Prediction meets causal inference: the role of treatment in clinical prediction models

Geloven, N. van; Swanson, S.A.; Ramspek, C.L.; Luijken, K.; Diepen, M. van; Morris, T.P.; ... ; Cessie, S. le

European Journal of Epidemiology (2020) 35:619–630
https://doi.org/10.1007/s10654-020-00636-1

METHODS

Prediction meets causal inference: the role of treatment in clinical prediction models

Nan van Geloven[1] · Sonja A. Swanson[2,3] · Chava L. Ramspek[4] · Kim Luijken[4] · Merel van Diepen[4] · Tim P. Morris[5] · Rolf H. H. Groenwold[1,4] · Hans C. van Houwelingen[1] · Hein Putter[1] · Saskia le Cessie[1,4]

Received: 10 January 2020 / Accepted: 18 April 2020 / Published online: 22 May 2020

Abstract

In this paper we study approaches for dealing with treatment when developing a clinical prediction model. Analogous to the estimand framework recently proposed by the European Medicines Agency for clinical trials, we propose a 'predictimand' framework of different questions that may be of interest when predicting risk in relation to treatment started after baseline. We provide a formal definition of the estimands matching these questions, give examples of settings in which each is useful and discuss appropriate estimators including their assumptions. We illustrate the impact of the predictimand choice in a dataset of patients with end-stage kidney disease. We argue that clearly defining the estimand is equally important in prediction research as in causal inference.

Keywords Clinical prediction model · Treatment · Censoring · Estimands · Predictimands

Introduction

Clinical prediction models provide individualized prognostic information that can be used to counsel patients about the likely course of their disease. Prediction models can also support treatment decisions. For instance, if a patient's risk of a poor outcome in the next year is relatively low, then she or he may not need treatment. If the risk of a poor outcome is high, additional preventive or curative treatments should be considered [1]. But what is exactly meant by "a patient's prognosis" or "a patient's risk" here? Do we mean the risk assuming that no treatment is given, the risk under current standard treatment, the risk of experiencing the event in the time period before being treated, or something else? An increasing number of clinical prediction models are becoming available through websites and apps. However, for many of these models it is unclear how the risk they aim to capture relates to treatment.

The data sources used for the development of clinical prediction models often contain data on a mix of patients who did and did not receive treatments that affect the risk of the event of interest. This holds both for observationally collected data and for trial data. Completely untreated cohorts can sometimes be found in historical data collections, but these cohorts are rare and may not be relevant to current practice in other important ways. Using data solely from untreated patients can also lead to highly selected cohorts that are not generalizable. Dealing with treated patients is thus a challenge that needs to be addressed when developing a clinical prediction model. Baseline treatments can be included as predictors when developing, or externally validating, a prediction model, meaning that for a new patient the model can be used to predict the outcome, provided information is available about baseline treatment status and other predictors [2]. However, for treatments that

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10654-020-00636-1) contains supplementary material, which is available to authorized users.

✉ Nan van Geloven
n.van_geloven@lumc.nl

1 Department of Biomedical Data Sciences, Leiden University Medical Center, Zone S5-P, PO Box 9600, 2300 RC Leiden, The Netherlands

2 Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands

3 Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, USA

4 Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

5 MRC Clinical Trials Unit, UCL London, London, UK

are initiated after baseline, there is no easy solution, nor is there a single question of interest [3]. We do not know at baseline what treatments will be initiated later on and we cannot use future values as predictors. Moreover, the way treatments are dealt with during model development will have impact on how predictions can be used in future patients. Currently, ad hoc approaches are used where treatment initiation may be ignored, patients are censored at the moment they start or switch treatment, or even excluded completely from the development cohort. Such analysis choices are often reported as mere technical analysis issues [4], but they may have a major impact on interpretation. The predictions resulting from such approaches might end up targeting a different risk than researchers intended. For example, censoring patients when they receive a transplantation when estimating survival among patients listed for liver transplantation has shown to overestimate the actual observed waiting list mortality [5, 6]. The TRIPOD reporting guideline on development and validation of prediction models offers hardly advice on the issue. The only related point on the TRIPOD checklist is that one should describe the treatments that patients received, if relevant (item 5c) [7].

The European Medicines Agency (EMA) has recently released a new guideline that provides a framework to deal with additional treatments started after baseline and other so-called intercurrent (post-baseline but pre-outcome) events in the context of clinical trials: the addendum to the ICH E9 guideline Statistical Principles for Clinical Trials [8]. The guideline distinguishes five possible strategies: a 'treatment policy' strategy that follows the intention-to-treat principle and ignores any change in treatment after baseline, a 'composite' strategy that includes the start of an additional treatment as part of the outcome definition, a 'hypothetical' strategy that aims to assess the outcome in a scenario where no additional treatment would be given, a 'principal stratum' strategy where the outcome is considered in a certain subset of patients who would never start the additional treatment independent of their allocated treatment, and a 'while on treatment' strategy where the outcome is assessed in the time period up until the additional treatment is started. Together with the definition of the population, the outcome of interest and the effect measure, each strategy defines an estimand which is the target quantity that the trialists aim to estimate. Each estimand represents a different causal question. In the trial context, these questions relate to treatment effects. In prediction models, the aim is to assess patients' expected outcomes conditional on certain patient characteristics measured at baseline. We argue that, just like in trials, the quantity that a clinical prediction model targets should be unequivocally defined. Therefore we map the E9 estimand framework to the context of prediction models, proposing a predictimand[1] framework.

Our focus in this manuscript is on prognosis over time, so on time-to-event or failure-time outcomes. We consider treatments that are initiated during follow-up. Previous studies on estimands in prediction research considered point (time-invariant) treatments [2] and binary outcomes [3]. Recently, Pajouheshnia and colleagues discussed analysis methods for time-to-event prediction of untreated risk in the presence of time-dependent treatment [9]. Here we will extend that work by considering different questions that can be assessed by prediction models and that will be of interest in different applied settings. We formulate these strategies analogous to the trial estimand framework from the E9 addendum. We provide a formal definition of the prediction estimands matching these questions. For each of the estimands, we discuss key assumptions for estimation and list common estimators. We illustrate the impact of the predictimand choice in a dataset of patients with end-stage kidney disease.

Notation

To simplify, we consider only one treatment (A) that is related to the event of interest. Patients are all event-free and without this treatment at time zero. At that moment we collect baseline covariates $X(0)$, which will be used to determine the prognosis of the patient. T is the time to the event of interest. Some patients will start treatment A over time, with V the time to treatment start and $A(t)$ the time dependent treatment indicator. In principle, $A(t)$ could switch between 0 (no treatment) and 1 (treatment) multiple times over the follow up, but, to enhance readability, in the following we will assume that once patients initiated treatment, they stay in the treated condition throughout. For patients who experience the event of interest before treatment start, V is latent. Both T and V can be censored by end of study or loss to follow up which we assume for simplicity to be non-informative censoring mechanisms, conditional on baseline covariates $X(0)$. In some studies, patients are no longer followed for the event of interest after treatment initiation, for instance in a registry of patients on dialysis that no longer follows patients after a kidney transplantation. This situation is depicted in Fig. 1a. If follow up on the event of interest does continue after a new treatment is initiated, we are in the setting depicted in Fig. 1b. The two figures are not causal directed acyclic graphs, but could be viewed as state transition diagrams of

[1] We chose the term 'predictimand' as a portmanteau of 'prediction' and 'estimand'. Another term that could be used in this context is predictant, whereas predicand or predicend might better follow the rules of Latin grammar.

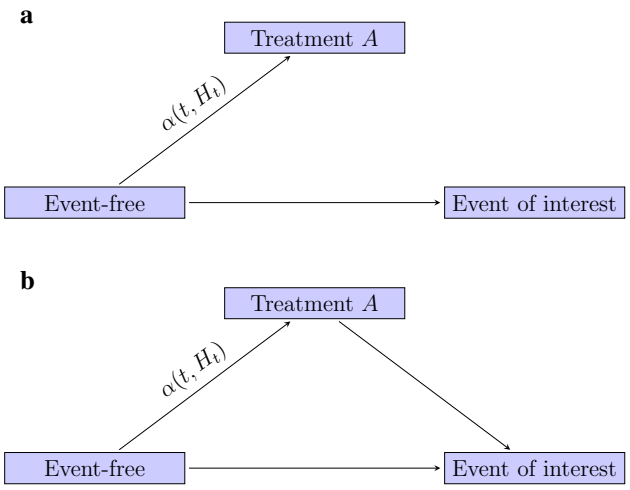

Fig. 1 Graphical representation of the studied situation. Follow up on the event of interest may stop (**a**) or continue (**b**) after treatment initiation

a multistate model [10]. Figure 1a is similar to a competing risks setting and Fig. 1b is similar to that of an illness death model, with the intermediate state in our case not disease but treatment. The pace at which patients initiate treatment is denoted by the transition intensity $\alpha(t, H_t)$, where H_t is the history of the patient up to time point t. Later on, we will distinguish between situations where the treatment decisions are only based on patients' baseline prognostic covariates, i.e., $H_t = \{X(0)\}$ and situations where the treatment decisions are also based on prognostic markers that evolve over time, i.e., $H_t = \{X(s); s \leq t\}$.

Predictimands

Below we describe four strategies for how to deal with treatment initiation after baseline in the development of a prediction model. We formulate the interpretation of the resulting prediction estimands, give examples of settings where they are useful and discuss how they apply to new patients that were not used for development of the predictions, i.e., their generalizability [11]. The four strategies described have analogous estimands in the E9 addendum. The fifth -principle stratum- estimand described in the addendum was not mapped to the prediction setting as it refers to a counterfactual subpopulation that has no immediate analogue to the prediction setting. In the "Estimators and their assumptions" section we focus on assumptions needed for estimation of the predictimands and list some common estimators. An overview is presented in Table 1.

Table 1 Overview of the four strategies of dealing with treatment initation after baseline

Strategy	Estimand	Example	Estimators	Key assumptions
Ignore treatment	Risk of the event, regardless of treatment	Risk of cardiovascular events where some patients will initiate statins according to routine-care prescriptions	Survival model for T, do not censor at V	Treatment assignment policy in application setting similar to development data
Composite	Risk of the event or treatment initiation	Risk of a composite of cardiovascular death, myocardial infarction and treatment with revascularisation (PCI or CABG)	Survival model for min(T, V)	Treatment assignment policy in application setting similar to development data
While untreated	Risk of the event occurring before treatment is started	Risk of dying while on the waiting list for a liver transplant	Competing risks methods	Treatment assignment policy in application setting similar to development data
Hypothetical	Risk of the event if treatment were never started	Risk of a natural pregnancy without IVF treatment	Survival model for T, censor at V or include treatment as time-dependent covariate in the model and set to 0 when predicting	Exchangeability, consistency and positivity

T time to event of interest; *V* time to start of treatment; *PCI* percutaneous coronary intervention; *CABG* coronary artery bypass grafting; *IVF* in vitro fertilization

"Ignore treatment" strategy

In this strategy, treatment initiation is considered part of standard practice. The value for T, the time to event of interest, is used regardless of whether patients start treatment or not. So V, the time to treatment, is not used in any way. This is analogous to the "treatment policy" strategy in the E9 guideline [8]. It has also been described as "simply ignore treatment" [2]. The risk that is estimated equals:

$$P(T \leq t_{\text{hor}}|X(0)), \tag{1}$$

i.e., the risk of the event of interest occurring before a time horizon t_{hor} under the treatment practice inherent to the development dataset. An example of where this strategy was used, is in the development of QRISK3, a risk prediction algorithm that targets a person's risk of a heart attack or stroke over the next 10 years [12]. The algorithm was developed using individuals who did not use statins at baseline, but statin use after baseline was ignored, as discussed in [3]. The calculated risks therefore belong to a population where some individuals will receive statins during follow up. The algorithm will only be generalizable to new patient groups if in those groups the same treatment assignment policy is used as in the development cohort. This implies that for all subgroups defined by the predictors in the prediction model, a similar proportion of patients should initiate treatment as in the development set. Provided the treatment is effective in reducing the risk of the outcome, the risk calculated with the "ignore treatment" strategy will be lower than the untreated risk, that is, the risk for a patient who will not be treated during follow-up. If the ignore treatment predictimand is falsely interpreted as untreated risk and used for future decisions on prescribing (statins), it will underestimate the true untreated risk and this could lead to undertreatment in new patients [13]. The "ignore treatment" strategy requires continued follow up after treatment initiation, so it cannot be used in the study design depicted in Fig. 1a.

"Composite" strategy

In the second strategy, treatment is combined into a composite outcome together with the event of interest. In such a "composite" strategy we target

$$P(\min(T, V) \leq t_{\text{hor}}|X(0)), \tag{2}$$

i.e., the risk of the event of interest or the treatment occurring before time t_{hor}. In this strategy the treatment is integrated in the clinical outcome. An example is predicting the composite of cardiovascular death, myocardial infarction and treatment with revascularisation (PCI or surgery) [14]. Including the treatment in the outcome may seem a somewhat artificial way of dealing with treatment and may lead to a less well interpretable outcome, but some use cases exist. A "composite" strategy can for instance be used when the treatment has very likely prevented an imminent occurrence of the event of interest and treatment can be viewed as a proxy of the event (i.e., without PCI or surgery a patient would develop a myocardial infarction). Other settings where a composite outcome can be useful is when a poor outcome is more clearly captured by its consequence of needing treatment than by giving a precise description of poor health status. For example, Von Dadelszen et al. [15] used a composite outcome including amongst others receiving infusion of a third parental hypertensive drug, intubation, transfusion with any blood product and dialysis, when predicting severe maternal outcomes in pre-eclampsia. A third use case for the composite strategy is predicting the chances of a good outcome (one minus composite) defined as staying event-free without requiring additional treatment. The "composite" strategy does not need continued follow up after treatment initiation. Applying a composite prediction estimand to new patients requires similar treatment assignment policies as in the development cohort. For instance, if in the new setting where the predictions are applied, patients in a certain subgroup are treated more often than similar patients in the development cohort, then the predictions of the combined outcome will be lower than the true probabilities in that subgroup (miscalibration).

"While untreated" strategy

In the "while untreated" strategy, we are only interested in the event of interest if it happens before treatment is started. Events occurring after treatment start do not count as events. We target the risk of the event of interest occurring before time t_{hor} and before treatment is started.

$$P(T \leq t_{\text{hor}}, T < V|X(0)). \tag{3}$$

This prediction estimand is well-known from competing risks analysis. It is often referred to as cumulative incidence. Starting treatment is then considered a competing event that precludes observing the untreated event of interest. Cumulative incidence has also been referred to as the absolute risk, actual risk, crude probability, crude cumulative incidence function, absolute cause-specific risk or subdistribution function [16, 18, 19]. It has recently been conceptualized as the "risk without elimination of competing events" [17]. The analogous name for this strategy in the E9 addendum is "while on treatment", referring to the response of the patient during the period where patients are still on their originally assigned treatment. Note that what distinguishes this strategy from the "ignore treatment" strategy is that at the moment treatment is started, the event of interest will by definition not occur anymore. An example is estimating the

risk of dying while on the waiting list for a liver transplant [5]. Typically this strategy will be of interest if the treatment is not freely available, but limited due to waiting lists or other logistical constraints. Also for this strategy, predictions on new patients are only well calibrated if the assignment policy of treatment is similar to the development cohort. If in the application setting a subgroup is treated more often or sooner than in the development set, the predictions will overestimate the true "while untreated" risk in this subgroup.

"Hypothetical" strategy

In this strategy we envision a world where treatment does not exist. We aim to estimate the untreated risk before time t_{hor}:

$$P(T^{v=\infty} \leq t_{hor}|X(0)), \tag{4}$$

where $T^{v=\infty}$ represents the counterfactual time to the event of interest if V is set to infinity, i.e., in a hypothetical world where treatment A is eliminated. Like the "while untreated" strategy, this strategy too has an analogue in the competing risk literature. Young et al. [17] refer to this risk as the "risk under elimination of competing events". It has been referred to as the marginal cumulative incidence, net risk, or pure risk [18, 19]. The risk quantifies how likely the event of interest would be if nobody were to receive treatment. Since we are only interested in the risk up to t_{hor}, we could similarly have used $T^{v>t_{hor}}$ instead of $T^{v=\infty}$. In the E9 addendum this strategy is similarly referred to as the "hypothetical" strategy.

An example application is estimation of the 'risk' of a natural pregnancy without use of assisted reproductive techniques such as IVF [20]. Other hypothetical scenarios (e.g., what if treatment is started after 1 year) could in principle also be targeted. In fact, if we could set v exactly according to the function that clinicians used in the development cohort to determine when to start treatment, this estimand would reduce to the "ignore treatment" estimand in (1). But here we only discuss the hypothetical untreated risk further. Estimating the hypothetical untreated risk is challenging (more on this later), but, once constructed, it is readily generalisable to new patients when posing the question what would happen if the new patient is never treated. This untreated/baseline risk is useful to inform decisions on treatment A. Note that the three other strategies (ignore treatment, composite and while untreated) cannot be used to inform the decision to start treatment A, as this might lead to something that has been described as the 'prediction paradox': predictions influencing behaviours (i.e., treatment decisions) that in turn invalidate predictions [13]. These three predictimands will be miscalibrated if the treatment decisions made in new patients differ from those in the development cohort.

Estimators and their assumptions

In this section we focus on key assumptions and design elements that are necessary for estimating each predictimand. Without being exhaustive, we link the estimands to common estimators.

"Ignore treatment" strategy

Since in this strategy starting treatment after baseline is ignored, standard time to event regression methods may be used to relate the event of interest to the covariates $X(0)$. For instance, one could use Cox regression models combined with the nonparametric Breslow (or Efron) estimator for the baseline hazard or flexible parametric survival models. The main assumption here is that of non-informative censoring for reasons like loss to follow up or end of study. Particular methods may additionally assume proportional hazards for the covariates. Note that since we do not censor at the moment of treatment start, this strategy requires continued follow up after treatment initiation.

"Composite" strategy

Also with this strategy the analysis is relatively straightforward and can be done with any chosen survival regression technique. Either occurrence of the event of interest or the occurrence of treatment counts as an event, whichever comes first. When one would use a Cox-like model, the assumption of proportionality of covariate effects should hold for the composite outcome, which may be less likely than for single outcomes. Also, the non-informative censoring by loss to follow up or end of study should hold in relation to the composite outcome. Continued follow up after treatment initiation is not needed.

"While untreated" strategy

Estimation of cumulative incidence as expressed in (3) can be done with competing risks methods. Without covariates and without censoring for other reasons like loss to follow up or end of study, cumulative incidence is estimable by the number of patients with the event of interest divided by the total number of patients at baseline. With covariates and censoring for reasons like loss to follow up or end of study, the estimation can be done in various ways, see for instance [10, 21–24]. Continued follow up after treatment initiation is not needed.

"Hypothetical" strategy

Estimation of (4) is challenging and relies on strong assumptions regarding the treatment assignment policy in the development data. These assumptions are similar to those that are needed for identifying the causal treatment effect of A, which makes sense since the strategy implicitly imposes a counterfactual or potential outcome version of A into the estimand of interest. Three key assumptions are required: exchangeability, consistency and positivity [25]. The first one, exchangeability, is often the most challenging. It is sometimes called the 'no unmeasured confounding' assumption and requires that we have measured and appropriately corrected for (a sufficient subset of) the variables that both influenced the treatment decisions and are prognostic for the event of interest [25]. This typically requires measuring time-dependent covariates since updated measurements of risk factors are very likely to influence treatment decisions. The second assumption, consistency, is often described as observed outcomes being equal to counterfactual outcomes. It means that in the hypothetical world where treatment is eliminated, a patient's untreated risk is the same as her or his untreated risk in the real world. If knowledge of the unavailability of treatment changes the risk behaviour of patients, this assumption does not hold. This second assumption is typically not prohibitive; in some causal frameworks such changing risk behaviour is taken as a definition of the patient population [26]. The third assumption, positivity, means that we have observed a non-zero number of treated and untreated patients in our data for all covariate patterns during the time horizon that we want to use in our predictions (or have observed a sufficient number so as to smooth over gaps with modelling assumptions). For instance if all patients with a certain characteristic are treated after 1 year, then we don't have information for estimating untreated outcomes beyond 1 year for such patients. It may well be that a shorter prediction horizon t_{hor} has to be chosen to fulfil the positivity assumption.

The analysis approach for the "hypothetical" strategy depends on how treatment decisions were made for the patients in the development data and on whether or not post-treatment follow up is used. Below, we sketch four analysis approaches. We first discuss two settings where it is assumed that in the development dataset treatment decisions were only based on baseline covariates $X(0)$, i.e., prognostic patient characteristics that are known at the moment we want to make the prediction. In most healthcare settings it is quite implausible that risk factor progression after baseline doesn't influence treatment decisions, but we include these options to indicate the limitations of common estimation approaches. Then, we discuss two settings where risk factor progression is accounted for. There are two analysis approaches to target the hypothetical risk: we may stop follow up when treatment is started, by censoring the time to event of interest at the moment treatment starts (Fig. 1a), or follow up data after the start of treatment may be included (Fig. 1b). As noted before, in some studies no follow up information is collected after treatment initiation, in which case only the censoring option remains. In the censoring approach positivity is only needed for untreated individuals, i.e., if for some covariate patterns no patients are treated, this is not per se a problem for estimating the untreated risk.

Several other estimation approaches than the ones described below have been proposed for the hypothetical untreated risk, each with their own assumptions. For instance using g-formula [17], copulas [27, 28] or multiple imputation [29, 30]. We refer to the mentioned references for further details on these methods.

Baseline covariates, censoring

In the situation sketched in Fig. 1a, time to event is censored at treatment start. The main assumption of this approach is that censoring by treatment is non-informative, conditional on the baseline prognostic covariates in the prediction model ($X(0)$). In other words, censoring the follow up at treatment start only gives a valid estimate of the hypothetical untreated risk if baseline prognostic factors that relate to treatment are included in the prediction model and treatment start is independent of changing values of prognostic markers during follow-up, i.e., $H_t = \{X(0)\}$. This is a strong and often implausible assumption that cannot be tested on the data. Such an assumption can only be verified with those who were responsible for the treatment decisions.

Baseline covariates, modelling

When follow up after treatment start is available, this follow-up information can be used in the development of the prediction model. Treatment A can be added as an additional, time dependent, covariate to the prediction model. Then a prediction under the "hypothetical" strategy of no treatment can be obtained by setting $A(t) = 0$ for all t in the prediction horizon ($t \leq t_{\text{hor}}$). This approach is valid under the same strong and often implausible assumption as the first approach that we described. Only in case treatment choices were solely based on collected prognostic baseline factors that are included in the prediction, the model including $X(0)$ is sufficient to get an unconfounded estimate of the effect of A. There is a caveat to this approach of modelling the treatment effect. The advantage of using a longer period of follow up is at the expense of having to model the effect of the treatment on the event of interest. Therefore, the functional form of the treatment effect should be carefully chosen: the assumption of proportional hazards between the treated and untreated patients over time and the presence of possible

interactions between patient characteristics and treatment should be checked.

Time varying covariates, censoring

We now turn to the more realistic situation where time-varying prognostic patient characteristics have additionally influenced the treatment decisions, i.e., $H_t = \{X(s), s \leq t\}$. For example when the condition of patients has been monitored by repeatedly measuring their blood values and these measurements have influenced the decisions about treatment initiation. In the censoring case (Fig. 1a), inverse probability of censoring weighting (IPCW) could be used. This approach assumes that treatment start is independent of the future untreated risk conditional on baseline covariates $X(0)$ and time dependent covariates $X(t)$. $X(t)$ is used to estimate time-varying conditional probabilities of starting treatment. By assigning weights to patients that are inversely proportional to their conditional probability of not yet being treated, a weighted population is created that (under the assumptions of exchangeability, consistency and positivity) mirrors the pseudo-population that would have been observed in the absence of treatment [17, 31–34]. Applying inverse probability weighting when estimating a survival model for the event of interest with censoring at treatment start thus corrects for the informative censoring related to $X(t)$.

Time varying covariates, modelling

When post-treatment follow up is used to model the effect of treatment on the event of interest, adding both $X(t)$ and A in the survival model similar to the "Baseline covariates, modelling" approach, will in the presence of time varying covariates not yield a useful prediction model, because for a new patient $X(t)$ will not be known when predicting at baseline. The hypothetical untreated risk to estimate from Fig. 1b is similar to what is called the 'controlled direct effect' in mediation analysis, when setting treatment as the mediator at a fixed zero level (no treatment) [35]. A potential solution is to fit a marginal structural model with inverse probability of treatment weighting (IPTW) to break the link between $X(t)$ and $A(t)$ [36]. Using these weights, a model containing $X(0)$, $A(t)$ and potential interactions as predictors can be fitted and one can estimate the hypothetical risk from this model setting $A(t) = 0$. Details on implementing this approach for prediction modelling in both logistic and time-to-event models can be found in [3]. We again have to assume correct specification of the treatment effect.

Data application

In this section, we study the four proposed prediction estimands in data from the Netherlands Cooperative Study on the Adequacy of Dialysis (NECOSAD) [37]. This study is a multicenter cohort study in which patients with end stage renal disease were included at dialysis initiation if they were 18 years or older and had no previous kidney transplantation and no previous dialysis. The NECOSAD study was approved by the local medical ethics committees and all patients gave informed consent. Patients were followed until renal transplantation, death or end of study. Here we consider death the event of interest and renal transplantation is the treatment that may be initiated at some point in time after baseline. As in NECOSAD patients were not followed after renal transplantation, information on death after transplantation was retrieved by linking the NECOSAD data to the Dutch registry of renal replacement therapy, RENINE (Registratie Nierfunctievervanging Nederland) [38]. Patients were included between 1997 and 2007, and followed until February 1, 2015. Patients who were not coded as deceased or lost to follow up in RENINE were assumed to be alive at end of follow up. Our initial data set contained n = 2051 patients. We removed 6 patients with missing information on age, yielding 2045 patients in our analysis. For 43 patients who were still alive at the last follow up visit of the NECOSAD study, no link to the registry could be made and we censored time to death and, where applicable, time to transplantation for these patients at their last follow up in NECOSAD. The median time in follow up of the 2045 patients was 5.1 years. In this period, 749 patients received a kidney transplant, 1470 patients died of whom 248 after transplantation. Age and baseline dialysis type (hemodialysis (HD) versus peritoneal dialysis (PD)) were used as baseline predictors of mortality ($X(0)$). With HD, blood is pumped out of the patient's body and filtered by an artificial kidney machine. With PD, cleansing fluid is pumped into the patient's abdominal cavity and the lining of the abdomen acts as a natural filter to wash out waste and toxins. As this example is used for illustration purpose, we did not include more baseline variables in the prediction model. Additionally, for estimating the hypothetical prediction estimand, we used the following time dependent covariates $X(t)$ as predictors of treatment: Charlson comorbidity score, BMI and calcium blood values, which were measured at 6 months intervals. We estimated the mortality risk over a time span of 10 years, given age (as a continuous variable) and baseline dialysis type. We used the packages `survival`, `mstate`, and `ipw` of the R statistical software [39]. Our analysis code along with a simulated dataset can be found in the Supplementary Materials. The different predictimands were estimated as follows:

- The "ignore treatment" strategy targets the total mortality risk, regardless of whether patients did or did not receive a transplantation. For estimation, we used a Cox proportional hazards model with the non-parametric baseline hazard estimated using the approach proposed by Efron (further on referred to as Cox–Efron) [40]. Death was defined as the event, age and dialysis type as baseline covariates and we censored the patients alive at the moment of last follow up. Note that for estimating the "ignore treatment" risk, follow up for death after transplantation is needed, which was retrieved by linking the NECOSAD data to the Dutch Renal Registry.
- With the "composite" strategy, we estimate the risk of either dying or receiving a transplantation. To this end, transplantation and death were combined as composite event in a Cox–Efron analysis, again with age and dialysis type as baseline covariates and censoring those alive at the moment of last follow up. Studying a composite outcome in this situation can be informative, e.g., for policy makers, to know how long patients will likely stay alive and without transplantation and thus remain on dialysis treatment. For estimating the "composite" risk, follow up after transplantation is not needed.
- The "while untreated" strategy assesses the risk of dying before receiving a transplantation. To estimate this risk, we fitted two cause-specific Cox–Efron models: one model with death as event, age and dialysis type as baseline covariates and censoring at time of transplant or at moment of last follow up alive, and one model with transplant as event, age and dialysis type as baseline covariates and censoring at death and at last follow up alive. The two cause specific hazard models were used to obtain the cumulative incidence for death [10]. Follow up after transplantation is not needed for the "while untreated" risk.
- In the "hypothetical" strategy we estimate the risk of dying if no transplantation is performed. We followed the four different estimation methods described in ""Hypothetical" strategy" section.

    - First, we fitted a Cox–Efron model for death, with age and dialysis type as baseline covariates and where event times were censored when the patient received a transplantation or at the end of follow up. This model was then used to predict the mortality risk over time. This approach assumes that the decisions on transplantation were based only on on age and dialysis type. Follow up after transplantation is not needed in this case.
    - Second, transplantation was included as a time-dependent covariate in a Cox–Efron model with age and dialysis type as baseline covariates, again assuming that only these two baseline covariates drove the transplantation decisions. To model the effect of transplantation correctly, we explored whether adding interactions between transplantation and baseline covariates improved the model, but it did not. Since the transplantation effect seemed to change over time, we used a time varying coefficient for treatment according to a step function (with jumps at 3 and 8 years, chosen by visual inspection of the Schoenfeld residual plot). The hypothetical untreated risk was then estimated by setting $A(t) = 0$. For this second approach where the effect of transplantation is modelled, we needed the additional follow up of death after transplantation.
    - Third, we repeated the first analysis where we censored at treatment start, now applying inverse probability weighting to correct for time-dependent covariates that might have additionally influenced the transplantation decisions. Stabilized weights were estimated based on two Cox–Efron models with transplantation as event: a denominator model including Charlson comorbidity score, BMI and calcium blood values as covariates and a numerator model with only an intercept. Some missings occurred in the time dependent covariates and we performed a single imputation method for each using a linear mixed model with follow up time, age and dialysis type as fixed factors and a random intercept. For patients who did not have any measurement of these covariates (1 for Charlson score, 17 for BMI, 83 for calcium), we imputed the median of the other patients.
    - Fourth, we repeated the second analysis where we modelled the effect of transplantation, applying the same inverse probability weights as in the third approach.

In Fig. 2 the predicted 10 year mortality curves are presented for a patient of age 50 and for a patient of age 70, both starting on hemodialysis. Each curve represents a different type of mortality risk. Several observations can be made from the curves. The risks obtained from the "composite" strategy are highest while the curves from the "while untreated" strategy were lowest for most of the follow up times. This is according to expectation as the "composite" strategy counts every transplanted patient as event, while the "while untreated" counts transplantation as non-event. The other two strategies infer that part of the transplanted patients will reach the event, either according to observed deaths after transplantation in the "ignore treatment" strategy, or according to what would be expected if these patients were not transplanted in the

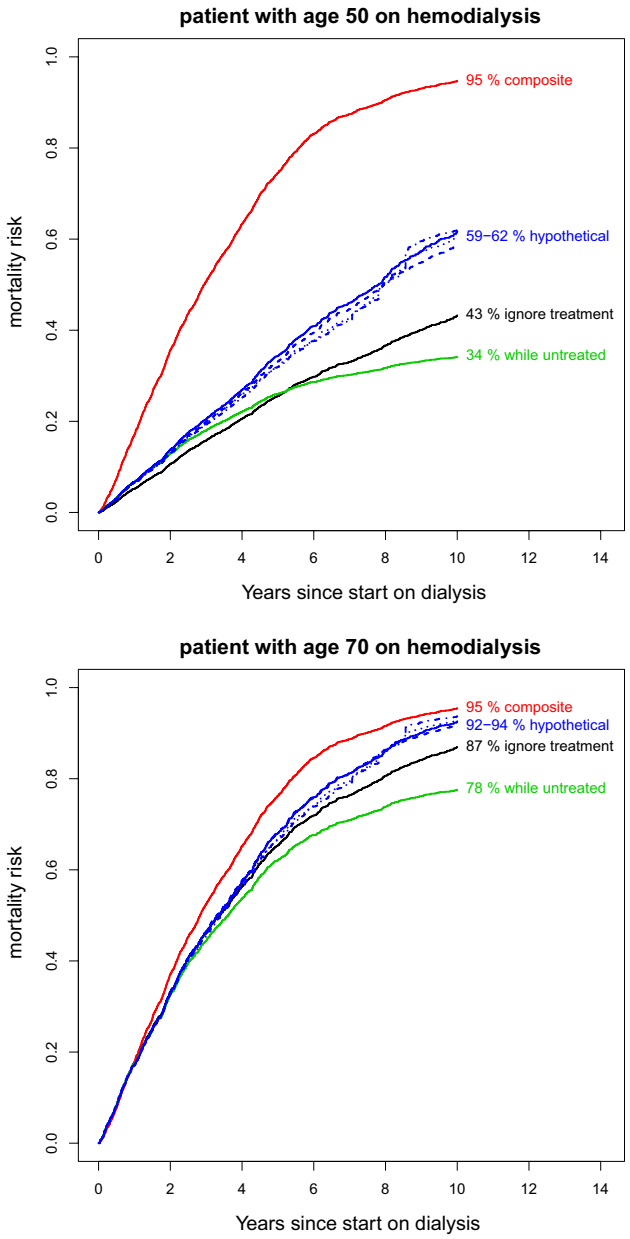

Fig. 2 Predicted mortality curves and 10 year mortality risks for patients aged 50 and 70 on hemodialysis. red: composite, green: while untreated/cumulative incidence, black: ignore treatment, solid blue: hypothetical—censor at treatment, dashed blue: hypothetical—modelling treatment, dotted blue: hypothetical—censor at treatment + IPW, dotdash blue: hypothetical—modelling treatment + IPW

"hypothetical strategy". The fact that the "while untreated" strategy does not yield the lowest predictions at all times can be explained by the modelling assumptions (proportionality of covariates is acting on different scales, composite hazard versus cause specific hazard versus marginal hazard). The composite curves for 50 and 70 years-old were very similar because younger patients have a lower probability of dying but a higher probability of getting a transplantation. The four curves belonging to the "hypothetical" strategy are higher than those from the "ignore treatment" strategy, indicating that the current transplantation policy reduces mortality compared to a hypothetical scenario where nobody would receive a transplant. This is more apparent at age 50, since more patients are transplanted at that age. The curves for patients starting on peritoneal dialysis were very similar to the curves for hemodialysis and are therefore not shown.

Our focus is on the interpretation of the different prediction estimands. Below we sketch how our results could be used in a fictitious conversation between a doctor and a patient of age 50 starting on hemodialysis.

- Doctor: You have progressed to end stage renal disease, meaning your kidneys no longer function sufficiently. My advice would be to start hemodialysis
- Patient: What is my prognosis on hemodialysis?
- Doctor: If we did not perform kidney transplantations, our best estimate is that 59–62% of patients your age would die within 10 years. (**hypothetical**)
- Patient: Ok, but what about my prognosis given that I may receive a kidney transplantation?
- Doctor: With availability and allocation of transplants like in recent years, about 43% of patients dies within 10 years. (**ignore treatment**)
- Patient: So, will I get a transplant in time?
- Doctor: I cannot say, we need a matching donor and there is a waiting list. Again assuming that availability and allocation of transplants does not change, in the next 10 years, you have about 34% chance of dying before getting a transplant. (**while untreated**)
- Patient: What are the chances I will survive for 10 years and still be on dialysis?
- Doctor: With unchanged transplant availability and allocation, you have a 5% chance to still be alive and without transplant in 10 years. (1 minus **composite**)

We note that our simplified model with only two predictors is not meant for use in clinical practice. Also for simplicity, we omitted to report uncertainty intervals around the predictions. This example serves as an illustration that the different strategies of handling treatment start after baseline answer different risk questions.

Discussion

Starting treatment after baseline is very common in risk prediction settings. We argue that the way treatment is dealt with should not be degraded to 'just a technical analysis choice'. In fact, different strategies may yield very different risk predictions. If not dealt with up-front, the choice may sneak in by the choice of analysis rather than being identified

intentionally as the prediction estimand of interest. Decisions about how one wants to handle treatment initiation should be prespecified, based on which interpretation of risk is most appropriate. In some cases, multiple questions may be of interest and therefore investigators may choose to estimate more than one of the four predictimands we described here.

Being clear about the prediction question of interest in the context of post-baseline occurrences is not only important for treatment initiation. Any post-baseline behaviour or event that may be modifiable could be considered in light of our proposed framework. For example, in the transplantation setting, another modifiable post-baseline event that could be considered is patients who stop dialysis due to recovery. (In our analyses patients were censored in case of ceasing dialysis due to recovery, implying we used the "hypothetical" strategy with respect to this event.)

A recent systematic review on prediction models for on-dialysis mortality identified 16 models studying time to death [37]. Five of these used the Cox model with censoring at transplantation, implicitly targeting the "hypothetical" prediction estimand. None of these five studies explained the consequences of censoring on treatment start for the interpretation of the calculated risk, and none paid attention to the non-informative censoring assumption. Three other models included death after transplantation in their outcome ("ignore treatment" strategy). Three studies excluded patients who received transplants after baseline from their analyses, leading to predictions that are not generalizable. Five did not write anything about how they dealt with transplantation, essentially rendering risk numbers that cannot be interpreted.

Whereas causal inference research is typically strictly distinguished from prediction research [41], we show in our paper that when predicting in the presence of modifiable events after baseline such as treatment initiation, methods from both domains are needed. A causal inference model aims to quantify what the counterfactual or potential outcomes of patients would be with and without an intervention and infers a causal effect of intervention from that. A prediction model aims to provide correct predictions of an outcome given a set of prognostic factors that do not have to be causally related to the outcome. The "hypothetical" prediction estimand could be classified as a type of counterfactual prediction, since we predict potential outcomes 'if the world were different', namely, if no one receives treatment [41]. Several untestable assumptions are needed here. The other three prediction estimands give 'real world' predictions and can be estimated from a development dataset without untestable assumptions. However, when predicting for new patients using these three strategies, we assume that similar treatment assignment policies apply as in the development cohort. This is also a very strong assumption that cannot be tested upfront. The prognostic factors used in a clinical prediction model do not have to be causally related to the outcome, however the predictions they render will only be valid if the treatment policies that patients to whom the prediction model is applied are (1) clearly defined in the prediction estimand and (2) similar as in the development cohort (except for the "hypothetical" strategy).

Our focus in this paper is on the role of treatment in clinical prediction models. We have sketched how different strategies of handling treatment lead to different prediction estimands. The definition of the role of treatment and other intercurrent events is necessary but not sufficient to define a prediction estimand. Other aspects that need to be defined are the target population/setting (e.g., patients with end stage renal disease), the relevant outcome with an appropriate time horizon (e.g, 10-year mortality) and a time-point at which the prediction will be made (e.g., at start of dialysis) [42].

Throughout the paper we have referred to a single treatment and assumed that treated patients remained treated throughout follow up. Usually many types of treatment are relevant for patients. For instance, data used for the development of a cardiovascular risk model may contain information on patients who start using statins, patients who start using antihypertensives or lipid lowering drugs, patients following a particular diet etc. Depending on the goal of the risk prediction, a choice should be made as to how each is handled. Typically a mixture of approaches will be used. Many treatments will be considered 'care as usual' with assignment policies that are considered stable over time and can be handled as background according to the 'ignore treatment' strategy. However, if for example the prediction model is aimed to input to the question of whether new patients should or should not be given statins, then a different strategy should be used for statins. For each treatment, an explicit choice should be made as to which prediction estimand is targeted and appropriate attention should be given to the necessary assumptions for estimating this. When only few patients are treated (e.g., due to a short prediction horizon) or when treatment effects are small relative to the effect of the other prognostic factors in the model, the numerical differences between the strategies will be less pronounced than in our transplantation example, but may still be relevant.

In case patients switch between 'off' and 'on' treatment multiple times during follow up, the definition and analysis approach of the "ignore treatment" and "hypothetical" strategy can stay unchanged. An application of the "hypothetical" strategy in such an 'on' and 'off' switching situation is presented in [9]. In the "composite" strategy it seems most sensible to count the first occurrence of a treatment episode as an event, but recurrent event approaches could also be considered [43]. The definition of the "while untreated" strategy could be extended to represent the risk of the event

of interest during all untreated episodes, so not only up to the first treatment episode as in the current definition.

An aspect of prediction modelling that was not addressed in our paper is assessment of predictive performance, i.e., model validation. Standard methods to validation of predictions apply to the "ignore treatment" and "composite" predictimands. For the "while untreated" strategy, methods suitable for competing risks analyses are needed, see for instance [44–46]. Validating predictions generated with the "hypothetical" strategy is more involved since also in a validation dataset there will likely be patients who start treatment after baseline. Validation in such a setting may require similar assumptions and estimation techniques as during estimation of the hypothetical predictions [47]. This warrants further research.

There might be a trade-off between relevance of an estimand and the assumptions that one is willing to make in order to produce it. Due to its strong and untestable assumptions some authors have argued against using a "hypothetical" estimand saying one can better 'stick to this world' [48]. We argue that the future use of the prediction model should drive the predictimand choice. One should start by defining a clear estimand before considering how to compute it. When there is much uncertainty on the used assumptions, sensitivity analyses could be performed to assess the degree of uncertainty in the predictions [49].

In any case, when using a prediction model in clinical care, the meaning of predictions presented to patients should be unequivocally clear. Our predictimand framework can help researchers explicating what risk is targeted by their model.

Funding Tim Morris was supported by the Medical Research Council (Grant Nos. MC_UU_12023/21 and MC_UU_12023/29)

References

1. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595. https://doi.org/10.1136/bmj.e5595.
2. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, Reitsma JB, Riley RD, Peelen LM. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. J Clin Epidemiol. 2016;78:90–100. https://doi.org/10.1016/j.jclinepi.2016.03.017.
3. Sperrin M, Martin GP, Pate A, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. Stat Med. 2018;37:4142–54. https://doi.org/10.1002/sim.7913.
4. Pajouheshnia R, Damen JAAG, Groenwold RHH, Moons KGM, Peelen LM. Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. Diagn Progn Res. 2017;1:15.
5. Kim WR, Therneau TM, Benson JT, Kremers WK, Rosen CB, Gores GJ, Dickson ER. Deaths on the liver transplant waiting list: an analysis of competing risks. Hepatology. 2006;43(2):345–51.
6. Staplin ND, Kimber AC, Collett D, Roderick PJ. Dependent censoring in piecewise exponential survival models. Stat Methods Med Res. 2015;24(3):325–41. https://doi.org/10.1177/0962280214544018.
7. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1–73. https://doi.org/10.7326/M14-0698.
8. ICH E9 working group. ICH E9 (R1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. EMA/CHMP/ICH/436221/2017. 2020 https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf. Accessed 24 Feb 2020.
9. Pajouheshnia R, Schuster NA, Groenwold RHH, Rutten FH, Moons KGM, Peelen LM. Accounting for time-dependent treatment use when developing a prognostic model from observational data: a review of methods. Stat Neerlandica. 2020;74(1):38–51. https://doi.org/10.1111/stan.12193.
10. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26:2389–430.
11. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130(6):515–24.
12. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099.
13. Peek N, Sperrin M, Mamas M, van Staa T, Buchan I. Hari Seldon, QRISK3, and the prediction paradox. BMJ. 2017;357:2099.
14. Hicks KA, Mahaffey KW, Mehran R, Nissen SE, Wiviott SD, Dunn B, Solomon SD, Marler JR, Teerlink JR, Farb A, Morrow DA, Targum SL, Sila CA, Hai MTT, Jaff MR, Joffe HV, Cutlip DE, Desai AS, Lewis EF, Gibson CM, Landray MJ, Lincoff AM, White CJ, Brooks SS, Rosenfield K, Domanski MJ, Lansky AJ, McMurray JJV, Tcheng JE, Steinhubl SR, Burton P, Mauri L, O'Connor CM, Pfeffer MA, Hung HMJ, Stockbridge NL, Chaitman BR, Temple RJ. Standardized data collection for cardiovascular trials initiative (SCTI). 2017 Cardiovascular and Stroke Endpoint Definitions for Clinical Trials. Circulation. 2018;137:961–72. https://doi.org/10.1161/CIRCULATIONAHA.117.033502.
15. von Dadelszen P, Payne B, Li J, Ansermino JM, Broughton Pipkin F, Côté AM, Douglas MJ, Gruslin A, Hutcheon JA, Joseph KS, Kyle PM, Lee T, Loughna P, Menzies JM, Merialdi M, Millman AL, Moore MP, Moutquin JM, Ouellet AB, Smith GN, Walker JJ, Walley KR, Walters BN, Widmer M, Lee SK, Russell JA, Magee LA, PIERS Study Group. Prediction of adverse maternal outcomes in pre-eclampsia: development and validation of the

fullPIERS model. Lancet. 2011;377(9761):219–27. https://doi.org/10.1016/S0140-6736(10)61351-7.
16. Grunkemeier GL, Jin R, Eijkemans MJ, Takkenberg JJ. Actual and actuarial probabilities of competing risks: apples and lemons. Ann Thorac Surg. 2007;83:1586–92.
17. Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernán MA. A causal framework for classical statistical estimands in failure-time settings with competing events. Stat Med. 2020;39:1199–236. https://doi.org/10.1002/sim.8471.
18. Geskus RB. Data analysis with competing risks and intermediate states. New York: Chapman and Hall/CRC; 2015.
19. Pfeiffer RM, Gail MH. Absolute risk: methods and applications in clinical management and public health. New York: Chapman and Hall/CRC; 2017.
20. van Geloven N, Geskus RB, Mol BW, Zwinderman AH. Correcting for the dependent competing risk of treatment using inverse probability of censoring weighting and copulas in the estimation of natural conception chances. Stat Med. 2014;33:4671–80. https://doi.org/10.1002/sim.6280.
21. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94:496–509.
22. Scheike TH, Zhang MJ, Gerds TA. Predicting cumulative incidence probability by direct binomial regression. Biometrika. 2008;95:205–20.
23. Nicolaie MA, van Houwelingen JC, Putter H. Vertical modelling: a pattern mixture approach for competing risks modelling. Stat Med. 2010;29:1190–205.
24. Sachs MC, Discacciati A, Everhov ÅH, Olén O, Gabriel EE. Ensemble prediction of time-to-event outcomes with competing risks: a case-study of surgical complications in Crohn's disease. J R Stat Soc Ser C (Appl Stat). 2019;68:1431–46.
25. Hernán MA, Robins JM. Causal inference: what if. Boca Raton: Chapman and Hall/CRC; 2020.
26. Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? Epidemiology. 2010;21:872–5.
27. Zheng M, Klein P. Estimates of marginal survival for dependent competing risks based on an assumed copula. Biometrika. 1995;82:127–38.
28. Escarela G, Carriere JF. Fitting competing risks with an assumed copula. Stat Methods Med Res. 2003;12:333–49.
29. Hsu CH, Taylor JMG. Nonparametric comparison of two survival functions with dependent censoring via nonparametric multiple imputation. Stat Med. 2009;28:462–75.
30. Jackson D, White IR, Seaman S, et al. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. Stat Med. 2014;33:4681–94.
31. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log rank tests. Biometrics. 2000;56:779–88.
32. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol. 2008;168:656–64.
33. Matsuyama Y, Yamaguchi T. Estimation of the marginal survival time in the presence of dependent competing risks using inverse probability of censoring weighted (IPCW) methods. Pharm Stat. 2008;7:202–14.
34. Howe CJ, Cole SR, Chmiel JS, et al. Limitation of inverse probability of censoring weights in estimating survival in the presence of strong selection bias. Am J Epidemiol. 2011;173:569–77.
35. Goetgeluk S, Vansteelandt S, Goetghebeur E. Estimation of controlled direct effects. J R Stat Soc B. 2009;70:1049–66.
36. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11:550–60.
37. Ramspek CL, Voskamp PW, van Ittersum FJ, Krediet RT, Dekker FW, van Diepen M. Prediction models for the mortality risk in chronic dialysis patients: a systematic review and independent external validation study. Clin Epidemiol. 2017;9:451–64. https://doi.org/10.2147/CLEP.S139748.
38. Hoekstra T, Hemmelder MH, van Ittersum FJ. RENINE annual report 2015. https://www.nefrovisie.nl/wp-content/uploads/2017/03/RENINE-year-report_08032017.pdf. Accessed 15 Oct 2019.
39. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017. https://www.R-project.org/.
40. Efron B. The efficiency of Cox's likelihood function for censored data. J Am Stat Assoc. 1977;72:557–65.
41. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. Chance. 2019;32:42–9. https://doi.org/10.1080/09332480.2019.1579578.
42. Pajouheshnia R. Prognostic research in treated populations (Doctoral dissertation); 2018. Retrieved from http://dspace.library.uu.nl/handle/1874/371548.
43. Cook RJ, Lawless JF. The statistical analysis of recurrent events. New York: Springer; 2007.
44. Saha P, Heagerty PJ. Time-dependent predictive accuracy in the presence of competing risks. Biometrics. 2010;66(4):999–1011.
45. Schoop R, Beyersmann J, Schumacher M, Binder H. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biom J. 2011;53(1):88–112.
46. Zhang Z, Cortese G, Combescure C, Marshall R, Lee M, Lim HJ, Haller B. written on behalf of AME big-data clinical trial collaborative group. Overview of model validation for survival regression model with competing risks using melanoma study data. Ann Transl Med. 2018;6(16):325.
47. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. BMC Med Res Methodol. 2017;17(1):103.
48. Andersen PK, Keiding N. Interpretability and importance of functionals in competing risks and multistate models. Stat Med. 2011;31:1074–88.
49. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. Berlin: Springer; 2011.