Uncovering hidden influences: Impact of omitted covariates on the estimation of treatment effects using Cox regression in randomized and propensity score matched trials

Thesis
to obtain the academic degree of
Doctor rerum medicarum (Dr. rer. medic.)
in the field of Biostatistics

submitted to the Faculty of Medicine of
Martin Luther University Halle-Wittenberg

by Alexandra Strobel

Supervisor: apl. Prof. Dr. Andreas Wienke

Reviewer:

1. Prof. Antje Jahn, Darmstadt
2. Prof. Annika Hoyer, Bielefeld

Date of defense: 23.04.2025

Referat

Hazard ratios (HRs) sind das am häufigsten verwendete Effektmaß in klinischen Studien, die einen Ereigniszeit-Endpunkt analysieren. In den letzten Jahren gab es jedoch zunehmend Kritik an der Verwendung von HRs, vor allem wegen ihrer Nicht-Kollabierbarkeit und den Herausforderungen bei ihrer (kausalen) Interpretation. Diese Arbeit beleuchtet einen weiteren Aspekt und beschäftigt sich mit den Verzerrungen, die entstehen können, wenn Kovariablen, welche sowohl die Ereigniszeit als auch die Behandlungszuweisung beeinflussen können, nicht in der Analyse beachtet werden. Fehlspezifikationen des Cox Modells aufgrund solcher Kovariablen können sowohl in randomisierten als auch in Propensity Score (PS) gematchten Studien zu verzerrten Behandlungsschätzungen führen. Oft bleibt jedoch unklar, ob (und ja, welche) Kovariablen für diese Verzerrungen verantwortlich sind. Die vorliegende Arbeit stellt einen neuen methodischen Ansatz namens *Dynamic Landmarking* vor, welcher einen visuellen Hinweis auf verzerrte Effektschätzungen bietet und hilft, Kovariablen, die diese Verzerrungen verursachen, zu identifizieren. Das Verfahren basiert auf einer sukzessiven Löschung sortierter Beobachtungen und der wiederholten Schätzung von Cox Modellen, bis die Anzahl der verbleibenden Ereignisse nicht mehr ausreicht, um eine valide Schätzung zu liefern. Ergänzend wird in jedem Schritt die Balance der beobachteten, aber unberücksichtigten Kovariablen anhand der Summe der quadratischen z-Differenzen bewertet. Durchgeführte Simulationsstudien zeigen, dass *Dynamic Landmarking* ein effektives Werkzeug ist, um verzerrte Behandlungsschätzungen in den beiden ausgewählten Studiendesigns zu erkennen. Während in randomisierten Studien relevante prognostische Faktoren identifiziert werden können, die einen Selektionsbias verursachen, ermöglicht die Methode in PS gematchten Studien die klare Unterscheidung zwischen prognostischen Faktoren und Confoundern. Darüber hinaus wurde die Methode genutzt, um den Selektionsbias in 27 randomisierten kontrollierten Studien (RCTs) zu untersuchen. Dabei konnten keine empirischen Hinweise auf den Selektionsbias gefunden werden, was darauf hindeutet, dass dieser Bias in vielen Fällen von geringer praktischer Bedeutung ist. Dies lässt sich vor allem durch kleine Behandlungseffekte und homogene Patientenpopulationen aufgrund strenger Ein- und Ausschlusskriterien erklären. Zusammenfassend zeigt sich, dass *Dynamic Landmarking* ein geeignetes Instrument zur Prüfung von Behandlungseffekten aus Cox Modellen ist. Es ermöglicht die Identifikation potenzieller Verzerrungen und ihrer Ursachen. Die empirische Analyse von RCTs legt nahe, dass HRs durch Selektionsbias kaum beeinträchtigt werden und daher in dieser Hinsicht als zuverlässiges Effektmaß genutzt werden können.

Strobel, Alexandra: Uncovering hidden influences: Impact of omitted covariates on the estimation of treatment effects using Cox regression in randomized and propensity score matched trials, Halle (Saale), Univ., Med. Fac., Diss., 28 pages, 2025

Abstract

Hazard ratios (HRs) are the most common treatment effect measures in clinical trials focusing on time-to-event outcomes. However, in recent years there has been increasing criticism of HRs, particularly regarding their non-collapsibility or with respect to their (causal) interpretation. This work addresses another critical aspect related to unobserved or omitted covariates that impact the survival outcome and/or the treatment allocation but are frequently disregarded. Misspecification of the Cox model due to such covariates could result in heavily biased treatment estimates, affecting both randomized and propensity score (PS) matched trials. However, researchers frequently lack clarity on whether (and, if so, which) covariates might induce this bias. Therefore, this work presents a methodological approach called *Dynamic Landmarking*, that provides a visual indication of potentially biased treatment effect estimates obtained from Cox models and identifies omitted covariates that could cause this bias. The approach is based on successive deletion of sorted observations and gradually refitting Cox models until no sufficient number of events is contained in the data. In addition, the balance of observed, but omitted covariates is assessed using the sum of squared z-differences. Using simulation studies, it was demonstrated that *Dynamic Landmarking* indeed serves as an effective visual tool for detecting biased treatment estimates in both study designs. In randomized settings, relevant omitted prognostic factors have been identified that cause so-called built-in selection bias. Regarding PS matched trials, *Dynamic Landmarking* successfully identified relevant omitted prognostic factors and confounders, making a clear distinction between them. Furthermore, the method was used to assess the built-in selection bias in individual patient data from 27 large randomized controlled trials (RCTs). No empirical evidence of this bias has been found in these studies, which leads to the conclusion that this type of bias is of limited practical relevance in the majority of cases. This is mainly due to small treatment effects and homogeneous patient populations resulting from strict inclusion and exclusion criteria. In summary, *Dynamic Landmarking* can be used to verify if estimated treatment effects obtained from a Cox model are biased and whether measured but omitted covariates cause this bias. The empirical investigation of RCTs suggests that HRs are not materially affected by the built-in selection bias and can therefore be safely used, at least concerning this aspect.

Strobel, Alexandra: Uncovering hidden influences: Impact of omitted covariates on the estimation of treatment effects using Cox regression in randomized and propensity score matched trials, Halle (Saale), Univ., Med. Fac., Diss., 28 pages, 2025

Contents

Abbreviations

RCT	Randomized controlled trial
PS	Propensity score
PSM	Propensity score matching
HR	Hazard ratio
w.l.o.g.	Without loss of generality
AFT	Accelerate failure time
$\mathbf{SSQ_{zDiff}}$	Sum of squared z-differences
$\mathbf{N(\mu, \sigma^2)}$	Normal distribution with expectation μ and variance σ^2
$\boldsymbol{\mathcal{X}^2_k}$	Chi-Squared distribution with k degrees of freedom

List of Figures

1 Introduction and objectives

1.1 Statistical setting

1.1.1 Covariate distribution in randomized and propensity score matched trials

Randomized controlled trials (RCTs) are considered the gold standard for evaluating research data and efficiently translating results into clinical practice. As highlighted in the CONSORT statement, they serve as the foundation of evidence-based medicine, providing the highest level of evidence[1]. The primary strength of this study design lies in its high internal validity resulting from the randomization process, which ensures an equal distribution of both observed and unobserved covariates[2–4]. In this sense, the treatment groups analyzed are often referred to as "exchangeable", defining a condition in which the treatment groups have comparable risks of an outcome, allowing for valid causal inferences in the counterfactual framework. This especially implies the possibility of estimating an average causal treatment effect in RCTs[5–7].

However, ethical and practical constraints increasingly require the use of non-randomized trials to estimate treatment effects. These face the challenge that treatment allocation may depend on (un-)observed covariates, leading to systematic differences in baseline characteristics between groups[8]. Disregarding such differences would lead to heavily biased treatment effect estimates[9]. A prominent way to deal with this covariate imbalance is through propensity score (PS) methods. Rosenbaum and Rubin introduced the PS in 1983 as a balancing score[10]. It describes the probability p_i for an individual i $(i = 1, ..., N)$ to receive a treatment Z_i conditional on a set of observed covariates X_i: $p_i = P(Z_i = 1|X_i)$. In the context of RCTs, the true PS equals 0.5, as all individuals have an equal chance of receiving treatment by design[11]. In contrast, individuals in a non-randomized trial generally have varying PS values, which are typically estimated using logistic regression. One important advantage of the PS lies in its ability to summarize all relevant confounding factors into a single numerical value. This not only simplifies the analysis, but also reduces the risk of overfitting in non-randomized trials[12]. Overall, four different PS methods are widely used in the literature: matching, stratification, adjustment and inverse probability of treatment weighting[13–15]. Under the assumptions of positivity, consistency and unconfoundedness, all of the PS methods mentioned provide an unbiased causal treatment effect estimate if all other assumptions of the model used for data analysis hold. Currently, propensity score matching (PSM) is especially popular for analyzing medical data, as ev-

idenced by two systematic reviews covering several medical fields[16,17]. PSM entails to create a sample of individuals who share a similar value of the PS. This process leads to, on average, well-balanced treatment groups, indicating that given the true PS, individuals who received treatment, and those who did not, have a similar distribution of observed baseline covariates. Therefore, the resulting sample after PSM can mimic that of an RCT, especially in terms of exchangeability[18–21]. Consequently, direct comparisons of outcomes between individuals in the treatment groups can be made, and treatment effects can be assessed using metrics identical to those in RCTs.

In recent years, several balance diagnostics have been introduced to examine the comparability of treatment groups according to PSM. One of the most prominent methods is the use of standardized differences[22], which can be calculated for both continuous and binary covariates by

$$d_{con} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}} \tag{1}$$

$$d_{bin} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)}{2}}} \tag{2}$$

with $\overline{x}_i,\ s_i^2$ and $\hat{p}_i$ being the estimated means, variances and proportions of the treatment group i $(i = 1, 2)$. The balance measures (1) and (2) compare differences in units of the pooled standard deviation[14]. However, Kuss (2013) mentioned some crucial disadvantages of the standardized difference, including the dependency on sample size in its distribution and the non-comparability on different scales[23]. More precisely, the large sample distribution of the standardized differences converges to a normal distribution with an expected value of zero and variance $\frac{n_1+n_2}{n_1 \cdot n_2}$, assuming that the true standardized difference equals zero and that the treatment groups are independent[22]. To address these limitations, the z-differences as alternative balance measure were introduced. For continuous and binary covariates they are defined as

$$z_{con} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3}$$

$$z_{bin} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{4}$$

where $\overline{x}_i,\ s_i^2,\ n_i$ and $\hat{p}_i$ denote estimated means, variances, sample sizes, and proportions in the treatment group i $(i = 1, 2)$. Additionally, a z-difference for ordinal covariates z_{ord} was also proposed by Kuss (2013) and in case of nominal covariates it is suggested to calculate

each binary z-difference for all nominal categories[23]. The z-differences are asymptotically $N(0,1)$-distributed in RCTs and follow a $N\left(0,\frac{1}{2}\right)$-distribution in a perfectly PS matched trial[11]. While the z-difference can be used for describing covariate-specific balance, it is also possible to construct a global measure for overall-balance of the entire set of covariates by the sum of squared z-differences (SSQ_{zDiff}). Assuming n continuous, m binary and p nominal covariates, the SSQ_{zDiff} is calculated as

$$SSQ_{zDiff} = \sum_{i=1}^{n} z_{i,con}^2 + \sum_{j=1}^{m} z_{j,bin}^2 + \sum_{k=1}^{p} z_{k,ord}^2. \tag{5}$$

In RCTs the SSQ_{zDiff} follows an approximate $\mathcal{X}^2_{m+n+p}$-distribution with an expected value of $m+n+p$ under the assumptions of no baseline differences, independent treatment groups and independence of all covariates. In contrast, the expected value of SSQ_{zDiff} equals $\frac{m+n+p}{2}$ under the same assumptions in a perfectly PS matched trial. Hence, the distribution of SSQ_{zDiff} will in general not depend on the sample size. As an aggregate measure it provides information on the overall balance of the covariates, making it suitable as an indicator for determining the success of randomization and PSM[24].

1.1.2 Hazards and Hazard Ratios

When analyzing a time-to-event outcome in clinical trials, the Cox model[25,26] is commonly used because it allows dealing with censored observations and provides the hazard ratio (HR) as a single-number summary of the treatment effect. The Cox model specifies the hazard of a time-to-event T as

$$\lambda(t|\mathbf{X}) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t, \mathbf{X})}{\Delta t} = \lambda_0(t) \cdot \exp(\beta^T \mathbf{X}) \tag{6}$$

where $\lambda_0(t)$ describes an unspecified baseline hazard function that is assumed to be common for all individuals i $(i = 1, ..., N)$, $\mathbf{X}$ is a $p \times 1$ vector of observed covariates with a corresponding $p \times 1$ vector of regression coefficients β. Importantly, the risk set at time t is composed of individuals that have not yet experienced the event of interest and have not yet been removed for other reasons, such as censoring. Suppose that model (6) holds for a covariate vector $\mathbf{X} = (X_{trt}, \mathbf{W})^T$, where X_{trt} is a binary treatment variable and $\mathbf{W}$ represents additional covariates, with corresponding regression coefficients $\beta = (\beta_{trt}, \beta_W)^T$. Within this framework, the true model is given by

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \cdot \exp(\beta_{trt} X_{trt} + \beta_W^T \mathbf{W}). \tag{7}$$

The hazard $\lambda(t|\mathbf{X})$ defined in model (7) is referred to as *conditional hazard* with β_{trt} summarizing the *conditional effect* of the treatment X_{trt}, providing a subject-specific interpretation, i.e., what treatment effect can be expected when moving an individual from one group to the other[27]. However, if $\mathbf{W}$ is either unobserved or omitted from the model the resulting *marginal hazard* would be

$$\lambda(t|X_{trt}) = \lambda_0(t) \cdot \exp(\beta_{trt} X_{trt}). \tag{8}$$

The estimated *marginal* treatment effect β_{trt} from model (8) lacks of an individual interpretation, but represents a population-average effect, moving a whole population from one group to the other[28–30]. Put differently, $\lambda(t|X_{trt})$ corresponds to a weighted average of the individual hazards for those in the risk set at time t, where the weights are determined by the distribution of $\mathbf{W}$ within this risk set. In clinical trials, its common to present and interpret the treatment effect as a HR that is directly derived from model (6). Assume therefore again $\mathbf{X} = (X_{trt}, \mathbf{W})^T$, then

$$\begin{aligned} \frac{\lim_{\Delta t \to 0} P(t < T \leq \Delta t | T > t, X_{trt} = 1, \mathbf{W} = \mathbf{w})}{\lim_{\Delta t \to 0} P(t < T \leq \Delta t | T > t, X_{trt} = 0, \mathbf{W} = \mathbf{w})} &= \frac{\lambda_0(t) \cdot \exp(\beta_{trt} \cdot 1 + \beta_W^T \mathbf{w})}{\lambda_0(t) \cdot \exp(\beta_{trt} \cdot 0 + \beta_W^T \mathbf{w})} \\ &= \frac{\lambda_0(t) \cdot \exp(\beta_{trt} \cdot 1) \cdot \exp(\beta_W^T \mathbf{w})}{\lambda_0(t) \cdot \exp(\beta_{trt} \cdot 0) \cdot \exp(\beta_W^T \mathbf{w})} \\ &= \exp(\beta_{trt}) \end{aligned} \tag{9}$$

yields a single-number summary for the treatment effect, assuming the covariates $\mathbf{W}$ to be equal and time-invariant between the compared individuals. However, $\exp(\beta_{trt})$ from (9) is deceptive for several reasons, which are explained in detail in the following.

1.2 Challenges in estimating Hazard Ratios

Several challenges arise when presenting a HR as main result of a clinical trial. Firstly, HRs are often misinterpreted as relative risk. Sutradhar and Austin (2018), have emphasized that the magnitude of the HR cannot be interpreted as the magnitude of the relative risk. Yet, it can still suggest the direction of the estimated effect[31]. As a consequence, wrongly interpreted HRs provide misleading context in the field of survival analysis and should therefore be avoided by researchers[32]. A second important issue arises due to non-collapsibility of HRs, indicating that the magnitude of the effect measure is changing when conditioning on a covariate that is associated with the time-to-event[28]. Bian et al. (2024) describe the non-collapsibility as a property of a coefficient in a statistical model, stating: "When focusing on the coefficient of the treatment and collapsing over covariate(s), the

conditional effect does not equal the marginal effect even in the absence of confounding and effect modification"[33]. Consequently, marginal and conditional HR will not provide the same estimate for a treatment effect[34–36]. This further implies that different studies investigating the same time-to-event and treatment but adjust for different covariates will generally yield diverse estimates for the HR of treatment, even if the theoretical causal effect will be identical[36,37]. Moreover, the Cox model relies on certain strong assumptions, which are often to restrictive for the data. For example, it assumes proportional hazards, however, there are some scenarios where this assumption does not seem reasonable, such as early, delayed, diminishing, cure or crossing effects[38]. In addition, even when the proportional hazards assumption is satisfied, the model must be correctly specified, meaning that all relevant covariates influencing treatment allocation and/or survival outcomes must be considered[39]. This implicitly goes along with the assumption of homogeneity concerning unobserved or omitted covariates, suggesting that all included individuals are essentially identical regarding all covariates which are not considered in the model. In other words, every individual is assumed to have the same baseline risk of experiencing an event with regard to unobserved or omitted covariates[40]. Only if this holds for the fitted Cox model, the conditional treatment effect will be estimated consistently[29,41]. Please note, in present work the term "omitted covariate" will from now on refer to a covariate that was measured during the trial but disregarded from data analysis.

1.2.1 Omitted covariates in randomized controlled trials

In RCTs the HR is typically reported without any adjustments, because authors argue that randomization ensures, on average, an identical distribution for all observed and unobserved covariates. However, as highlighted by Hernán (2010) and others, treatment effect estimates may be biased without adjustments due to built-in selection bias (also referred to as "the depletion of the susceptible" or "survival of the fittest")[29,42–44]. Assume again $\mathbf{X} = (X_{trt}, \mathbf{W})^T$ as previously described in Section 1.1.2. Since patients were randomized with respect to X_{trt}, $\mathbf{W}$ represents a set of covariates that are unrelated to treatment allocation but may serve as prognostic factors influencing patients' survival time. Researchers are usually interested in the individual-specific treatment effect, which can be obtained by the conditional Cox model (7). Under the proportional hazards assumption, this model would be correctly specified, providing a conditional HR estimated by $\exp(\beta_{trt})$. However, if any part of $\mathbf{W}$ is disregarded, one would employ only a marginal Cox model (marginal with respect to the omitted covariate), which will in general differ from the conditional one as the Cox model is non-collapsible[28,45]. As a direct consequence

of omitting any covariate of $\mathbf{W}$, "unobserved" heterogeneity will be induced, implying that the patients differ in their baseline risk concerning this omitted prognostic factor[43,44]. The effect is typically mentioned in frailty theory[27,46–49]. More precisely, without loss of generality (w.l.o.g.) let higher values of $\mathbf{W}$ signify higher baseline risk of getting an event of interest (indicating more frail or high-risk individuals). Then high-risk individuals tend to experience the event of interest earlier than low-risk individuals. In addition, having an effective treatment, which reduces, e.g. the risk of dying, will decrease the prevalence of high-risk individuals faster in the untreated group (e.g., placebo) than in the treated group, leading to a systematic selection process that results in non-exchangeable groups after randomization (see Figure 1).

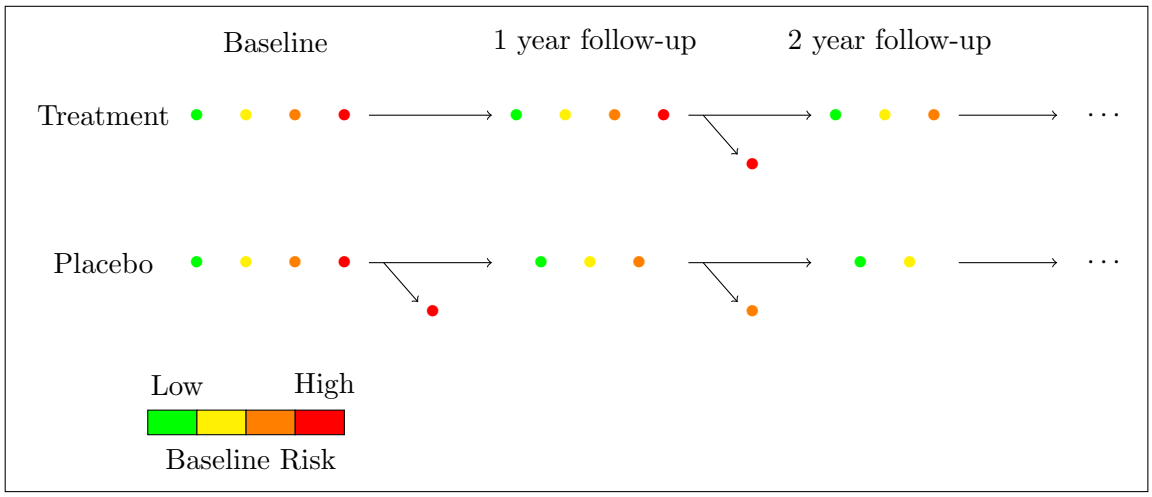

Figure 1. Built-in selection bias in RCTs due to an omitted prognostic factor. Patients with higher baseline risk regarding the omitted covariate tend to experience the event of interest earlier, leading to a systematic selection process during follow-up time. Consequently, patients from treatment and placebo group are non-exchangeable after randomization. Figure is based on Stensrud et al. (2019)[43].

Therefore, conditioning the hazard on having survived up to a specific time t will in general lead to a systematically different distribution of the omitted covariate in the treatment groups during follow-up[39,42]. This effect can also be derived from equation (9), showing that $\exp(\beta_{trt})$ contrasts the hazard functions with and without intervention for two separate groups of individuals: those who survive time $t > 0$ with treatment ($X_{trt} = 1$) and those who survive time $t > 0$ without treatment ($X_{trt} = 0$). These groups will in general fail to be comparable if unobserved heterogeneity is induced by omitted or unobserved covariates. In conclusion, the HR is obtained from groups that systematically differ in omitted characteristics, making them non-comparable. As a result, the HR cannot be interpreted in a causal manner because the counterfactual framework is disrupted by the

inherent selection bias[28,29,50]. It was pointed out by Stensrud et al. (2019), that if omitted covariates are assumed to be multiplicative on the hazard scale, the built-in selection bias will increase with the magnitude of the treatment effect, the heterogeneity in risk at baseline and the length of follow-up[43]. Additionally, higher (right) censoring rates have been shown to result in less biased treatment effect estimates[51].

1.2.2 Omitted covariates in propensity score matched trials

When interpreting treatment effects from PS matched trials, two important assumptions regarding model specification have to be made. Both, the PS model and the Cox model have to be correctly specified. Model misspecification in both cases lead to worries about wrong causal statements[52,53]. For the PS model (usually logistic regression), all relevant confounders have to be considered and correctly included. Drake (1993) found that omitting counfounders would lead to substantially biased treatment effect estimates, as there will be residual confounding bias[54]. This was also confirmed by Dehejia and Wahba (1999), who additionally found, that the causal estimates were not sensitive to the specification of the functional form of the PS, once all important covariates had been included[55]. The main issue when omitting a confounder from the PS model arises because the treatment groups are not comparable after PSM due to residual counfounding bias. This implies that the treatment groups differ on certain baseline characteristics and are therefore not exchangeable. Hence, the estimated HR does not yield a causal interpretation. However, even if the PS model is correctly specified and no unobserved or omitted confounders are present, there may be prognostic factors that, while not affecting treatment allocation, could influence patients' survival outcome. The PS model does not consider such covariates, and neglecting them could introduce built-in selection bias (see Section 1.2.1). As well as in RCTs, prognostic factors would be equally distributed between the treatment groups after PSM as they are assumed to be independent of the treatment allocation. However, if the treatment is effective and the prognostic factor influences the survival outcome, the consequence would be a systematic elimination of individuals during follow-up, resulting in the built-in selection bias[51,56]. Hence, only after PSM the treatment groups would be exchangeable and estimating a marginal Cox model in a PS matched trial would therefore not yield a treatment effect estimate with individual-specific interpretation[56]. Overall, the main idea of PSM (and randomization respectively) is to ensure that patients in the treatment groups do not differ in any characteristic except for their treatment allocation. Additionally, the analyzed population is often assumed to be homogeneous with respect to omitted covariates, meaning that patients do not differ in their individual baseline risk

of getting an event of interest regarding unobserved or omitted covariates (see Figure 2A). However, two issues could arise when a covariate is omitted from data analysis. On the one hand, omitting a true confounder from the PS model, would lead to residual confounding bias, resulting in treatment groups that differ in one (or even more) baseline characteristics even after PSM. Treatment groups are then non-comparable at baseline as w.l.o.g patients with higher baseline risk regarding the omitted covariate are more likely to be allocated into the treatment group (see Figure 2B). On the other hand, omitting a prognostic factor from the Cox model would result in the previously described built-in selection bias, i.e. the treatment groups are comparable after PSM, however patients differ in their individual baseline risk regarding the covariate leading to systematic selection during the follow-up period (see Figure 2C and Figure 1).

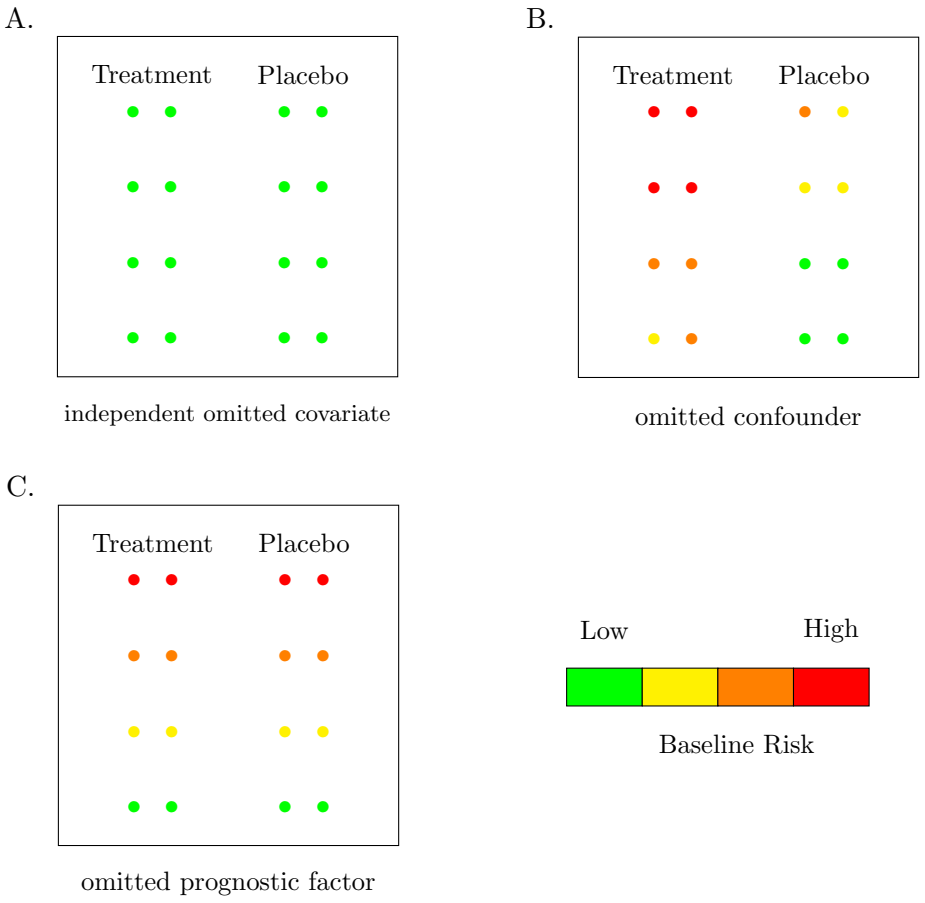

Figure 2. Distribution of an omitted covariate after PSM. **A.** Independent omitted covariate that does neither influence the time-to-event nor the treatment allocation. Groups are exchangeable and patients do not differ in their baseline risk of getting an event. **B.** Independent omitted confounder influencing both, treatment allocation and time-to-event. Groups are non-exchangeable after PSM as, w.l.o.g., patients with higher baseline risk are more likely to receive treatment. **C.** Independent omitted prognostic factor only influencing patients time-to-event. Groups are exchangeable after PSM, but patients differ in their baseline risk of getting an event of interest.

In both RCTs and PS matched trials, usually a marginal Cox model (that is, model (8) described in Section 1.1.2) is fitted with treatment as the only variable. Although one has to specify two models in the non-randomized setting, this also comes with one main advantage. In case the omitted covariate is correlated with one or more considered confounders from the PS model, residual confounding and/or built-in selection can be minimized[56–58]. Rubin and Thomas (1996) stated that "excluding potentially relevant variables should be done only [...] when the excluded variables are highly correlated with variables already in the propensity score model"[59]. Indeed, recent work found that replacing a highly correlated covariate instead of the true confounder in the PS model would result in a relative bias less than 5%[57]. Due to the correlation, the omitted covariate will indirectly be considered by the PS model and will consequently be matched by design.

1.3 Research question

HRs have been criticized during recent years and this criticism is underlined by many theoretical contributions. One major point involves the omission of covariates that are disregarded during data analysis. In RCTs such covariates are referred to as prognostic factors, which induce the built-in selection bias. In PS-matched trials the bias due to omitted covariates has two potential sources. On the one hand, omitted confounders are an issue, when covariates determine the treatment allocation and the survival outcome. Then misspecification of the PS-model might be a result. On the other hand, omitted prognostic factors in non-randomized studies can also introduce the built-in selection bias if the Cox model is misspecified. However, researchers often do not know whether and, if so, which covariates might cause the bias. This is mainly because the HR, as a single-number summary, provides no indication of whether bias is present or not. Therefore, the work aims to

1. present a methodological framework, *Dynamic Landmarking*, designed for RCTs and PS-matched trials, which visualises whether an estimated HR is subject to built-in selection or confounding bias and identifies omitted covariates causing it.

2. conduct an empirical investigation of individual patient data from 27 large RCTs, to assess the magnitude and clinical relevance of the built-in selection bias.

2 Discussion

2.1 *Dynamic Landmarking* and its application

Dynamic Landmarking is a methodological approach that provides a visual tool for identifying whether treatment effect estimates in RCTs and PS matched trials are subject to built-in selection or confounding bias. Additionally, omitted covariates which are measured during the trial but omitted from data analysis are investigated whether they induce these biases. The main idea of *Dynamic Landmarking* is based on the Landmarking approach of Van Houwelingen[60–62]. In a first step, the balanced data (either by randomization or PSM) is sorted by observation time and a univariable Cox model with treatment as the only variable is fitted to the full data set. Afterwards, the earliest M ($M > 0$) observations are deleted regardless of the event status (censored or time-to-event) and a new Cox model is fitted to the smaller data set. This procedure of deleting earliest observations and refitting Cox models is continued until the data set no longer contains a sufficient number of observations for convergence. In parallel, the SSQ_{zDiff} measures the balance of observed but omitted covariates in each step.

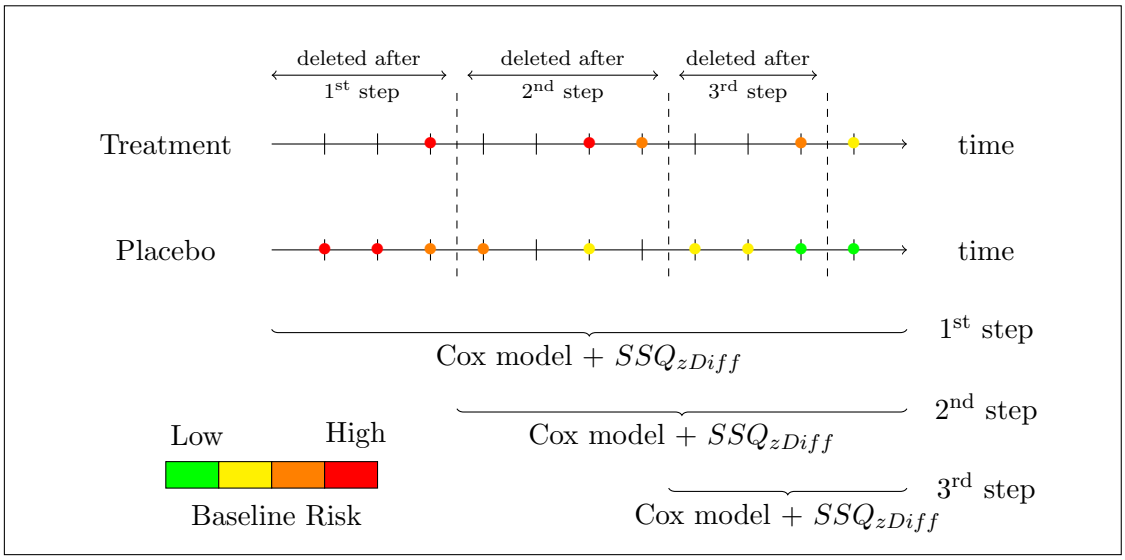

Figure 3. *Dynamic Landmarking* procedure for the omission of an independent prognostic factor, which causes built-in selection bias. Approach is based on successive deletion of M sorted observations (in Figure $M = 4$ is assumed in each step) and refitting Cox models for the smaller data set.

Assume again, w.l.o.g., that higher values of the omitted covariates indicate higher baseline risk. It then seems reasonable that high-risk individuals would generally have shorter observation times than low-risk patients, as they tend to experience the event earlier.

Consequently, individuals with higher risk will appear first in the list of observation times sorted in ascending order and are resultantly deleted at first during *Dynamic Landmarking*[63,64]. A graphical illustration of the procedure is given in Figure 3. By collecting the estimated treatment effects as log(HR) and the SSQ_{zDiff} after each deletion step, two trajectories as functions of remaining observations are obtained (see Figure 4 for an exemplary illustration of the graphical output from *Dynamic Landmarking* and Publications for detailed explanations). Within different simulation scenarios, it was shown that *Dynamic Landmarking* is able to visualize treatment effect estimates underlying either built-in selection or confounding bias resulting from omitted covariates in both study designs. Precisely, a systematic shift in the treatment effect trajectory can be seen, resulting from the non-random successive deletion of individuals. The simulation study confirmed previous findings, demonstrating that larger treatment effects and greater influence of omitted covariates on patients' survival and/or treatment allocation result in stronger bias[41,43]. Additionally, it was also validated that an omission of a covariate, which is correlated with one included in the PS model, would lead to less biased treatment effect estimates - at least if the correlation is assumed to be strong[51]. Furthermore, the method distinguishes between omitted prognostic factors and omitted confounders. While omitted prognostic factors are balanced in the initial estimation procedure (in both RCTs and PS-matched trials), omitted confounders exhibit significant imbalance even before the first deletion step, as they are associated with treatment allocation. Hence, the initial value of SSQ_{zDiff} will give a first hint on the causal direction of the omitted covariate. Through this differentiation based on the balance, it is not only possible to identify HRs subjected to bias, but also potential variables causing it. Depending on whether it is a prognostic factor or a confounder, the statistical analysis must be adjusted (see Figure 5 for recommendations). In addition, if the trajectory of the treatment effect indicates the presence of bias, but the omitted covariates did not cause it, other statistical models should be considered (see Section 2.3).

2.2 Evidence of built-in selection bias in randomized controlled trials

Dynamic Landmarking was used to assess the built-in selection bias in a large sample of RCTs. Concretely, publicly available individual data sets from 32 RCTs, which were already used for methodological investigation by Kent et al. (2016)[65], were considered. Each trial had a time-to-event outcome; however, to avoid problems with competing risks, the empirical investigation was restricted to all-cause mortality, which reduced the assessed sample of trials to 27 RCTs, each with more than 1,000 individual observations.

The RCTs were originally conducted between 1980 and 2010. Most trials (74.1%) were carried out in the field of cardiovascular research, including the evaluation of interventions for atrial fibrillation[66], acute myocardial infarct[67–69], acute stroke[70], heart failure[71,72] or hypertension[73]. Moreover, the sample also included some other conditions, such as pre-diabetes[74,75], chronic hepatitis C[76] or acute kidney failure[77,78]. *Dynamic Landmarking* was applied to each RCT while considering a marginal Cox model, which only includes treatment as variable. Age and sex were used as baseline covariates to measure balance by SSQ_{zDiff}, as these are the only two prognostic factors that were equally collected in all RCTs. Furthermore, age and sex can reasonably be assumed to be independent which is a main assumption for the distribution of SSQ_{zDiff} (that is $\mathcal{X}_2^2$ for two covariates in each RCT). Additionally, both age and sex showed a small to medium effect in a univariable Cox model in each trial, leading to the conclusion that both covariates are prognostic factors influencing the survival outcome and may therefore induce heterogeneity. An exemplary illustration of *Dynamic Landmarking* obtained from the Action to Control Cardiovascular Risk in Diabetes (ACCORD) blood pressure (BP) trial[74] is shown in Figure 4.

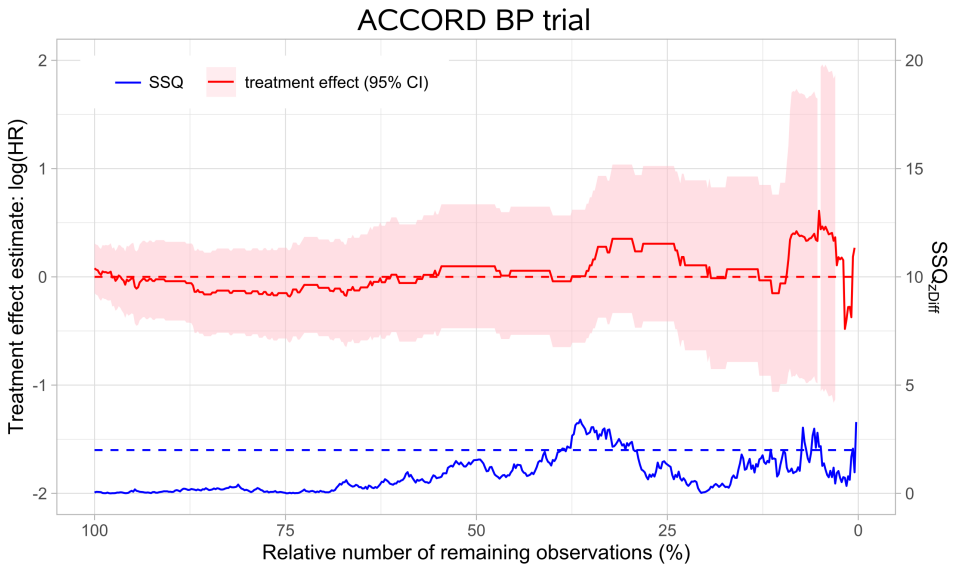

Figure 4. Graphical illustration of *Dynamic Landmarking* in the ACCORD BP trial ($M = 10$). Trajectories of treatment effect estimates (solid red line) and SSQ_{zDiff} (solid blue line) as function of remaining observations are shown. Dashed lines symbolize no treatment effect (red), i.e. $\log(HR) = 0$, and balanced prognostic factors (blue), i.e. $SSQ_{zDiff} = 2$. The Cox model only includes treatment as variable and balance was measured using two omitted covariates, namely age and sex.

This RCT aimed to investigate whether therapy targeting normal systolic pressure (i.e., below 120 mm Hg) reduces major cardiovascular events in participants with type 2 diabetes at high risk for cardiovascular event. Overall survival was examined as secondary

endpoint[74]. The study consists of 4,733 patients with $M = 10$ individuals being deleted at each step of *Dynamic Landmarking*. During the deletion process, no systematically changing treatment effect trajectory was found and only random fluctuations were observed. In addition, the omitted prognostic factors stayed balanced as indicated by SSQ_{zDiff}. Hence, following the interpretation scheme (see Figure 5), there is no indication that the estimated hazard ratio is subject to built-in selection bias.

Overall, no empirical evidence of the built-in selection bias was found in the 27 RCTs investigated, including the ACCORD BP study presented in Figure 4. The absence can be explained in a simple way: the built-in selection bias would only be notable, if both, large treatment effects and large heterogeneity (i.e., omitted covariates have high impact on survival outcome) are present in the data[27,43,79]. However, very large treatment effects are uncommon in RCTs because equipoise is expected when conducting an RCT. This is according to a large review by Djulbegovic et al. (2012), which compared new, experimental treatments with established ones from 743 RCTs to determine the extent to which the newer treatments were more effective. The authors showed, that treatment effects are in general symmetrically distributed, resulting in unpredictability of new treatment effects. Precisely, only slightly more than one half of new treatments performed better than the established treatments. Importantly, the time at which the study was conducted did not affect these results, allowing the conclusion to apply to the studies examined in the this work[80]. In addition, the simulation study that was performed showed that there will be a notable systematic shift in the trajectory of the treatment effect if the true hazard ratio would be 2 or greater.[63] However, the equipoise of treatment effects leads to the fact, that treatment effects with a magnitude of $\log(HR) = 1$ are only seen in about 3% and a magnitude of $\log(HR)$ larger than 1.5 essentially never occurs[80]. Even in the case the treatment effect would be large enough, there still has to be large heterogeneity present in the data. However, the study population is generally well selected due to strong exclusion and inclusion criteria and heterogeneity between patients is thus minimized in RCTs. Indeed, a systematic review comparing 305 trials of treatment for physical condition found, that more than one half of the studies excluded 75% or more patients due to exclusion criteria[81]. This is accompanied by an additional analysis that estimates the variance in frailty using a univariable gamma frailty model of the 27 RCTs and finds that this variance is usually close to zero, reflecting a negligible amount of unobserved heterogeneity between patients in the data[63]. Furthermore, as mentioned above, the prognostic factors age and sex showed only a small to moderate effect on the time-to-event, which also indicates low heterogeneity. A third reason for the absence of the built-in selection bias might be

the remarkably high censoring rates of the analyzed RCTs. Almost three quarters of the investigated trials showed an event rate less than 20%. Assuming independent censoring during the follow-up, less biased treatment effect estimates would be expected, as censored patients will have no major impact on the built-in selection bias[51]. Overall, low treatment effects, low heterogeneity and high censoring rates in the RCTs result in the conclusion of no empirical evidence of the built-in selection bias in RCTs. Thus, HRs are not as hazardous as announced in literature, at least with respect to this issue. The warnings about the built-in selection bias in RCTs are mainly of theoretical nature and have only little practical relevance in most cases.

2.3 Methods to address biased treatment effect estimates

Dynamic Landmarking provides a visual tool for identifying whether treatment effect estimates from time-to-event analyses in RCTs and PS matched trials are subject to built-in selection or confounding bias. In addition, the methodological approach could identify omitted covariates inducing these kinds of biases.

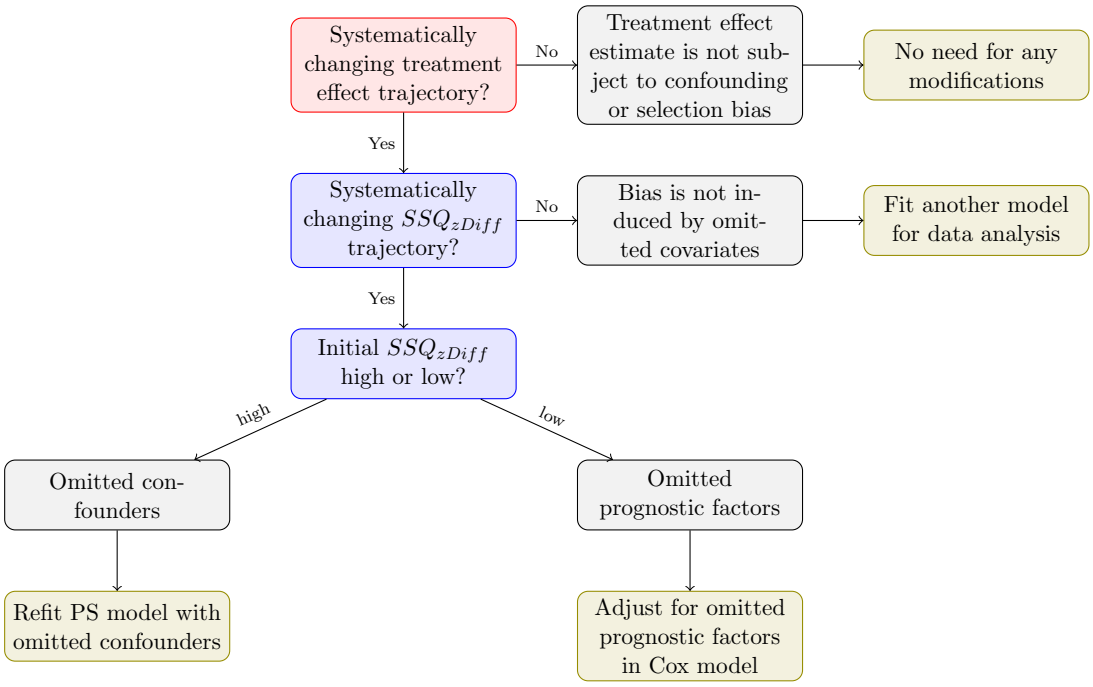

Figure 5. Interpretation and recommendation for *Dynamic Landmarking* results. Red box is related to treatment effect trajectories, blue boxes are related to SSQ_{zDiff}-trajectories. Grey boxes give possible interpretations for course of trajectories and green boxes provide recommendations for further data analysis.

Although *Dynamic Landmarking* identifies the problem, it will not correct for it. Therefore, the user has to decide how to deal with potentially biased treatment estimates in such cases. Importantly, one should consider different correcting methods while taking into account clinical expertise. Overall, the interpretation and recommendation scheme given in Figure 5 provides an inside in how to deal with the graphical output of *Dynamic Landmarking*. In the following, a small selection of correcting methods that might address the case of systematically changing treatment effect trajectories are is presented.

2.3.1 *Dynamic Landmarking* identified omitted covariates

In case *Dynamic Landmarking* identifies omitted covariates that induce built-in selection or residual confounding bias, then addressing this issue becomes quite straightforward. Omitting a prognostic factor (i.e., systematically changing treatment effect trajectory and low initial SSQ_{zDiff}-value that increase during the deletion process) can be handled by simply adjusting the Cox model for the omitted covariate[82]. This approach can be applied to both, RCTs and PS matched trials. In PS matched trials residual confounding bias becomes an issue when omitting a confounder from the PS model. In that case, *Dynamic Landmarking* shows high initial SSQ_{zDiff}-values and two possible corrections can be considered. First, adjusting the Cox model for the counfounder might be an option. A second possibility would be to re-estimate the PS model while considering previously omitted confounders. Stürmer et al. (2005) found in a systematic review that "there is little evidence that these methods yield substantially different estimates compared with conventional multivariable methods"[83]. This was also confirmed by Elze et al. (2017) who stated: "PS methods are not necessarily superior to conventional covariate adjustment, and care should be taken to select the most suitable method"[12]. Hence, both approaches are suitable for addressing residual confounding bias caused by omitted covariates[84–87].

2.3.2 *Dynamic Landmarking* did not identify omitted covariates

In case *Dynamic Landmarking* provides a systematically changing treatment effect trajectory but no systematically changing SSQ_{zDiff} can be observed, then omitted covariates do not cause built-in selection or confounding bias and other statistical models for data analysis must be considered. The reasons for such a behavior of *Dynamic Landmarking* can be numerous. It is possible that the true treatment effect is actually time-dependent, or that unobserved covariates could cause true unobserved heterogeneity, resulting in unmeasurable built-in selection bias. Furthermore, the PS model could be wrongly specified or the multiplicative assumption of the Cox model is too restrictive for the analyzed data.

There are several methods in the literature to address each of these issues, however, the main challenge is to decide for the correct one depending on the clinical setting. Assume, e.g., a violation of the proportional hazards assumption, which might be observed in case of a delayed treatment effect or if a cure effect arises[88–90]. Then, a more flexible Cox model with time-varying effects or covariates would be an option for modeling the treatment effect[91–93]. Suppose therefore again a covariate vector $\mathbf{X} = (X_{trt}, \mathbf{W})^T$ with corresponding regression coefficients $\beta = (\beta_{trt}, \beta_W)^T$ as previously described in Section 1.1.2. A general time-dependent treatment effect can be modeled within a Cox model as

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \cdot \exp(g(\beta_{trt}, t) \cdot X_{trt} + \beta_W^T \cdot \mathbf{W}) \tag{10}$$

with β_{trt} the regression coefficient of the treatment X_{trt}, $g(\beta_{trt}, t)$ being a specific function of time t that must be specified by the investigator and $\mathbf{W}$ representing a set of other (time-invariant) covariates with corresponding regression coefficients β_W^T. If $g(\beta_{trt}, t)$ is a simple function, it can be written as $g(\beta_{trt}, t) = \beta_{trt} \cdot g(t)$ and model (10) can be rewritten by

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \cdot \exp(\beta_{trt} \cdot X_{trt}(t) + \beta_W^T \cdot \mathbf{W}) \tag{11}$$

with $X_{trt}(t) = g(t) \cdot X_{trt}$. This property shows that a time-varying treatment effect can be modeled using a time-varying covariate $X_{trt}(t)$[91]. Time-dependent effects provide a flexible method for assessing non-proportionality. However, this approach should be used with caution. Indeed, if the selected time function is misspecified, the final model will not be appropriate, leading to biased treatment effect estimates[92,93].

Moreover, true unobserved prognostic factors (often referred to as "frailty"), resulting in unmeasurable heterogeneity, might be an issue. Here, the proportional hazard assumption holds conditional on the frailty variable, which is usually modeled as random variable Z, acting in a multiplicative way on the individual hazard[27,46,47]. More precisely, consider the above notation and assume $\mathbf{W}$ was not measured during the trial. Then the random variable $Z = \exp(\beta_W^T \mathbf{W})$ can be defined and equation (7) can be rewritten as follows

$$\lambda(t|X_{trt}, Z) = Z \cdot \lambda_0(t) \cdot \exp(\beta_{trt} \cdot X_{trt}). \tag{12}$$

Unfortunately, proportional hazards frailty models (12) and non-proportional hazards cannot be distinguished in a univariate setting of survival data, making the decision on the right model even more complicated[94].

Another opportunity for addressing systematically changing treatment effect trajectories

detected by *Dynamic Landmarking* would be to choose a statistical model other than the Cox model, such as accelerate failure time (AFT) models[95] or additive hazard models[96]. Both models differ from the Cox model in the assumption of how the effects of the individual covariates act in the model. While the AFT model assumes that these have a multiplicative (proportional) effect with respect to the survival time, the additive hazards model assumes an additive effect on the hazards. A general AFT model can be described in terms of the survival function by

$$S(t|\mathbf{X}) = S_0(\exp(-(\beta_{trt} \cdot X_{trt} + \beta_W^T \cdot \mathbf{W})) \cdot t) \tag{13}$$

where S_0 represents the baseline survival function and $\exp(-(\beta_{trt} \cdot X_{trt} + \beta_W^T \cdot \mathbf{W}))$ is known as acceleration factor depending on a treatment X_{trt} and a set of covariates $\mathbf{W}$. Model (13) assumes that the effects are multiplicative by the accelerated factor on the time scale of t, making the interpretation of the estimated treatment effect more intuitive[97]. In contrast the additive hazards model by Aalen (1989) can be written as

$$\lambda(t|\mathbf{X}) = \lambda_0(t) + \beta_{trt}(t) \cdot X_{trt} + \beta_W^T(t) \cdot \mathbf{W} \tag{14}$$

where the notation corresponds to the previously described models. Because regression functions may vary with time, their analysis may reveal changes in the influence of the covariates over time, which is one of the main advantages of model (14)[98]. Furthermore, there are several approaches in case (independent) unmeasured confounders are assumed. The literature provides a large sample of sensitivity approaches[99–103] or suggests to use instrumental variable approaches[104,105] for addressing unmeasured confounders. In case the PS model is not correctly specified, due to interaction or a time-dependent structure of the considered covariates, more flexible PS approaches like time-depending PS models[51], PS Calibration[83] or large-scale PS[106] should be considered. No matter which correction is chosen, the new model should be confirmed regarding clinical and statistical plausibility. Hence, in cases where the original Cox model has been modified, *Dynamic Landmarking* could help to check whether the modified Cox model still provides treatment effect estimates that are subject to a specific source of bias. In summary, there is a wide selection of methods for estimating HRs. *Dynamic Landmarking* as a visual tool assesses whether the use of a Cox model is appropriate in the specific situation and, if necessary, can help to identify omitted covariates that still need to be taken into account. However, it does not provide a specific solution, if systematically changing trajectories are found. The model adaptation is thus left to the user.

2.4 Strengths and Limitations

The present work contributes methodological and empirical insights into the field of survival analysis by introducing an innovative method designed to examine HRs calculated from a Cox model for potential biases. This approach stands out as a valuable tool for researchers engaged in clinical trials, offering a systematic technique to identify and understand sources of bias in treatment effects. One of the notable strengths lies in the method's ability to identify omitted covariates that might induce built-in selection or confounding bias into the estimated treatment effects. Omitted covariates represent a common concern in both RCTs and PS matched trials, as typically more covariates are measured than used in the final analysis. *Dynamic Landmarking* helps in identifying covariates that should be considered for the PS model or for adjustment in the Cox model. Moreover, even in scenarios where no omitted covariates induce bias, the method remains relevant by still visualizing treatment effect estimates underlying any different bias. This flexibility empowers researchers to adapt their survival models, considering factors such as non-proportional hazards or true unobserved heterogeneity as potential sources of bias. An additional, noteworthy aspect of the research is the empirical investigation of the clinical relevance of built-in selection bias in a large sample of RCTs. Since *Dynamic Landmarking* could hardly provide visual indications of such bias, it was concluded that this issue is not practically relevant, and RCTs are not substantially distorted by built-in selection bias. Thus, this work also provides one of the first empirical insights into this, so far, predominantly theoretical topic, highlighting a main strength.

Some limitations also have to be mentioned. Firstly, *Dynamic Landmarking*, while sufficient in visually indicating built-in selection and confounding bias, lacks in providing a specific solution if necessary. The responsibility for correcting the statistical model is thus shifted to users, who have to decide on the appropriate adaptations to address potentially biased treatment effect estimates. Secondly, the simulation studies were conducted under assumptions of proportional hazards and non-informative censoring, which introduces a potential limitation. In real-world scenarios, where these assumptions may not hold, the results may deviate, raising questions about the generalizability of the findings. Furthermore, the empirical investigation, although extensive, is predominantly derived from studies in cardiovascular research, offering valuable insights into a particular domain. Yet, the applicability of the findings to other fields might be restricted. Additionally, the focus on the balance measurements involving two omitted covariates, namely age and sex, acknowledges the possibility of other relevant covariates that were not considered in the

analysis. Lastly, SSQ_{zDiff} as an aggregated balance measure neglects correlation among covariates and should therefore used with caution. This emphasizes the need for further investigations in this scientific field.

2.5 Conclusion

This work has presented a methodological approach aimed at visually highlighting treatment effect estimates subject to built-in selection or confounding bias in RCTs and PS matched trials. The method has the main benefit to identify covariates responsible for such biases. To make valid statements in both study types and, consequently, running causal inferences, it is imperative to consider relevant confounders and/or prognostic factors in the data analysis. Unfortunately, in reality, it is not always feasible to collect data on all covariates. In contrast, during the study, typically more covariates are collected than are ultimately used in the final analysis. The selection of confounders for a PS model and the choice of factors to be adjusted for in a Cox model heavily depend on current scientific knowledge and the responsible analyzing authority. Therefore, it is plausible that important covariates may be overlooked in the analysis and bias is induced. The *Dynamic Landmarking* approach presented here addresses precisely this problem and aims to recognize whether the calculated HR as a single number is subject to potential bias. Furthermore, the empirical investigation involving 27 RCTs revealed that RCTs rarely present biased effect estimates. The two main reasons for this result are the small treatment effects due to equipoise and the low heterogeneity due to strict inclusion and exclusion criteria. While this is true for most RCTs, it may not hold for every single study. In such cases, *Dynamic Landmarking* gives a visual indication by systematically changing treatment effect trajectories and numerous methods are available to address the bias. Moreover and to conclude, in agreement with many researchers, it is essential, especially in survival analysis, to present not just a single HR, but a comprehensive result consisting of several relative and absolute values, including Kaplan-Meier curves[107–109]. However, there is no need to entirely dismiss the HR, especially when the possibility of distortion can now be examined, at least on a visual basis.

.

3 References

1. Cuschieri, S. The CONSORT statement. *Saudi J Anaesth* **13,** S27–S30 (2019).

2. Saturni, S. *et al.* Randomized controlled trials and real life studies. Approaches and methodologies: a clinical point of view. *Pulm Pharmacol Ther* **27,** 129–138 (2014).

3. Spieth, P. *et al.* Randomized controlled trials - a matter of design. *Neuropsychiatr Dis Treat* **10,** 1341–1349 (2016).

4. Kabisch, M., Ruckes, C., Seibert-Grafe, M. & Blettner, M. Randomized controlled trials: part 17 of a series on evaluation of scientific publications. *Dtsch Ärztebl Int* **108,** 663–668 (2011).

5. Eichler, H. G. *et al.* Threshold-crossing: A Useful Way to Establish the Counterfactual in Clinical Trials? *Clin Pharmacol Ther* **100,** 699–712 (2016).

6. Hernán, M. A. & Robins, J. M. *Causal Inference: What If* (Chapman & Hall/CRC, 2020).

7. Bours, M. A nontechnical explanation of the counterfactual definition of confounding. *J Clin Epidemiol* **121,** 91–100 (2020).

8. Garcia-Huidobro, D. & Michael-Oakes, J. Squeezing observational data for better causal inference: Methods and examples for prevention research. *Int J Psychol* **52,** 96–105 (2017).

9. Pan, W. & Bai, H. *Propensity score analysis: Fundamentals and developments* (Guilford Press, 2015).

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

11. Kuss, O., Blettner, M. & Börgermann, J. Propensity Score: an Alternative Method of Analyzing Treatment Effects. *Dtsch Ärztebl Int* **113,** 597–603 (2016).

12. Elze, M. C. *et al.* Comparison of Propensity Score Methods and Covariate Adjustment: Evaluation in 4 Cardiovascular Studies. *J Am Coll Cardiol* **69,** 345–357 (2017).

13. Austin, P. C. & Mamdani, M. M. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med* **25,** 2084–2106 (2006).

14. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* **46,** 399–424 (2011).

15. Ebrahim-Valojerdi, A. & Janani, L. A brief guide to propensity score analysis. *Med J Islam Repub Iran* **32,** 122 (2018).

16. Stürmer, T. *et al.* A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J. Clin. Epidemiol.* **59,** 437–447 (2006).

17. Granger, E., Watkins, T., Sergeant, J. & Lunt, M. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Med Res Methodol* **20,** 132 (2020).

18. Austin, P. C. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med* **26,** 3078–3094 (2007).

19. Austin, P. C. The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol* **61,** 537–545 (2008).

20. Austin, P. C. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med* **29,** 2137–2148 (2010).

21. Austin, P. C. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med* **32,** 2837–2849 (2013).

22. Austin, P. C. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Making* **29,** 661–677 (2009).

23. Kuss, O. The z-difference can be used to measure covariate balance in matched propensity score analyses. *J Clin Epidemiol* **66,** 1302–1307 (2013).

24. Kuss, O. & Strobel, A. Quality measures and criteria for the application of propensity scores. *GMS Med Inform Biom Epidemiol* **20,** Doc1 (2024).

25. Cox, D. R. Regression Models and Life-Tables. *J. R. Stat.* **34,** 187–220 (1972).

26. Cox, D. R. Partial Likelihood. *Biometrika* **62,** 269–276 (1975).

27. Balan, T. A. & Putter, H. A tutorial on frailty models. *Stat Methods Med Res* **29,** 3424–3454 (2020).

28. Martinussen, T. & Vansteelandt, S. On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Anal* **19,** 279–296 (2013).

29. Martinussen, T., Vansteelandt, S. & Andersen, P. K. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Anal* **26,** 833–855 (2020).

30. Pearl, J. An introduction to causal inference. *Int J Biostat* **6,** Article 7 (2010).

31. Sutradhar, R. & Austin, P. C. Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Ann Epidemiol* **28,** 54–57 (2018).

32. Sashegyi, A. & Ferry, D. On the Interpretation of Hazard Ratio and Communication of Survival Benefit. *The Oncologist* **22,** 484–486 (2017).

33. Bian, H., Pang, M., Wang, G. & Lu, Z. Non-collapsibility and built-in selection bias of period-specific and conventional hazard ratio in randomized controlled trials. *BMC Med Res Methodol* **24,** 292 (2024).

34. Sjoelander, A., Dahlqwist, E. & Zetterqvist, J. A Note on the Noncollapsibility of Rate Differences and Rate Ratios. *Epidemiology* **27,** 356–359 (2016).

35. Samuelsen, S. O. Cox regression can be collapsible and Aalen regression can be non-collapsible. *Lifetime Data Anal* **29,** 403–419 (2023).

36. Daniel, R., Zhang, J. & Farewell, D. Making apples from oranges: Comparing non-collapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biom J* **63,** 528–557 (2021).

37. Schuster, N., Twisk, J., Riet, G., Heymans, M. & Rijnhart, J. Noncollapsibility and its role in quantifying confounding bias in logistic regression. *BMC Med Res Methodol* **21,** 136 (2021).

38. Bardo, M. *et al.* Methods for non-proportional hazards in clinical trials: A systematic review. *Stat. Methods Med. Res.* **33,** 1069–1092 (2024).

39. De Neve, J. & Gerds, T. A. On the interpretation of the hazard ratio in Cox regression. *Biom J* **62,** 742–750 (2020).

40. Riffenburgh, R. H. in *Statistics in Medicine (Third Edition)* 581–591 (Academic Press, 2012).

41. Aalen, O. O., Cook, R. J. & Røysland, K. Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Anal* **21,** 579–593 (2015).

42. Hernán, M. A. The hazards of hazard ratios. *Epidemiology* **21,** 13–15 (2010).

43. Stensrud, M. J., Aalen, J. M., Aalen, O. O. & Valberg, M. Limitations of hazard ratios in clinical trials. *Eur Heart J* **40,** 1378–1383 (2019).

44. Bartlett, J. W. *et al.* The Hazards of Period Specific and Weighted Hazard Ratios. *Stat Biopharm Res* **12,** 518–519 (2020).

45. Steenland, K., Karnes, C., Darrow, L. & Barry, V. Attenuation of exposure-response rate ratios at higher exposures: a simulation study focusing on frailty and measurement error. *Epidemiology* **26,** 395–401 (2015).

46. Hougaard, P. Frailty models for survival data. *Lifetime Data Anal* **1,** 255–273 (1995).

47. Wienke, A. *Frailty Models in Survival Analysis* (Chapman & Hall/CRC, 2010).

48. Duchateau, L. and Janssen, P. *The frailty model* (Springer Verlag, 2008).

49. Hanagal, D. D. *Modeling Survival Data Using Frailty Models* (Chapman & Hall/CRC, 2011).

50. Martinussen, T. Causality and the Cox Regression Model. *Annu Rev Stat Appl.* **9,** 249–259 (2022).

51. Wyss, R. *et al.* Use of Time-Dependent Propensity Scores to Adjust Hazard Ratio Estimates in Cohort Studies with Differential Depletion of Susceptibles. *Epidemiology* **31,** 82–89 (2020).

52. Lenis, D., Ackerman, B. & Stuart, E. A. Measuring Model Misspecification: Application to Propensity Score Methods with Complex Survey Data. *Comput Stat Data Anal* **128,** 48–57 (2018).

53. Wallin, G. & Wiberg, M. Model Misspecification and Robustness of Observed-Score Test Equating Using Propensity Scores. *JEBS* **48,** 603–635 (2023).

54. Drake, C. Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biometrics* **49,** 1231–1236 (1993).

55. Dehejia, R. H. & Wahba, S. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association* **94,** 1053–1062. (2023) (1999).

56. Fireman, B. *et al.* Consequences of Depletion of Susceptibles for Hazard Ratio Estimators Based on Propensity Scores. *Epidemiology* **31,** 806–814 (2020).

57. Gayat, E., Resche-Rigon, M., Mary, J. Y. & Porcher, R. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharm Stat* **11,** 222–229 (2012).

58. Hansen, B. B. The prognostic analogue of the propensity score. *Biometrika* **95,** 481–488 (2008).

59. Rubin, D. B. & Thomas, N. Matching using estimated propensity scores: relating theory to practice. *Biometrics* **52,** 249–264 (1996).

60. Van Houwelingen, H. C. Dynamic Prediction by Landmarking in Event History Analysis. *SJS* **34,** 70–85 (2007).

61. Van Houwelingen, H. C. & Putter, H. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Anal* **14,** 447–463 (2008).

62. Putter, H. & van Houwelingen, H. C. Landmarking 2.0: Bridging the gap between joint models and landmarking. *Stat Med* **41,** 1901–1917 (2022).

63. Strobel, A., Wienke, A. & Kuss, O. How hazardous are hazard ratios? An empirical investigation of individual patient data from 27 large randomized clinical trials. *Eur J Epidemiol* **38,** 859–867 (2023).

64. Strobel, A., Wienke, A., Gummert, J., Bleiziffer, S. & Kuss, O. Built-in selection or confounder bias? Dynamic Landmarking in matched propensity score analyses. *BMC Med Res Methodol* **24,** 316 (2024).

65. Kent, D. M. *et al.* Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. *Int J Epidemiol* **45,** 2075–2088 (2016).

66. Wyse, D. G. *et al.* A comparison of rate control and rhythm control in patients with atrial fibrillation. *N Engl J Med* **347,** 1825–1833 (2002).

67. Magnesium in Coronaries (MAGIC) Trial Investigators. Early administration of intravenous magnesium to high-risk patients with acute myocardial infarction in the Magnesium in Coronaries (MAGIC) Trial: a randomised controlled trial. *Lancet* **360,** 1189—1196 (2002).

68. Berkman, L. F. *et al.* Effects of treating depression and low perceived social support on clinical events after myocardial infarction: the Enhancing Recovery in Coro-

nary Heart Disease Patients (ENRICHD) Randomized Trial. *JAMA* **289,** 3106–3116 (2003).

69. TIMI Study Group. Comparison of invasive and conservative strategies after treatment with intravenous tissue plasminogen activator in acute myocardial infarction. Results of the thrombolysis in myocardial infarction (TIMI) phase II trial. *N Engl J Med* **320,** 618–627 (1989).

70. International Stroke Trial Collaborative Group. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *Lancet* **349,** 1569–1581 (1997).

71. Eichhorn, E. J., Domanski, M. J., Krause-Steinrauf, H., Bristow, M. R. & Lavori, P. W. A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. *N Engl J Med* **344,** 1659–1667 (2001).

72. Digitalis Investigation Group. The effect of digoxin on mortality and morbidity in patients with heart failure. *N Engl J Med* **336,** 525–533 (1997).

73. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA* **288,** 2981–2997 (2002).

74. Gerstein, H. C. *et al.* Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med* **358,** 2545–2559 (2008).

75. Nathan, D. M. *et al.* The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* **329,** 977–986 (1993).

76. Di Bisceglie, A. M. *et al.* Prolonged therapy of advanced chronic hepatitis C with low-dose peginterferon. *N Engl J Med* **359,** 2429–2441 (2008).

77. Eknoyan, G. *et al.* Effect of dialysis dose and membrane flux in maintenance hemodialysis. *N Engl J Med* **347,** 2010–2019 (2002).

78. Bulger, E. M. *et al.* Out-of-hospital hypertonic resuscitation following severe traumatic brain injury: a randomized controlled trial. *JAMA* **304,** 1455–1464 (2010).

79. Stensrud, M. Interpreting Hazard Ratios: Insights from Frailty Models. *arXiv* (2018).

80. Djulbegovic, B. *et al.* New treatments compared to established treatments in randomized trials. *Cochrane Database Syst Rev* **10,** MR000024 (2012).

81. He, J., Morales, D. R. & Guthrie, B. Exclusion rates in randomized controlled trials of treatments for physical conditions: a systematic review. *Trials* **21,** 228 (2020).

82. Kahan, B. C., Jairath, V., Doré, C. J. & Morris, T. P. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials* **15,** 139 (2014).

83. Stürmer, T., Schneeweiss, S., Avorn, J. & Glynn, R. J. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol* **162,** 279–289 (2005).

84. Wilkinson, J. D., Mamas, M. A. & Kontopantelis, E. Logistic regression frequently outperformed propensity score methods, especially for large datasets: a simulation study. *J Clin Epidemiol* **152,** 176–184 (2022).

85. Amoah, J. *et al.* Comparing Propensity Score Methods Versus Traditional Regression Analysis for the Evaluation of Observational Data: A Case Study Evaluating the Treatment of Gram-Negative Bloodstream Infections. *Clin Infect Dis* **71,** e497–e505 (2020).

86. Jager, K. J., Zoccali, C., Macleod, A. & Dekker, F. W. Confounding: what it is and how to deal with it. *Kidney Int* **73,** 256–260 (2008).

87. Van Lancker, K., Dukes, O. & Vansteelandt, S. Ensuring valid inference for Cox hazard ratios after variable selection. *Biometrics* **79,** 3096–3110 (2023).

88. Xie, L. S. & Lu, H. A change point-based analysis procedure for improving the success rate of decision-making in clinical trials with delayed treatment effects. *Front Pharmacol* **14,** 1186456 (2023).

89. Gregson, J. *et al.* Nonproportional Hazards for Time-to-Event Outcomes in Clinical Trials: JACC Review Topic of the Week. *J Am Coll Cardiol* **74,** 2102–2112 (2019).

90. Ristl, R. *et al.* Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology. *Pharm Stat* **20,** 129–145 (2021).

91. Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E. & Groothuis-Oudshoorn, C. G. M. Time-varying covariates and coefficients in Cox regression models. *Ann Transl Med* **6,** 121 (2018).

92. Fisher, L. D. & Lin, D. Y. Time-dependent covariates in the Cox proportional-hazards regression model. *Annu Rev Public Health* **20,** 145–157 (1999).

93. Bellera, C. A. *et al.* Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med Res Methodol* **10,** 20 (2010).

94. Balan, T. A. & Putter, H. Nonproportional hazards and unobserved heterogeneity in clustered survival data: When can we tell the difference? *Stat Med* **38,** 3405–3420 (2019).

95. Wei, L. J. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat Med* **11,** 1871–1879 (1992).

96. Aalen, O. O. A linear regression model for the analysis of life times. *Stat Med* **8,** 907–925 (1989).

97. Faruk, A. The comparison of proportional hazards and accelerated failure time models in analyzing the first birth interval survival data. *J. Phys.: Conf. Ser.* **974,** 012008 (2018).

98. Cao, H. *A Comparison Between the Additive and Multiplicative Risk Models.* PhD thesis (University of Laval, 2005).

99. McCandless, L. C., Gustafson, P. & Levy, A. R. A sensitivity analysis using information about measured confounders yielded improved uncertainty assessments for unmeasured confounding. *J Clin Epidemiol* **61,** 247–255 (2008).

100. Lin, D. Y., Psaty, B. M. & Kronmal, R. A. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54,** 948–963 (1998).

101. Lin, N. X., Logan, S. & Henley, W. E. Bias and sensitivity analysis when estimating treatment effects from the cox model with omitted covariates. *Biometrics* **69,** 850–860 (2013).

102. Vanderweele, T. J. & Arah, O. A. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* **22,** 42–52 (2011).

103. Comment, L., Coull, B. A., Zigler, C. & Valeri, L. Bayesian data fusion: Probabilistic sensitivity analysis for unmeasured confounding using informative priors based on secondary data. *Biometrics* **78,** 730–741 (2022).

104. Uddin, M. J. *et al.* Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *Int J Clin Pharm* **38,** 714–723 (2016).

105. Nguyen, T. T. *et al.* Instrumental variable approaches to identifying the causal effect of educational attainment on dementia risk. *Ann Epidemiol* **26,** 71–76 (2016).

106. Zhang, L., Wang, Y., Schuemie, M. J., Blei, D. M. & Hripcsak, G. Adjusting for indirectly measured confounding using large-scale propensity score. *J Biomed Inform* **134,** 104204 (2022).

107. Uno, H. *et al.* Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* **32,** 2380–2385 (2014).

108. Kropko, J. & Harden, J. J. Beyond the Hazard Ratio: Generating Expected Durations from the Cox Proportional Hazards Model. *B J Pol S* **50,** 303–320 (2020).

109. Chen, D. *et al.* Beyond the Cox Hazard Ratio: A Targeted Learning Approach to Survival Analysis in a Cardiovascular Outcome Trial Application. *Stat Biopharm Res* **15,** 524–539 (2023).

4 Theses

1. *Dynamic Landmarking* provides a post-hoc diagnosis tool for visualizing whether an estimated hazard ratio could be distorted by special forms of bias.

2. If a study contains measured but omitted covariates, *Dynamic Landmarking* can identify if any of these omitted covariates induce built-in selection or confounding bias.

3. If omitted covariates cause bias, the methodological approach is able to distinguish between hazard ratios underlying either confounding or built-in selection bias by using a global balance measure.

4. In RCTs, *Dynamic Landmarking* is best at detecting built-in selection bias if the treatment effect is strong, heterogeneity is large, follow-up is long, and censoring rate is low.

5. In PS matched trials, the approach works best in identifying residual confounding bias if the omitted covariates do show both, strong impact on treatment allocation and strong impact on patients' survival.

6. If an omitted covariate is highly correlated with an included covariate, both, built-in selection and confounding bias are less pronounced.

7. Applying *Dynamic Landmarking* to a large sample of 27 RCTs yields no visually apparent evidence of built-in selection bias.

8. The absence of the built-in selection bias in RCTs is mainly due to treatment effects being small and patient populations being homogeneous.

9. Refitting the PS model again, including the identified omitted confounder, provides one way to address residual confounding bias identified through *Dynamic Landmarking*, as shown by an empirical example from cardiac surgery.

10. If the methodological approach does not identify omitted covariates inducing built-in selection or confounding bias, other statistical models need to be considered.

Publications

Publication 1:

Strobel, A., Wienke, A., & Kuss, O. (2023). How hazardous are hazard ratios? An empirical investigation of individual patient data from 27 large randomized clinical trials. *European Journal of Epidemiology*, 38(8), 859–867.

Publication 2:

Strobel, A., Wienke, A., Gummert, J., Bleiziffer S. & Kuss, O (2024). Built-in selection or confounder bias? *Dynamic Landmarking* in matched propensity score analyses. *BMC Medical Research Methodology*, 24, 316.

Publication 1

Strobel, A., Wienke, A., & Kuss, O. (2023). How hazardous are hazard ratios? An empirical investigation of individual patient data from 27 large randomized clinical trials. *European Journal of Epidemiology*, 38(8), 859–867.

Contribution to Publication 1:

Conception, data cleaning and data preparation, programming method, conducting simulation study and analyzing data, writing and revision of manuscript.

European Journal of Epidemiology (2023) 38:859–867
https://doi.org/10.1007/s10654-023-01026-z

METHODS

How hazardous are hazard ratios? An empirical investigation of individual patient data from 27 large randomized clinical trials

Alexandra Strobel[1] · **Andreas Wienke**[1] · **Oliver Kuss**[2,3]

Received: 11 November 2022 / Accepted: 19 June 2023 / Published online: 6 July 2023

Abstract
The use of hazard ratios as the standard treatment effect estimators for randomized trials with time-to-event outcomes has been the subject of repeated criticisms in recent years, e.g., for its non-collapsibility or with respect to (causal) interpretation. Another important issue is the built-in selection bias, which arises when the treatment is effective and when there are unobserved or not included prognostic factors that influence time-to-event. In these cases, the hazard ratio has even been termed "hazardous" because it is estimated from groups that increasingly differ in their (unobserved or omitted) baseline characteristics, yielding biased treatment estimates. We therefore adapt the Landmarking approach to assess the effect of ignoring a gradually increasing proportion of early events on the estimated hazard ratio. We propose an extension called "Dynamic Landmarking". This approach is based on successive deletion of observations, refitting Cox models and balance checking of omitted but observed prognostic factors, to obtain a visualization that can indicate built-in selection bias. In a small proof-of-concept simulation, we show that our approach is valid under the given assumptions. We further use "Dynamic Landmarking" to assess the suspected selection bias in the individual patient data sets of 27 large randomized clinical trials (RCTs). Surprisingly, we find no empirical evidence of selection bias in these RCTs and thus conclude that the supposed bias of the hazard ratio is of little practical relevance in most cases. This is mainly due to treatment effects in RCTs being small and the patient populations being homogeneous, e.g., due to inclusion and exclusion criteria.

Keywords Cox model · Survival analysis · Hazard ratio · Bias · RCT

Introduction

Randomized controlled trials (RCTs) are the gold standard for evaluating treatments and interventions in medical research. Randomization guarantees the balance of known and, more importantly, unknown or unobserved prognostic factors between the intervention and control group. In particular, randomization ensures that all potential confounding factors are distributed equally between groups [14]. For describing treatment effects with time-to-event data, the Cox model [6, 7] is commonly used because it avoids parametric assumptions about the baseline distribution, allows dealing with censored observations and provides the hazard ratio as an effect measure that is easy to interpret. Due to random treatment allocation in RCTs, the hazard ratio is commonly estimated without adjustments. Notably, the parameters can be estimated consistently and interpreted causally only if the Cox model is correctly specified. However, the hazard ratio has also been criticized recently because it is non-collapsible [18, 21] or only comprehensible when it is (wrongly) interpreted and communicated as a relative risk [8, 24].

We focus here on another problem of the hazard ratio, which arises when prognostic factors are not observed or are measured but not included in the model; in this case, selection bias can arise, even in randomized trials (e.g., [1, 13]). If there are such prognostic factors that influence time-to-event, as well as an effective treatment, patients in the control group tend to experience the event faster, leading

✉ Alexandra Strobel
alexandra.strobel@uk-halle.de

[1] Institute of Medical Epidemiology, Biostatistics, and Informatics, Interdisciplinary Center for Health Sciences, Medical Faculty, Martin-Luther-University Halle-Wittenberg, Halle, Germany

[2] German Diabetes Center, Leibniz Center for Diabetes Research, Institute for Biometrics and Epidemiology, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

[3] Centre for Health and Society, Faculty of Medicine, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

to unbalanced treatment arms with increasing follow-up time. Consequently, the hazard ratio is estimated from groups that differ more and more in their unobserved/omitted baseline characteristics [21]. Such selection effects are also well known from frailty models [2, 10, 26]. Several authors showed that the hazard ratio does not have a causal interpretation if unobserved or omitted prognostic factors are present. Additionally, an attenuation of the true conditional hazard ratio during the follow-up period can be seen in such cases [e.g., 1, 4, 22].

While these issues have been well described in theory (e.g., [23, 26]) and demonstrated with large simulation studies (e.g., [5, 19]), we are not aware of empirical evidence of their practical relevance for effect estimation in RCTs. To address this limitation, the present investigation uses an empirical reanalysis of data from a number of large RCTs with survival outcomes to quantify the size of this bias and to derive practical implications to address it. In particular, this paper focuses on the magnitude of the bias and not on the causal interpretation of hazard ratios. First, we introduce the Cox model and its built-in selection bias due to measured but omitted prognostic factors. Afterwards, we describe the method we use to quantify the selection bias and offer a small "proof-of-concept" simulation to assess its validity. In Sect. "Empirical investigation", the results from the RCTs are presented, followed by a discussion including practical recommendations for the use of hazard ratios in RCTs.

The Cox model and its built-in selection bias

The proportional hazard model is given by

$$\lambda\left(t|X_i\right) = \lambda_0(t)e^{\beta' X_i}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function that depends on time t and is assumed to be common across all individuals i ($i = 1, \ldots, N$). Furthermore, X_i is a vector of observed covariates, and β' denotes the vector of the corresponding regression coefficients. Assuming proportional hazards, the hazard ratio for given values of a single binary treatment covariate X_{trt}, is denoted by

$$\frac{\lambda(t|X_{trt} = 1)}{\lambda(t|X_{trt} = 0)} = e^{\beta}$$

and thus constant over time. However, this only holds for a correctly specified model. Misspecification as a result of omitted prognostic factors can lead to heavily biased treatment estimates [20]. This bias is more severe with more effective treatments and omitted prognostic factors with stronger impacts on the outcome. An omitted variable will in general induce "unobserved" heterogeneity; that is, individuals will differ in characteristics that could be explained by the omitted variable. Assuming an association between the omitted prognostic factor and the probability of an event, some individuals will then be more susceptible to the event than others are and experience it earlier. However, due to effective treatment, frail individuals in the treatment group remain longer in the study than those in the control group, leading to increasing heterogeneity and increasing bias in the hazard ratio as the follow-up continues [13]. More precisely, assuming a binary treatment X_{trt} and an omitted covariate X_1, the true conditional model would be

$$\lambda\left(t|X_i\right) = \lambda_0(t)e^{\beta_{trt}X_{trt,i}+\beta_1 X_{1,i}}.$$

However, as X_1 presents an omitted prognostic factor (observed, but not included or even unmeasured), we would only fit the marginal model $\lambda(t|X_{trt,i})$, which does not provide a conditional, subject-specific interpretation for the treatment effect. Instead, we estimate marginal, population-averaged hazard ratios [4]; not accounting for this result in differing treatment estimates.

Measuring selection bias via "Dynamic Landmarking" and balance checking

The idea behind our approach is quite simple and based on the Landmarking approach by Van Houwelingen et al. [24, 25]. First, the data set is sorted according to follow-up duration, and a univariable Cox model is fitted to the full data set. Then, gradually, the earliest M observations regardless of the event status (time-to-event or censoring) are deleted, and a Cox model is refitted for the new, smaller data set. After each deletion step, the start of the follow-up interval for the new Cox model is moved forwards. More precisely, the new time zero for the new Cox model corresponds to the follow-up time of the latest of the M deleted individuals in the previous step. This procedure of deleting M observations and refitting the Cox model (which we propose to call "Dynamic Landmarking") is repeated until the data set no longer contains a sufficient number of events for convergence in the parameter estimation procedure. All estimated Cox models contain only the treatment variable and are not adjusted for other measured covariates. Collecting estimated regression parameters (as log hazard ratios) that result from successive deletion and refitting yields a trajectory in the log hazard ratio as a function of the relative number of remaining observations.

In addition, the balance of omitted prognostic factors is calculated in each deletion step. To measure the balance in

the treatment group (henceforth indexed by T) and the control group (indexed by C), we use z-differences [16]. These are $N(0,1)$-distributed in a RCT and can be calculated for continuous and binary variables as follows:

$$z_{cont} = \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\frac{\hat{\sigma}_T^2}{N_T} + \frac{\hat{\sigma}_C^2}{N_C}}}$$

$$z_{bin} = \frac{\hat{p}_T - \hat{p}_C}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T)}{N_T} + \frac{\hat{p}_C(1-\hat{p}_C)}{N_C}}}$$

Here $\bar{x}_T, \bar{x}_C, \hat{\sigma}_T^2, \hat{\sigma}_C^2, \hat{p}_T, \hat{p}_C, N_T, N_C$ denote the respective estimated means, variances, proportions, and sample sizes of the two groups. In a data set with k independent observed (binary and continuous) prognostic factors, the sum of k squared z-differences (SSQ_{zDiff}) follows an approximate χ_k^2-distribution with expectation k.

As we consider observed but omitted prognostic factors here, the imbalance in these prognostic factors can be measured by the SSQ_{zDiff}. We expect that individuals with higher baseline risk (e.g., due to higher values of an omitted prognostic factor) have a shorter time-to-event and are thus deleted from the data set earlier. This results in a systematic change in the trajectories of both the treatment effect estimate and the SSQ_{zDiff} as the "Dynamic Landmarking" proceeds. Assessing balance in measured but omitted prognostic factors has two main benefits. First, it is possible to check whether randomization succeeded or failed for the full data set. Second, important prognostic factors can be identified if an increased imbalance is observed from the $SSQ_{zDiff}-$ trajectory. On the other hand, achieving balance in all omitted prognostic factors indicates that these are not associated with the time-to-event outcome.

A proof-of-concept simulation

In this section, we present the results of a small simulation study, which shows that "Dynamic Landmarking" can indeed give a visual indication of selection bias due to the omission of a prognostic factor. For this task, we simulate different randomized trail scenarios by varying the impact of treatment and the impact of the omitted prognostic factor on the time-to-event. In addition, we considered various censoring rates.

Data generation and comparison

For all scenarios, we generated 20 data sets with 5,000 subjects each and the following specifications:

- A Weibull baseline hazard function with scale $\alpha = 0.1$ and shape $\gamma = 1.5$
- One binary treatment variable $X_{trt} \sim Bin(1, 0.5)$ and corresponding regression parameter $\beta_{trt} \in \{\log(1.25), \log(1.5), \log(3)\}$
- A normally distributed covariate $X_1 \sim N(0, 10)$ representing an observed but omitted covariate with different influences on survival: $\beta_1 \in \{0, \log(1.25), \log(3)\}$
- X_{trt} and X_1 are independent
- Different censoring rates, with 10%, 50% or 80% of the individuals being right-censored

Furthermore, we assumed proportional hazards and non-informative censoring for the data generation process. In each scenario and for each generated data set, we stepwise deleted the earliest M observations ($M = 10$), fitted Cox models with treatment as the only variable and measured the balance with respect to the omitted prognostic factor X_1 via the SSQ_{zDiff}. In the scenario in which there is no influence of X_1 on survival, e.g., $\beta_1 = 0$, $E\left(SSQ_{zDiff}\right) = 1$ holds. Otherwise, that is, in scenarios with $\beta_1 \neq 0$, we would expect values of $SSQ_{zDiff} > 1$ as "Dynamic Landmarking" proceeds.

Results

In Fig. 1, we give the results of our proof-of-concept simulation for the medium censoring rate of 50%. In the first column, the results are shown with no influence of the omitted covariate X_1 on time-to-event. Therefore, correctly specified Cox models including all relevant prognostic factors (that is, only treatment) were fitted, and as expected, the trajectories of the treatment effect estimates show only random fluctuations around the true conditional hazard ratio. Similarly, SSQ_{zDiff} shows no systematic deviations from the expected value of one, indicating good balance and the absence of selection bias.

In the second and third columns, we present data under the assumption that the omitted covariate X_1 has a weak (column 2) or strong (column 3) influence on the time-to-event outcome. We observe that a weak effect of treatment on survival, as well as a weak impact of the omitted covariate on survival, does not affect the treatment estimates considerably. However, if both have a strong impact on the time-to-event, the "Dynamic Landmarking" results can show a visible indication of selection bias. We find that the hazard ratio is already biased for the full data set, which is caused by the difference in the marginal and conditional hazard ratios due to the omitted covariate. As "Dynamic Landmarking" proceeds, the treatment effect estimates systematically change, and correspondingly, the SSQ_{zDiff} shows an increasingly compromised balance, indicating the presence of selection bias. This is also true for other censoring rates, although the power of the method

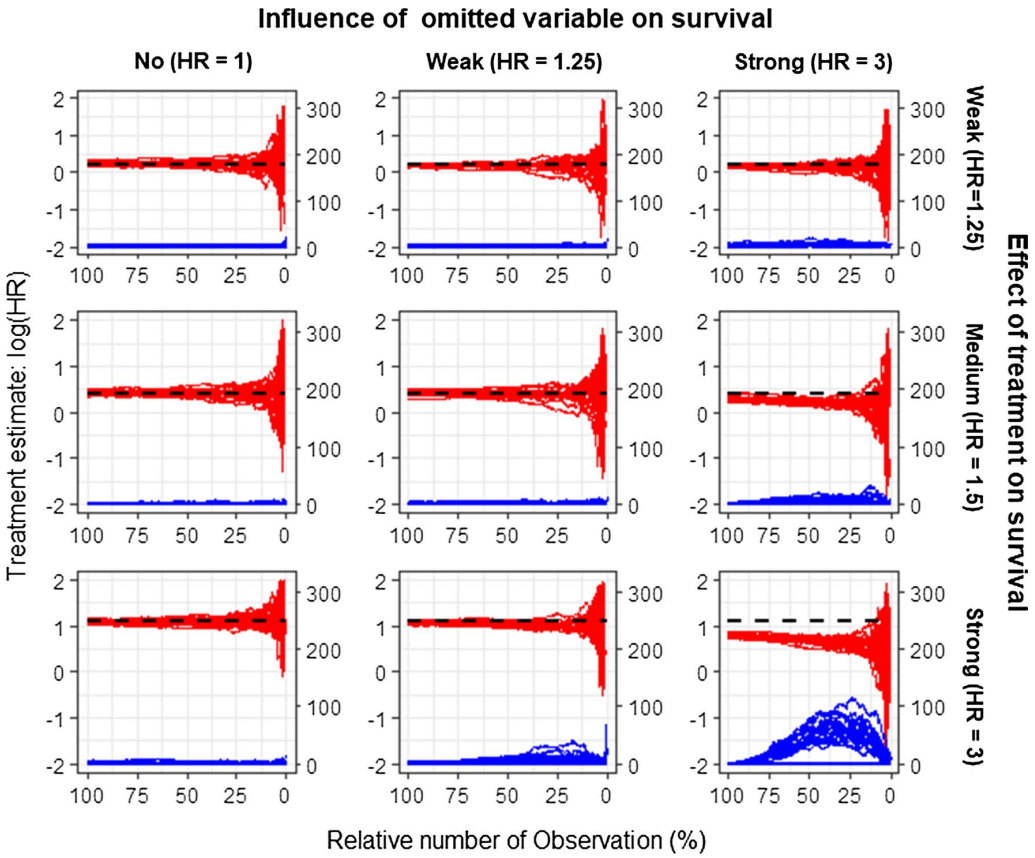

Fig. 1 Trajectories of treatment effect estimate (left y-axis, red) on the log(HR) scale and sum of squared z-differences (right y-axis, blue) for balance measurement of the omitted covariate X_1 for the 20 simulated data sets. The thick blue line represents the expected value for the squared z-difference balance under randomization, i.e., $SSQ_{zDiff} = 1$. Dashed black lines show the true simulated treatment estimate. The results are based on a medium censoring rate, i.e., 50% of all individuals were censored

is somewhat limited for higher censoring rates since censored individuals have less material impact on the estimates (see supplementary information). In summary, the simulation shows that a stronger influence of the omitted prognostic factor on time-to-event and a stronger impact of the treatment on time-to-event, the more visible the potential built-in selection bias and the less balanced the omitted prognostic factors. "Dynamic Landmarking" is able to detect and visualize the presence of selection bias if all other requirements of the Cox model (proportional hazards, no time-varying treatment, etc.) are met.

Notably, we do not necessarily expect the trajectories of treatment effect estimates to converge towards the null because we show the percentage of remaining observations rather than the original observation time on the x-axis. In particular, fitting a new Cox model after each deletion should not be confused with fitting a single full data set and the corresponding follow-up time.

Empirical investigation

Data and methods

We used publicly available individual data sets from 32 large RCTs already analysed for methodological investigations by Kent et al. [15] and described in more detail therein. After having approval from the Institutional Review Board of the Medical Faculty of the Heinrich-Heine-University Duesseldorf from March 2019 (Study No.: 5986R, Registration-ID: 201,707 4356) we contacted the three clinical study registries [NHLBI2018, NIDDK2018, GlaxoSmithKline2018] and were allowed to reuse data sets that were already used for a previous project [17]. To avoid problems with competing risks, we restricted our analysis to all-cause mortality as outcome and were able to include 27 trials (see Table 1), each with more than 1,000 individual observations. To unify measurement of covariate balance across trials, we

Table 1 Description of trial characteristics

Acronym	Year	Patients	Events	Mean age (±sd)	Females	HR (95%CI) for age*	HR (95%CI) for sex* (Ref: women)	Condition	Intervention	Comparator
ACCORD P	2008	5518	7.7%	62.8 (±6.6)	30.7%	1.07 (1.06; 1.09)	1.54 (1.22; 1.93)	Type 2 diabetes	Intensive strategy	Standard treatment
ACCORD BP	2008	4733	6.2%	62.7 (±6.7)	47.7%	1.07 (1.05; 1.09)	1.28 (1.01; 1.61)	Type 2 diabetes	Intensive strategy	Standard treatment
AFFIRM	2002	4060	16.4%	69.5 (±8.1)	39.3%	1.06 (1.05; 1.07)	1.11 (0.95; 1.30)	Atrial fibrillation	Rate control therapy	Rythm control therapy
ALLHAT CA	2002	24,303	15.1%	66.9 (±7.7)	47.1%	1.07 (1.06; 1.07)	1.37 (1.29; 1.47)	Hypertension	Amlodipine or lisinopril	Chlorthalidone
ALLHAT CL	2002	24,309	15.3%	66.9 (±7.7)	46.7%	1.08 (1.07; 1.08)	1.37 (1.28; 1.46)	Hypertension	Amlodipine or lisinopril	Chlorthalidone
ALLHAT LLT	2002	10,355	13.0%	66.3 (±7.6)	48.8%	1.07 (1.07; 1.08)	1.38 (1.24; 1.54)	Hypertension	Pravastatin	Usual care
AMIS	1980	4524	10.3%	54.8 (±8.0)	11.1%	1.02 (1.004; 1.03)	1.85 (1.30; 2.63)	Myocardial infarction	Aspirin	Placebo
ATN	2008	1124	52.6%	59.6 (±15.3)	29.4%	1.01 (1.01; 1.02)	0.98 (0.82; 1.18)	AKI/Sepsis	Intensiv RRT	Less intensive therapy
BEST	2001	2707	31.7%	60.2 (±12.3)	21.9%	1.02 (1.01; 1.03)	1.22 (1.03; 1.45)	Advanced/Congestive HF	Bucindolol hydrocholoride	Placebo
BHAT	1982	3826	8.5%	54.8 (±8.4)	15.7%	0.91 (0.68; 1.22)	1.05 (1.03; 1.06)	Acute MI	Propranolol	Placebo
CAST	1991	1497	6.5%	61.1 (±9.3)	17.6%	1.04 (1.02; 1.06)	0.58 (0.37; 0.92)	Myocardial infarction	Class I and Ib antiarrhythmic agents	Placebo
DCCT	1993	1441	4.2%	26.8 (±7.1)	47.2%	1.03 (0.95; 1.12)	–	Type 1 diabetes	Intensive therapy	Conventional therapy
DIG	1997	7788	33.5%	63.9 (±10.9)	24.7%	1.03 (1.02; 1.03)	1.21 (1.10; 1.32)	Heart failure	Digoxin	Placebo
ENRICHD	2003	2481	13.7%	60.8 (±12.4)	43.7%	1.06 (1.05:; 1.07)	0.73 (0.59; 0.91)	Acute MI	Cognitive behaviour therapy	Usual medical care
FAVORIT	2011	4108	10.5%	51.9 (±9.4)	37.2%	1.06 (1.05;1.07)	1.04 (0.86; 1.27)	Stable kidney transplant	Multivitamin + folic acid, vitamin B12 B6	identical multivitamin alone
HALTC	2008	1050	0.02%	50.6 (±7.2)	29.0%	1.01 (0.95; 1.07)	1.28 (0.47; 2.13)	Chronic hepatitis C	pegylated interferon alpha-2a	Same, discontinue
HEMO FLUX	2002	1846	42.2%	57.1 (±14.0)	56.2%	1.04 (1.03; 1.05)	0.99 (0.87; 1.13)	Haemodialysis	High-flux dialysis	Low-flux dialysis
IST HEPDOS	1997	9716	22.9%	71.7 (±11.6)	46.0%	1.06 (1.06; 1.07)	0.80 (0.74; 0.87)	Acute stroke	Unfractionated heparin, aspirin	Placebo
IST HEP	1997	19,433	22.5%	71.7 (±11.6)	46.5%	1.06 (1.06; 1.07)	0.78 (0.73; 0.83)	Acute stroke	Unfractionated heparin, aspirin	Placebo
MAGIC	2002	6211	15.2%	67.8 (±10.4)	44.8%	1.06 (1.05; 1.07)	0.50 (0.44; 0.57)	Acute MI	IV magnesium Sulphate	Placebo
OAT	2006	2166	7.9%	58.6 (±11.0)	22.0%	1.04 (1.02; 1.05)	0.82 (0.58; 1.16)	Congestive heart failure	PCI + stenting with optimal therapy	Optimal therapy alone
PEACE	2004	8290	7.6%	64.3 (±8.2)	18.0%	1.07 (1.06; 1.08)	1.20 (0.96; 1.48)	Coronary artery disease	Trandolapril	Placebo
SHEP	1991	4648	9.6%	72.2 (±6.7)	56.8%	1.07 (1.05; 1.08)	1.52 (1.27; 1.83)	Hypertension	Chlorthalidone/atenolol antihypertensiva	Placebo
SOLVD INT	1992	2568	37.4%	60.4 (±9.9)	19.6%	1.02 (1.01; 1.02)	1.11 (0.95; 1.31)	Congestive heart failure	Enalapril	Placebo
SOLVD PRE	1992	4225	15.1%	58.6 (±10.3)	11.3%	1.03 (1.02; 1.04)	1.03 (0.80; 1.32)	Congestive heart failure	Enalapril	Placebo
TIMI B	1989	1434	3.6%	55.0 (±10.4)	14.7%	1.06 (1.03; 1.09)	0.37 (0.20; 0.66)	Acute MI	Invasive strategy	Conservative strategy
TIMI	1989	3339	4.9%	56.8 (±10.2)	17.9%	1.08 (1.06; 1.09)	0.44 (0.31; 0.60)	Acute MI	Invasive strategy	Conservative strategy

*Calculated by univariable Cox model

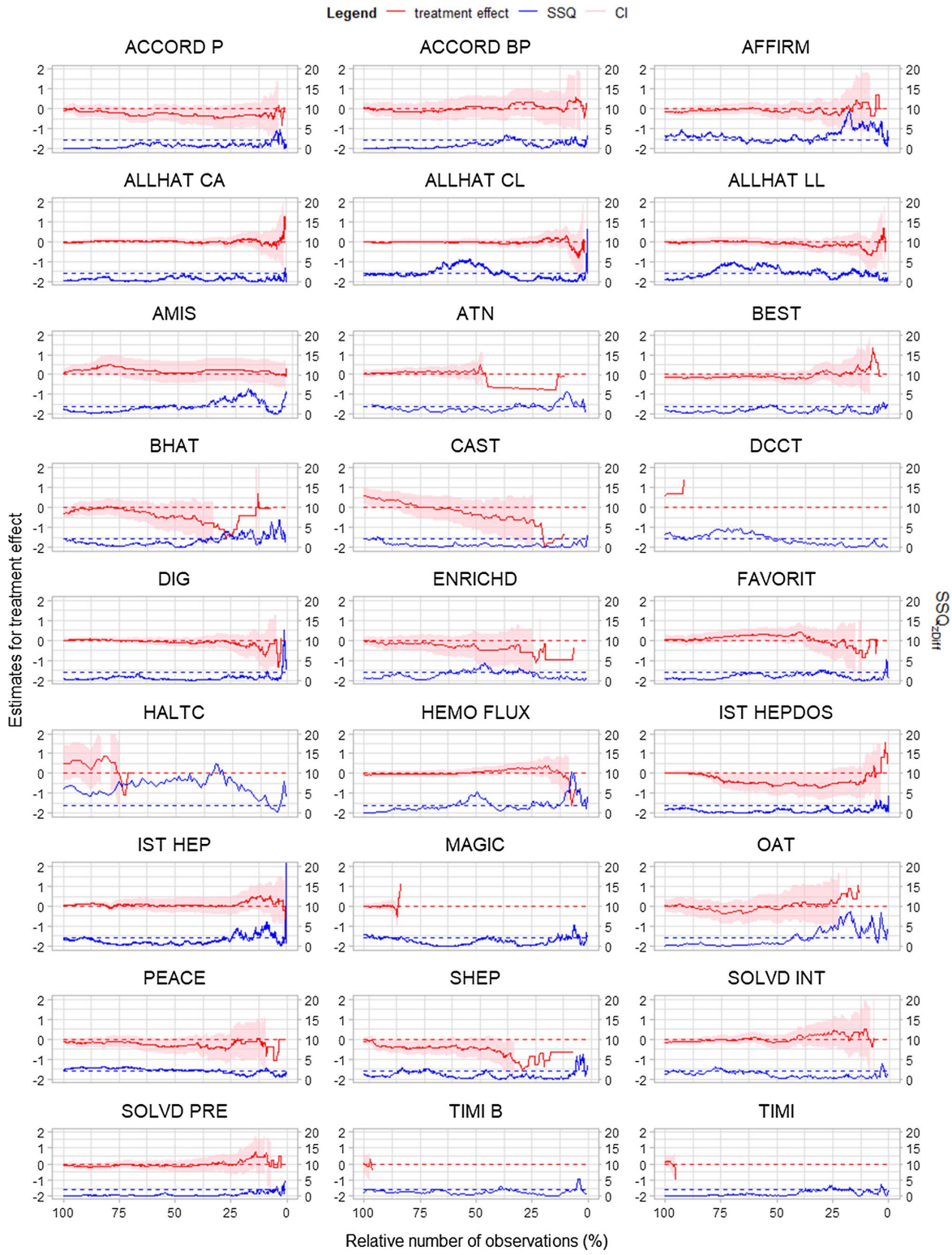
Legend
treatment effect
SSQ
CI
ACCORD P
ACCORD BP
AFFIRM
ALLHAT CA
ALLHAT CL
ALLHAT LL
AMIS
ATN
BEST
BHAT
CAST
DCCT
DIG
ENRICHD
FAVORIT
HALTC
HEMO FLUX
IST HEPDOS
IST HEP
MAGIC
OAT
PEACE
SHEP
SOLVD INT
SOLVD PRE
TIMI B
TIMI
Estimates for treatment effect
SSQdiff
Relative number of observations (%)

Fig. 2 Trajectories of treatment effect estimates (red) and SSQ_{zDiff} (blue) over the course of "Dynamic Landmarking" for the 27 RCTs. Treatment effect estimates on the log(HR) scale and corresponding confidence intervals are shown on the left y-axis. Sums of squared z-differences $\left(SSQ_{zDiff}\right)$ for balance measuring of age and sex in all 28 RCTs are given on the right y-axis. Dashed lines symbolize the case of no treatment effect (red) and balanced prognostic factors (blue). The x-axis shows the relative number of remaining observations in sorted data sets

used patients' baseline age and sex because these are the only two prognostic factors that were available in all trials and are accepted to have an influence on survival. We then applied the "Dynamic Landmarking" process of stepwise deletion ($M = 10$), fitting a Cox model and balance checking to obtain a trajectory of treatment effect estimates and SSQ_{zDiff} for each data set. For the two prognostic factors age and sex that were assumed to be independent, SSQ_{zDiff} has an expected value of 2 in a randomized setting.

Results

Overall, 18,095 Cox models with treatment as the only variable were fitted across all trials. In Fig. 2, we show the trajectories of the treatment effect estimates with their pointwise 95% confidence intervals and the corresponding SSQ_{zDiff} for age and sex for each trial. For most trials, we did not see a systematic change in parameter estimates over the course of the "Dynamic Landmarking" process. In contrast, most trajectories of treatment effect estimates stayed remarkably constant and merely showed random fluctuations. In addition, a relevant imbalance in prognostic factors was rarely seen with SSQ_{zDiff} being close to the expected value of 2. Some trials, e.g., DCCT or MAGIC, showed a completely erratic behaviour, which might be due to the low overall number of deaths in these trials. However, the covariate balance of age and sex could still be calculated, as SSQ_{zDiff} is computed from all observations of the remaining risk set, irrespective of their event status. In conclusion, we found no empirical evidence of relevant selection bias in the considered trials.

Discussion

There is no visually apparent evidence of selection bias in randomized controlled trials; therefore, we conclude that hazard ratios are not hazardous in this respect—at least not as hazardous as announced in the respective body of literature. This is the simple and, in our view, rather surprising result of our analysis of the original data of 27 large RCTs. To arrive at this result, we proposed "Dynamic Landmarking" as a method to visualize the suspected selection, and considered scenarios in which the heterogeneity between patients is explained by observed but omitted prognostic factors. In a small proof-of-concept simulation, we demonstrated that this method gives an indication of selection bias caused by an omitted prognostic factor with a strong influence on survival. For both, the simulation and the empirical data, we fitted Cox models with treatment as the only variable and measured the covariate balance for observed but omitted prognostic factors. The simulation showed that a stronger influence of the omitted covariate on the survival outcome indeed causes a more visible systematic change in the treatment effect estimate and an increasing imbalance in the omitted covariate itself. Our empirical investigation, however, yielded no evidence that this also happens in real RCTs. Considering measured but omitted prognostic factors has two main advantages. First, it is possible to identify measured prognostic factors, which would induce selection bias if one would not account for them. Second, the case in which no measured covariate shows an increased imbalance but a systematic changing trajectory of treatment effect estimate is still seen may be explained by unobserved heterogeneity or a truly time-dependent treatment effect. Depending on which of these cases occurs, the data should be handled differently. In the first case, adjusting for the omitted factor might be sufficient to avoid built-in selection bias. In the second case, frailty models [26] or more flexible time-dependent extensions of the Cox model [3, 11, 12] may be used for data analysis. In any case, "Dynamic Landmarking" cannot distinguish between true unobserved heterogeneity and time-dependent treatment effects.

The explanation for the absence of built-in selection bias in the empirical data is likely simple. This bias would only occur in RCTs with both very large treatment effects and in the presence of a high influence of the omitted covariate on survival (i.e., large "unobserved" heterogeneity) [21, 23]. However, very large treatment effects are rarely seen in RCTs, because we expect equipoise of treatments before the trial. Indeed, and as shown in a large Cochrane review [9], only slightly more than half of new experimental treatments perform better than established treatments when tested in RCTs. In particular, log(HR)s with a magnitude larger than 1 are seen in only about 3% of trials, and log(HR)s with a magnitude larger than 2 essentially never occur [9]. In addition, with respect to patient heterogeneity, study populations in RCTs are generally careful selected to increase internal validity but at the cost of external validity, thus minimizing patient heterogeneity. Indeed, in an additional analysis of the RCT data with gamma-frailty models (results shown in the supplementary information, see Fig. S3), we saw that the estimated frailty variance is usually close to zero, pointing to negligible amounts of unobserved heterogeneity.

We must also acknowledge some limitations of our study. The RCTs of Kent et al. [15] mainly originate from the field

of cardiovascular medicine, and thus, the results in other clinical disciplines might diverge from those reported here. We restricted our analysis to only two prognostic factors (age and sex) for imbalance measurement because only these two were available in all data sets. We are aware that, in addition to these two selected prognostic factors, there may be other (unobserved) prognostic factors, which have more influence on the survival outcome. Furthermore, we cannot draw any conclusions for studies that show only very few events. Due to the large number of censored patients in the data set, no meaningful trajectory can be drawn, and no potential systematic change can be identified. We are also aware that the selected prognostic factors (age and sex) mostly show a weak association with the survival outcome; therefore, one would only expect a very slight increase in the imbalance. Related to this, the SSQ_{zDiff} as an aggregated balance measure, could produce less conclusive results if one includes prognostic factors without impact on the survival outcome for its computation. Last, "Dynamic Landmarking" only gives a visual indication of potential selection bias and is only one approach to assess whether the treatment estimate changes during the trial. Alternative methods that could be considered are already mentioned above [3, 11, 12]. Further research is necessary to assess which of these approaches is most efficient and accurate in detecting built-in selection bias and whether the relative performance of these approaches depends on trial characteristics.

To summarize, we feel that the warnings of the built-in selection bias in RCTs when using hazard ratios are of little practical relevance in most cases and that the hazard ratios from most analysed RCTs are not materially affected by this bias. The empirical evidence we provide suggests that the built-in selection bias does not materially affect the results of most RCTs we considered. Therefore, we feel that hazard ratios can be safely used to analyse RCTs with time to event outcomes, at least with respect to built-in selection bias. However, although the hazard ratio does not suffer from the announced selection bias, its problems in terms of interpretability, causality or non-collapsibility remain virulent and should be carefully considered by the biostatistical community.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10654-023-01026-z.

Acknowledgements The studies ATN, DCCT/EDIC, FAVORIT, HEMO, and HALT-C were conducted by the ATN, DCCT/EDIC, FAVORIT, HEMO and HALT-C investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The NIDDK Central Repositories supplied the data from the ATN, DCCT/EDIC, FAVORIT, HEMO and HALT-C studies reported here. The manuscript was not prepared in collaboration with these investigators and does not necessarily reflect the opinions or views them. The manuscript was using material from the study groups that performed the ACCORD,AFFIRM, ALLHAT, AMIS, BEST, BHAT, CAST, DIG, ENRICHD, MAGIC, OAT, PEACE, SHEP, SOLVD and the TIMI2 studies, which were obtained from the NHLBI Biologic Specimen and Data Repository Information Coordination Centre. The manuscript does not necessarily reflect the opinions or views of these study groups or the NHLBI.

Author contributions All authors contributed to the study conception and design. AS, OK and AW performed material preparation, data collection and analysis. AS wrote the first draft of the manuscript and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support was received during the preparation of this manuscript.

Data availability Data and code are available for replication upon request.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics approval After having received approval from the Institutional Review Board of the Medical Faculty of the Heinrich-Heine-University Duesseldorf from March 2019 (Study No.: 5986R, Registration-ID: 201707 4356) we contacted the three clinical study registries [NHLBI2018, NIDDK2018, GlaxoSmithKline2018] and were allowed to reuse the data sets.

References

1. Aalen OO, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Anal. 2015;21(4):579–93. https://doi.org/10.1007/s10985-015-9335-y.
2. Aalen OO, Borgan Ø, Gjessing HK. Survival and event history analysis: a process point of view. New York: Springer-Verlag; 2008.
3. Abrahamowicz M, Mackenzie T, Esdaile JM. Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. J Am Stat Assoc. 1996;91(436):1432–9. https://doi.org/10.1080/01621459.1996.10476711.
4. Balan TA, Putter H. A tutorial on frailty models. Stat Methods Med Res. 2020;29(11):3424–54. https://doi.org/10.1177/0962280220921889.
5. Cecilia-Joseph E, Auvert B, Broët P, Moreau T. Influence of trial duration on the bias of the estimated treatment effect in clinical trials when individual heterogeneity is ignored. Biom J. 2015;57(3):371–83. https://doi.org/10.1002/bimj.201400046.
6. Cox DR. Partial likelihood. Biometrika. 1975;62(2):269–76. https://doi.org/10.1093/biomet/62.2.269.
7. Cox DR. Regression models and life tables. J R Stat Soc Se B (Methodol). 1972;34(2):187–220.
8. De Neve J, Gerds TA. On the interpretation of hazard ratio in Cox regression. Biom J. 2020;62(7):742–50.
9. Djulbegovic B, Kumar A, Glasziou PP, Perera R, Reljic T, Dent L, Raftery J, Johansen M, Di Tanna GL, Miladinovic B, Soares HP, Vist GE, Chalmers I. New treatments compared to established treatments in randomized trials. Cochrane Database Syst Rev. 2012. https://doi.org/10.1002/14651858.MR000024.pub3.
10. Duchateau L, Janssen P. The frailty model. 2008. New York: Springer Verlag.

11. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–26.
12. Hastie T, Tibshirani R. Varying-coefficient models. J R Stat Soc. 1993;55:757–96.
13. Hernán MA. The hazards of hazard ratios. Epidemiology. 2008;21(1):13–5. https://doi.org/10.1097/EDE.0b013e3181c1ea43.
14. Kabisch M, Ruckes C, Seibert-Grafe M, Blettner M. Randomized controlled trials: part 17 of a series on evaluation of scientific publications. Deutsches Ärzteblatt Int. 2011. https://doi.org/10.3238/arztebl.2011.0663.
15. Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. Int J Epidemiol. 2016;45(6):2075–88. https://doi.org/10.1093/ije/dyw118.
16. Kuss O. The z-difference can be used to measure covariate balance in matched propensity score analyses. J Clin Epidemiol. 2013;66(11):1302–7. https://doi.org/10.1016/j.jclinepi.2013.06.001.
17. Kuss O, Miller M. Unknown confounders did not bias the treatment effect when improving balance of known confounders in randomized trials. J Clin Epidemiol. 2020;126:9–16. https://doi.org/10.1016/j.jclinepi.2020.06.012.
18. Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. Lifetime Data Anal. 2013;19(3):279–96. https://doi.org/10.1007/s10985-013-9242-z.
19. McName R. How serious is bias in effect estimation in randomised trial with survival data given risk heterogeneity and informative censoring? Stat Med. 2017;36(21):3315–33. https://doi.org/10.1002/sim.7343.
20. Sjölander A, Dahlqwist E, Zetterqvist J. A note on the noncollapsibility of rate differences and rate ratios. Epidemiology. 2016;27(3):356–9. https://doi.org/10.1097/EDE.0000000000000433.
21. Stensrud MJ. Interpreting hazard ratios: insights from frailty models. arXiv: Methodology. 2018.
22. Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. Eur Heart J. 2019;40(17):1378–83. https://doi.org/10.1093/eurheartj/ehy770.
23. Sutradhar R, Austin PC. Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. Ann Epidemiol. 2018;28(1):54–7. https://doi.org/10.1016/j.annepidem.2017.10.014.
24. Van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. Scand J Stat. 2007;34:78–85.
25. Van Houwelingen HC, Putter H. Dynamic prediction in clinical survival analysis. Boca Raton: Chapmann & Hall/CRC; 2012.
26. Wienke A. Frailty models in survival analysis. Boca Raton: Chapmann & Hall/CRC; 2010.

Publication 2

Strobel, A., Wienke, A., Gummert, J., Bleiziffer S. & Kuss, O (2024). Built-in selection or confounder bias? *Dynamic Landmarking* in matched propensity score analyses. *BMC Medical Research Methodology*, 24, 316.

Contribution to Publication 2:

Conception, data cleaning and data preparation, programming method, conducting simulation study and analyzing data, writing and revision of manuscript.

Strobel *et al. BMC Medical Research Methodology* (2024) 24:316
https://doi.org/10.1186/s12874-024-02444-7

BMC Medical Research Methodology

RESEARCH Open Access

Built-in selection or confounder bias? *Dynamic Landmarking* in matched propensity score analyses

Alexandra Strobel[1*], Andreas Wienke[1], Jan Gummert[2], Sabine Bleiziffer[2] and Oliver Kuss[3,4]

Abstract

Background Propensity score matching has become a popular method for estimating causal treatment effects in non-randomized studies. However, for time-to-event outcomes, the estimation of hazard ratios based on propensity scores can be challenging if omitted or unobserved covariates are present. Not accounting for such covariates could lead to treatment estimates, differing from the estimate of interest. However, researchers often do not know whether (and, if so, which) covariates will cause this divergence.

Methods To address this issue, we extended a previously described method, *Dynamic Landmarking*, which was originally developed for randomized trials. The method is based on successively deletion of sorted observations and gradually fitting univariable Cox models. In addition, the balance of observed, but omitted covariates can be measured by the sum of squared z-differences.

Results By simulation we show, that *Dynamic Landmarking* provides a good visual tool for detecting and distinguishing treatment effect estimates underlying built-in selection or confounding bias. We illustrate the approach with a data set from cardiac surgery and provide some recommendations on how to use and interpret *Dynamic Landmarking* in propensity score matched studies.

Conclusion *Dynamic Landmarking* is a useful post-hoc diagnosis tool for visualizing whether an estimated hazard ratio could be distorted by confounding or built-in selection bias.

Keywords Cox model, Hazard ratio, Built-in selection bias, Confounding bias

Background

Randomized controlled trials (RCTs) are the gold standard for evaluating treatment effects in medical research, because random treatment allocation should guarantee balanced known and unknown covariates in the compared groups, resulting in the absence of confounding (for terminology used in manuscript see Tab. S1). However, even if confounding is minimized after randomization, prognostic factors (i.e. covariates that are associated with the outcome but not with treatment allocation) may still be present. For time-to-event data, the Cox model [7, 8] is commonly used for statistical analysis, providing

*Correspondence:
Alexandra Strobel
alexandra.strobel@uk-halle.de
[1]Institute of Medical Epidemiology, Biostatistics, and Informatics, Interdisciplinary Center for Health Sciences, Medical Faculty, Martin-Luther-University Halle Wittenberg, Halle, Germany
[2]Heart and Diabetes Center North Rhine-Westphalia, Ruhr-University Bochum, Bad Oeynhausen, Germany
[3]German Diabetes Center, Leibniz Center for Diabetes Research, Institute for Biometrics and Epidemiology, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany
[4]Centre for Health and Society, Faculty of Medicine, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

the hazard ratio as the generic effect measure. Typically, in RCTs the Cox model does not include prognostic factors as covariates. Instead, a marginal Cox model with only the treatment as a single covariate is estimated, yielding a marginal hazard ratio that is interpreted as a population-averaged treatment effect. However, there is often interest in understanding treatment effects at a subject-specific level. A subject-specific (conditional) interpretation of the hazard ratio can only be made when conditioning the Cox model on all prognostic factors. This particularly means that if a single prognostic factor (whether observed or unobserved) is omitted from the Cox model, it would prevent the hazard ratio from being interpreted on a subject-specific level. More precise, assume a proportional hazards model (1)

$$\lambda\ (t|Z,U) = \lambda_0\,(t)\exp(\beta_Z Z + \beta_U U) \qquad (1)$$

where $\lambda_0\,(t)$ is an unspecified baseline hazard function, depending on time t and is assumed to be common across all individuals. Furthermore, Z and U are some observed covariates with their corresponding regression coefficients β_Z and β_U. Then $\lambda\ (t|\,Z,U)$ defines the conditional hazard with β_Z summarizing the conditional effect of Z, yielding a subject-specific interpretation. On the other hand, if U will be omitted, one would estimate model (2), i.e.:

$$\lambda\ (t|Z) = \lambda_0\,(t)\exp\,(\beta_Z Z) \qquad (2)$$

with $\lambda\ (t|\,Z)$ reflecting the marginal hazard, yielding an population-averaged interpretation. Importantly, conditional and marginal Cox models will not provide the same estimates for a treatment effect if additional prognostic factors are associated with the time-to-event outcome [9, 29, 30]. This circumstance is referred to as "non-collapsibility", indicating that the magnitude of the effect measure is changing when conditioning on a prognostic factor [10]. This is often accompanied by the term "built-in selection bias", which can be seen as result of conditioning on previous survival within hazard rates. More precise, assume an omitted prognostic factors (i.e., measured during the trial but omitted from the Cox model), which introduces heterogeneity, causing individuals at higher baseline risk (regarding omitted prognostic factors) to expect the event earlier than those at lower risk [1, 17]. Given an effective treatment, this would result in higher-risk individuals surviving longer in the treated group than in the control group. This results in a deviation from the marginal and conditional hazard ratio, due to conditioning on prior survival. Depending on the magnitude of the treatment effect, the influence of the omitted prognostic factor on the time-to-event outcome and the follow-up time, the magnitude of the built-in selection bias changes [5, 28, 31]. Therefore, when aiming for a conditional treatment effect (more precise, conditional on all prognostic factors) in RCTs, all prognostic factors have to be included in the Cox model. Please note: In the case where treatment is the only prognostic factor influencing time-to-event and there are no other prognostic factors, the marginal model and the conditional model would give the same value for the marginal and the conditional hazard ratio. This is because the Cox model would then include all relevant prognostic factors, that is, only the treatment allocation, and no other adjustments are needed for estimating a conditional treatment effect. As a result, non-collapsibility would not be an issue and thus built-in selection bias would not occur.

In non-randomized trails, the situation might be more complex because confounding becomes an additional issue. Here, treatment allocation is generally determined by baseline characteristics, leading to systematic differences between treatment groups [25]. One prominent way to address these baseline differences is balancing the data by Propensity Score (PS) matching [26, 27]. Here, in a first step the PS for each individual is usually estimated via a logistic regression model. In a second step the PS is used for estimating the treatment effect of interest (that is, in our case the hazard ratio) [21]. Under the assumptions of positivity, consistency, and unconfoundedness for the PS, valid causal statements about treatment effects can be made. Misspecification of the PS model due to the omission of relevant confounders would lead to confounding bias, resulting in a biased treatment effect estimate. However, even if the PS model includes all confounders, non-collapsibility (and the corresponding built-in selection bias) plays a role when fitting a Cox model in the PS matched trial. Usually, as in RCTs, a marginal Cox model with the treatment effect as the single covariate is fitted to the data, yielding a marginal (population-averaged) treatment effect estimate. However, when aiming for a conditional (subject-specific) treatment effect, the Cox model needs to be conditional on all relevant prognostic factors. Note that prognostic factors cannot be taken into account by PS models, as the PS addresses the association between a covariate and the treatment allocation, which (by definition) is not present in prognostic factors. Therefore, when estimating a treatment effect in PS matched trials, two potential issues could arise when covariates are omitted from the analysis. First, omitting a prognostic factor from the Cox model would lead to the built-in selection bias. Second, omitting a confounder from the PS model would entail confounding bias. Both issues have the consequence that the final treatment effect estimate differs from the estimate of interest (that is, a conditional and unbiased treatment effect) [6, 14]. For an overview of concepts and comparison in RCTs and PS-matched trials please see Tab. S2.

The choice of covariates for the PS model and the subsequent outcome model relies on scientific understanding and clinical expertise. This especially introduces the possibility of omission of covariates that were measured during the trial, but not included in the PS model or, after PS matching, in the Cox model. It is therefore of interest to investigate whether an estimated treatment effect is subject to confounding bias or built-in selection bias. Unfortunately, the hazard ratio provides the effect in a single number, not giving a hint for any of these issues. Therefore, a recent article introduced a new method, *Dynamic Landmarking*, for diagnosing whether an estimated treatment effect from a Cox model was subject to built-in selection bias in RCTs [32]. The original methodological approach was designed to detect potential prognostic factors that are measured but omitted from the Cox model and could therefore induce built-in selection bias.

The aim of the present work is to extend the existing *Dynamic Landmarking* approach to PS matched trials. More precisely, we want to use *Dynamic Landmarking* as a post-hoc diagnosing tool in order to check if the estimated hazard ratio could be distorted by confounding or built-in selection bias. Moreover, we are interested in detecting covariates that were observed (e.g., are present in the data set), but omitted from the analysis, which could either induce potential built-in selection or confounding bias.

First, we describe the extension of *Dynamic Landmarking* to the PS matched case. Second, we give the results of a simulation study to examine how the approach performs in a PS matched trial. Third, we apply the extended procedure to a real data set from cardiac surgery and finally discuss the results.

Methods

The original *Dynamic Landmarking* is a methodological approach, which provides a visual tool for diagnosing if an estimated treatment effect is subject to built-in selection bias. Furthermore, omitted prognostic factors that are measured during the trial but omitted from the Cox model, are investigated whether they induce built-in selection bias. The idea of *Dynamic Landmarking* is quite simple: First, the dataset is sorted by observation time and a univariable Cox model only including the treatment is fitted to the full data set. Afterwards, the earliest M $(M > 0)$ observations are deleted regardless of the event status (observed or censored) and a new univariable Cox model is fitted to the smaller data set. After each deletion step, the start of the follow-up interval for the new Cox model is moved forwards. More precisely, the new time zero for the new Cox model corresponds to the follow-up time of the latest of the M deleted individuals in the previous step. This procedure of deleting the earliest M observations and refitting univariable Cox models is continued until the data set no longer contains a sufficient number of observations for convergence. In general, high-risk individuals will have shorter observation times than low-risk individuals, as they tend to expect the event of interest earlier. Consequently, individuals with higher baseline risk (regarding the omitted prognostic factors) will be deleted earlier during *Dynamic Landmarking*.

In parallel, the balance of omitted prognostic factors is measured in each step by the sum of squared z-differences (SSQ_{zDiff}) [19], with $SSQ_{zDiff} = \sum z^2_{con} + \sum z^2_{bin} + \sum z^2_{ord} + \sum z^2_{nom}$, whereby e.g.,

$$z_{cont} = \frac{\bar{x}_T - \bar{x}_C}{\sqrt{\frac{\widehat{\sigma}^2_T}{N_T} + \frac{\widehat{\sigma}^2_C}{N_C}}} \quad \text{and} \quad z_{bin} = \frac{\widehat{p}_T - \widehat{p}_C}{\sqrt{\frac{\widehat{p}_T(1-\widehat{p}_T)}{N_T} + \frac{\widehat{p}_C(1-\widehat{p}_C)}{N_C}}}.$$

Here $\bar{x}_T$, $\bar{x}_C$, $\widehat{\sigma}^2_T$, $\widehat{\sigma}^2_C$, $\widehat{p}_T$, $\widehat{p}_C$, N_T, N_C denote the respective estimated means, variances, proportions, and sample sizes of the two groups (formula for all z-differences can be found in Formula S1). The SSQ_{zDiff} is a global balance measure and follows a chi-squared-distribution with expectation k for k independent covariates.

After each deletion-and-refitting step, the point estimator for the treatment effect and the SSQ_{zDiff} is saved, yielding a trajectory depending on the remaining number of individuals. Through the systematic removal of individuals, treatment effects are gradually estimated within a population of lower-risk patients, potentially leading to a systematic shift in the effect trajectory due to the presence of built-in selection bias. Moreover, a potential imbalance in omitted prognostic factors arises, manifesting as a systematic shift in the SSQ_{zDiff} trajectory [32].

To apply *Dynamic Landmarking* in non-randomized trials, a balancing procedure, e.g. PS matching, has to be applied prior to sorting the data regarding the observation time. Afterward, the original *Dynamic Landmarking* is carried out. However, note that omitted variables in RCTs (by design) can only be prognostic factors. In PS matched studies, however, they can be both prognostic factors and confounders. This potentially creates two problems, first built-in selection bias due to omission of prognostic factors and, second, confounding bias due to omitted confounders, and of course, both should be addressed separately by *Dynamic Landmarking*. This distinction between omitted prognostic factors and omitted confounders can be made by looking at the definition of SSQ_{zDiff}: Omitting a observed confounder from the PS model would result in unbalanced groups after PS matching. This is because the association of the omitted confounder with the treatment allocation is still present, resulting in large values of SSQ_{zDiff} already at the

beginning of *Dynamic Landmarking*, that is, before the first deletion step. Omitting a prognostic factor from the Cox model on the other hand would still yield balanced groups after PS matching resulting in lower initial values of SSQ_{zDiff}. Hence, initial SSQ_{zDiff} -values for the full data set will give a first hint on whether the omitted variable is a confounder or a prognostic factor.

The following preconditions must be met in order to achieve valid results from *Dynamic Landmarking*: First, independent censoring has to be assumed. Second, the conditional hazard ratio for treatment is assumed to be constant across the population and over time, i.e. proportional hazards hold and treatment effect is time-invariant. Third, for measuring the balance by SSQ_{zDiff} at least one available covariate has to be omitted from either the PS or the Cox model.

Results from a simulation study

Data generation process

We simulated a non-randomized intervention trial with Z denoting the treatment, Y the time-to-event outcome, X a known and measured confounder and U an omitted covariate, see Fig. 1 for the corresponding graphical illustration of the data generation process. Both, X and U, follow a standard normal distribution. First, we simulated the probability of treatment allocation for each subject i from the logistic model

$$logit\left(p_i\right) = \alpha_0 + \alpha_X \cdot X_i + \alpha_U \cdot U_i.$$

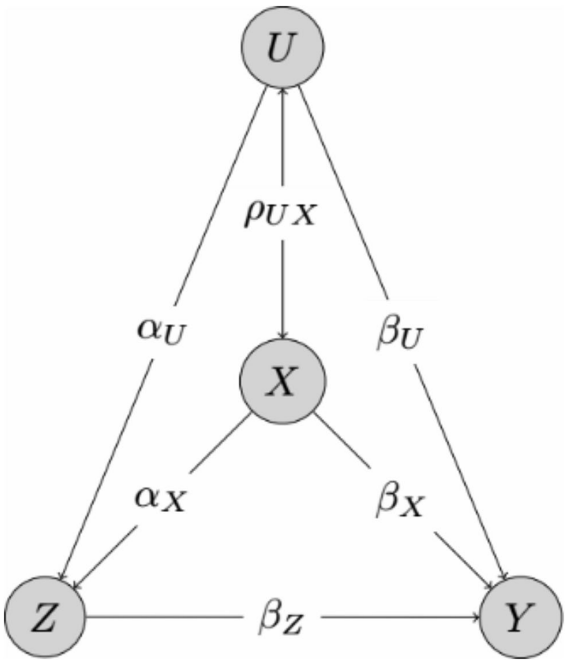

Fig. 1 Graphical illustration for data generation process

For the intercept, $\alpha_0 = -1.21$ was chosen in order to obtain approximately 24% treated individuals, which was motivated by the empirical example in Section "Illustration of the procedure with an example from cardiac surgery". The parameter α_X was set to $\log(3)$. This denotes a strong impact of the confounder X on the treatment assignment. Afterwards, we generated the actual treatment status Z_i from a Bernoulli distribution with subject-specific probability p_i. We then simulated the time-to-event outcome Y_i for each individual using a Weibull baseline hazard with parameters $\lambda = 0.01$ and $\gamma = 1.5$. The final hazard function used was:

$$h\left(t|Z,X,U\right) = \gamma\,\lambda\,t^{\gamma-1} \cdot e^{\beta_Z Z + \beta_X X + \beta_u U}.$$

For the regression parameter β_X we used the value $\log(3)$, which was intended to denote a strong impact of X on the time-to-event outcome. We considered different effects of U on treatment allocation ($\alpha_U \in \{\log(0.5), \log(0.66), \log(0.8), \log(1), \log(1.25), \log(2), \log(3)\}$). We further varied the effect of U on the time-to-event outcome by using the following regression coefficients: $\beta_U \in \{\log(0.5), \log(0.66), \log(0.8), \log(1), \log(1.25), \log(1.5), \log(2), \log(3)\}$. Furthermore, we assumed various correlations between U and X: $\rho_{XU} \in \{0,\ 0.2,\ 0.6,\ 0.9\}$. Moreover, we considered different values for the conditional treatment effect: $\beta_Z \in \{\log(1.25), \log(1.5), \log(2), \log(3)\}$ and assumed censoring proportions of approximately 10%, 40% and 80% which were generated using a exponential distribution with parameter $\lambda \in \{0.2, 0.6, 0.9\}$For each scenario, we simulated 500 data sets with 5,000 individuals each. Please be aware that U is classified differently based on the values of α_U and β_U U is considered an independent covariate when both $\alpha_U = 0$ and $\beta_U = 0$ a prognostic factor when $\alpha_U = 0$ and $\beta_U \neq 0$, an instrumental variable when $\alpha_U \neq 0$ and $\beta_U = 0$, and finally, a confounder when both $\alpha_U \neq 0$ and $\beta_U \neq 0$.

Data analyses

For each scenario, we estimated the PS by logistic regression, including the known confounder X, but excluding the covariate U: $logit\left(p_i\right) = \alpha_0 + \alpha_X \cdot X_i$. We then performed a 1:1 PS-matching without replacement. Each treated individual was matched with the greedy nearest available neighbour with a caliper width of 0.2 of the standard deviation of the logit of the propensity score [2, 3]. In a second step, we applied *Dynamic Landmarking* to the PS-matched data set. Therefore, we fitted stratified (for the matching stratum) Cox models with treatment as the only covariate:

$$h_j\left(t|Z\right) = h_{0,j}\left(t\right) \cdot \; e^{\beta_{\,Z} Z} \tag{3}$$

Here, $h_{0,j}$ refers to the baseline hazard function for matching stratum j. These stratified (for matching stratum) Cox model will be referred to "stratified Cox model" from now on. Please note, that U was omitted from both, the PS model and the Cox model, whereas X was considered in the PS model in each scenario.

Results

Omitting a prognostic factor – detecting induced built-in selection bias

In Fig. 2 we give the results for an omitted prognostic factor U (i.e., $a_U = 0$), a highly effective treatment ($\beta_{\,Z} = \log(3)$) and a censoring proportion of 10%. Results for smaller treatment effects and higher censoring proportions are given in the supplementary information (see Fig. S1 – Fig. S5). Two important things should be noted: First, in these scenarios, the PS model was correctly specified and built-in selection bias is induced by the omission of a prognostic factor. Second, the treatment effect trajectory will not be equal to the true simulated effect $\beta_{\,Z}$ at the beginning of *Dynamic Landmarking*. This is because we show the percentage of remaining individuals on the x-axis and not the original observation time. As a result, the initial treatment effect estimate derived from *Dynamic Landmarking* corresponds to the estimate one would obtain at the end of a study using a stratified Cox model. However, since a relevant prognostic factor has been excluded, this initial estimate is already subject to built-in selection bias, leading to a discrepancy between the estimated and the true simulated effect from the beginning on.

The mean sample size of the PS matched data was 2,402 in the simulation. In the first column of Fig. 2, U is independent of the confounder X ($\rho_{\,UX} = 0$). We observe that a higher impact of U on the time-to-event outcome causes a more visible systematic shift in the treatment effect trajectory. Additionally, all scenarios show low initial SSQ_{zDiff}-values indicating the omission of a prognostic factor that is still balanced between the treatment groups after PS matching. Moreover an increase of the SSQ_{zDiff}-trajectory is observed during the deletion of the first 50% of observations. Similar results were obtained for smaller treatment effects and higher censoring rates. However, as highlighted by serveral authors [e.g. 31, 35], the built-in selection bias occurs less prominent in case of smaller treatment effects and smaller

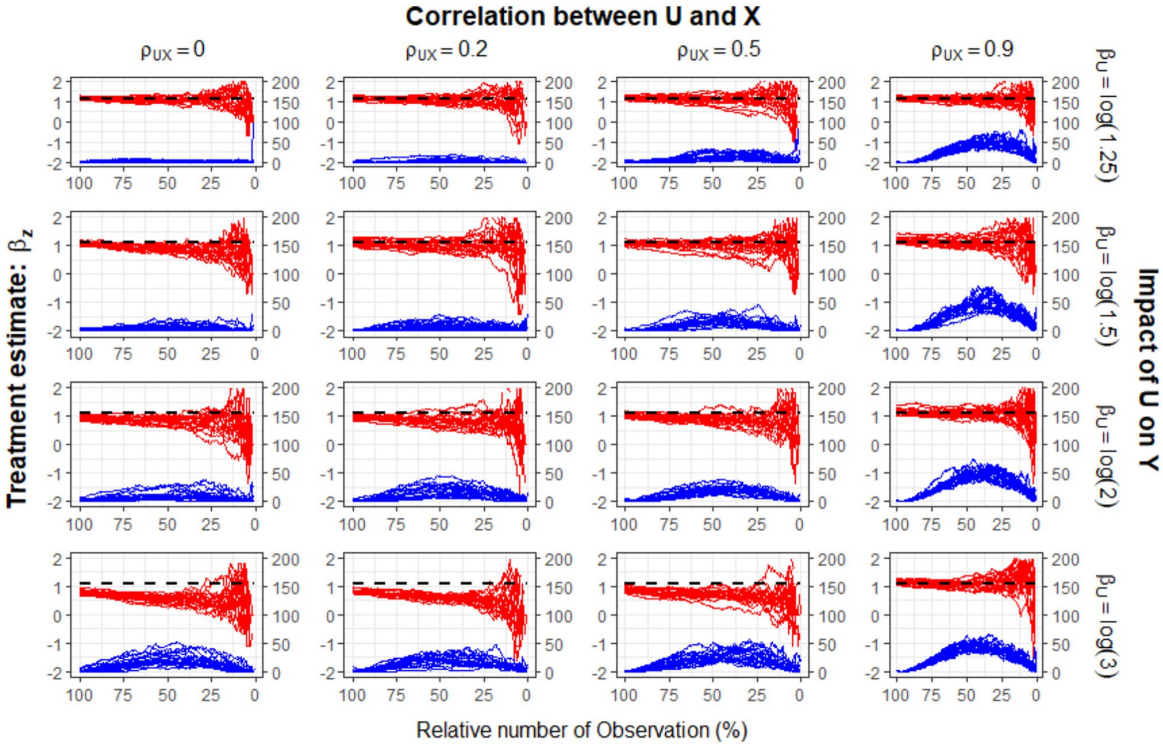

Fig. 2 Trajectories of treatment effect (left y-axis, red) on the log(HR) scale and sum of squared z-differences (right y-axis, blue) for balance measuring of the omitted covariate U for 500 simulated data sets. Dashed black lines show the true, simulated conditional treatment effect estimate $\beta_{\,Z} = \log(3)$. All scenarios assume the omission of a prognostic factor U, i.e. $\alpha_{\,U} = 0.$, and a censoring rate of 10%

prognostic effects. Consequently, in such cases, Dynamic Landmarking would identify a less pronounced decline in treatment effect trajectories. In the remaining columns, we simulated a non-zero correlation between X and U varying it from weak to strong. Here we find that the estimated treatment effect moves closer to the true simulated one if the correlation gets stronger. Importantly, less systematic changes in the treatment effect trajectory can be observed. This is because the omitted prognostic factor U is indirectly accounted for by including X in the PS model, allowing a correction towards the true treatment effect. And of course, the stronger the correlation, the closer will the estimated hazard ratio be to the true, simulated one [14].

Omitting a confounder – detecting confounding bias

The results of the simulation when omitting a true confounder (i.e., $\alpha_U \neq 0$) from the PS model are shown in Fig. 3. We present the results for a true, simulated treatment effect of $\beta_Z = \log(3)$ and a censoring proportion of 10%. Results for smaller treatment effects can be found in the supplementary material (see Fig. S5 and Fig. S6). Moreover, negative values of α_U and β_U (and combinations) are considered in Fig. S8 - Fig. S10. Note, that all these scenarios cover the case when the PS model is misspecified as a relevant confounder is omitted. In addition, there are no (omitted) prognostic factors simulated in this scenario. In the first column, we again assume that an independent confounder has been omitted ($\rho_{UX} = 0$). As in the first simulation (Section "Omitting a prognostic factor – Detecting induced built-in selection bias"), we observe a more visible systematic shift in the trajectory of the treatment effects while the influence of U on the time-to-event outcome increases. Moreover, the systematic shift can be observed more clearly when the omitted confounder is strongly associated with treatment allocation (see the first column of Fig. 3A compared to first column of Fig. 3B and C). In other words, *Dynamic Landmarking* better detects confounding bias if the association with the treatment allocation is strong (i.e., $|\alpha_u| \gg 0$). The SSQ_{zDiff}-trajectories behave in an expected way, i.e., achieving extremely high values at the beginning of *Dynamic Landmarking*. Referring to the formula of the z-differences, we would expect that w.l.o.g. $\overline{x}_T > \overline{x}_C$ or $\widehat{p}_T > \widehat{p}_C$ respectively. It follows, that $z_{con} > 0$ (or $z_{bin} > 0$ reps.) and consequently large initial values of SSQ_{zDiff} are observed at the beginning of *Dynamic Landmarking*, that is, before the first deletion step.

When adding a correlation between U and X, we find that the estimated treatment effects becomes closer to the true, simulated treatment effect, the stronger the correlation. In addition, the SSQ_{zDiff} come closer to being balanced after PS matching as correlation increases. This is because the omitted covariate U will be matched in parallel with the true confounder X, if U and X are correlated [e.g., 33, 37].

Illustration of the procedure with an example from cardiac surgery

We now apply the *Dynamic Landmarking* approach to individual patient data from a non-randomized trial on aortic valve implantation in cardiac surgery [12]. Here, the effect of transcatheter (either transapical (TA) or transfemoral (TF)) aortic valve implantation (TAVI) in comparison to a conventional surgical treatment (minimally invasive aortic valve replacement (MIC-AVR)) in patients with moderate surgical risk was investigated. In the original analysis, the authors used 23 baseline covariates and a 1:1:1 PS-matching algorithm for the three treatments TA-TAVI, TF-TAVI, and MIC-AVR to evaluate treatment effects by fitting stratified Cox models to the matched data set. For our investigation here, we will concentrate on the two-group comparison of MIC-AVR vs. TA-TAVI. Comparing a catheter-based intervention versus a surgical approach is of special methodological interest, because the treatments are applied to distinctly different patient populations. Unlike surgical interventions, catheter-based aortic valve implantation does not require opening the chest (sternotomy), making it suitable for much more medically compromised patients, often referred to as "high-risk patients". For this reason, strong confounding is to be expected. Indeed, in the original analysis we already noted that the overlap of the logit-transformed PS is very small before PS matching and covariates are heavily imbalanced between intervention groups. Additionally, a univariable Cox model with treatment as the only covariate and overall survival as outcome, showed an extremely strong effect of a hazard ratio of 6.40 (95%CI: 5.33; 7.69) for the MIC-AVR group in comparison to the TA-TAVI group. After PS matching with 13 randomly selected covariates (see Table 1 for details) the hazard ratio reduced to 2.13 (95%CI 1.31; 3.45) indicating a strong influence of confounding in the crude model. Moreover, considering all 23 covariates from the original article yielded a hazard ratio of 1.64 (95%CI: 1.23; 2.19).

Given this strong degree of confounding, we use the dataset for illustrative purposes and assess it in three different ways. First, a raw model (without any prior PS-matching or any other confounder adjustment) was fitted to the data set, which means that we omitted all 28 covariates from data analysis. Second, a partially PS-matched data set with 13 (out of 28) randomly included covariates was used for *Dynamic Landmarking*. Hence, 15 randomly selected covariates were omitted from data analysis. We assessed whether the selected covariates for PS matching have an influence on the results and

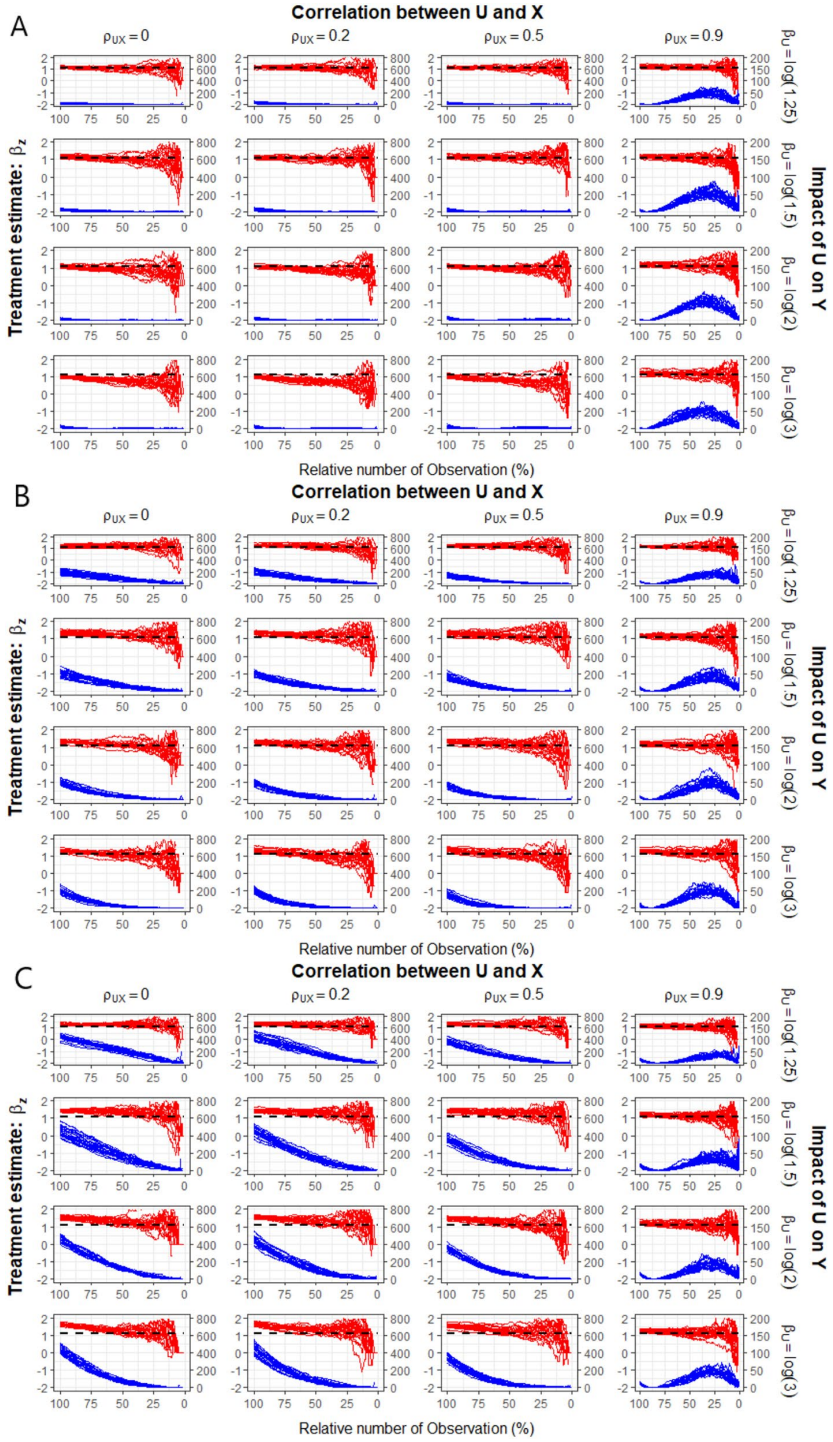

Fig. 3 Trajectories of treatment effect (left y-axis, red) on the log(HR) scale and sum of squared z-differences (right y-axis, blue) for balance measuring of the omitted covariate U for 500 simulated data sets. Dashed black lines show the true, conditional treatment estimate $\beta_z = \log(3)$. All scenarios assume the omission of a true confounder U with **A**: low impact on treatment allocation, i.e., $\alpha_u = \log(1.25)$**B**: moderate impact on treatment allocation, i.e., $\alpha_U = \log(2)$. **C**: high impact on treatment allocation, i.e. $\alpha_U = \log(3)$

Table 1 Notation for scenarios

Scenario	Description	Matched covariates	Notation scenario	Notation SSQ_{zDiff}
I	No (0) covariates are included in the PS model (raw analysis without any PS-matching); 28 covariates are omitted from data analysis	-	0/28-scenario	- /SSQ_{zDiff}(28)
II	13 covariates are included in the PS model; 15 covariates are omitted from the data analysis	Gender, weight, euroSCORE II, German Aortic valve score, STS score, Hypertension, pulmonary hypertension, Stroke, PAOD, Cerebrovascular disease, Atrial fibrillation, Previous MI, NYHA class	13/15-scenario	SSQ_{zDiff}(13) /SSQ_{zDiff}(15)
III	23 covariates are included in the PS model; 5 covariates are omitted from the data analysis	Covariates from scenario II, Age, year of surgery, height, LVEF, GFR, Previous aortic valve surgery, DM, COPD, CAD, priority urgent	23/5-scenario	SSQ_{zDiff}(23) /SSQ_{zDiff}(5)

therefore repeated the partially matching various times using different sets of randomly selected/omitted covariates. All scenarios showed similar results regarding the trajectories of *Dynamic Landmarking*; therefore, we present only one representative example in the paper (chosen covariates can be found in Table 1). In a third scenario, we reproduced the PS matching analysis from the original publication, including the 23 original and omitting the remaining five covariates. For all scenarios we used greedy nearest neighbour procedure with a caliper of width, equal to 0.2 of the standard deviation of the logit of the propensity score. Actually, the idea of *Dynamic Landmarking* is to measure the balance of omitted covariates; however, for a real data set it is also important to check the balance of the PS matched covariates. Therefore, we present the SSQ_{zDiff} in Section "Patients' characteristics before and after PS matching" for both, included and omitted covariates. For better clarity, we introduce a special notation to separate included and omitted covariates for each scenario: An x/y-scenario describes a scenario were 'x' covariates are included in the PS model and 'y' covariates are omitted from the data analysis but are used for balance measuring during *Dynamic Landmarking*. Analogously, $SSQ_{zDiff}(x)/SSQ_{zDiff}(y)$ describes the sum of squared z-differences for the ('x' included)/('y' omitted) covariates. Table 1 summarizes the three scenarios.

Patients' characteristics before and after PS matching

Table 2 summarizes the preoperative patient characteristics for each scenario. Unsurprisingly, most of the characteristics are extremely imbalanced without PS matching (0/28-scenario), as both groups strongly differ in their baseline characteristics (SSQ_{zDiff}: - / 6,538.44). In the 13/15-scenario, 240 pairs could be matched based on the following covariates: gender, weight, euroSCORE II, German aortic valve score, STS score, hypertension, pulmonary hypertension, stroke, PAOD, cerebrovascular disease, atrial fibrillation, previous MI, and NYHA class. Interestingly, the 13/15-scenario improved the balance of both, the included and omitted covariates (SSQ_{zDiff}: 62.20 / 476.63); however, the balance of the included covariates is still unsatisfactory, as the expected value for a perfect matching would be 6.5 for 13 matched covariates [20]. In the 23/5-scenario we utilized the same covariates as in the 13/15-scenario and additionally included age, year of surgery, height, LVEF, GFR, previous aortic valve surgery, diabetes mellitus, COPD, CAD, and priority status as covariates in the PS model. This resulted in 177 pairs hardly differing in terms of preoperative covariates and their balance (SSQ_{zDiff}: 27.14 / 4.66). It can be seen that the variables that were not used for PS matching in the 13/15- and 23/5-scenario nevertheless show a decreasing imbalance. This is due to the

Table 2 Patients' characteristics (italic numbers are matched characteristic in each scenario)

	0/28-model (*N* = 2536)			**13/15-model (*N* = 480)**			**23/5-model (*N* = 354)**		
Variable	**MIC-AVR (*n* = 1929)**	**TA-TAVI (*n* = 607)**	**z-Diff/SMD**	**MIC-AVR (*n* = 240)**	**TA-TAVI (*n* = 240)**	**z-Diff/SMD**	**MIC-AVR (*n* = 177)**	**TA-TAVI (*n* = 177)**	**z-Diff/ SMD**
Female	836 (43.3%)	328 (54.0%)	-4.62/-0.21	*133 (55.4%)*	*118 (49.2%)*	*1.37/0.13*	*88 (49.7%)*	*87 (49.2%)*	*-0.11/-0.01*
Weight	81.04 (± 16.12)	73.66 (± 16.06)	-9.86/-0.45	*76.17 (± 16.11)*	*76.68 (± 15.58)*	*0.35/0.03*	*76.47 (± 15.89)*	*77.17 (± 14.86)*	*-0.43/-0.04*
euroSCORE II	1.62 (± 1.44)	8.77 (± 8.87)	19.78/1.13	*3.87 (± 2.70)*	*6.80 (± 11.62)*	*3.80/0.33*	*5.42 (± 9.5)*	*3.58 (± 2.69)*	*2.48/0.21*
German Aortic Valve score	1.32 (± 0.73)	3.81 (± 3.38)	18.02/1.02	*2.35 (± 1.13)*	*3.40 (± 4.53)*	*3.48/0.31*	*3.26 (± 4.71)*	*2.32 (± 1.19)*	*2.59/0.28*
STS score	1.84 (± 1.37)	7.56 (± 5.89)	23.73/1.34	*4.01 (± 2.17)*	*5.81 (± 7.25)*	*3.69/0.31*	*5.49 (± 7.46)*	*3.97 (± 2.41)*	*2.58/0.25*
Hypertension	1447 (75.0%)	549 (90.4%)	-9.90/-0.42	*217 (90.4%)*	*213 (88.8%)*	*1.15/0.06*	*156 (88.1%)*	*157 (88.7%)*	*0.17/0.02*
Pulmonary hypertension	177 (9.2%)	202 (33.3%)	-11.9/-0.61	*56 (23.3%)*	*56 (25.3%)*	*0.00/0.00*	*42 (23.7%)*	*42 (23.7%)*	*0.00/0.00*
Stroke	37 (1.9%)	51 (8.4%)	-5.55/-0.30	*18 (7.5%)*	*23 (9.6%)*	*-1.56/-0.03*	*9 (5.1%)*	*11 (6.2%)*	*0.46/-0.04*
PAOD	60 (3.1%)	193 (31.8%)	-14.85/-0.81	*38 (15.8%)*	*40 (16.7%)*	*-0.48/-0.02*	*30 (16.9%)*	*26 (14.7%)*	*-0.58/-0.05*
Cerebrovascular disease	89 (4.6%)	140 (23.1%)	-10.39/-0.55	*36 (15.0%)*	*27 (11.3%)*	*2.47/0.09*	*22 (12.4%)*	*30 (16.9%)*	*1.20/0.11*
Atrial fibrillation	36 (1.9%)	167 (27.5%)	-13.95/-0.78	*26 (10.8%)*	*33 (13.8%)*	*-1.86/0.07*	*22 (12.4%)*	*20 (11.3%)*	*-0.33/0.03*
Previous MI	58 (3.0%)	100 (16.5%)	-8.66/-0.46	*12 (5.0%)*	*14 (5.8%)*	*-0.78/-0.06*	*15 (8.5%)*	*15 (8.5%)*	*0.00/0.00*
NYHA class	219 (11.3%)	20 (3.3%)	-14.34/0.47	*6 (2.5%)*	*12 (5.0%)*	*-0.72/0.04*	*12 (6.8%)*	*6 (3.4%)*	*-0.44/0.01*
I	983 (51.0%)	174 (28.7%)		*97 (40.4%)*	*79 (32.9%)*		*58 (32.8%)*	*73 (41.2%)*	
II	700 (36.3%)	345 (56.8%)		*119 (49.6%)*	*131 (54.6%)*		*97 (54.8%)*	*86 (48.6%)*	
III	27 (1.4%)	68 (11.2%)		*18 (7.5%)*	*18 (7.5%)*		*10 (5.6%)*	*12 (6.8%)*	
IV									
Age	67.85 (± 10.98)	81.28 (± 6.08)	38.24/1.51	76.78 (± 6.42)	80.59 (± 6.07)	6.68/0.61	*79.38 (± 6.46)*	*78.29 (± 5.53)*	*1.71/0.18*
Year of surgery	74 (3.8%)	16 (2.6%)	0.10/0.03	4 (1.6%)	9 (3.8%)	0.32/0.18	*7 (4.0%)*	*7 (4.0%)*	*-0.81/0.09*
2009	146 (7.6%)	41 (6.8%)		22 (9.2%)	11 (4.6%)		*7 (4.0%)*	*18 (10.2%)*	
2010	168 (8.7%)	49 (8.1%)		23 (9.6%)	20 (8.3%)		*15 (8.5%)*	*14 (7.9%)*	
2011	218 (11.3%)	76 (12.5%)		27 (11.3%)	28 (11.7%)		*24 (13.6%)*	*19 (10.7%)*	
2012	273 (14.2%)	97 (16.0%)		29 (12.1%)	42 (17.5%)		*30 (16.9%)*	*27 (15.3%)*	
2013	352 (18.3%)	113 (18.6%)		51 (21.3%)	48 (20.0%)		*29(16.4%)*	*37 (20.9%)*	
2014	323 (16.7%)	121 (19.9%)		38 (15.8%)	56 (23.3%)		*43 (24.3%)*	*29 (16.4%)*	
2015	236 (12.2%)	53 (8.7%)		31 (12.9%)	12 (5.0%)		*12 (6.8%)*	*17 (9.6%)*	
2016	139 (7.2%)	41 (6.8%)		15 (6.3%)	14 (5.8%)		*10 (5.6%)*	*9 (5.1%)*	
2017									
Height	170.53 (± 9.51)	165.49 (± 9.45)	-11.45/-0.53	166.75 (± 9.43)	167.29 (± 9.67)	0.61/0.06	*166.98 (± 10.07)*	*167.47 (± 8.96)*	*-0.49/-0.05*
LVEF	60.94 (± 9.29)	51.25 (± 12.16)	-18.03/-089	58.01 (± 10.23)	53.83 (± 11.42)	-4.22/-0.39	*55.95 (± 9.93)*	*56.15 (± 10.78)*	*-0.18/-0.02*
GFR	78.74 (± 20.25)	55.83 (± 22.81)	-22.12/-1.06	60.45 (± 23.16)	64.64 (± 20.81)	2.09/0.19	*63.78 (± 22.63)*	*64.77 (± 23.43)*	*-0.41/-0.04*
Previous aortic valve surgery	1 (0.1%)	13 (2.1%)	-3.54/-0.20	1 (0.4%)	3 (1.3%)	-1.76/-0.09	*1 (0.5%)*	*1 (0.5%)*	*0.00/0.00*
Diabetes mellitus	362 (18.8%)	214 (35.3%)	-7.73/-0.38	72 (30.0%)	61 (25.4%)	2.23/0.10	*53 (29.9%)*	*50 (28.2%)*	*-0.35/-0.04*
COPD	88 (4.6%)	105 (17.3%)	-7.93/-0.47	34 (14.2%)	27 (11.3%)	1.93/0.09	*21 (11.8%)*	*21 (11.8%)*	*0.00/0.00*
CAD	171 (8.9%)	99 (16.3%)	-25.94/-0.22	37 (15.4%)	46 (19.2%)	-6.66/-0.10	*27 (15.3%)*	*32 (18.1%)*	*-0.67/-0.08*
1-vessel	75 (3.9%)	83 (13.7%)		17 (7.1%)	29 (12.1%)		*20 (11.3%)*	*20 (11.3%)*	
2-vessel	46 (2.4%)	214 (35.3%)		13 (5.4%)	57 (23.8%)		*32 (18.1%)*	*24 (13.6%)*	
3-vessel									
Priority urgent (emergency)	9 (0.5%)	14 (2.3%)	-2.93/-0.16	3 (1.3%)	8 (3.3%)	-2.70/-0.14	*5 (2.8%)*	*3 (1.7%)*	*-0.72/-0.08*
MELD-Score	7.54 (± 2.16)	8.27 (± 4.98)	-3.51/-0.27	9.41 (± 3.77)	9.75 (± 4.53)	-0.87/-0.16	9.69 (± 4.36)	9.05 (± 3.37)	1.50/0.23
Diameter of aortic valve	23.47 (± 1.89)	25.88 (± 2.07)	-25.52/-1.72	22.79 (± 1.81)	26.02 (± 2.06)	-18.24/-2.36	26.04 (± 2.07)	25.89 (± 1.83)	0.06/0.11
Drainage quantity	420.83 (± 328.31)	486.38 (± 429.62)	-2.56/-0.24	458.30 (± 391.12)	489.45 (± 415.36)	-0.84/-0.11	462.20 (± 357.03)	471.13 (± 366.32)	-0.23/-0.03

Table 2 (continued)

	0/28-model (*N* = 2536)			13/15-model (*N* = 480)			23/5-model (*N* = 354)		
Variable	**MIC-AVR (*n* = 1929)**	**TA-TAVI (*n* = 607)**	**z-Diff/SMD**	**MIC-AVR (*n* = 240)**	**TA-TAVI (*n* = 240)**	**z-Diff/SMD**	**MIC-AVR (*n* = 177)**	**TA-TAVI (*n* = 177)**	**z-Diff/SMD**
preoperative haemoglobin level	13.77 (± 1.51)	12.26 (± 1.69)	7.27/1.33	12.80 (± 1.77)	12.46 (± 1.78)	2.06/0.27	12.5 (± 1.72)	12.77 (± 1.71)	-1.47/-0.22
preoperative creatinine level	0.99 (± 0.49)	1.45 (± 1.08)	-10.92/-0.78	1.34 (± 0.99)	1.14 (± 0.47)	2.74/0.36	1.24 (± 0.85)	1.20 (± 0.85)	0.44/0.07

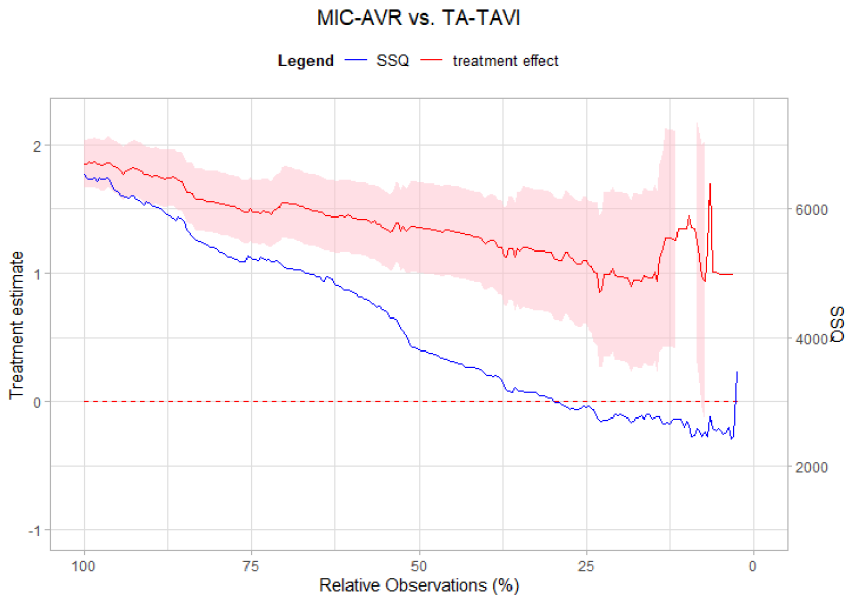

Fig. 4 Trajectories of treatment effect (left y-axis, red) on the log(HR) scale and sum of squared z-differences (right y-axis, blue) for Scenario I (0/28)

anticipated association between included and omitted covariates, which results in a parallel matching also for the omitted covariates.

Dynamic Landmarking for scenario I (0/28)

In the first scenario, we applied *Dynamic Landmarking* for the raw model without performing any PS matching prior to fitting a univariable Cox model with treatment as the only covariate. The results can be found in Fig. 4. Not surprisingly, we observe a consistently shifting treatment effect trajectory. Upon analysing the balance of the 28 omitted covariates, we notice the very high initial values of SSQ_{zDiff} (concrete: 6,538.44). Consequently, *Dynamic Landmarking* indicates that these omitted covariates might induce confounding bias. This results in a biased treatment effect estimate for this model (expressed as a hazard ratio of 6.40) due to confounding. One approach to rectify this bias would be to employ a PS model, taking into account the omitted covariates, before fitting the stratified Cox model.

Dynamic Landmarking for Scenario II (13/15)

After PS matching with 13 covariates, we applied the *Dynamic Landmarking* approach and collected the regression parameters to draw a trajectory depending on the remaining number of observations (see Fig. 5). We still observe a systematic shift in the treatment effect estimates, at least for the first 50% of deleted patients, and correspondingly a decreasing SSQ_{zDiff} during the procedure. Therefore, as expected from the simulation results, a still biased treatment effect estimate is obtained in the 13/15-scenario, pointing to confounding bias which is induced by the 15 omitted covariates. We further observe that the omitted 15 covariates also improve their balance after PS matching, indicating that included and omitted covariates are correlated. However, this correlation does not appear to be strong enough to obtain a treatment effect that is not influenced by confounding bias. Consequently, the user either needs to adjust the Cox model for the omitted confounders or must include them in the initial PS-matching. *Dynamic Landmarking* should be repeated for the enlarged confounder set to

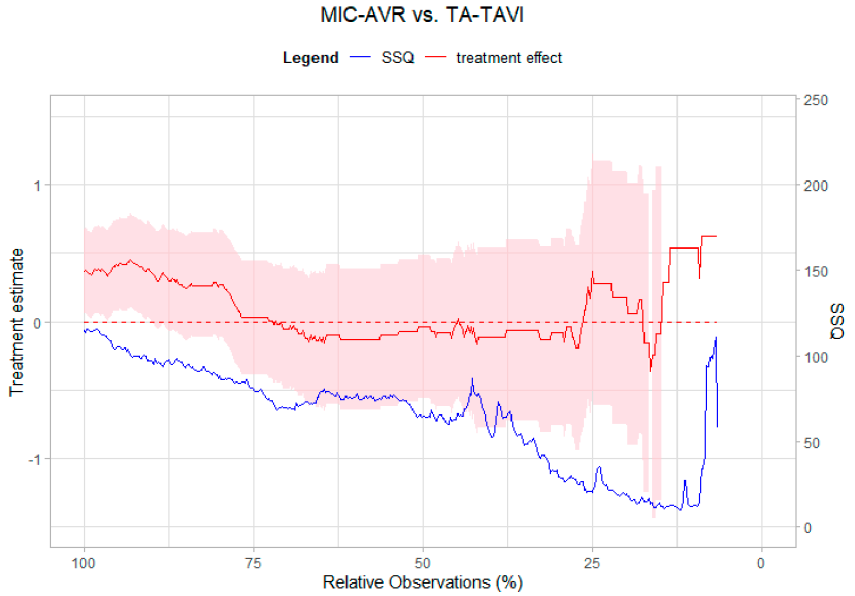

Fig. 5 Trajectories of treatment effect (left y-axis, red) on the log(HR) scale and sum of squared z-differences (right y-axis, blue) for Scenario II (13/15)

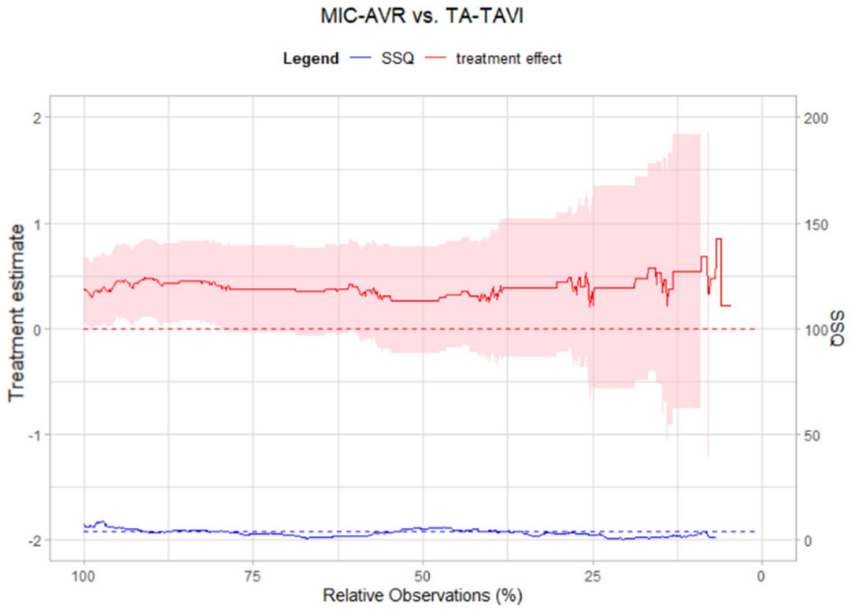

Fig. 6 Trajectories of treatment effect (left y-axis, red) on the log(HR) scale and sum of squared z-differences (right y-axis, blue) for Scenario III (23/5)

check whether the treatment effect estimate is still influenced by confounding or built-in selection bias.

Dynamic Landmarking for scenario III (23/5)

In the last scenario, all original 23 confounders were included as covariates in the PS model. *Dynamic Landmarking* shows a treatment effect trajectory with only random fluctuations and no systematic change in the SSQ_{zDiff}-trajectory (see Fig. 6) in this data set. For balance fitting we used five additional covariates (MELD-Score, diameter of aortic valve, drainage quantity, haemoglobin and, creatinine level) which were measured during the trial, but not included in the original analysis by Furukawa (2018). We observe balanced covariates

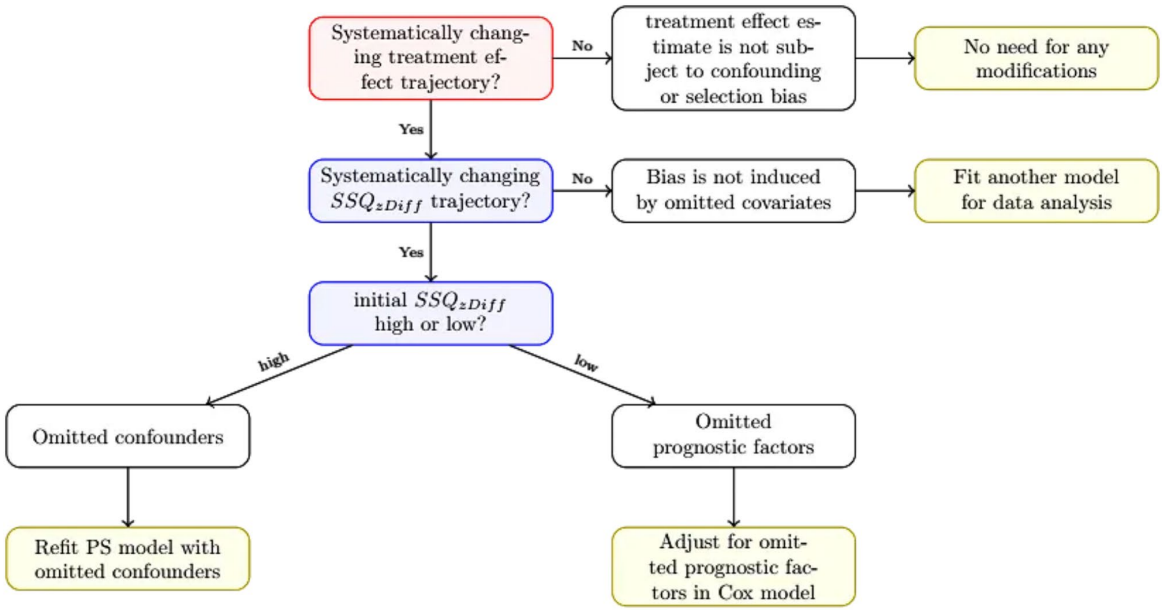

Fig. 7 Interpretation and recommendation for Dynamic Landmarking results under the assumption of uncorrelated omitted covariates. Red boxes are related to treatment effect trajectories, blue boxes are related to SSQ_{zDiff}-trajectories. Grey boxes give possible interpretations for course of trajectories and green boxes are recommendations for data analysis

during the whole *Dynamic Landmarking* process, which indicates that these five covariates do not have a relevant impact on the treatment effect estimate. To summarize, we would conclude that the estimated treatment effect in the 23/5-scenario might not be subject to confounding or built-in selection bias, as no systematic shift in the treatment effect estimate can be observed.

Discussion

Dynamic Landmarking can be used in PS matched analysis as a post-hoc diagnosing tool to visualize if the estimated treatment effects from a Cox model thread to confounding or built-in selection bias. Furthermore, the approach can give a hint on whether prognostic factors or confounders have been omitted from the data analysis. Depending on the causal direction of the omitted covariate, different issues could arise. While an omitted prognostic factor would induce built-in selection bias, resulting in a difference between conditional and marginal treatment effect, the omssion of confounders would result in confounding bias. We showed by simulation that *Dynamic Landmarking* indeed is able to visualize and distiguish between both issues, at least in case of independent omitted covariates. More precisely, both built-in selesction bias and confounding bias show systematically changing treatment effect trajectories during *Dynamic Landmarking*. Furthermore, omitted confounders tend to be heavily unbalanced between the groups yielding high initial SSQ_{zDiff}- values for the full PS matched data set. On the other hand, prognostic are still balanced after PS-matching, yielding small SSQ_{zDiff}-values at the Beginning of *Dynamic Landmarking*, but showing an increasing imbalance for the first 50% of deleted observations while the procedure continues. This is what previous work also showed for RCTs [32]. Please note that, while an inspection of the initial SSQ_{zDiff}-values give a first hint on the causal direction of the omitted covariate, it is important to consider both. This is because omitted instrumental variables (i.e., $\beta_U = 0, \alpha_U \neq 0$) would show high intial SSQ_{zDiff}-values. However, in such cases the treatment effect trajectory will remain stable with only random fluctuations (see supplement, Fig. S11).

For omitted covariates, that were independent from included ones, we provide an interpretation- and decision-scheme for *Dynamic Landmarking* (see Fig. 7). We suggest to analyse the visual output of *Dynamic Landmarking* in a two-step-algorithm: First the treatment effect trajectory has to be regarded. Only if a systematic shift is observed in the treatment effect trajectory the SSQ_{zDiff}-trajectory should be involved and interpreted as mentioned. Moreover, to differentiate correctly between built-in selection and confounding bias, the user has to run the Dynamic Landmarking with each omitted covariate seperatly. Please note, that it might be possible to observe a systematically changing treatment effect tajectory, but no change in the SSQ_{zDiff}-trajectory. In such cases we would conclude, that the treatment effect still cannot be interpreted as time-invariant effect, but it

is not possible to identify omitted covariates causing this (e.g., there might be some true unobserved/unmeasured confounders or prognostic factors [17, 36] which have to be accounted for).

In case the omitted covariate(s) are correlated with one or more considered confounders from the PS model, confounder bias or built-in selection bias can be minimized [11, 14, 15]. Rubin and Thomas (1996) stated that "excluding potentially relevant variables should be done only [.] when the excluded variables are highly correlated with variables already in the propensity score model" [27]. Indeed, recent work found that replacing a highly correlated (namely, 0.8) covariate instead of the true confounder in the PS model would result in a relative bias less than 5% [14]. Due to the correlation, the omitted covariate will indirectly accounted for in the PS model. This result is reflected in the observed behaviour of the SSQ_{zDiff} trajectories: The stronger the correlation between matched confounder and omitted covariate, the more balanced is the omitted covariate – at least at the initial state of the *Dynamic Landmarking* procedure.

The primary focus of *Dynamic Landmarking* is on assessing the estimated treatment effect, which is why the treatment effect trajectories should be examined first when using this approach. Additionally, it can provide insights into omitted covariates that might need to be included in the analysis. However, the approach should not be compared or equated with variable selection methods. While variable selection aims to identify an appropriate set of covariates before data analysis [e.g. 13, 16] *Dynamic Landmarking* serves as a post-hoc tool to verify whether the model assumptions and corresponding effect estimates are valid. We believe that our approach should be viewed as a complement to, rather than a replacement for, such analyses.

By our empirical example we showed how induced confounder bias impacted both, treatment effect and SSQ_{zDiff}-values. Indeed, the omission of true confounders led to a systematically changing treatment effect trajectory and a high intial SSQ_{zDiff}- values. Additionally, it is important to note that although the omitted confounders are correlated with the matched confounders, this correlation alone is insufficient for obtaining an estimate of the treatment effect that is not subject to confounding bias, as showed in Fig. 5. In practice, one should estimate the PS again, including the omitted confounders in the PS model and check by a repeated run of *Dynamic Landmarking*, whether the estimates are still biased (results see Fig. 6). Of course, in real life the user would not intentionally induce bias by omitting confounders, but would immediately assess a well-specified PS model using *Dynamic Landmarking*. If no constant treatment effect trajectory can be obtained by our approach we would conclude, that other assumptions (e.g., real unobserved covariates or a time-dependent treatment effect) might be an explanation for the systematic shift. In that case, a more flexible model, e.g., time-dependent propensity score [35] or frailty modelling [36], may be used for data analysis.

We have to acknowledge some limitations of our work. First, *Dynamic landmarking* is based on the assumption that the conditional treatment (conditional on all relevant prognotic factors) is constant over time, implying proportional hazards in the data. If this is true, then the method is a good diagnostic tool for identifying whether a treatment estimate from the Cox model underlies confounding or built-in selection bias. In practice, however, time-dependent treatment effects may be observed. It is already known that it is not possible to distinguish between time-dependent treatment estimates (i.e. non-proportional hazards) and induced heterogeneity (built-in selection bias) [4, 10, 24]. In fact, this is also true for our method. Therefore, as with other methods, an assumption about the true effect (here, being constant over time and across the population) has to be made.

Second, the SSQ_{zDiff} is an aggregated balance measure summarizing the global balance of all omitted covariates. We showed that the intial SSQ_{zDiff} can be used to distinguish between built-in selection bias and confounding bias. We analyzed these two issue by separate simulation scenarios. In pratice, however, both issue can occur at the same time and consequently the SSQ_{zDiff} may be estimated for prognostic factors as welll as confounders and summarized in one number. It should then be noted that the z-difference of confounders dominates the value of the SSQ_{zDiff}, as it is naturally larger than the z-difference of a prognostic factor. This can complicate the interpretation of the approach in such scenarios. One way to correctly distinguish the two effects would be to separately perform *Dynamic Landmarking* for each omitted covariate.

Third, we focused here on a specific PS method (PS-matching). Generally, PS-matching has some limitations per se, which have been discussed previously in literature [18, 34] and could also be present in our work. Related to that, we believe that recent results for optimal and matching weights will lead to increasing use of PS-weighting techniques at the cost of PS-matching [22, 23]. It seems of further interest to investigate how *Dynamic Landmarking* will perform in such situations.

Conclusion

Overall and to summarize, we feel that *Dynamic Landmarking* is a good visual tool to verify if a Cox model used provides a treatment estimate that is not subject to confounding or built-in selection bias in PS matched trials. One substantial assumption for a valid interpretation of the resulting hazard ratio is that all relevant

confounders are considered and no prognostic factors is omitted. In practice, however, it will hardly be possible to efficiently collect all covariates, confounders as well as prognostic factors. While the literature suggests that PS-matching can yield valid results in the presence of omitted variables if they are correlated with the matched confounders, this assertion is applicable only in cases of exceptionally strong correlations, which are uncommon in practical scenarios [20]. Furthermore, data collection often involves gathering more variables than those used in the final analysis. The choice of covariates for PS matching and subsequent analysis relies on current scientific understanding and clinical expertise, but it is also influenced by the researcher. Consequently, there is a possibility that omitted covariates, which were measured but not considered, may introduce built-in selection bias or confounding bias. This is precisely where *Dynamic Landmarking* comes into play, providing an opportunity to examine whether (and if so, which) covariates could distort the treatment effect estimate.

Abbreviations

RCT	Randomized controlled trials
PS	Propensity score
SSQ_{zDiff}	Sum of squared z-difference
TA	Transapical
TF	Transfemoral
TAVI	Transcatheter aortic valve implantation
MIC-AVR	Minimally invasive aortic valve replacement

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02444-7.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
All authors contributed to the study conception and design. A.S., O. K. and A. W. performed material preparation and analysis. A.S. wrote the first draft of the manuscript. J.G. and S.B. were part of data collection. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. The authors declare that no funds, grants, or other support was received during the preparation of this manuscript.

Data availability
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Ethics approval was obtained from the Martin-Luther-University Halle-Wittenberg ethics committee (Number: 2023 – 128).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 June 2024 / Accepted: 13 December 2024
Published online: 21 December 2024

References

1. Aalen OO, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Anal. 2015;21(4):579–93. https://doi.org/10.1007/s10985-015-9335-y.
2. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivar Behav Res. 2011a;46(3):399–424. https://doi.org/10.1080/00273171.2011.568786.
3. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011b Mar-Apr;10(2):150 – 61. https://doi.org/10.1002/pst.433
4. Balan TA, Putter H. Nonproportional hazards and unobserved heterogeneity in clustered survival data: When can we tell the difference? Stat Med. 2019;38(18):3405–20. https://doi.org/10.1002/sim.8171.
5. Bartlett JW, Morris TP, Stensrud MJ, Daniel RM, Vansteelandt SK, Burman CF. The Hazards of Period Specific and Weighted Hazard Ratios. Stat Biopharm Res. 2020;12(4):518–9. https://doi.org/10.1080/19466315.2020.1755722.
6. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149–56. https://doi.org/10.1093/aje/kwj149.
7. Cox DR. Partial likelihood. Biometrika. 1975;62(2):269–76. https://doi.org/10.1093/biomet/62.2.269.
8. Cox DR. Regression models and life tables. J Royal Stat Soc Ser B (Methodological). 1972;34(2):187–220. http://www.jstor.org/stable/2985181.
9. Daniel R, Zhang J, Farewell D. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biom J. 2021;63(3):528–57. https://doi.org/10.1002/bimj.201900297.
10. De Neve J, Gerds TA. On the interpretation of the hazard ratio in Cox regression. Biom J. 2020;62(3):742–50. https://doi.org/10.1002/bimj.201800255.
11. Fireman B, Gruber S, Zhang Z, et al. Consequences of Depletion of Susceptibles for Hazard Ratio Estimators Based on Propensity Scores. Epidemiology. 2020;31(6):806–14. https://doi.org/10.1097/EDE.0000000000001246.
12. Furukawa N, Kuss O, Emmel E, et al. Minimally invasive versus transapical versus transfemoral aortic valve implantation: A one-to-one-to-one propensity score-matched analysis. J Thorac Cardiovasc Surg. 2018;156(5):1825–34. https://doi.org/10.1016/j.jtcvs.2018.04.104.
13. Garcia RI, Ibrahim JG, Zhu H. Variable selection in the cox regression model with covariates missing at random. Biometrics. 2010;66(1):97–104. https://doi.org/10.1111/j.1541-0420.2009.01274.x.
14. Gayat E, Resche-Rigon M, Mary JY, Porcher R. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. Pharm Stat. 2012;11(3):222–9. https://doi.org/10.1002/pst.537.
15. Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008;95(2):481–8. https://doi.org/10.1093/biomet/asn004.
16. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J. 2018;60(3):431–49. https://doi.org/10.1002/bimj.201700067.
17. Hernán MA. The hazards of hazard ratios. Epidemiology. 2010;21(1):13–5. https://doi.org/10.1097/EDE.0b013e3181c1ea43.
18. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. Political Anal. 2019;27(4):435–54. https://doi.org/10.1017/pan.2019.11.
19. Kuss O. The z-difference can be used to measure covariate balance in matched propensity score analyses. J Clin Epidemiol. 2013;66(11):1302–7. https://doi.org/10.1016/j.jclinepi.2013.06.001.
20. Kuss O, Miller M. Unknown confounders did not bias the treatment effect when improving balance of known confounders in randomized trials. J Clin Epidemiol. 2020;126:9–16. https://doi.org/10.1016/j.jclinepi.2020.06.012.
21. Kuss O, Blettner M, Börgermann J. Propensity Score: an Alternative Method of Analyzing Treatment Effects. Dtsch Arztebl Int. 2016;113(35–36):597–603. https://doi.org/10.3238/arztebl.2016.0597.

22. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. Int J Biostat. 2013;9(2):215–34. https://doi.org/10.1515/ijb-2012-0030 . Published 2013 Jul 31.
23. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Am Stat Assoc. 2018;113(521):390–400. https://doi.org/10.1080/01621459.2016.1260466.
24. Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. Lifetime Data Anal. 2013;19(3):279–96. https://doi.org/10.1007/s10985-013-9242-z.
25. Pan W, Bai H. (2015). Propensity Score Analysis Concepts and Issues.
26. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55. https://doi.org/10.1093/biomet/70.1.41.
27. Rubin DB, Thomas N. Matching using estimated propensity scores: Relating theory to practice. Biometrics. 1996;52(1):249–6. https://doi.org/10.2307/2533160.
28. Steenland K, Karnes C, Darrow L, Barry V. Attenuation of exposure-response rate ratios at higher exposures: a simulation study focusing on frailty and measurement error. Epidemiology. 2015;26(3):395–401. https://doi.org/10.1097/EDE.0000000000000259.
29. Samuelsen SO. Cox regression can be collapsible and Aalen regression can be non-collapsible. Lifetime Data Anal. 2023;29(2):403–19. https://doi.org/10.1007/s10985-022-09578-0.
30. Sjölander A, Dahlqwist E, Zetterqvist J. A Note on the Noncollapsibility of Rate Differences and Rate Ratios. Epidemiology. 2016;27(3):356–9. https://doi.org/10.1097/EDE.0000000000000433.
31. Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. Eur Heart J. 2019;40(17):1378–83. https://doi.org/10.1093/eurheartj/ehy770.
32. Strobel A, Wienke A, Kuss O. How hazardous are hazard ratios? An empirical investigation of individual patient data from 27 large randomized clinical trials. Eur J Epidemiol. 2023;38(8):859–67. https://doi.org/10.1007/s10654-023-01026-z.
33. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution–a simulation study. Am J Epidemiol. 2010;172(7):843–54. https://doi.org/10.1093/aje/kwq198.
34. Wang J. To use or not to use propensity score matching? Pharm Stat. 2021;20(1):15–24. https://doi.org/10.1002/pst.2051.
35. Wyss R, Gagne JJ, Zhao Y, Zhou EH, Major JM, Wang SV, Desai RJ, Franklin JM, Schneeweiss S, Toh S, Johnson M, Fireman B. Use of Time-Dependent Propensity Scores to Adjust Hazard Ratio Estimates in Cohort Studies with Differential Depletion of Susceptibles. Epidemiology. 2020;31(1):82–9. https://doi.org/10.1097/EDE.0000000000001107.
36. Wienke A. Frailty models in survival analysis. Boca Raton: Chapmann&Hall/CRC; 2010.
37. Zhang L, Wang Y, Schuemie MJ, Blei DM, Hripcsak G. Adjusting for indirectly measured confounding using large-scale propensity score. J Biomed Inf. 2022;134:104204. https://doi.org/10.1016/j.jbi.2022.104204.

Publisher's note

Declaration of previous attempts

(1) I declare that I have not undergone a doctoral procedure or started a doctoral program at any other university.

(2) I declare that the information I have provided is true and that I have not submitted the scientific work to any other scientific institution for the purpose of obtaining an academic degree.

Halle (Saale),

(Original signature)

Declaration of independence

I declare in lieu of an oath that I have written this thesis independently and without outside help.

All rules of good scientific practice have been observed; no sources and aids other than those indicated by me have been used and passages taken verbatim or in terms of content from the works used have been identified as such. I assure that I have not used the paid help of mediation and consulting services (PhD advisors or other persons) for the preparation of the content of this thesis. No one has directly or indirectly received monetary benefits from me for work related to the content of the submitted dissertation.

Halle (Saale),

(Original signature)

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Andreas Wienke, for his invaluable support, guidance, and encouragement. The insightful discussions, the useful advice, and the assistance with various projects and endeavors beyond the scope of my doctoral studies have all been instrumental in shaping my academic development.

I would also like to thank Prof. Oliver Kuß for his extensive input, constructive feedback, and ongoing support. His guidance and recommendations have contributed significantly to my personal and professional growth.

I am grateful to Dr. Katharina Hennig for her commitment during this time. Her contributions have been essential in advancing my scientific career, providing valuable input on various projects, and offering supportive advice whenever needed.

Thank you to the entire IMEBI team for being such wonderful colleagues. I am especially grateful to Sophie Diexer for her support, thoughtful discussions and the friendship that has grown during this time.

To my family and friends I'm thankful for their encouragement and understanding throughout this journey.

Finally, my deepest gratitude goes to my partner, Clemens, who has been by my side through every challenge and triumph. Your unwavering support and belief in me have been indispensable. Without you, I may never have pursued this academic path, and I certainly would not be where I am today. Thank you for everything.