time-to-event outcome of interest. We have implemented a data-adaptive estimation method based on generalized random forests, where inverse probability weights are constructed to make the forest implementation directly applicable to the time-to-event setting. Our method makes no parametric model restrictions and really benefits from the flexibility of the generalized random forest which adaptively adjusts the propensity of treatment for covariates. This altogether makes it highly applicable to drug discovery studies with many candidate drug treatments

and not much prior subject matter knowledge. As an illustration, we have considered a particular application where it was of interest to rank a list of treatments according to their effect on depression.

We have discussed two different target parameters in the presence of competing risks, defined in terms of net and crude probabilities, respectively, with different interpretations. Particularly, net probabilities allow us to make inference for treatment directly on the outcome of interest, irrespective of that treatment's effect on the competing risks. We argue for the utility of net probabilities when looking for new active substances as part of a drug discovery study, but emphasize, in accordance with earlier criticism,[27] that they are not sensible interpreting the size of the effect, for example, when counseling a patient. Crude probabilities should always be considered if interest is in the real world and the aim is to predict for a given patient. The methods proposed recently by Reference 28 provide an alternative route for isolating direct effects on the event of interest.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Helene C. W. Rytgaard https://orcid.org/0000-0003-2096-9823

REFERENCES

1. Sudeep P, Francesco I, Eyers Patrick A, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov*. 2019;18(1):41-58.
2. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Stat*. 2019;47(2):1148-1178.
3. Neyman J. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). English translation by DM Dabrowska and TP Speed (1990). *Stat Sci*. 1923;5:465-480.
4. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688.
5. Young Jessica G, Stensrud Mats J, Tchetgen Tchetgen Eric J, Hernán MA. A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat Med*. 2020;39(8):1199-1236.
6. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
7. Ishwaran H. Variable importance in binary regression trees and forests. *Electron J Stat*. 2007;1:519-537.
8. Strobl C, Boulesteix A, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinform*. 2008;9(1):307.
9. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. *J Am Stat Assoc*. 2010;105(489):205-217.
10. Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random survival forests for high-dimensional data. *Stat Anal Data Min*. 2011;4(1):115-132.
11. Tuglus C, Laan MJ. Targeted methods for biomarker discovery, the search for a standard; 2008.
12. Bembom O, Petersen ML, Rhee S, et al. Biomarker discovery using targeted maximum-likelihood estimation: application to the treatment of antiretroviral-resistant HIV infection. *Stat Med*. 2009;28(1):152-172.
13. Wang H, Laan MJ. Dimension reduction with gene expression data using targeted variable importance measurement. *BMC Bioinform*. 2011;12(1):312.
14. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*. 2018;113(523):1228-1242.
15. Yifan C, Kosorok Michael R, Erik S, Stefan W, Ruoqing Z. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv preprint arXiv:2001.09887*; 2020.
16. Chiang CL. A stochastic study of the life table and its applications. III. The follow-up study with the consideration of competing risks. *Biometrics*. 1961;17(1):57-78.
17. Gray RJ. A class of *K*-sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat*. 1988;16(3):1141-1154.
18. Hernan MA, Robins JM. *Causal Inference*. Boca Raton, Fl: Chapman & Hall/CRC; 2020.
19. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860.
20. Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Stat Med*. 2017;36(8):1272-1284.
21. Joseph S, Petter L. Standard errors for bagged and random forest estimators. *Comput Stat Data Anal*. 2009;53(3):801-811.
22. Stefan W, Trevor H, Bradley E. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *J Mach Learn Res*. 2014;15(1):1625-1651.
23. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2005;24:1713-1723.
24. Jason F, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496-509.
25. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics*. 1990;46(3):813-826.
26. Brice O, Lyngholm SA, Thomas S, Christian T-P, Alexander GT. Risk regression: predicting the risk of an event using Cox regression models. *R J*. 2017;9(2):440-460.
27. Kragh AP, Niels K. Interpretability and importance of functionals in competing risks and multistate models. *Stat Med*. 2012;31(11-12):1074-1088.

28. Stensrud Mats J, Young Jessica G, Vanessa D, Robins James M, Hernán MA. Separable effects for causal inference in the presence of competing events. *J Am Stat Assoc*. 2022;117(537):175-183.
29. Gill RD, Laan MJ, Robins JM. Coarsening at random: characterizations, conjectures, counter-examples. Proceedings of the First Seattle Symposium in Biostatistics, Springer; 1997:255–294.
30. Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer Science and Business Media; 2003.
31. Tsiatis A. *Semiparametric Theory and Missing Data*. New York: Springer Science and Business Media; 2007.
32. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am Stat Assoc*. 1999;94(448):1096-1120.
33. Rosenblum M, Laan MJ. Simple examples of estimating causal effects using targeted maximum likelihood estimation; 2011.
34. R Core Team. *R: A Language and Environment for Statistical Computing*. Austria: R Foundation for Statistical ComputingVienna; 2021.

SUPPORTING INFORMATION

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

How to cite this article: Rytgaard HCW, Ekstrøm CT, Kessing LV, Gerds TA. Ranking of average treatment effects with generalized random forests for time-to-event outcomes. *Statistics in Medicine*. 2023;42(10):1542-1564. doi: 10.1002/sim.9686

APPENDIX A

We here detail the identifiability assumptions for the effect on net probabilities and the effect on crude probabilities, respectively. For simplicity of presentation we restrict the setting in all appendices to a single treatment variable A_k and consider the other treatment variables as part of the covariate vector $\boldsymbol{X}$.

A.1 Identifiability assumptions for the effect on net probabilities

Identification of $\theta_{\text{net}}(\boldsymbol{x})$ in terms of the observed data distribution depends on three untestable causal assumptions: Consistency, coarsening at random and positivity.

First, the assumption of consistency entails that the counterfactual event time $T^{1,a}$ ($T^{2,a}$) corresponds to the observed event time for those subjects who were actually uncensored, free of event type $j = 2$ ($j = 1$) and were exposed to the treatment level $A_k = a$. Particularly, consistency provides the counterfactual variables as follows:

$$T = \min(T^{1,A_k}, T^{2,A_k}), \text{ and } T^{2,a} = T^{2,A_k} \text{ and } T^{1,a} = T^{1,A_k} \text{ on the event that } A_k = a \text{ for } a = 0, 1. \tag{1a}$$

Here T^{2,A_k} is the uncensored counterfactual event time of type $j = 2$ under the observed treatment.

The second assumption of coarsening at random is characterized as follows. The full data we would have liked at observe are $(\boldsymbol{X}, T^{1,0}, T^{1,1})$. These are not fully observed due to censoring, the competing event and the treatment decision A_k, and we observe only the coarsened data $(\boldsymbol{X}, A_k, \tilde{T}, \tilde{\Delta})$.[29-31] To identify $(\boldsymbol{X}, T^{1,0}, T^{1,1})$ from the data, we need coarsening at random (CAR),[29,30] i.e., that the coarsening mechanism only depends on the full data structure $(\boldsymbol{X}, T^{1,0}, T^{1,1})$ through the observed data structure $(\boldsymbol{X}, A_k, \tilde{T}, \tilde{\Delta})$. Coarsening at random is implied by the following conditional independence conditions:

$$\begin{aligned} &T^{1,a} \perp\!\!\!\perp A_k \mid \boldsymbol{X}, \\ &T^{1,A_k} \perp\!\!\!\perp (C, T^{2,A_k}) \mid A_k, \boldsymbol{X}, \end{aligned} \tag{1b}$$

for $a = 0, 1$, also refer to as "no unmeasured confounding".

The last assumption of positivity requires for the coarsening mechanism that

$$P(\min(C, T^{2,A_k}) \geq t_0 \mid A_k, \boldsymbol{X}) \, (\pi_k(\boldsymbol{X}))^{A_k} (1 - \pi_k(\boldsymbol{X}))^{1-A_k} > \eta > 0, \tag{1c}$$

almost surely.

Under Assumptions 1a,1b and 1c, we can link the distribution of the counterfactual variables to the observed data distribution as follows:

$$\begin{aligned}
&P(\tilde{T} \in dt, \tilde{\Delta} = 1, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\
&\quad = P(\tilde{\Delta} = 1 \mid T^{1,A_k} = t, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T^{1,A_k} \in dt, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\
&\quad = P(\min(C, T^{2,A_k}) \geq t \mid T^{1,A_k} = t, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T^{1,A_k} \in dt \mid A_k = a, \boldsymbol{X} = \boldsymbol{x}) \\
&P(A_k = a \mid \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} \in d\boldsymbol{x}) \\
&\quad = P(\min(C, T^{2,A_k}) \geq t \mid A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T^{1,a} \in dt \mid \boldsymbol{X} = \boldsymbol{x}) P(A_k = a \mid \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} \in d\boldsymbol{x}). \qquad \text{(A1)}
\end{aligned}$$

Particularly, the first line of Assumption 1b together with the Assumption 1a of consistency implies that

$$\begin{aligned}
P(T^{1,A_k} \in dt \mid A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) &\overset{1a}{=} P(T^{1,a} \in dt \mid A_k = a, \boldsymbol{X} = \boldsymbol{x}) \\
&\overset{1b}{=} P(T^{1,a} \in dt \mid \boldsymbol{X} = \boldsymbol{x}),
\end{aligned}$$

whereas the second line of Assumption 1b yields that

$$P(\min(C, T^{2,A_k}) \geq t \mid T = t, \Delta = 1, A_k = a, \boldsymbol{X} = \boldsymbol{x}) = P(\min(C, T^{2,A_k}) \geq t \mid A_k = a, \boldsymbol{X} = \boldsymbol{x}).$$

Assumption 1c ensures that the right hand side of (A2) is non-zero and well-defined.

A.2 Identifiability assumptions for the effect on crude probabilities

The assumptions needed to identify $\theta_{\text{crude}}(\boldsymbol{x})$ are less restrictive than those needed for $\theta_{\text{net}}(\boldsymbol{x})$ and correspond to the standard setting for right-censored survival times. The consistency assumption for $\theta_{\text{crude}}(\boldsymbol{x})$ can be expressed as

$$T = T^a \text{ and } \Delta = \Delta^a \text{ on the event that } A = a, \text{ for } a = 0, 1. \qquad \text{(2a)}$$

The full data we would have liked to observe are $(\boldsymbol{X}, T^0, T^1, \Delta^0, \Delta^1)$, but we observe only the coarsened data $(\boldsymbol{X}, A_k, \tilde{T}, \tilde{\Delta})$ due to censoring C and treatment decision A_k. The equivalent of Assumption 1b,

$$\begin{aligned}
&(T^a, \Delta^a) \perp\!\!\!\perp A_k \mid \boldsymbol{X}, \quad \text{for } a = 0, 1, \\
&(T, \Delta) \perp\!\!\!\perp C \mid A_k, \boldsymbol{X}, \qquad \text{(2b)}
\end{aligned}$$

yields coarsening at random. We further make the positivity assumption that,

$$P(C \geq t_0 \mid A_k = a, \boldsymbol{X}) \, (\pi_k(\boldsymbol{X}))^a (1 - \pi_k(\boldsymbol{X}))^{1-a} > \eta > 0, \qquad \text{(2c)}$$

almost surely, for $a = 0, 1$.

We can now express the observed data distribution as,

$$\begin{aligned}
&P(\tilde{T} \in dt, \tilde{\Delta} = 1, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\
&\quad = P(\tilde{\Delta} \geq 1 \mid T = t, \Delta = 1, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T \in dt, \Delta = 1, A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\
&\quad = P(C \geq t \mid T = t, \Delta = 1, A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T \in dt, \Delta = 1 \mid A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) \\
&P(A_k = a \mid \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} \in d\boldsymbol{x}) \\
&\quad = P(C \geq t \mid A_k = a, \boldsymbol{X} = \boldsymbol{x}) P(T^a \in dt, \Delta^a = 1 \mid \boldsymbol{X} = \boldsymbol{x}) P(A_k = a \mid \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} \in d\boldsymbol{x}), \qquad \text{(A2)}
\end{aligned}$$

relying on Assumptions 2a,2b and 2c. Particularly, the first line of Assumption 2b together with the Assumption 2a of consistency implies that

$$\begin{aligned}
P(T \in dt, \Delta = 1 \mid A_k = a, \boldsymbol{X} \in d\boldsymbol{x}) &\overset{2a}{=} P(T^a \in dt, \Delta^a = 1 \mid A_k = a, \boldsymbol{X} = \boldsymbol{x}) \\
&\overset{2b}{=} P(T^a \in dt, \Delta^a = 1 \mid \boldsymbol{X} = \boldsymbol{x}),
\end{aligned}$$

whereas the second line of Assumption 2b yields that

$$P(C \geq t \mid T = t, \Delta = 1, A_k = a, \boldsymbol{X} = \boldsymbol{x}) = P(C \geq t \mid A_k = a, \boldsymbol{X} = \boldsymbol{x}).$$

Assumption 2c ensures that the right hand side of (A2) is non-zero and well-defined.

APPENDIX B

B.1 Weighted outcome for net probabilities

Define the weighted outcome:

$$\tilde{Y}' = \frac{\mathbb{1}\{\tilde{T} \leq t_0, \tilde{\Delta} = 1\}}{G'(\tilde{T} - \mid A_k, \boldsymbol{X})},$$

with weights given by

$$G'(\tilde{T} - \mid A_k, \boldsymbol{X}) = P(\min(T^{2,A_k}, C) \geq t \mid A_k, \boldsymbol{X}).$$

For this weighted outcome we have that,

$$\begin{aligned}\theta_{\text{net}}(\boldsymbol{x}) &= P(T^{1,1} \leq t_0 \mid \boldsymbol{X} = \boldsymbol{x}) - P(T^{1,0} \leq t_0 \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}\left[\tilde{Y}' \,\middle|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 1\right] - \mathbb{E}\left[\tilde{Y}' \,\middle|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 0\right]. \end{aligned} \tag{B1}$$

This follows straightforwardly by the identification in, and just after, Equation (A1); indeed, we note that

$$\begin{aligned}\mathbb{E}[\tilde{Y}' \mid \boldsymbol{X}, A_k] &= \mathbb{E}\left[\frac{\mathbb{1}\{\tilde{T} \leq t_0, \tilde{\Delta} = 1\}}{G'(\tilde{T} - \mid \boldsymbol{X}, A_k)} \,\middle|\, \boldsymbol{X}, A_k\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\{T^{1,A_k} \leq t_0\}\mathbb{1}\{\tilde{\Delta} = 1\}}{G'(\tilde{T} - \mid \boldsymbol{X}, A_k)} \,\middle|\, T^{1,A_k}, \boldsymbol{X}, A_k\right] \,\middle|\, \boldsymbol{X}, A_k\right] \\ &= \mathbb{E}\left[\mathbb{1}\{T^{1,A_k} \leq t_0\} \mid \boldsymbol{X}, A_k\right] \mathbb{E}\left[\frac{\mathbb{E}[\mathbb{1}\{\tilde{\Delta} = 1\} \mid T^{1,A_k}, \boldsymbol{X}, A_k]}{G'(T^{1,A_k} - \mid \boldsymbol{X}, A_k)}\right] \\ &= \mathbb{E}\left[\mathbb{1}\{T^{1,A_k} \leq t_0\} \mid \boldsymbol{X}, A_k\right] \mathbb{E}\left[\frac{G'(T^{1,A_k} - \mid \boldsymbol{X}, A_k)}{G'(T^{1,A_k} - \mid \boldsymbol{X}, A_k)}\right] \\ &= \mathbb{E}[\mathbb{1}\{T^{1,A_k} \leq t_0 \mid \boldsymbol{X}, A_k],\end{aligned}$$

and

$$\mathbb{E}[\mathbb{1}\{T^{1,A_k} \leq t_0 \mid \boldsymbol{X}, A_k = a] = \mathbb{E}[\mathbb{1}\{T^{1,a} \leq t_0\} \mid \boldsymbol{X}, A_k = a] = \mathbb{E}[\mathbb{1}\{T^{1,a} \leq t_0\} \mid \boldsymbol{X}],$$

for $a = 0, 1$, which yields (B1).

B.2 Weighted outcome for crude probabilities

For the weighted outcome,

$$\tilde{Y} = \frac{\mathbb{1}\{\tilde{T} \leq t_0, \tilde{\Delta} = 1\}}{G(\tilde{T} - \mid A_k, \boldsymbol{X})},$$

we have that,

$$\begin{aligned}\theta_{\text{crude}}(\boldsymbol{x}) &= P(T^1 \leq t_0, \Delta^1 = 1 \mid \boldsymbol{X} = \boldsymbol{x}) - P(T^0 \leq t_0, \Delta^0 = 1 \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}\left[\tilde{Y} \,\middle|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 1\right] - \mathbb{E}\left[\tilde{Y} \,\middle|\, \boldsymbol{X} = \boldsymbol{x}, A_k = 0\right]. \end{aligned} \tag{B2}$$

This follows straightforwardly by the identification in, and just after, Equation (A2); indeed, we note that

$$\begin{aligned}
\mathbb{E}[\tilde{Y} \mid \boldsymbol{X}, A_k] &= \mathbb{E}\left[\frac{\mathbb{1}\{\tilde{T} \le t_0, \tilde{\Delta} = 1\}}{G(\tilde{T} - \mid \boldsymbol{X}, A_k)} \;\middle|\; \boldsymbol{X}, A_k\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\{T \le t_0, \Delta = 1\}\mathbb{1}\{\tilde{\Delta} \ge 1\}}{G(T - \mid \boldsymbol{X}, A_k)} \;\middle|\; T, \Delta, \boldsymbol{X}, A_k\right] \;\middle|\; \boldsymbol{X}, A_k\right] \\
&= \mathbb{E}\left[\mathbb{1}\{T \le t_0, \Delta = 1\} \frac{\mathbb{E}\left[\mathbb{1}\{\tilde{\Delta} \ge 1\} \mid T, \Delta, \boldsymbol{X}, A_k\right]}{G(T - \mid \boldsymbol{X}, A_k)} \;\middle|\; \boldsymbol{X}, A_k\right] \\
&= \mathbb{E}\left[\mathbb{1}\{T \le t_0, \Delta = 1\} \frac{G(T - \mid \boldsymbol{X}, A_k)}{G(T - \mid \boldsymbol{X}, A_k)} \;\middle|\; \boldsymbol{X}, A_k\right] \\
&= \mathbb{E}\left[\mathbb{1}\{T \le t_0, \Delta = 1\} \mid \boldsymbol{X}, A_k\right]
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\mathbb{1}\{T \le t_0, \Delta = 1\} \mid \boldsymbol{X}, A_k = a] &= \mathbb{E}[\mathbb{1}\{T^a \le t_0, \Delta^a = 1\} \mid \boldsymbol{X}, A_k = a] \\
&= \mathbb{E}[\mathbb{1}\{T^a \le t_0, \Delta^a = 1\} \mid \boldsymbol{X}],
\end{aligned}$$

for $a = 0, 1$, which yields (B2).

APPENDIX C

To explain the general idea of GRFs, we use a generic (uncensored) random variable $Y \in \mathbb{R}$ and a corresponding generic parameter of interest,

$$\theta(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}],$$

representing the treatment effect of A_k on Y conditional on $\boldsymbol{X} = \boldsymbol{x}$.[2] consider a conditional average partial effect estimation problem which they formulate in terms of a structural model. Below we demonstrate the equivalence of their setting with the counterfactual formulation and show that the conditional average treatment effect estimation problem considered here is a special case. In particular, we show that the parameter $\theta(\boldsymbol{x})$ can be identified in terms of

$$\theta(\boldsymbol{x}) = \frac{\mathrm{cov}(A_k, Y \mid \boldsymbol{X} = \boldsymbol{x})}{\mathrm{Var}(A_k \mid \boldsymbol{X} = \boldsymbol{x})}. \tag{C1}$$

This means that $\theta(\boldsymbol{x})$ can be estimated by providing estimators for $\mathrm{cov}(A_k, Y \mid \boldsymbol{X} = \boldsymbol{x})$ and $\mathrm{Var}(A_k \mid \boldsymbol{X} = \boldsymbol{x})$, respectively. The forest outputs weights that can be used to define such estimators as follows. First, forest weights are obtained by averaging over the neighborhoods $L_b(\boldsymbol{x})$ defined by the trees, $b = 1, \ldots, B$,

$$\alpha_i(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{b,i}(\boldsymbol{x}), \qquad \text{where,} \quad \alpha_{b,i}(\boldsymbol{x}) = \frac{\mathbb{1}\{\boldsymbol{X}_i \in L_b(\boldsymbol{x})\}}{\sum_{k=1}^{n} \mathbb{1}\{\boldsymbol{X}_k \in L_b(\boldsymbol{x})\}}. \tag{C2}$$

Then, the forest estimator $\hat{\theta}_\alpha(\boldsymbol{x})$ is given by,

$$\hat{\theta}_\alpha(\boldsymbol{x}) = \left(\sum_{i=1}^{n} \alpha_i(\boldsymbol{x})\left(A_i - \bar{A}_{k,\alpha}\right)^2\right)^{-1} \left(\sum_{i=1}^{n} \alpha_i(\boldsymbol{x})\left(A_i - \bar{A}_{k,\alpha}\right)\left(Y_i - \overline{Y}_\alpha\right)\right). \tag{C3}$$

Here, $\bar{A}_{k,\alpha} = \sum_{i=1}^{n} \alpha_i(\boldsymbol{x}) A_i$ and $\overline{Y}_\alpha = \sum_{i=1}^{n} \alpha_i(\boldsymbol{x}) Y_i$ are estimators for the propensity score $\pi_k(\boldsymbol{x}) = \mathbb{E}[A_k \mid \boldsymbol{X} = \boldsymbol{x}]$ and for $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$, respectively. Theorem 5 and Section 6 of Athey et al 2019[2] provide conditions under which $\hat{\theta}_\alpha$ converges in distribution to a normal distribution centered around the true $\theta(\boldsymbol{x})$. They further propose an estimator $\hat{\sigma}_n(\boldsymbol{x})$ for the standard deviation of the asymptotic distribution.

A key part of the generalized random forest algorithm is the splitting rule that targets specifically estimation of the quantity of interest $\theta(\boldsymbol{x})$. Each split starts with a mother node $M \subset \mathcal{X}$, corresponding to a subset of $\mathcal{X}$, that is to be split into two daughter nodes $D_1 \cup D_2 = M$. For $l = 1, 2$, let $\hat{\theta}_{D_l}$ be the daughter node local estimate of $\theta(\boldsymbol{x})$ given by (C3) with $\alpha_i(\boldsymbol{x}) = \mathbb{1}\{\boldsymbol{X}_i \in D_l\}$ that simply gives weight one to all samples falling in the respective daughter node. To derive their approximate criterion for picking good splits,[2] use a gradient-based approximation of the mother node estimator $\hat{\theta}_M$. In the setting without censoring and competing risks, as we demonstrate below, it can be seen that the "pseudo-outcomes" used in the "labeling step" of the splitting rule correspond to mother node specific estimates of the efficient influence function for the target parameter. Specifically, the split criterion is based on,

$$\rho_i = W_M^{-1}(A_{k,i} - \bar{A}_M)\left(Y_i - \overline{Y}_M - \left(A_{k,i} - \bar{A}_M\right)\hat{\theta}_M\right), \tag{C4}$$

where,

$$W_M = \frac{1}{\#\{i \ : \ \boldsymbol{X}_i \in M\}} \sum_{\{i \ : \ X_i \in M\}} (A_{k,i} - \bar{A}_M)^2,$$

and $\bar{A}_M$, $\overline{Y}_M$ are mother node averages. Each split of a mother node M into daughter nodes D_1, D_2 is carried out such as to maximize,

$$\tilde{\mathcal{L}}(D_1, D_2) = \sum_{l=1}^{2} \frac{1}{\#\{i \ : \ \boldsymbol{X}_i \in D_l\}} \left(\sum_{i \in \{i \ : \ X_i \in D_l\}} \rho_i \right)^2,$$

with ρ_i as defined in (C4).

C.1 Equivalence between counterfactual formulation and structural model formulation

We demonstrate the equivalence of the setting of section 6 in Athey et al 2019[2] with the counterfactual formulation and show that the conditional average treatment effect estimation problem considered in the main paper (Section 4) is a special case hereof.

Accordingly, we here consider observed data $O = (\boldsymbol{X}, A_k, Y), \boldsymbol{X} \in \mathcal{X}, A_k \in \{0, 1\}$ and $Y \in \mathbb{R}$ (uncensored). Further, let Y^1 be the counterfactual outcome that would have been observed under $A_k = 1$, and Y^0 be the counterfactual outcome that would have been observed under $A_k = 0$. The consistency assumption states that

$$Y = A_k Y^1 + (1 - A_k) Y^0, \tag{C5}$$

and the exogeneity assumption (no unmeasured confounding) that $(Y^1, Y^0) \perp\!\!\!\perp A_k \mid \boldsymbol{X}$. The conditional treatment effect is defined as,

$$\theta(\boldsymbol{x}) = \mathbb{E}[Y^1 - Y^0 \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}].$$

The second equality follows under the exogeneity assumption together with the consistency assumption.

Assume on the other hand that,

$$Y_i = a_i + b_i A_k + \varepsilon_i, \tag{C6}$$

in correspondence with section 6 of Athey et al 2019[2] with our $a_i + \varepsilon_i$ collapsed into just ε_i.

We show that (C6) imposes no restriction when A_k is binary. Under consistency, we can express Y as,

$$\begin{aligned}
Y &= A_k Y^1 + (1 - A_k) Y^0 \\
&= A_k Y^1 + (1 - A_k) Y^0 + A_k \left(\mathbb{E}[Y^1 \mid \boldsymbol{X}] - \mathbb{E}[Y^0 \mid \boldsymbol{X}]\right) - A_k \mathbb{E}[Y^1 \mid \boldsymbol{X}] - (1 - A_k)\mathbb{E}[Y^0 \mid \boldsymbol{X}] \\
&\quad + \mathbb{E}[Y^0 \mid \boldsymbol{X}] \\
&= \mathbb{E}[Y^0 \mid \boldsymbol{X}] + A_k \left(\mathbb{E}[Y^1 \mid \boldsymbol{X}] - \mathbb{E}[Y^0 \mid \boldsymbol{X}]\right) + A_k \left(Y^1 - \mathbb{E}[Y^1 \mid \boldsymbol{X}] + (1 - A_k)\left(Y^0 - \mathbb{E}[Y^0 \mid \boldsymbol{X}]\right)\right).
\end{aligned}$$

So if we let,

$$\begin{aligned} a_i &:= \mathbb{E}[Y^0 \mid \boldsymbol{X}_i], \\ b_i &:= \mathbb{E}[Y^1 \mid \boldsymbol{X}_i] - \mathbb{E}[Y^0 \mid \boldsymbol{X}_i], \quad \text{and,} \\ \varepsilon_i &:= (1 - A_{k,i}) \left(Y^0 - \mathbb{E}[Y^0 \mid \boldsymbol{X}_i]\right) + A_{k,i} \left(Y^1 - \mathbb{E}[Y^1 \mid \boldsymbol{X}_i]\right), \end{aligned}$$

we are back on the form in (C6).

Further note that,

$$\mathbb{E}[\varepsilon_i \mid A_k, \boldsymbol{X}] = (1 - A_{k,i}) \left(\mathbb{E}[Y^0 \mid \boldsymbol{X}_i] - \mathbb{E}[Y^0 \mid \boldsymbol{X}_i]\right) + A_{k,i} \left(\mathbb{E}[Y^1 \mid \boldsymbol{X}_i] - \mathbb{E}[Y^1 \mid \boldsymbol{X}_i]\right) = 0,$$

so that,

$$\mathbb{E}[Y_i \mid A_k, \boldsymbol{X}] = \mathbb{E}[Y^0 \mid \boldsymbol{X}_i] + \theta(\boldsymbol{x}) A_k.$$

C.2 Identification of the target parameter

We demonstrate that,

$$\theta(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] = \frac{\operatorname{cov}(A_k, Y \mid \boldsymbol{X} = \boldsymbol{x})}{\operatorname{Var}(A_k \mid \boldsymbol{X} = \boldsymbol{x})}. \tag{C7}$$

This follows since $A_k \in \{0, 1\}$, so that we have:

$$\begin{aligned} \operatorname{cov}(A_k, Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}[A_k Y \mid \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[A_k \mid \boldsymbol{X} = \boldsymbol{x}]\, \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] \\ &= \mathbb{E}[A_k Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x}) \\ &\quad + \mathbb{E}[A_k Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\, (1 - \pi_k(\boldsymbol{x})) - \pi_k(\boldsymbol{x})\, \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] \\ &= \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x}) - \pi_k(\boldsymbol{x}) \Big(\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x}) \\ &\quad + \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\, (1 - \pi_k(\boldsymbol{x})) \Big) \\ &= \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x})(1 - \pi_k(\boldsymbol{x})) \\ &\quad - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\, \pi_k(\boldsymbol{x})(1 - \pi_k(\boldsymbol{x})) \\ &= (\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}])\ \pi_k(\boldsymbol{x})(1 - \pi_k(\boldsymbol{x})), \\ &= (\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}])\ \operatorname{Var}(A_k \mid \boldsymbol{X} = \boldsymbol{x}), \end{aligned}$$

which yields (C7).

C.3 Influence function used for splitting

The influence function used to split in the GRF algorithm for estimation of treatment effects in section 6 of Athey et al (2019)[2] is,

$$\rho_i = W_M^{-1} (A_{k,i} - \bar{A}_M) \left(Y_i - \overline{Y}_M - \left(A_{k,i} - \bar{A}_M\right) \hat{\theta}_M\right), \tag{C8}$$

where,

$$W_M = \frac{1}{\#\{i \ : \ \boldsymbol{X}_i \in M\}} \sum_{\{i \ : \ X_i \in M\}} (A_{k,i} - \bar{A}_M)^2,$$

and $\bar{A}_M$, $\overline{Y}_M$ are mother node averages. Note that ρ_i in (C8) is a mother node specific estimator for,

$$\phi(Y, A_k) = (\operatorname{Var}(A_k \mid \boldsymbol{x}))^{-1} (A_k - \pi_k(\boldsymbol{x})) (Y - \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - (A_k - \pi_k(\boldsymbol{x}))\, \theta(\boldsymbol{x})). \tag{C9}$$

We here demonstrate that $\phi(Y, A_k)$ in (C9) can also be written,

$$\phi(Y, A_k) = \left(\frac{A_k}{\pi_k(\boldsymbol{x})} - \frac{1 - A_k}{1 - \pi_k(\boldsymbol{x})} \right) (Y - \mathbb{E}[Y \mid A_k, \boldsymbol{X} = \boldsymbol{x}]), \tag{C10}$$

which we recognize as the efficient influence function for estimation of the parameter $\theta(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]$.[32,33]

First note that $\text{Var}(A_k \mid \boldsymbol{x}) = (1 - \pi_k(\boldsymbol{x}))\pi_k(\boldsymbol{x})$ since A_k is binary. Next, by iterated expectations, we have that,

$$\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]\ \pi_k(\boldsymbol{x}) + \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\ (1 - \pi_k(\boldsymbol{x})).$$

Moreover, we can write $(A_k - \pi_k) = A_k\ (1 - \pi_k) + (1 - A_k)\ \pi_k$. Also recall that $\theta_{\text{net}}(\boldsymbol{x}) = \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]$.

Now rewrite,

$$\begin{aligned} \frac{A_k}{\pi_k(\boldsymbol{x})} - \frac{(1 - A_k)}{1 - \pi_k(\boldsymbol{x})} &= \frac{A_k(1 - \pi_k(\boldsymbol{x}))}{\pi_k(\boldsymbol{x})(1 - \pi_k(\boldsymbol{x}))} - \frac{(1 - A_k)\pi_k(\boldsymbol{x})}{(1 - \pi_k(\boldsymbol{x}))\pi_k(\boldsymbol{x})} \\ &= (A_k - \pi_k(\boldsymbol{x}))\ (\text{Var}(A_k \mid \boldsymbol{x}))^{-1}, \end{aligned}$$

and,

$$\begin{aligned} &\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] \\ &\quad = \pi_k(\boldsymbol{x})\ \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] + (1 - \pi_k(\boldsymbol{x}))\ \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] \\ &\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}](1 - \pi_k(\boldsymbol{x})) + (1 - \pi_k(\boldsymbol{x}))\ \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] \\ &\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - (1 - \pi_k(\boldsymbol{x}))\,(\mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]) \\ &\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] + (1 - \pi_k(\boldsymbol{x}))\,(\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]), \end{aligned}$$

and likewise,

$$\begin{aligned} &\mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] \\ &\quad = \pi_k(\boldsymbol{x})\ \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] + (1 - \pi_k(\boldsymbol{x}))\ \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] \\ &\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] + \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]\ \pi_k(\boldsymbol{x}) - \pi_k(\boldsymbol{x})\ \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] \\ &\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - \pi_k(\boldsymbol{x})\,(\mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}]) \\ &\quad = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] + (0 - \pi_k(\boldsymbol{x}))\,(\mathbb{E}[Y \mid A_k = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid A_k = 0, \boldsymbol{X} = \boldsymbol{x}]). \end{aligned}$$

Collecting the above, we rewrite (C10) as,

$$\begin{aligned} \phi(Y, A_k) &= \left(\frac{A_k}{\pi_k(\boldsymbol{x})} - \frac{1 - A_k}{1 - \pi_k(\boldsymbol{x})} \right) (Y - \mathbb{E}[Y \mid A_k, \boldsymbol{X} = \boldsymbol{x}]) \\ &= (A_k - \pi_k(\boldsymbol{x}))\ (\text{Var}(A_k \mid \boldsymbol{x}))^{-1}\,(Y - \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - (A_k - \pi_k(\boldsymbol{x}))\ \theta(\boldsymbol{x})), \end{aligned}$$

which yields (C9).

APPENDIX D

We here collect descriptive statistics for our data analysis.

Table D1 shows the number of subjects in each age group and in each comorbidity group. Table D2 shows the number of subjects exposed to the different drug groups in the exposure window. Table D3 shows the number of relapse with depression within 5 years, along with number of subjects who die without depression.

TABLE D1 Comorbidities and demographics of the Danish population-based registry study

	Male ($n = 28\,748$)	Female ($n = 49\,952$)	Total ($n = 78\,700$)
Infections	6429 (22.4)	14 183 (28.4)	20 612 (26.2)
Neoplasms	3775 (13.1)	9419 (18.9)	13 194 (16.8)
Diseases of blood	674 (2.3)	1403 (2.8)	2077 (2.6)
Diseases of the nervous system	6560 (22.8)	11 836 (23.7)	18 396 (23.4)
Diseases of the circulatory or respiratory system	9446 (32.9)	16 408 (32.8)	25 854 (32.9)
Nutritional and metabolic diseases	6863 (23.9)	11 752 (23.5)	18 615 (23.7)
Diseases of the skin and subcutaneous tissue	2040 (7.1)	3752 (7.5)	5792 (7.4)
Diseases of the musculoskeletal system	7856 (27.3)	15 212 (30.5)	23 068 (29.3)
Diseases of the genitourinary system and pregnancy, childbirth and the puerperium	4002 (13.9)	21 003 (42.0)	25 005 (31.8)
Age in (0,18]	1900 (6.6)	4595 (9.2)	6495 (8.3)
Age in (18,25]	3437 (12.0)	7466 (14.9)	10 903 (13.9)
Age in (25,30]	2213 (7.7)	4347 (8.7)	6560 (8.3)
Age in (30,40]	4668 (16.2)	8457 (16.9)	13 125 (16.7)
Age in (40,50]	5216 (18.1)	7360 (14.7)	12 576 (16.0)
Age in (50,60]	4349 (15.1)	5321 (10.7)	9670 (12.3)
Age in (60,70]	2689 (9.4)	3486 (7.0)	6175 (7.8)
Age in (70,80]	2308 (8.0)	4054 (8.1)	6362 (8.1)
Age > 80	1968 (6.8)	4866 (9.7)	6834 (8.7)

Note: Shown are counts (%).

TABLE D2 Number (%) of subjects purchasing treatments in the Danish population-based registry study

	Male ($n = 28\,748$)	Female ($n = 49\,952$)	Total ($n = 78\,700$)
N06	18 740 (65.2)	33 327 (66.7)	52 067 (66.2)
N05	10 049 (35.0)	16 784 (33.6)	26 833 (34.1)
N02	3384 (11.8)	7467 (14.9)	10 851 (13.8)
A02	2509 (8.7)	4486 (9.0)	6995 (8.9)
J01	2118 (7.4)	5739 (11.5)	7857 (10.0)
B01	2687 (9.3)	3484 (7.0)	6171 (7.8)
N03	1713 (6.0)	2965 (5.9)	4678 (5.9)
C03	1610 (5.6)	3450 (6.9)	5060 (6.4)
G03	42 (0.1)	7352 (14.7)	7394 (9.4)
R03	1254 (4.4)	2381 (4.8)	3635 (4.6)
C09	2219 (7.7)	3079 (6.2)	5298 (6.7)
M01	1503 (5.2)	3082 (6.2)	4585 (5.8)
C10	1822 (6.3)	2356 (4.7)	4178 (5.3)
A10	1236 (4.3)	1339 (2.7)	2575 (3.3)
C07	1401 (4.9)	2075 (4.2)	3476 (4.4)
S01	866 (3.0)	2149 (4.3)	3015 (3.8)

TABLE D1 (Continued)

	Male ($n = 28\,748$)	Female ($n = 49\,952$)	Total ($n = 78\,700$)
C08	1170 (4.1)	1897 (3.8)	3067 (3.9)
A12	815 (2.8)	1907 (3.8)	2722 (3.5)
A06	756 (2.6)	1503 (3.0)	2259 (2.9)
C01	695 (2.4)	1082 (2.2)	1777 (2.3)
G04	1107 (3.9)	328 (0.7)	1435 (1.8)
H03	235 (0.8)	1390 (2.8)	1625 (2.1)
D07	633 (2.2)	1234 (2.5)	1867 (2.4)
N07	897 (3.1)	666 (1.3)	1563 (2.0)
B03	485 (1.7)	1076 (2.2)	1561 (2.0)
R05	370 (1.3)	951 (1.9)	1321 (1.7)
R06	442 (1.5)	1143 (2.3)	1585 (2.0)
A03	334 (1.2)	1010 (2.0)	1344 (1.7)
M05	158 (0.5)	1011 (2.0)	1169 (1.5)
H02	374 (1.3)	755 (1.5)	1129 (1.4)
N04	261 (0.9)	431 (0.9)	692 (0.9)
D01	458 (1.6)	708 (1.4)	1166 (1.5)
R01	357 (1.2)	672 (1.3)	1029 (1.3)

TABLE D3 Number of relapse events, censoring and competing events after 5 years

Δ	Event type	Number of subjects	Percent of total
0	Censoring	67 794	86.14%
1	Depression relapse	4613	5.861%
2	Competing event	6293	7.996%

APPENDIX E

We here provide further details of how we simulated the data for the empirical study presented in Section 5. The simulated data include three latent event times, one for each of two causes, and one for the right-censoring time. The minimum of the three latent times is the observed time $\tilde{T}$. The event type $\tilde{\Delta}$ has the value 0 for right-censored, 1 for event of cause one and for event of cause two. The distributions of the latent times depends on five binary covariates, $X_1, \ldots, X_5$, two continuous Gaussian covariates, X_6, X_7, and ten binary treatment variables, $A_1, \ldots, A_{10}$. Different scenarios with variations over the simulation parameters were used to parametrize the cause-specific hazard functions of both causes, and the hazard function of the distribution of the censoring time (to generate covariate dependent and covariate independent censoring, respectively). For example, in the majority of the simulation settings, the parametrization for the cause-specific hazard function of cause 1 is defined by the Cox-Weibull hazard model[23] with `shape=0.01`, and `scale=2`:

$$\lambda_1(t|\boldsymbol{A},\boldsymbol{X}) = 0.02\ t\ \exp(1.25A_1 + A_2 + 0.3A_4 + 0.7A_5 + X_1 + 0.3X_2 - 0.5X_6), \tag{E1}$$

whereas in the setting addressing model misspecification it is defined as

$$\lambda_1(t|\boldsymbol{A},\boldsymbol{X}) = 0.02\ t\ \exp(1.25A_1 + A_2 + 0.3A_4 + 0.7A_5 + X_1 + 0.3X_2 + 0.3X_6 + 0.4X_6^2); \tag{E2}$$

TABLE E1 Additional simulation results for estimation of crude and net effects, varying the effect that X_1 has on the cause one specific hazard ('A1_T1') and the effect that X_1 has on the cause two specific hazard ('A1_T2')

Method	Net/crude	True.ate	P(C < 5)	A1_T1	A1_T2	Bias	SD	SE	Coverage
Causal_forest	Net_effect	−5.40	13.10	0.80	0.80	0.10	2.21	2.16	94.60
Causal_forest	Net_effect	0.00	13.10	1.00	0.80	0.39	2.21	2.24	95.10
Causal_forest	Net_effect	5.90	13.10	1.25	0.80	0.43	2.30	2.29	94.60
Causal_forest	Crude_effect	6.10	13.10	1.25	0.80	−0.41	1.95	2.00	94.90
Causal_forest	Crude_effect	5.40	13.10	1.25	1.00	−0.41	2.00	1.99	93.20
Causal_forest	Crude_effect	4.40	13.10	1.25	1.25	−0.32	2.01	1.98	93.40

the difference between the two settings really amounts to the squared effect of X_6 included in the setting addressing misspecification. The parametrization for the covariate dependent hazard function of the censoring distribution is defined as

$$\lambda^c(t|\boldsymbol{A},\boldsymbol{X}) = 0.02\ t\ \ \exp(0.1A_1 - 0.3A_3 - 0.4X_2 + 0.1X_6). \tag{E3}$$

In the covariate independent censoring setting the parametrization is simply $\lambda^c(t|\boldsymbol{A},\boldsymbol{X}) = 0.02\ t$. Across all settings, the cause-specific hazard function for cause 2 is parametrized as

$$\lambda_2(t|\boldsymbol{A},\boldsymbol{X}) = 0.02\ t\ \ \exp(A_1 + \beta A_2 + -0.3A_5 - 0.1X_1 + 0.6X_2 + 0.1X_6), \tag{E4}$$

with variations over β used to investigate ranking. The distribution of the treatment variable A_1 is generated from the logistic regression model

$$\mathbb{E}[A_1|\boldsymbol{A},\boldsymbol{X}] = \text{expit}(-X_1 + 0.7X_6 + 0.2A_7).$$

The total of all simulated scenarios can be obtained from the material at https://github.com/helenecharlotte/forestCausalSearch by loading the matrix called `VARYING` into the software R.[34] The site also provides all R-codes used and examples of how to apply the function `causalhunter()` which implements the methods described in this article. The specific simulation settings that were used to obtain the figures shown in Section 5 are the following. Figure 1 shows results from the setting with the cause one specific hazard defined by (E1), with the cause two specific hazard defined by (E4) (with $\beta = 1.25$), and with censoring being either independent of covariates (left panel) or on the form (E3) (right panel). Figure 2 shows results from the setting with the cause one specific hazard defined by (E2), with the cause two specific hazard defined by (E4) (with $\beta = 1.25$), and with covariate independent censoring. Figure 3 shows results from the setting with the cause one specific hazard defined by (E1), with the cause two specific hazard defined by (E4) (with $\beta = 0.2, 1, 2$), and with covariate independent censoring. Table E1 provides additional results for estimation of both crude and net effects varying the effect that X_1 has on the cause one and the cause two specific hazards.

APPENDIX F

Figure F1 shows together rankings based on the forest approach, and the cause-specific hazard regression (G-formula) approach. Due to convergence issues, the latter only adjusts for covariates and not all the large number of (other) treatment variables. We note that the two approaches produce quite different results; besides the mentioned lack of adjustment issue with the cause-specific hazard approach, this is likely also partly due to the strict model assumptions imposed by the cause-specific hazard regression approach.

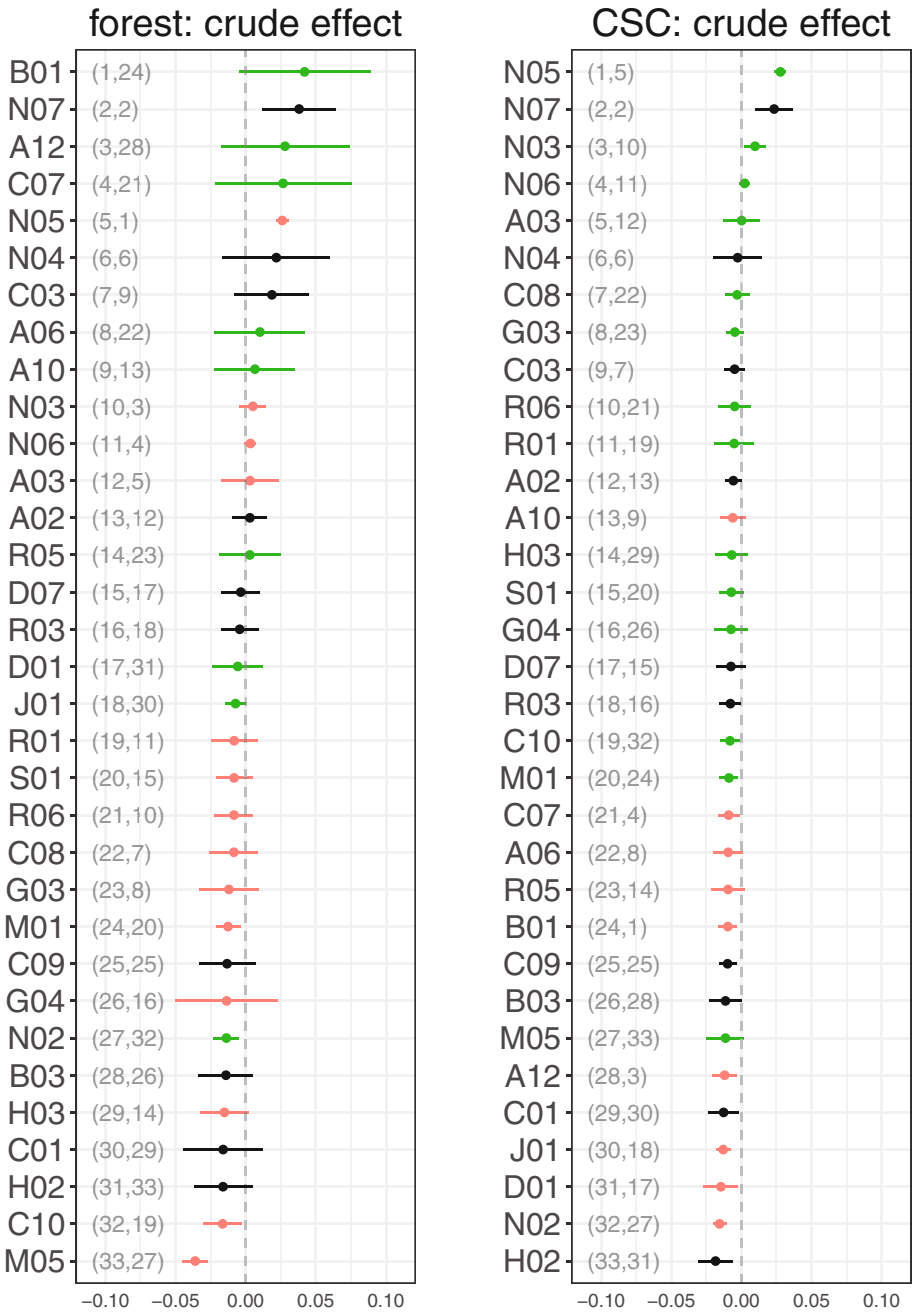

FIGURE F1 Comparison of rankings based on the forest, and the cause-specific hazard regression approach without adjustment for all other treatments