

Evaluating treatment effects on longitudinal outcomes with attrition due to death: Methods for a two-dimensional estimand with a case study in Quality of Life

Dries Reynders¹, Doranne Thomassen², Satrajit Roychoudhury³,
Cecilie Delphin Amdal^{4,5}, Jammbe Z. Musoro⁶, Willi Sauerbrei⁷,
Saskia le Cessie^{1,2,8}, Els Goetghebeur¹;

on behalf of the SISAQOL-IMI Work Package 3 members listed in Appendix

¹ Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium.

² Department of Biomedical Data Sciences, Leiden University Medical Center, P.O. box 9600, Postzone S-05-S, Leiden 2300 RC, The Netherlands.

³ Pfizer Inc, New York, NY, USA.

⁴ Research Support Services, Oslo University Hospital, Oslo, Norway.

⁵ Department of Oncology, Oslo University Hospital, Oslo, Norway.

⁶ European Organisation for Research and Treatment of Cancer (EORTC) Headquarters, Brussels, Belgium.

⁷ Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany.

⁸ Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

Abstract

When longitudinal outcomes are evaluated in mortal populations, their non-existence after death complicates the analysis and its causal interpretation. Where popular methods often merge longitudinal outcome and survival into one scale or otherwise try to circumvent the problem of mortality, some highly relevant questions require survival to be acknowledged as a unique condition. ”*What are my chances of survival*” and ”*What can I expect for my condition while still alive*” reflect the intrinsically two-dimensional outcome of survival and longitudinal outcome while-alive. We define a two-dimensional causal while-alive estimand for a point exposure and compare two methods for estimation in an observational setting. Regression-Standardization models survival and the observed longitudinal outcome before standardizing the latter to a target population weighted by its estimated survival. Alternatively, Inverse Probability of Treatment and Censoring Weighting weights the observed outcomes twice, to account for censoring and differences in baseline-case-mix. Both approaches rely on the same causal identification assumptions, but require different models to be correctly specified. With its potential to extrapolate,

Regression-Standardization is more efficient when all assumptions are met. We show finite sample performance in a simulation study and apply the methods to a case study on quality of life in oncology.

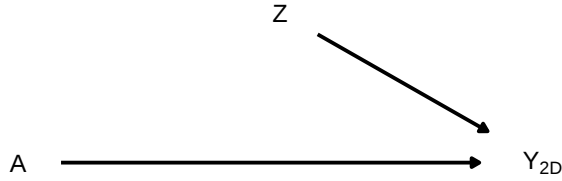
1 Introduction

In many clinical studies, the outcome of interest is measured repeatedly over time. The analysis becomes particularly complex when terminal events such as death occur during the study period. For example, consider the impact on Quality of Life (QoL) of an oncology treatment, the evaluation of a weight loss drug in patients with late stage diabetes, or physical and mental functioning of care home residents. The longitudinal outcomes do not exist after death, necessitating careful consideration of the non-negligible proportion of patients that die, first in the estimand and then in the analysis targeting relevant questions. Arguably, key questions for patients and caregivers are, under each possible treatment assignment: *'If I start that treatment, what are my chances to remain alive?'* and *'What can I expect regarding my quality of life while I am alive'*. These questions illustrate how the longitudinal measure and survival are inherently tied together in a two-dimensional outcome, where the former is contingent upon the latter. While we focus on death, similar reasoning may hold for other terminal events.

The estimand-framework described in the ICH-E9(R1) addendum [1] considers different strategies to address death, when estimating treatment effects in clinical trials: while-alive, composite endpoint, hypothetical and principal stratum strategies. In what follows, as shorthand notation, we will name the estimand according to the strategy applied to handle death - i.e. a while-alive estimand considers an estimand applying a while-alive strategy to handle death and similarly for the other strategies. In general, a while-alive estimand evaluates outcomes in every trial-participant as long as they are still alive. A composite endpoint estimand integrates death with the longitudinal outcome into one scale, e.g. time to QoL deterioration or death. A hypothetical estimand targets an effect in a world where death is avoided and a principal stratum estimand targets an effect on the longitudinal outcome in a latent sub-population that would remain alive under either treatment.

A while-alive estimand, summarising the outcome of interest at different times among those still alive, informs on the situation as-is, answering the key questions presented earlier. Given the intrinsic two-dimensional nature, survival probabilities are required to describe the complete outcome. Combined, we see them as the outcome following treatment and possibly influenced by other baseline covariates.

Some argue that a while-alive estimand cannot provide causal treatment effects because those still alive under the experimental treatment may be inherently different from those still alive under control [2–6]. Such selection may of course be present. It is, however, an intrinsic part of the natural course of the disease in combination with treatment choices. Moving away from this would generate bias on the joint distribution of the longitudinal and survival outcomes at different times t following either treatment assignment. To illustrate this, we present a causal Directed Acyclic Graph (DAG) of a randomized experiment with treatment assignment A , prognostic factors Z and a two-dimensional outcome Y_{2D} consisting of survival time (T) and the history of longitudinal outcomes while alive ($\bar{Y}(T)$) = $Y(t) : t \leq T$). For $Y_{2D} = (T, \bar{Y}(T))$:



This simple form does not suggest conditioning on a post baseline event (remaining alive), but rather represents the treatment impact on the two-dimensional outcome. This combined impact on the longitudinal outcome and survival is what informs patient-clinician discussions on our main questions: ‘*What are the chances of survival?*’ and ‘*if I survive, what can I expect for my condition?*’

While-alive estimands acknowledge the two-dimensional nature of the longitudinal outcome and survival and provide information on patient-experienced results. On the other hand, composite endpoint estimands ignore the very distinct nature of death, principal stratum estimands ignore patients who (would) die and hypothetical estimands, ignore that patients will die. Nevertheless, hypothetical estimands are often (implicitly) targeted by applying standard techniques like marginal effects from (Generalized) Linear Mixed Models that implicitly impute values after death[7–9] and theoretical guidance to handle truncation by death often focuses on composite[3] or principal stratum[4, 6, 10, 11] approaches.

In this article, we make a case for the primary importance of a causal while-alive estimand and present two methods to estimate it from observational data, assuming no unmeasured baseline confounding and independent censoring conditional on (time-varying) covariates. The first method applies Regression-Standardization[12, 13] starting from modelling both dimensions of the outcome on baseline covariates and then standardizing to a target population of interest. The second method builds on ‘regression conditional on being alive’ [8] and applies inverse probability weighting[13, 14] twice to obtain baseline covariate-balance between treatment groups and to account for censoring.

We start by introducing notation and formally define a causal two-dimensional while-alive estimand for a target population of interest in Section 2. In Section 3 we develop two corresponding estimators for an observational setting. The main ideas are developed by targeting a single timepoint and then extended to a fully longitudinal analysis. In Section 4 a small simulation study supports the methods’ performance in finite samples. The methods are then applied in a case study in Section 5, where the effect of an anticancer drug on QoL is evaluated by comparing single arm trial data with an external control.

2 Preliminaries

2.1 Notation

Consider two samples S_1 and S_0 of patients taking the treatment of interest and standard of care, respectively. For $i \in S_1 \cup S_0$ we define random variables:

- A_i : treatment-indicator. 0 (standard of care) / 1 (experimental treatment) prescribed or otherwise assigned at time $t = 0$
- T_i : time to death, measured from a well-defined time origin $t = 0$ (e.g. disease onset or study enrollment)

- $Y_i(t)$: continuous outcome of interest at time t where measurements are scheduled to be taken at pre-planned time points $t_j, j = 1, \dots, J$ prior to T_i . After T_i , Y_i is no longer defined.
- $\overline{\mathbf{Y}}_i(t)$: the vector of repeated measures $Y_i(t_j)$ for times $t_j \leq t$
- C_i : time to censoring, measured on the same time scale as T_i
- $X_i = \min(T_i, C_i)$, the observed follow-up time for survival
- δ_i : failure indicator 0 ($C_i < T_i$) / 1 ($C_i \geq T_i$)
- $\mathbf{Z}_i$: a vector of baseline covariates potentially including outcome at baseline

Consider further

- t^* : fixed timepoint of interest
- $\mathcal{P}^*$: target population, the population for which we want to estimate the average treatment effect
- S^* : a representative sample of $\mathcal{P}^*$
- $\mathcal{P}_a$: population for which S_a is representative
- $E_{\mathcal{P}}$ and $f_{\mathcal{P}}$ the population $\mathcal{P}$ -specific expected value and density, respectively

To formalize causal effects we introduce potential outcomes [15] under possible treatments $a = 0, 1$ (assuming consistency and no interference as defined in Section 3):

$$(T_i^a, \overline{\mathbf{Y}}_i^a(T_i^a))$$

individual i 's set of potential outcomes under either treatment assignment $a = 0, 1$ of which only $a = a_i$ is observed.

2.2 While-Alive Estimand

We define the components of our two-dimensional causal while-alive estimand as follows:

1. Population: target population $\mathcal{P}^*$ (e.g. corresponding to the treated population $\mathcal{P}_1$)
2. Treatment: e.g. experimental treatment as defined ($A = 1$) vs. standard of care ($A = 0$)
3. Outcome: continuous outcome of interest $Y(t)$ and survival time T , possibly at or up to a chosen time t^*
4. intercurrent events: treatment-policy for all intercurrent events apart from death. We are interested in the outcome regardless of whether patients stopped the treatment or had disease progression.[1]

5. Population Summary: the difference in expected value of $Y(t^*)$ among those still alive at t^* in $\mathcal{P}^*$ under the respective treatments as the first dimension, and difference in survival probabilities as the second dimension:

$$E_{\mathcal{P}^*}(Y^1(t^*)|T^1 > t^*) - E_{\mathcal{P}^*}(Y^0(t^*)|T^0 > t^*) \quad (1)$$

$$P_{\mathcal{P}^*}(T^1 > t^*) - P_{\mathcal{P}^*}(T^0 > t^*) \quad (2)$$

3 Estimation

First focusing on a fixed time t^* , we start from the sample as described in Section 2, without missing observations on $Y_i(t^*)$ except after censoring (hence at times t^* with $T_i > t^* > X_i$ when $\delta_i = 0$). To estimate the causal effect in target population $\mathcal{P}^*$, we require the following causal identification assumptions[15]:

A1. Consistency: $(T^a, \bar{\mathbf{Y}}^a(T^a)) = (T, \bar{\mathbf{Y}}(T))$ when $A = a$

A2. No Interference $(T_i^a, \bar{\mathbf{Y}}_i^a(T_i^a)) \perp\!\!\!\perp A_j \forall j \neq i$

A3. Positivity: $\forall z \in \mathcal{P}^* : P(A = a|Z = z) > 0$ for $a = 0$ and $a = 1$ or as it will be used equivalently $\forall z \in \mathcal{P}^* : f_{\mathcal{P}_0}(z) > 0$ and $f_{\mathcal{P}_1}(z) > 0$

A4. Conditional Exchangeability: $(T^a, \bar{\mathbf{Y}}^a(T^a)) \perp\!\!\!\perp A|\mathbf{Z}$

Under these assumptions, the treatment-specific terms of the while-alive estimand in Equation 1 at a specific time t^* can be estimated from the $\mathbf{Z}$ -specific expected Y -while-alive averaged over the $\mathbf{Z}$ -distribution of those still alive. The latter being a combination of the $\mathbf{Z}$ -distribution at baseline and their survival probabilities.

$$\begin{aligned} & E_{\mathcal{P}^*}(Y^a(t^*)|T^a > t^*) \\ &= \int_{\mathcal{P}^*} E_{\mathcal{P}^*}(Y^a(t^*)|\mathbf{Z} = \mathbf{z}, T^a > t^*) \cdot f_{\mathcal{P}^*}(\mathbf{Z} = \mathbf{z}|T^a > t^*) \cdot d\mathbf{z} \end{aligned} \quad (3)$$

$$= \frac{\int_{\mathcal{P}^*} E(Y(t^*)|A = a, \mathbf{Z} = \mathbf{z}, T > t^*) \cdot P(T > t|A = a, \mathbf{Z} = \mathbf{z}) \cdot f_{\mathcal{P}^*}(\mathbf{z}) \cdot d\mathbf{z}}{\int_{\mathcal{P}^*} P(T > t|A = a, \mathbf{Z} = \mathbf{z}) \cdot f_{\mathcal{P}^*}(\mathbf{z}) \cdot d\mathbf{z}} \quad (4)$$

where the first step applies the law of iterated expectations and the step from the counterfactuals in (3) to the observables in (4) rests on assumptions A1-A4 as listed above.

Figure 4 in appendix illustrates how the while-alive estimand is impacted both by $\mathbf{Z}$ -conditional Y -while-alive data and survival under both treatments. When survival differs between observed treatments, whether starting from randomized patients or merely a common target $\mathbf{Z}$ -distribution at baseline, this may naturally generate surviving populations at later times t with different baseline- $\mathbf{Z}$ -distribution under each treatment. We hence have in different alive populations to average Y -while-alive over. We consider two estimators for the estimand in Equation 1.

The first estimator (Section 3.1.1) is based on Regression-Standardization: for each group we are estimating $E(Y(t^*)|A = a, \mathbf{Z} = \mathbf{z}, T > t^*)$, the mean of Y conditional on $\mathbf{Z}$ and A in those still alive at t^* and survival probabilities $P(T > t^*|A = a, \mathbf{Z} = \mathbf{z})$, the first

two components of Equation 4. Then standardization is performed where the integral in 4 is approximated by averaging over the $\mathbf{Z}$ -distribution of the target population $\mathcal{P}^*$.

The second estimator (Section 3.1.2) employs Inverse Probability weighting twice. 1) Inverse Probability of Censoring weighting (IPCW) is used to make the weighted observed outcome data representative of those actually alive at that time. 2) To adjust for baseline confounding, Inverse Probability of Treatment weighting (IPTW) is employed to ensure each treatment group reflects the $\mathbf{Z}$ -distribution at baseline of target population $\mathcal{P}^*$.

We first focus on a specific timepoint t^* in sections 3.1 and 3.2 and then expand these approaches to the longitudinal setting in Section 3.3.

3.1 Estimation Under No censoring

To introduce the fundamentals of the approaches, we start by assuming no one is censored prior to t^* and $Y(t^*)$ is observed for all individuals still alive. Both methods rely on exchangeability, positivity, consistency and non-interference, assumptions A1-A4, as listed at the top of Section 3.

3.1.1 Regression-standardization

In the absence of censoring, estimates of $E(Y(t^*)|\mathbf{Z} = \mathbf{z}, A = a, T > t^*)$, the first factor in the RHS of Equation 4, can be readily obtained from the observed data using e.g. a (flexible) parametric mean model. The second factor, $P(T > t^*|\mathbf{Z} = \mathbf{z}, A = a)$, can be estimated using Cox regression or other time-to-event methods. The integral in Equation 4 can then either involve the known $\mathbf{Z}$ -density, or be empirically estimated from a (weighted) sample representing the target population as shown in Equation 5. There, a weighted average of estimated Y -while-alive is calculated per treatment over the random sample, with treatment-specific estimated survival probabilities as weights.

$$\begin{aligned} & \hat{E}_{\mathcal{P}^*}(Y^1(t^*)|T^1 > t^*) - \hat{E}_{\mathcal{P}^*}(Y^0(t^*)|T^0 > t^*) \\ &= \frac{\sum_{i \in S^*} \hat{E}(Y(t^*)|A = 1, \mathbf{Z} = \mathbf{z}_i, T > t^*) \cdot \hat{P}(T > t^*|A = 1, \mathbf{Z} = \mathbf{z}_i)}{\sum_{i \in S^*} \hat{P}(T > t^*|A = 1, \mathbf{Z} = \mathbf{z}_i)} \\ & \quad - \frac{\sum_{i \in S^*} \hat{E}(Y(t^*)|A = 0, \mathbf{Z} = \mathbf{z}_i, T > t^*) \cdot \hat{P}(T > t^*|A = 0, \mathbf{Z} = \mathbf{z}_i)}{\sum_{i \in S^*} \hat{P}(T > t^*|A = 0, \mathbf{Z} = \mathbf{z}_i)} \end{aligned} \quad (5)$$

3.1.2 IPT Weighting

In the absence of censoring, those still observed equal those still alive, hence we can estimate the expected Y -while-alive from those observed:

$$E\left(\frac{\sum_{i \in S_a} I(T_i > t^*) \cdot Y_i(t^*)}{\sum_{i \in S_a} I(T_i > t^*)}\right) = E_{\mathcal{P}_a}(Y(t^*)|A = a, T > t^*) \quad (6)$$

with $I(\text{FALSE}) = 0$ and $I(\text{TRUE}) = 1$.

To transport this result towards the target population at baseline, one can use IPT-Weights for each observation in each treatment group: $w_{IPTi}^a = \frac{f_{\mathcal{P}^*}(\mathbf{z}_i)}{f_{S_a}(\mathbf{z}_i)}$. The IPT-weighted average outcome among those observed, targets the estimand:

$$\hat{E}_{\mathcal{P}^*}(Y^a(t^*)|T^a > t^*) = \frac{\sum_{i \in S_a} w_{IPTi}^a \cdot I(T_i > t^*) \cdot Y_i(t^*)}{\sum_{i \in S_a} w_{IPTi}^a \cdot I(T_i > t^*)} \quad (7)$$

These weights can be estimated, using for example logistic regression, from a dataset where S_a and S^* are combined (if S^* is defined as e.g. a subset of S_a , then these observations are in the combined dataset twice. Once assigned to S_a , once to S^*). The required weight are then the odds of belonging to S^* ($w_{IPTi}^a = \frac{P(S=s^*|\mathbf{Z}=\mathbf{z}_i)}{1-P(S=s^*|\mathbf{Z}=\mathbf{z}_i)}$).

3.2 Including Censoring of survival

When participants' survival time is censored during the study, those observed may no longer be representative of all those still alive. To estimate the while-alive estimand, additional assumptions on the censoring mechanism are required:

A5. Conditional Independent Censoring for

- a. Survival: $T \perp\!\!\!\perp C | \mathbf{Z}$
- b. Longitudinal Outcome: $E(Y(t^*)|A, \mathbf{Z}, T > t^*, C > t^*) = E(Y(t^*)|A, \mathbf{Z}, T > t^*)$

3.2.1 Regression Standardization

If conditional on $\mathbf{Z}$, censoring is independent from $Y(t)$ and T , then again $P(T > t^*|A = a, \mathbf{Z} = \mathbf{z})$ can be estimated from the observed data using standard time-to-event methods and

$$E(Y(t^*)|A = a, \mathbf{Z} = \mathbf{z}, T > t^*) = E(Y(t^*)|A = a, \mathbf{Z} = \mathbf{z}, X > t^*) \quad (8)$$

The remainder of the calculation is identical to the method presented in 3.1.1. The baseline covariate set may need to be broadened beyond the treatment confounders set for this assumption to hold.

3.2.2 IPTC Weighting

To translate the sample of those observed ($X_i > t^*$) to a sample of those alive ($T_i > t^*$), inverse probability of censoring weighting can be applied. Under independent censoring conditional on Z , IPC Weights $w_{IPCi}^a(t^*) = P(C > t| \mathbf{Z} = \mathbf{z}_i, A = a)^{-1}$ transport the observed sample at t^* to represent those alive. Therefore, multiplying treatment balancing weights in Equation 7 with these IPC weights, arrives at an Inverse Probability of Treatment and Censoring Weighted (IPTCW) estimator:

$$\hat{E}_{\mathcal{P}^*}(Y^a(t^*)|T^a > t^*) = \frac{\sum_{i \in S_a} w_{IPCi}^a(t^*) \cdot w_{IPTi}^a \cdot I(T_i > t^*) \cdot y_i}{\sum_{i \in S_a} w_{IPCi}^a(t^*) \cdot w_{IPTi}^a \cdot I(T_i > t^*)} \quad (9)$$

Standard time-to-event methods like Cox-regression can be applied within each treatment-cohort, to estimate these IPC-weights by defining 'being censored' as the event of interest among those observed.

3.3 Longitudinal Analysis

So far, we focused on a specific timepoint t^* to clarify the key features of a while-alive estimand and -estimators for a treatment effect of a point exposure in an observational setting. A longitudinal, joint analysis over all timepoints is a relevant next step.

Kurland and Heagerty developed an IPC-Weighted GEE-approach, using an independence working correlation (IPC-IEE) to avoid the implicit imputation of outcomes after death[8]. Adding an IPTW-step to their approach for 'Regression Conditional on being Alive', makes for a longitudinal version of the presented time-specific IPTCW-approach (Equation 9).

The regression-standardization approach can be extended to the longitudinal setting similarly by using IEE to estimate $E(Y(t)|A = a, \mathbf{Z} = \mathbf{z}, T > t)$ at different times. The resulting estimates then serve as input for Equation 5.

3.4 Censoring dependent on time-varying covariates

Up to now, we assumed censoring to depend on baseline characteristics only. In practice, censoring may be related to post-baseline characteristics including the outcome history or intercurrent events like treatment discontinuation. If censoring is independent of residual survival time, given a set of time-varying covariates, this can be handled by including IPC-weights that incorporate this time-varying information. Hence the IPTCW approach as such remains unaltered, but more complex IPC-weights are used. The Regression-Standardization approach requires these IPTC weights to be incorporated in the regression models for asymptotically unbiased estimates for both $P_{\mathcal{P}*}(T > t^*|A = a, \mathbf{Z} = \mathbf{z})$ and $E_{\mathcal{P}*}(Y(t^*)|A = a, \mathbf{Z} = \mathbf{z}, T > t^*)$ [16].

3.5 Estimating Uncertainty

For the IPTCW-approach, robust standard errors can be used[17]. These tend to be conservative when estimating ATE. When estimating the ATT however, they may be conservative or anticonservative[18]. For regression-standardization, the survival component impedes such a direct approach, but for both methods bootstrapping can be applied.

4 Simulation

To illustrate finite sample performance of the methods presented above, we set up a simple simulation with a binary actionable treatment A , one baseline confounder Z , a longitudinal outcome $Y(t)$, survival T and censoring C . $Y(t)$ is impacted by A , Z , time and a person-specific intercept. Both T and C are impacted by A , Z and the longitudinal $Y(t)$. R code to replicate the simulation is available in supplemental material. The data generating mechanism is described in more detail in appendix section E.

In this scenario, $Y(t)$ tends to be higher for higher Z at baseline and is higher for $A = 1$ compared to $A = 0$, conditional on Z . Y tends to increase over time, with a sharper increase for $A = 1$. Further, survival is better for $A = 1$, conditional on Z and $Y(t)$. For $A = 0$, higher values of Z and $Y(t)$ are better, while higher $Y(t)$ are slightly detrimental

for $A = 1$. When $A = 0$, censoring is lower for higher $Y(t)$, while the opposite holds for $A = 1$.

As naive methods, the difference in observed average Y (*As Observed*), the difference in IPC (censoring model including $Y(t)$ as time varying covariate) weighted average Y (*IPCW only*) and difference in IPT weighted average Y (*IPT only*) as estimated with IEE, are presented.

IEE-IPTCW and IEE-Regression Standardization (*IEE-RegStand*) are presented once wrongly assuming censoring to depend only on baseline-covariates and once incorporating $Y(t)$ in the censoring model. For the regression standardization, this implies that in the second case, both the Y – and survival-model are IPC weighted.

Results are presented in Figure 1 and in Table 7. In the given scenario, IPCW-only, by failing to account for baseline confounding, over-represents lower Y values in $A = 0$, leading to a positive bias for the while-alive-treatment effect. Since the Z -distributions at baseline differ considerably between treatment groups and the impact of Z on Y is large, this positive bias is consistently high over time. IPT-only, by failing to account for the censoring, artificially improves the results for $A = 0$ (since mainly low Y -patients are censored) while doing the opposite for $A = 1$, resulting in a negative bias for the treatment effect. Early on, with limited impact of censoring, the estimates are not far off, but bias increases over time with the increasing censoring. The as-observed analysis lies in between these two extremes, starting close to IPC, but ending up closer to IPT and is also biased.

The proposed methods that account for both confounding and censoring, perform much better than the more naive methods, but only if the impact of $Y(t)$ on the censoring is properly accounted for. In this simulation, both IPTCW and Regression Standardization have very similar empirical standard deviations. In the next section, we apply these methods to a case study presented in the context of SISAQOL-IMI (Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints in Cancer Clinical Trials – IMI)[19].

5 case study: Health Related QoL in Single Arm Trial vs External control

5.1 Setting

We apply the methodology for the two-dimensional outcome to estimate the impact of an experimental treatment on Health Related QoL in a Single Arm Trial (SAT) in late stage oncology with external control. Such SATs play an increasingly important role in cancer clinical research. With RCTs sometimes deemed impractical or unethical, an increasing number of FDA and EMA approvals rest on pivotal evidence of SATs[20, 21].

In such trials, a lack of concurrent randomized control, makes comparative effectiveness conclusions challenging[22]. It is therefore helpful to carefully apply causal inference principles and draw conclusions under well understood assumptions, which may be questioned and be subject to sensitivity analyses in a second stage. The question we study below was put forward and discussed in a large consortium of stakeholders collaborating in the context of SISAQOL-IMI.

Simulation Study: Estimated ΔY While-Alive

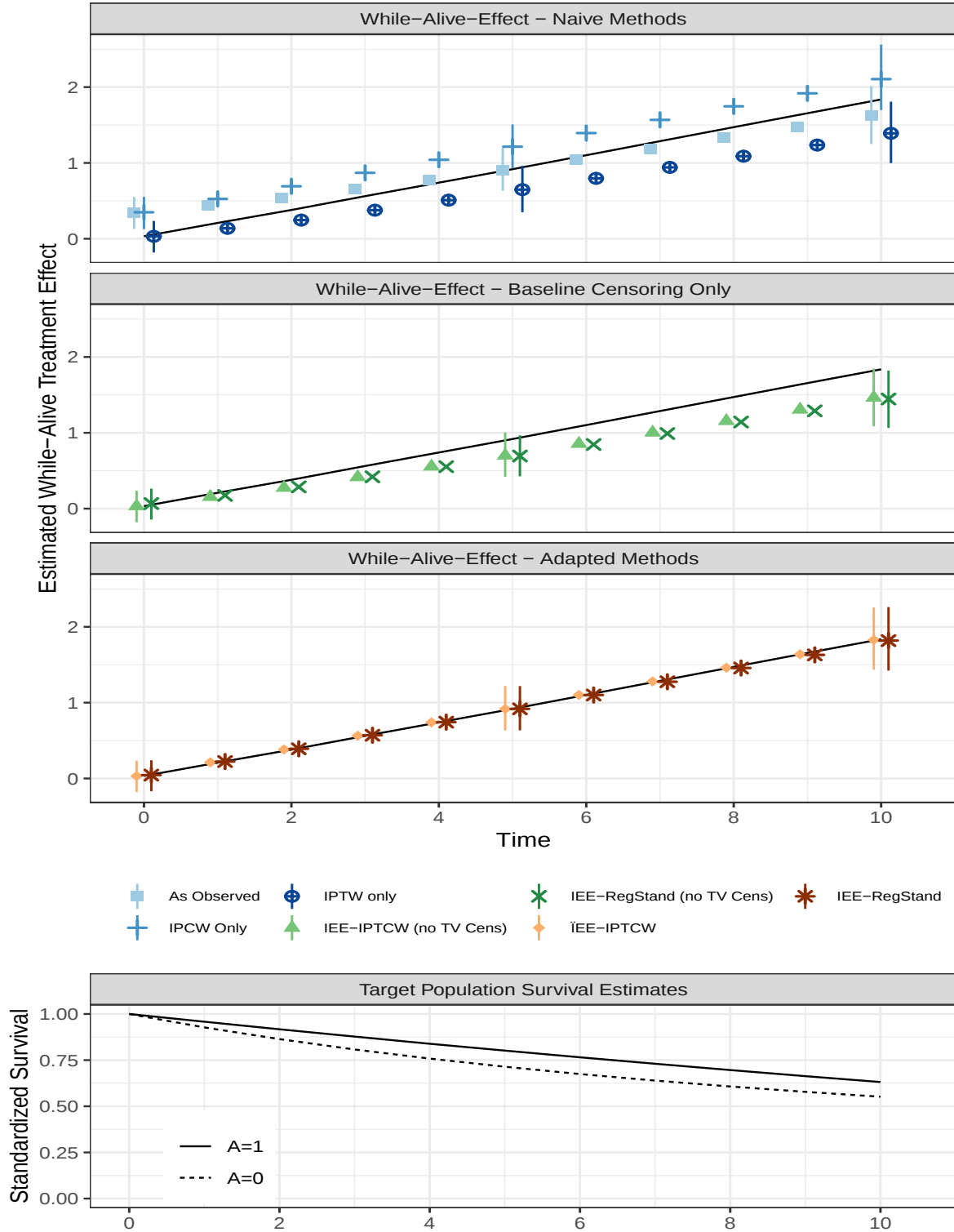


Figure 1: Mean estimated and true treatment effects over time by method (solid line represent the true values, calculated from a large N dataset). To reduce clutter, CIs are presented for selected timepoints only and are not presented for the naive approaches. The bottom panel provides target population survival estimates

5.2 Data sources

In an SAT of patients with non-small cell lung cancer[23, 24] we analyse QoL measured with the EORTC QLQ-C30 global QoL scale[25], which was a secondary outcome in the trial.

The same experimental treatment was also studied in an RCT with chemotherapy in the control arm[26]. Both trials used the EORTC QLQ-C30 questionnaire to measure QoL and the same three-weekly measurement schedule, which makes the control arm of the RCT a suitable comparator for the SAT, even though enrollment in the RCT was restricted to patients with fewer comorbidities and fewer prior lines of treatment than in the SAT. Additionally, the experimental RCT-arm can serve as a negative control[27]. With equal treatment regimen and fairly similar settings, we do not expect structural differences in outcomes with the SAT after adjustment for confounding in overlapping (sub)populations. If results contradict this, it would be a clear sign that the adjustment for confounding was insufficient or that other issues (e.g. informative drop out, differences in surrounding care) were not properly addressed.

The sponsor anonymized both datasets before a subset was shared for analysis. A single imputed version of the data is used for our purpose. QoL was imputed until time of death or censoring, using time-varying information on treatment discontinuation, progression and time to death as described in [28]. This imputed dataset has no intermittent missingness (missingness on the longitudinal outcome inbetween two observed timepoints or while still known to be alive) and hence allows the case study analysis to focus on while-alive, confounding and censoring.

5.3 Analysis

We consider QoL at baseline, sex, ECOG performance status at baseline and the number of prior treatments to be a sufficient set of baseline confounders for conditional exchangeability. For the models assuming only baseline-dependent censoring, the same variables are used for the censoring model. For the others, the observed (or imputed) QoL over time and an indicator of progression are added. Here, the treated SAT population best represents the eventually intended population to be treated and therefore is a relevant choice for the target population. However, some confounder values observed in the SAT (e.g. ECOG performance status 3), were not observed in the RCT (Table 12 in appendix). Therefore, defining the SAT-population as the target population would result in positivity issues. To ensure sufficient overlap between SAT and external control populations for the considered baseline confounders, the target population is restricted to the SAT-sub-population with $\text{ECOG} \leq 2$ and less than 3 prior drug therapies (the very limited number of patients with 3 prior therapies in the RCT would lead to unstable estimates).

Differences between the treatment groups in QoL-while-alive are estimated at three-week intervals using the methods presented in sections 3.2 and 3.3 and bootstrapping is used to derive the corresponding confidence intervals. Cox regression models are applied to create the standardized survival curves[29]. Separate Cox models are fitted per treatment and resulting predicted survival curves averaged over the same target population. This was performed both for the primary analysis comparing the experimental treatment in the SAT with the control arm in the RCT and for the sensitivity (negative control) analysis, comparing the experimental treatment groups in SAT and RCT.

5.4 Results

The results show better QoL-while-alive in the experimental treatment group in SAT vs control arm in RCT early on (Figure 2). This favorable outcome is sustained up till 30 weeks (10 cycles), while survival outcomes are similar. Results for all models are in line with each other, indicating that here the impact of the observed/imputed QoL on censoring is limited.

The small confidence interval at $t = 0$ for the timepoint specific regression-standardization in Figure 2 is an artefact due to the inclusion of the baseline QoL in the regression model to estimate itself. The timepoint is included to assess whether the methods manage to balance baseline QoL between the treatment arms. For the other timepoints, the IPTCW-methods have wider confidence intervals than the regression-standardization approaches.

In the sensitivity analysis ('negative control analyses') where the experimental SAT group was compared with the experimental RCT-arm, estimated differences in QoL while-alive are mostly close to zero with CIs overlapping zero. As such there is no strong evidence against the assumed 'no unmeasured confounders'.

6 Discussion

In this paper, we made a case for the primary importance of the two-dimensional outcome Y_{2D} with derived causal interpretation of while-alive estimands. Viewing longitudinal outcomes as inherently two-dimensional, accommodates a causal interpretation of a while-alive estimand as the total effect of the treatment in a mortal population. The developed methods can be equally applied in randomized experiments and for other terminal events, e.g. treatment discontinuation in a while-on-treatment estimand. While we started from a continuous outcome, the ideas presented apply equally to other types of outcomes and the methods are easily extended to e.g. binary or ordinal outcomes. One could also consider doubly robust versions, but this is out of scope for this article.

While-alive estimands acknowledge the separate status of death and provide insight into patients' lived experiences. This differs from alternative, more popular approaches to handle death. Combining longitudinal outcomes with death into a composite outcome, focusing on hypothetical scenarios where death doesn't occur, or looking at those who are guaranteed to survive, all result in very different estimands that either don't fully distinguish quality of life from not being alive, or target a world that does not inform all patients.

The simulation study illustrates the need for targeted estimators that come with the estimand. Ignoring treatment confounding or covariate-dependent censoring of survival, may result in considerable bias. We present an inverse probability weighting- and a Regression-Standardization approach to estimate while-alive causal effects in RCTs or observational studies. Both approaches rely on the same set of causal identification assumptions to account for confounding and dependent censoring, possibly depending on time-varying characteristics. Yet, they assume different models to be correctly specified. In the case study, both methods provided similar estimates. The smaller confidence intervals for the regression-standardization approaches, reflect their ability to exploit their modeling assumptions (e.g. linear or smooth trends, lack of interactions,...) and to extrapolate estimates to profiles that completely dropped out of the study at t^* . The proposed methods build on existing

Case Study: Estimated Δ HRQoL While–Alive

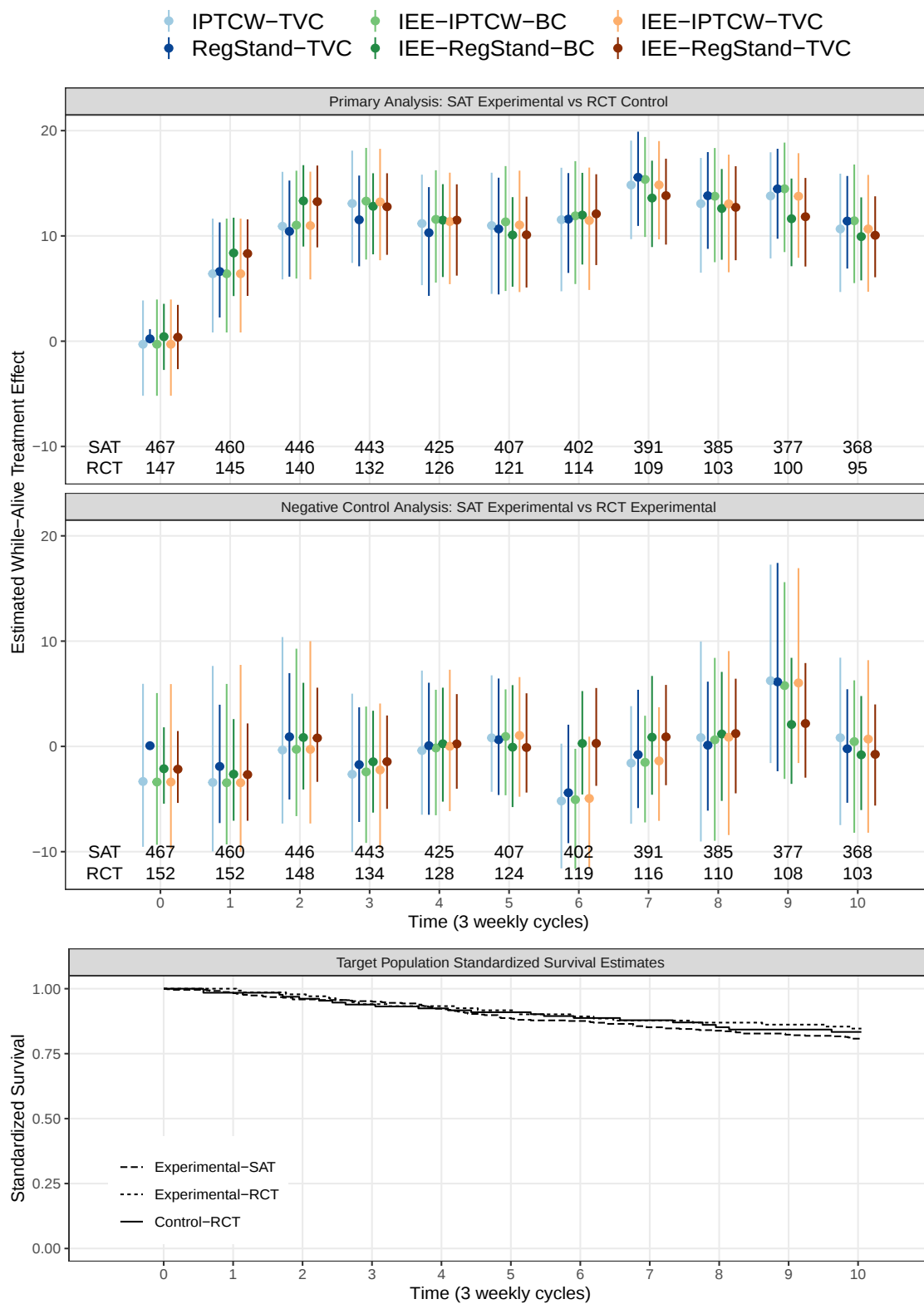


Figure 2: Estimated differences in QoL-while-alive between SAT and the respective arms of the RCT (Top panel: main analysis for efficacy. Dots are the point estimates, vertical lines represent bootstrapped 95% confidence intervals. Middle panel: negative control analysis comparing SAT with the experimental RCT-arm).

methods and hence can be implemented by combining existing software.

Our case study results show a clear positive treatment impact on average QoL-while-alive with comparable expected survival as the standard of care, under the assumptions made. These averages may cover complex mechanisms, but that also holds for clinical trials in general, that primarily target average treatment effect in mortal and non-mortal settings. If the estimated treatment effect on the two dimensions goes in opposite directions, this still reflects the tangible reality, but (shared) decision making is less straightforward. Some may prefer the prospect of improved survival, others prioritize quality of life over length of life. For individual patients, $\mathbf{Z}$ -specific estimates may provide additional insights, to help make these decisions.

As always, derived estimates of the treatment effect depend on often untestable assumptions. This ranges from the causal identification assumptions, over the censoring assumptions to the parametric assumptions implied by the fitted models. As argued by Dang et al.[30] and many others, sensitivity analyses investigating their potential impact, are an inherent part of any causal analysis. Here, the case study data lends itself perfectly for a negative control analysis. If all assumptions are reasonably met, we would not expect systematic differences between the treated in the SAT and the treated in the RCT, after adjustment for known baseline confounding and accounting for censoring. If this sensitivity analysis would show clear differences, it would point to shortcomings in the assumptions made or the study design.

On top of the causal assumptions, the censoring assumptions are a key component of the analysis. In the analysis of the case study, we assume censoring to be independent given baseline characteristics. However, in many settings censoring can rely heavily on time-varying characteristics. This information then needs to be included in the analysis, at the cost of increased complexity and more stringent positivity assumptions. In the case study, patients had very limited follow up for QoL after progression or treatment discontinuation. Although somewhat obscured by the single imputation performed prior to analysis, this means that post-progression and post-discontinuation estimates are based on limited available information, inhibiting a thorough analysis of potentially time-varying covariate-dependent censoring. When feasible, a trial design and data collection tailored to the targeted estimand, may mitigate this problem.

In addition to investigating potential biases with sensitivity analyses, providing honest estimates for the uncertainty is essential. We presented bootstrap-based confidence intervals, incorporating uncertainties in estimating the weights and the expected outcomes. It can be argued, however, that the choice of the specific external control data out of a collection of potential studies, is an additional source of heterogeneity-induced uncertainty that one would need to account for[31]. The between-study variability may result from a wide range of sources, both baseline patient mix and post-baseline features (such as surrounding care, measurement tools and schedules, adherence,...). Meta-analysis may provide an estimate for this between-study variability. In our case study, by selecting studies with similar protocols and measurement tools, some of these sources are recognized. Since our analyses explicitly adjust for case-mix, the between-study variability estimated from an unconditional meta-analysis, may be seen as an upper bound for this additional uncertainty on the estimated treatment effect.

To conclude: the difficulties arising in mortal cohorts, should not stop us from asking questions that inform all patients on what they may expect. While-alive causal estimands

are often a key component of these questions.

References

- [1] European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, 2020. URL <https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistics-en.pdf>.
- [2] Elizabeth Colantuoni, Daniel O. Scharfstein, Chenguang Wang, Mohamed D. Hashem, Andrew Leroux, Dale M. Needham, and Timothy D. Girard. Statistical methods to compare functional outcomes in randomized controlled trials with high mortality. *BMJ*, 360:j5748, 01 2018. doi: 10.1136/bmj.j5748. URL <https://www.bmj.com/content/360/bmj.j5748>. Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting PMID: 29298779.
- [3] Chenguang Wang, Daniel O. Scharfstein, Elizabeth Colantuoni, Timothy D. Girard, and Ying Yan. Inference in randomized trials with death and missingness. *Biometrics*, 73(2):431–440, 2017. doi: 10.1111/biom.12594. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12594>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12594>.
- [4] Michelle Shardell, Gregory E. Hicks, and Luigi Ferrucci. Doubly robust estimation and causal inference in longitudinal studies with dropout and truncation by death. *Biostatistics*, 16(1):155–168, 01 2015. doi: 10.1093/biostatistics/kxu032. URL <https://doi.org/10.1093/biostatistics/kxu032>.
- [5] Tyler J VanderWeele. Principal stratification — uses and limitations. *The International Journal of Biostatistics*, 7(1):28, 07 2011. doi: 10.2202/1557-4679.1329. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3154088/>. PMID: 21841939 PMCID: PMC3154088.
- [6] Brian L. Egleston, Daniel O. Scharfstein, Ellen E. Freeman, and Sheila K. West. Causal inference for non-mortality outcomes in the presence of death. *Biostatistics*, 8(3):526–545, 07 2007. doi: 10.1093/biostatistics/kxl027. URL <https://doi.org/10.1093/biostatistics/kxl027>.
- [7] Anaïs Rouanet, Catherine Helmer, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Interpretation of mixed models and marginal models with cohort attrition due to death and drop-out. *Statistical Methods in Medical Research*, 28(2):343–356, 02 2019. doi: 10.1177/0962280217723675. URL <https://doi.org/10.1177/0962280217723675>. Publisher: SAGE Publications Ltd STM.
- [8] Brenda F. Kurland and Patrick J. Heagerty. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 6(2):241–258, 04 2005. doi: 10.1093/biostatistics/kxi006. URL <https://doi.org/10.1093/biostatistics/kxi006>.

- [9] Brenda F. Kurland, Laura L. Johnson, Brian L. Egleston, and Paula H. Diehr. Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 24(2):211, 2009. doi: 10.1214/09-STS293. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2812934/>. PMID: 20119502 PMCID: PMC2812934.
- [10] Shanshan Luo, Wei Li, and Yangbo He. Causal inference with outcomes truncated by death in multiarm studies. *Biometrics*, 79(1):502–513, 2023. doi: 10.1111/biom.13554. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13554>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13554>.
- [11] Eric J. Tchetgen Tchetgen, M. Maria Glymour, Ilya Shpitser, and Jennifer Weuve. Rejoinder: To weight or not to weight?: On the relation between inverse-probability weighting and principal stratification for truncation by death. *Epidemiology*, 23(1):132, 01 2012. doi: 10.1097/EDE.0b013e31823b5081. URL https://journals.lww.com/epidem/fulltext/2012/01000/rejoinder__to_weight_or_not_to_weight__on_the.20.aspx.
- [12] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 01 1986. doi: 10.1016/0270-0255(86)90088-6. URL <http://www.sciencedirect.com/science/article/pii/0270025586900886>.
- [13] Miguel A Hernán and James M Robins. *Causal Inference: What If*. CRC Press, 2020. tex.ids: hernanCausalInferenceWhata.
- [14] James M. Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550, 09 2000. URL https://journals.lww.com/epidem/fulltext/2000/09000/marginal_structural_models_and_causal_inference_in.11.aspx.
- [15] Els Goetghebeur, Saskia le Cessie, Bianca De Stavola, Erica EM Moodie, Ingeborg Waernbaum, and “on behalf of” the topic group Causal Inference (TG7) of the STRATOS Initiative. Formulating causal questions and principled statistical answers. *Statistics in Medicine*, 39(30):4922–4948, 2020. doi: 10.1002/sim.8741. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8741>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8741>.
- [16] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995. doi: 10.2307/2291134. URL <https://www.jstor.org/stable/2291134>. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].
- [17] Peter C. Austin. Bootstrap vs asymptotic variance estimation when using propensity score weighting with continuous and binary outcomes. *Statistics in Medicine*, 41(22):4426–4443, 07 2022. doi: 10.1002/sim.9519. URL <http://dx.doi.org/10.1002/sim.9519>.

- [18] Sarah A Reifeis and Michael G Hudgens. On variance of the treatment effect in the treated when estimated by inverse probability weighting. *American Journal of Epidemiology*, 191(6):1092–1097, 05 2022. doi: 10.1093/aje/kwac014. URL <https://doi.org/10.1093/aje/kwac014>.
- [19] Madeline Pe, Ahu Alanya, Ragnhild Sorum Falk, Cecilie Delphin Amdal, Kristin Bjordal, et al. Setting international standards in analyzing patient-reported outcomes and quality of life endpoints in cancer clinical trials-innovative medicines initiative (sisaqol-imi): stakeholder views, objectives, and procedures. *The Lancet Oncology*, 24(6):e270–e283, 06 2023. doi: 10.1016/S1470-2045(23)00157-2. URL [https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045\(23\)00157-2/fulltext](https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(23)00157-2/fulltext). Publisher: Elsevier PMID: 37269858.
- [20] European Medicine Agency. Reflection paper on establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation application. 2024. URL <https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-establishing-efficacy-based-single-arm-trials-submitted-pivotal-evidence.pdf>.
- [21] Talal Hilal, Miguel Gonzalez-Velez, and Vinay Prasad. Limitations in clinical trials leading to anticancer drug approvals by the us food and drug administration. *JAMA internal medicine*, 180(8):1108–1115, 08 2020. doi: 10.1001/jamainternmed.2020.2250. PMID: 32539071 PMCID: PMC7296449.
- [22] Hans Ulrich Burger, Christoph Gerlinger, Chris Harbron, Armin Koch, Martin Posch, Justine Rochon, and Anja Schiel. The use of external controls: To what extent can it currently be recommended? *Pharmaceutical Statistics*, 20(6):1002–1016, 2021. doi: 10.1002/pst.2120. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.2120>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/pst.2120>.
- [23] Fiona Blackhall, Dong-Wan Kim, Benjamin Besse, Hiroshi Nokihara, Ji-Youn Han, Keith D. Wilner, Arlene Reisman, Shrividya Iyer, Vera Hirsh, and Alice T. Shaw. Patient-reported outcomes and quality of life in profile 1007: A randomized trial of crizotinib compared with chemotherapy in previously treated patients with alk-positive advanced non-small-cell lung cancer. *Journal of Thoracic Oncology*, 9(11):1625–1633, 11 2014. doi: 10.1097/JTO.0000000000000318. URL <https://www.sciencedirect.com/science/article/pii/S1556086415307292>.
- [24] Fiona Blackhall, D. Ross Camidge, Alice T. Shaw, Jean-Charles Soria, Benjamin J. Solomon, Tony Mok, Vera Hirsh, Pasi A. Jänne, Yuankai Shi, Pan-Chyr Yang, Tommaso De Pas, Toyoaki Hida, Javier Castro De Carpeño, Silvana Lenz, Anna Polli, Shrividya Iyer, Arlene Reisman, Keith D. Wilner, and Dong-Wan Kim. Final results of the large-scale multinational trial profile 1005: efficacy and safety of crizotinib in previously treated patients with advanced/metastatic alk-positive non-small-cell lung cancer. *ESMO Open*, 2(3):e000219, 01 2017. doi: 10.1136/esmoopen-2017-000219. URL <https://www.sciencedirect.com/science/article/pii/S205970292032425X>.

- [25] Neil K. Aaronson, Sam Ahmedzai, Bengt Bergman, Monika Bullinger, Ann Cull, Nicole J. Duez, Antonio Filiberti, Henning Flechtner, Stewart B. Fleishman, Johanna C. J. M. de Haes, Stein Kaasa, Marianne Klee, David Osoba, Darius Razavi, Peter B. Rofo, Simon Schraub, Kommer Sneeuw, Marianne Sullivan, and Fumikazu Takeda. The european organization for research and treatment of cancer qlq-c30: A quality-of-life instrument for use in international clinical trials in oncology. *JNCI: Journal of the National Cancer Institute*, 85(5):365–376, 03 1993. doi: 10.1093/jnci/85.5.365. URL <https://doi.org/10.1093/jnci/85.5.365>.
- [26] Alice T. Shaw, Dong-Wan Kim, Kazuhiko Nakagawa, Takashi Seto, Lucio Crinó, Myung-Ju Ahn, Tommaso De Pas, Benjamin Besse, Benjamin J. Solomon, Fiona Blackhall, Yi-Long Wu, Michael Thomas, Kenneth J. O’Byrne, Denis Moro-Sibilot, D. Ross Camidge, Tony Mok, Vera Hirsh, Gregory J. Riely, Shrividya Iyer, Vanessa Tassell, Anna Polli, Keith D. Wilner, and Pasi A. Jänne. Crizotinib versus chemotherapy in advanced alk-positive lung cancer. *New England Journal of Medicine*, 368(25): 2385–2394, 06 2013. doi: 10.1056/NEJMoa1214886. URL <https://www.nejm.org/doi/10.1056/NEJMoa1214886>. Publisher: Massachusetts Medical Society.
- [27] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383–388, 05 2010. doi: 10.1097/EDE.0b013e3181d61eeb. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3053408/>. PMID: 20335814 PMCID: PMC3053408.
- [28] Doranne Thomassen, Satrajit Roychoudhury, Cecilie Delphin Amdal, Dries Reynders, Jammbe Z. Musoro, Willi Sauerbrei, Els Goetghebeur, Saskia le Cessie, et al. The role of the estimand framework in the analysis of patient-reported outcomes in single-arm trials: a case study in oncology. *BMC Medical Research Methodology*, 24(1): 290, 11 2024. doi: 10.1186/s12874-024-02408-x. URL <https://doi.org/10.1186/s12874-024-02408-x>.
- [29] Per Kragh Andersen, Maja Pohar Perme, Hans C. van Houwelingen, Richard J. Cook, Pierre Joly, Torben Martinussen, Jeremy M. G. Taylor, Michal Abrahamowicz, and Terry M. Therneau. Analysis of time-to-event for observational studies: Guidance to the use of intensity models. *Statistics in Medicine*, 40(1):185–211, 2021. doi: <https://doi.org/10.1002/sim.8757>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8757>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8757>.
- [30] Lauren E. Dang, Susan Gruber, Hana Lee, Issa J. Dahabreh, Elizabeth A. Stuart, Brian D. Williamson, Richard Wyss, Iván Díaz, Debashis Ghosh, Emre Kıcıman, Demissie Alemayehu, Katherine L. Hoffman, Carla Y. Vossen, Raymond A. Huml, Henrik Ravn, Kajsa Kvist, Richard Pratley, Mei-Chiung Shih, Gene Pennello, David Martin, Salina P. Waddy, Charles E. Barr, Mouna Akacha, John B. Buse, Mark van der Laan, and Maya Petersen. A causal roadmap for generating high-quality real-world evidence. *Journal of Clinical and Translational Science*, 7(1):e212, 01 2023. doi: 10.1017/cts.2023.635. URL <https://www.cambridge.org/core/journals/journal-of-clinical-and-translational-science/>

- [31] Olivier Collignon, Anna Schritz, Stephen J Senn, and Riccardo Spezia. Clustered allocation as a way of understanding historical controls: Components of variation and regulatory considerations. *Statistical Methods in Medical Research*, 29(7):1960–1971, 07 2020. doi: 10.1177/0962280219880213. URL <https://doi.org/10.1177/0962280219880213>. Publisher: SAGE Publications Ltd STM.

Acknowledgements

We thank the wider SISAQOL-IMI consortium for many valuable discussions and the infrastructure that enabled this research.

We also thank the STRATOS Initiative for their valuable methodological input during this study, and their members’ involvement in SISAQOL-IMI Work Package 3.

This publication reflects the views of the individual authors and should not be construed to represent official views or policies of the European Medicines Agency (EMA), the US Food and Drug Administration (FDA), US National Cancer Institute (NCI), Medicines and Healthcare products Regulatory Agency (MHRA), Institute for Quality and Efficiency in Health Care (IQWiG), Health Canada, the Norwegian Medicines Agency (NOMA), the American Society of Clinical Oncology (ASCO) or the European Society for Medical Oncology (ESMO) or any other institution, organization, or entity. This publication reflects the authors’ view and neither IMI nor the EU, EFPIA are responsible for any use that may be made of the information contained therein. SISAQOL-IMI Work Package 3-members involved in the paper:

Dries Reynders¹, Doranne Thomassen², Satrajit Roychoudhury³, Cecilie Delphin Amdal^{4,5}, Jammbe Z. Musoro⁶, Willi Sauerbrei⁷, Saskia le Cessie^{1,2,8}, Els Goetghebeur¹, Ethan Basch⁹, Melanie Calvert^{10,11}, Joseph C. Cappelleri², Samantha Cruz Rivera¹⁰, Olalekan Lee Aiyegbusi^{10,11}, Geert Molenberghs¹², Khadija Rerhou Rantell¹³, Alexander Russell-Smith¹⁴, Anja Schiel¹⁵, Ragnild Sorum Falk³

⁹ American Society for Clinical Oncology, Alexandria, VA, USA

¹⁰ University of Birmingham, Birmingham, UK

¹¹ NIHR Birmingham Biomedical Research Centre, Birmingham, UK

¹² Katholieke Universiteit Leuven, Leuven, Belgium

¹³ Medicines and Healthcare products Regulatory Agency, London, UK

¹⁴ Pfizer Ltd., Sandwich, UK

¹⁵ Norwegian Medicines Agency, Oslo, Norway

A While-Alive

Figure 4 provides a graphical representation of selection through survival impacts the while-alive estimand and how censoring needs to be addressed to properly estimate it.

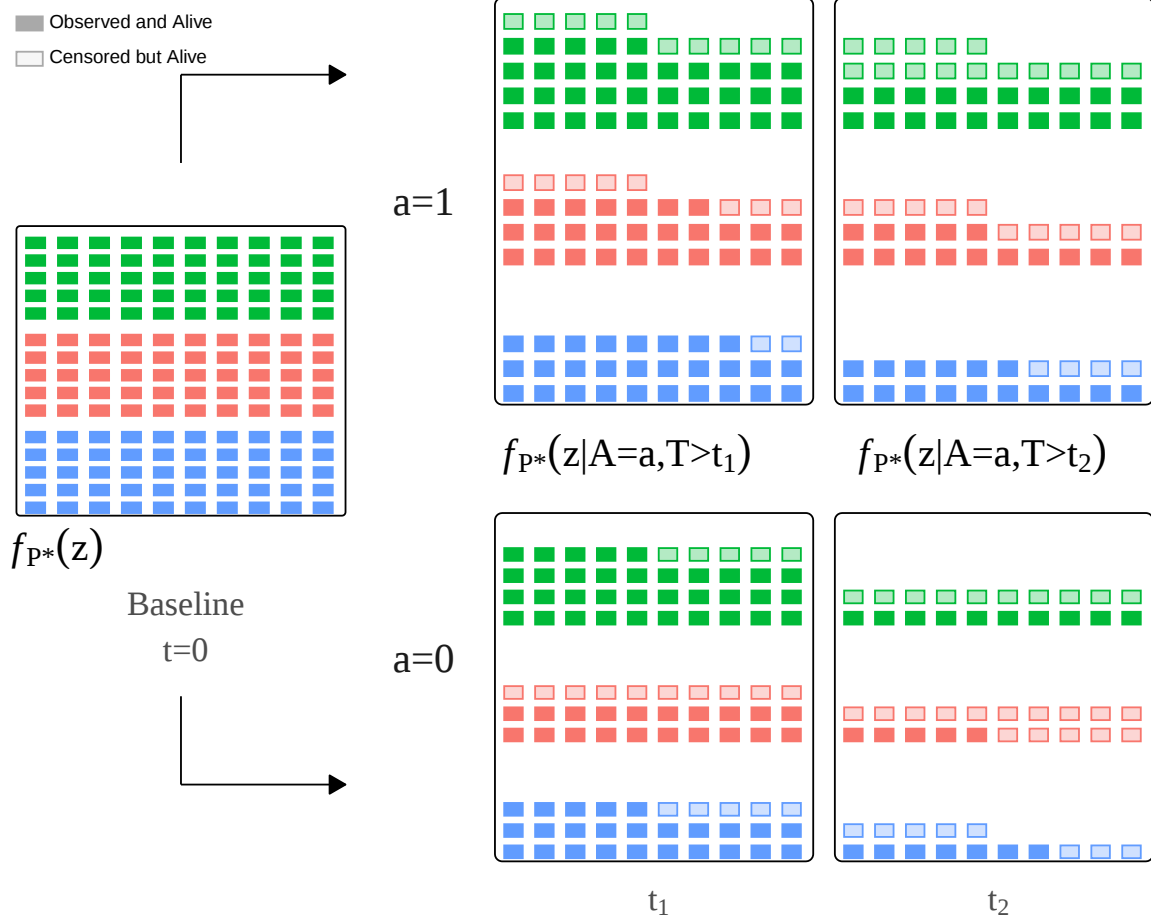


Figure 3: All blocks represent patients still alive, but shaded blocks means they are censored and hence in reality we do not know whether they are still alive or not. Colors represent different values for their baseline-confounders $\mathbf{Z}$. We can imagine that green patients tend to have better longitudinal outcomes than red patients who in turn tend to do better than blue. Observed results under both treatments $A = 0/1$, are transported to the baseline $\mathbf{Z}$ -distribution of target population $\mathcal{P}^*$. The different number of blocks of each color at times t_1 and t_2 between the treatments represents differential survival ($P(T > t^* | A = a, \mathbf{Z})$) which leads to divergent $\mathbf{Z}$ -distributions among those alive at t^* ($f_{\mathcal{P}^*}(z|A = a, T > t^*)$) under treatments $a = 0$ and $a = 1$. The value of our while-alive estimand does not only depend on (the trend of) the outcome within each color-profile, but also on the relative sizes of the green, red and blue group at each timepoint. As can be seen from the shaded blocks, due to censoring, the observed $\mathbf{Z}$ -distribution under those alive at t^* $f_{\mathcal{P}^*}(z|X > t^*)$ may differ from the actual distribution among survivors $f_{\mathcal{P}^*}(z|T > t^*)$, while it is the latter we are interested in.

B Simulation

R-code is available as supplementary material. 7 presents the empirical bias and SD for all evaluated methods.

The data generating mechanism for the results presented in the paper is:

- $Z \sim \mathcal{U}(-1; 1)$
- $\text{logit}(P(A = 1|Z)) = 0 + 1 \cdot Z$
- $E(Y|A, Z, t) = 0 \cdot A + 1 \cdot Z + 0.2 \cdot \text{time} + 0.2 \cdot A \cdot Z + 0.2 \cdot A \cdot \text{time}$ with a person-specific intercept $u_i \sim \mathcal{N}(0, 1)$ and $\epsilon_{ij} \sim \mathcal{N}(0, 1)$
- $P(T > t|A, Z, Y(t)) \sim \exp^{-\lambda_T(t) \cdot t}$ with $\lambda_T(t) = 0.0625 \cdot e^{-0.36 \cdot A - 0.36 \cdot Z - 0.22 \cdot Y(t) + 0.18 \cdot A \cdot Z + 0.26 \cdot A \cdot Y(t)}$ where $Y(t)$ switches values at integer values of t
- $P(C > t|A, Z, Y(t)) \sim \exp^{-\lambda_C(t) \cdot t}$ with $\lambda_C(t) = 0.05 \cdot e^{0 \cdot A - 0.11 \cdot Z - 0.69 \cdot Y(t) - 0.11 \cdot A \cdot Z + 0.92 \cdot A \cdot Y(t)}$ where $Y(t)$ switches values at integer values of t

The implemented methods are:

- **Naive Timepoint Specific**
 - **As Observed:** Difference in observed average Y between the two treatment groups as estimated with IEE.
 - **IPCW only:** Difference in IPC weighted average Y between the two treatment groups as estimated with IEE, with weights derived from arm specific Cox-models for censoring with $Y(t)$ as time-varying covariate.
 - **IPTW only:** Difference in IPT weighted average Y between the two treatment groups as estimated by IEE
- **Adapted Methods - Baseline Censoring only**
 - **IEE-IPTCW:** IPTC weighted IEE. One Cox regression model for censoring for all timepoints (separate per treatment group) - time considered as categorical variable in the IEE model. (Section 3.3). Censoring model with baseline-covariates only
 - **IEE-Regression Standardization:** One Cox regression model for death for all timepoints (separate per treatment group). IEE for $E(Y(t)|Z, T > t)$ for both arms combined (Section 3.3), with treatment-time and treatment- Z interaction and categorical time. No censoring weights included, as censoring is assumed to depend only on baseline-covariates already in the Y- and Survival-models.
- **Adapted Longitudinal - Time varying censoring**
 - **IEE-IPTCW:** IPTC weighted IEE. One Cox regression model for censoring with $Y(t)$ as time-varying covariate (separate per treatment group) - time considered as categorical variable in the IEE model. (Section 3.3)

- **IEE-Regression Standardization:** One Cox regression model for death for all timepoints (separate per treatment group). IEE for $E(Y(t)|Z, T > t)$ for both arms combined (Section 3.3), with treatment-time and treatment- Z interaction and categorical time. Both models are IPCWeighted with the same weights as the IPTCW-model.

Table 1: Bias and SD for the different estimators in the simulation study.

Label	M.	T 0	T 1	T 2	T 3	T 4	T 5	T 6	T 7	T 8	T 9	T 10
As Observed	Bias	0.32	0.23	0.15	0.09	0.04	-0.01	-0.05	-0.10	-0.14	-0.18	-0.22
	SD	0.11	0.11	0.12	0.13	0.14	0.14	0.15	0.16	0.17	0.17	0.20
IPCW Only	Bias	0.32	0.32	0.31	0.31	0.30	0.30	0.29	0.28	0.27	0.26	0.27
	SD	0.11	0.12	0.12	0.13	0.15	0.16	0.16	0.17	0.18	0.18	0.21
IPTW only	Bias	0.00	-0.07	-0.13	-0.19	-0.23	-0.27	-0.31	-0.35	-0.38	-0.42	-0.45
	SD	0.11	0.12	0.12	0.13	0.14	0.15	0.16	0.16	0.17	0.18	0.20
IEE-IPTCW (no TV Cens)	Bias	0.00	-0.05	-0.10	-0.15	-0.18	-0.22	-0.25	-0.28	-0.32	-0.35	-0.37
	SD	0.11	0.12	0.12	0.13	0.14	0.15	0.15	0.16	0.16	0.17	0.20
IEE-RegStand (no TV Cens)	Bias	0.03	-0.03	-0.09	-0.14	-0.19	-0.22	-0.26	-0.30	-0.33	-0.37	-0.39
	SD	0.10	0.11	0.12	0.13	0.13	0.14	0.15	0.15	0.16	0.17	0.19
IEE-IPTCW	Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.02	-0.01
	SD	0.11	0.12	0.12	0.13	0.14	0.15	0.16	0.16	0.17	0.18	0.20
IEE-RegStand	Bias	0.01	0.01	0.01	0.01	0.00	0.00	0.00	-0.01	-0.01	-0.03	-0.02
	SD	0.11	0.11	0.12	0.13	0.14	0.15	0.15	0.16	0.17	0.18	0.21

C Case Study

Table 12 provides an overview of the considered baseline distributions in the complete sample and after reduction for non-positivity.

For the analyses, data is prepared in the tstart-tstop-format required by the R-survival package to handle time-varying covariates. This allows the censoring model to be fitted regardless of whether time-varying covariates are included or not, from which then estimated probabilities of censoring are derived and the corresponding IPC-weights. For the single-timepoint analyses, only data up till that timepoint is included in the censoring-model. As for early timepoints there are few censoring-events, this sometimes results in issues where the survival package manages to fit a cox-model but does not provide derived censoring probabilities. As this is most likely to happen in a model with time varying covariates, first 'progression' is dropped from the model in an attempt to solve the issue. If that doesn't work, IPC-weights are set to 1 (this will only happen with very few censoring events). The results of the CoxPH-censoring models with time varying covariates are presented in Section F.1.

IPT-weights are estimated using the weightit-function of the WeightIt-package. Weights are trimmed at the 99th percentile.

Confidence intervals are based on 200 bootstraps.

C.1 Outcome Censoring Models

The tables below present the outcome of the CoxPH-models with time varying covariates for censoring in the primary analysis and the negative control analysis. *QoL* is the time varying observed Health Related Quality of Life at the start of each 3-weeks cycle. *Prog* is a time varying indicator for progression.

Note that $n=$ in the description below the tables refers to the number of rows in the dataset and not the number of patients, given the tstart-tstop format.

The censoring in both RCT-arms is around 20%, the censoring in the SAT is considerably lower.

C.1.1 Primary Analysis

Table 2: Fitting Proportional Hazards Regression Model: SAT Experimental

	coef	exp(coef)	se(coef)	z	p
sexMALE	-0.7271	0.4833	0.5901	-1.232	0.2179
Age	0.01839	1.019	0.0182	1.01	0.3123
ECOGBfactor1	-0.3376	0.7135	0.604	-0.5589	0.5762
ECOGBfactor2	0.7536	2.125	0.7105	1.061	0.2889
NPSTREGSfactor2	-0.6481	0.523	0.5126	-1.264	0.2061
QoL	-0.01885	0.9813	0.01302	-1.447	0.1478
Prog	1.429	4.174	0.5698	2.508	0.01215

Likelihood ratio test=17.26 on 7 df, p=0.0158042 n= 5099, number of events= 16

Table 3: Fitting Proportional Hazards Regression Model: RCT Control

	coef	exp(coef)	se(coef)	z	p
sexMALE	0.7306	2.076	0.3891	1.878	0.0604
Age	0.0213	1.022	0.01408	1.512	0.1305
ECOGBfactor1	-0.1611	0.8512	0.3785	-0.4256	0.6704
ECOGBfactor2	-0.2442	0.7833	1.067	-0.229	0.8189
NPSTREGSfactor2	-0.6433	0.5255	0.5458	-1.179	0.2385
QoL	-0.01266	0.9874	0.009441	-1.34	0.1801
Prog	-0.4276	0.6521	0.4015	-1.065	0.2868

Likelihood ratio test=7.89 on 7 df, p=0.3426893 n= 1522, number of events= 30

C.1.2 Negative Control Analysis

Table 4: Fitting Proportional Hazards Regression Model: SAT Experimental

	coef	exp(coef)	se(coef)	z	p
sexMALE	-0.7271	0.4833	0.5901	-1.232	0.2179
Age	0.01839	1.019	0.0182	1.01	0.3123
ECOGBfactor1	-0.3376	0.7135	0.604	-0.5589	0.5762
ECOGBfactor2	0.7536	2.125	0.7105	1.061	0.2889
NPSTREGSfactor2	-0.6481	0.523	0.5126	-1.264	0.2061
QoL	-0.01885	0.9813	0.01302	-1.447	0.1478
Prog	1.429	4.174	0.5698	2.508	0.01215

Likelihood ratio test=17.26 on 7 df, p=0.0158042 n= 5099, number of events= 16

Table 5: Fitting Proportional Hazards Regression Model: RCT Experimental

	coef	exp(coef)	se(coef)	z	p
sexMALE	0.03874	1.04	0.3768	0.1028	0.9181
Age	0.03844	1.039	0.01448	2.655	0.007942
ECOGBfactor1	-0.3257	0.722	0.3985	-0.8175	0.4137
ECOGBfactor2	0.1583	1.171	0.702	0.2255	0.8216
NPSTREGSfactor2	-0.1102	0.8956	0.562	-0.1962	0.8445
QoL	-0.005568	0.9944	0.009365	-0.5945	0.5521
Prog	0.2844	1.329	0.4761	0.5974	0.5502

Likelihood ratio test=11.86 on 7 df, p=0.105344 n= 1542, number of events= 34

D While-Alive

Figure 4 provides a graphical representation of selection through survival impacts the while-alive estimand and how censoring needs to be addressed to properly estimate it.

Table 6: Baseline demographics SAT vs Both RCT-arms. Complete study populations and remaining sample after aligning patient populations for positivity

Characteristic	Complete sample				Reduced for Positivity			
	Experimental - SAT, N = 879	Control - RCT, N = 158	Experimental - RCT, N = 162	Experimental - SAT, N = 467	Control - RCT, N = 147	Experimental - RCT, N = 152		
Sex								
FEMALE	503 (57%)	89 (56%)	88 (54%)	272 (58%)	82 (56%)	82 (54%)		
MALE	376 (43%)	69 (44%)	74 (46%)	195 (42%)	65 (44%)	70 (46%)		
Age	53 (44, 62)	49 (39, 59)	51 (41, 60)	53 (42, 63)	49 (41, 59)	51 (41, 59)		
Baseline ECOG Score								
0	249 (28%)	63 (40%)	72 (44%)	154 (33%)	57 (39%)	68 (45%)		
1	480 (55%)	84 (53%)	75 (46%)	248 (53%)	80 (54%)	70 (46%)		
2	119 (14%)	11 (7.0%)	15 (9.3%)	65 (14%)	10 (6.8%)	14 (9.2%)		
3	28 (3.2%)	0 (0%)	0 (0%)					
Unknown	3	0	0					
Number of Prior Drug Therapy Regimens								
1	181 (21%)	128 (81%)	142 (88%)	169 (36%)	121 (82%)	135 (89%)		
2	323 (37%)	29 (18%)	17 (10%)	298 (64%)	26 (18%)	17 (11%)		
3	189 (22%)	1 (0.6%)	3 (1.9%)					
≥4	183 (21%)	0 (0%)	0 (0%)					
Unknown	3	0	0					
baseline	50 (33, 75)	58 (42, 75)	58 (42, 75)	58 (33, 75)	58 (42, 75)	58 (42, 75)		
Unknown	41	10	7					
¹ n (%), Median (IQR)								

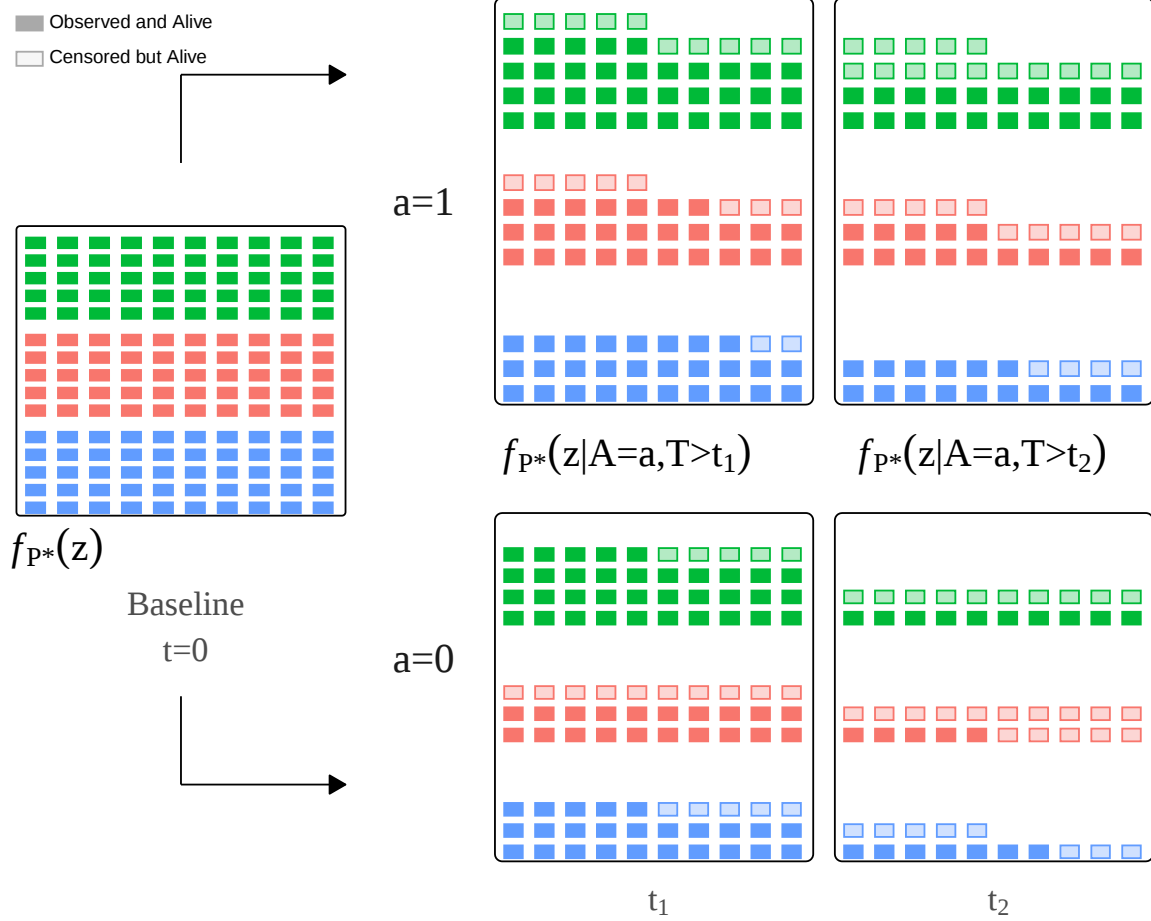


Figure 4: All blocks represent patients still alive, but shaded blocks means they are censored and hence in reality we do not know whether they are still alive or not. Colors represent different values for their baseline-confounders $\mathbf{Z}$. We can imagine that green patients tend to have better longitudinal outcomes than red patients who in turn tend to do better than blue. Observed results under both treatments $A = 0/1$, are transported to the baseline $\mathbf{Z}$ -distribution of target population $\mathcal{P}^*$. The different number of blocks of each color at times t_1 and t_2 between the treatments represents differential survival ($P(T > t^* | A = a, \mathbf{Z})$) which leads to divergent $\mathbf{Z}$ -distributions among those alive at t^* ($f_{\mathcal{P}^*}(z|A = a, T > t^*)$) under treatments $a = 0$ and $a = 1$. The value of our while-alive estimand does not only depend on (the trend of) the outcome within each color-profile, but also on the relative sizes of the green, red and blue group at each timepoint. As can be seen from the shaded blocks, due to censoring, the observed $\mathbf{Z}$ -distribution under those alive at t^* $f_{\mathcal{P}^*}(z|X > t^*)$ may differ from the actual distribution among survivors $f_{\mathcal{P}^*}(z|T > t^*)$, while it is the latter we are interested in.

E Simulation

R-code is available as supplementary material. 7 presents the empirical bias and SD for all evaluated methods.

The data generating mechanism for the results presented in the paper is:

- $Z \sim \mathcal{U}(-1; 1)$
- $\text{logit}(P(A = 1|Z)) = 0 + 1 \cdot Z$
- $E(Y|A, Z, t) = 0 \cdot A + 1 \cdot Z + 0.2 \cdot \text{time} + 0.2 \cdot A \cdot Z + 0.2 \cdot A \cdot \text{time}$ with a person-specific intercept $u_i \sim \mathcal{N}(0, 1)$ and $\epsilon_{ij} \sim \mathcal{N}(0, 1)$
- $P(T > t|A, Z) \sim \exp^{-\lambda_T(t) \cdot t}$ with $\lambda_T(t) = 0.0625 \cdot e^{-0.36 \cdot A - 0.36 \cdot Z - 0.22 \cdot Y(t) + 0.18 \cdot A \cdot Z + 0.26 \cdot A \cdot Y(t)}$ where $Y(t)$ switches values at integer values of t
- $P(C > t|A, Z) \sim \exp^{-\lambda_C(t) \cdot t}$ with $\lambda_C(t) = 0.05 \cdot e^{0 \cdot A - 0.11 \cdot Z - 0.69 \cdot Y(t) - 0.11 \cdot A \cdot Z + 0.92 \cdot A \cdot Y(t)}$ where $Y(t)$ switches values at integer values of t

The implemented methods are:

- **Naive Timepoint Specific**
 - **As Observed:** Difference in observed average Y between the two treatment groups as estimated with IEE.
 - **IPCW only:** Difference in IPC weighted average Y between the two treatment groups as estimated with IEE, with weights derived from arm specific Cox-models for censoring with $Y(t)$ as time-varying covariate.
 - **IPTW only:** Difference in IPT weighted average Y between the two treatment groups as estimated by IEE
- **Adapted Methods - Baseline Censoring only**
 - **IEE-IPTCW:** IPTC weighted IEE. One Cox regression model for censoring for all timepoints (separate per treatment group) - time considered as categorical variable in the IEE model. (Section 3.3). Censoring model with baseline-covariates only
 - **IEE-Regression Standardization:** One Cox regression model for death for all timepoints (separate per treatment group). IEE for $E(Y(t)|Z, T > t)$ for both arms combined (Section 3.3), with treatment-time and treatment- Z interaction and categorical time. No censoring weights included, as censoring is assumed to depend only on baseline-covariates already in the Y- and Survival-models.
- **Adapted Longitudinal - Time varying censoring**
 - **IEE-IPTCW:** IPTC weighted IEE. One Cox regression model for censoring with $Y(t)$ as time-varying covariate (separate per treatment group) - time considered as categorical variable in the IEE model. (Section 3.3)

- **IEE-Regression Standardization:** One Cox regression model for death for all timepoints (separate per treatment group). IEE for $E(Y(t)|Z, T > t)$ for both arms combined (Section 3.3), with treatment-time and treatment- Z interaction and categorical time. Both models are IPCWeighted with the same weights as the IPTCW-model.

Table 7: Bias and SD for the different estimators in the simulation study.

Label	M.	T 0	T 1	T 2	T 3	T 4	T 5	T 6	T 7	T 8	T 9	T 10
As Observed	Bias	0.32	0.23	0.15	0.09	0.04	-0.01	-0.05	-0.10	-0.14	-0.18	-0.22
	SD	0.11	0.11	0.12	0.13	0.14	0.14	0.15	0.16	0.17	0.17	0.20
IPCW Only	Bias	0.32	0.32	0.31	0.31	0.30	0.30	0.29	0.28	0.27	0.26	0.27
	SD	0.11	0.12	0.12	0.13	0.15	0.16	0.16	0.17	0.18	0.18	0.21
IPTW only	Bias	0.00	-0.07	-0.13	-0.19	-0.23	-0.27	-0.31	-0.35	-0.38	-0.42	-0.45
	SD	0.11	0.12	0.12	0.13	0.14	0.15	0.16	0.16	0.17	0.18	0.20
IEE-IPTCW (no TV Cens)	Bias	0.00	-0.05	-0.10	-0.15	-0.18	-0.22	-0.25	-0.28	-0.32	-0.35	-0.37
	SD	0.11	0.12	0.12	0.13	0.14	0.15	0.15	0.16	0.16	0.17	0.20
IEE-RegStand (no TV Cens)	Bias	0.03	-0.03	-0.09	-0.14	-0.19	-0.22	-0.26	-0.30	-0.33	-0.37	-0.39
	SD	0.10	0.11	0.12	0.13	0.13	0.14	0.15	0.15	0.16	0.17	0.19
IEE-IPTCW	Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.02	-0.01
	SD	0.11	0.12	0.12	0.13	0.14	0.15	0.16	0.16	0.17	0.18	0.20
IEE-RegStand	Bias	0.01	0.01	0.01	0.01	0.00	0.00	0.00	-0.01	-0.01	-0.03	-0.02
	SD	0.11	0.11	0.12	0.13	0.14	0.15	0.15	0.16	0.17	0.18	0.21

F Case Study

Table 12 provides an overview of the considered baseline distributions in the complete sample and after reduction for non-positivity.

For the analyses, data is prepared in the tstart-tstop-format required by the R-survival package to handle time-varying covariates. This allows the censoring model to be fitted regardless of whether time-varying covariates are included or not, from which then estimated probabilities of censoring are derived and the corresponding IPC-weights. For the single-timepoint analyses, only data up till that timepoint is included in the censoring-model. As for early timepoints there are few censoring-events, this sometimes results in issues where the survival package manages to fit a cox-model but does not provide derived censoring probabilities. As this is most likely to happen in a model with time varying covariates, first 'progression' is dropped from the model in an attempt to solve the issue. If that doesn't work, IPC-weights are set to 1 (this will only happen with very few censoring events). The results of the CoxPH-censoring models with time varying covariates are presented in Section F.1.

IPT-weights are estimated using the weightit-function of the WeightIt-package. Weights are trimmed at the 99th percentile.

Confidence intervals are based on 200 bootstraps.

F.1 Outcome Censoring Models

The tables below present the outcome of the CoxPH-models with time varying covariates for censoring in the primary analysis and the negative control analysis. *QoL* is the time varying observed Health Related Quality of Life at the start of each 3-weeks cycle. *Prog* is a time varying indicator for progression.

Note that $n=$ in the description below the tables refers to the number of rows in the dataset and not the number of patients, given the tstart-tstop format.

The censoring in both RCT-arms is around 20%, the censoring in the SAT is considerably lower.

F.1.1 Primary Analysis

Table 8: Fitting Proportional Hazards Regression Model: SAT Experimental

	coef	exp(coef)	se(coef)	z	p
sexMALE	-0.7271	0.4833	0.5901	-1.232	0.2179
Age	0.01839	1.019	0.0182	1.01	0.3123
ECOGBfactor1	-0.3376	0.7135	0.604	-0.5589	0.5762
ECOGBfactor2	0.7536	2.125	0.7105	1.061	0.2889
NPSTREGSfactor2	-0.6481	0.523	0.5126	-1.264	0.2061
QoL	-0.01885	0.9813	0.01302	-1.447	0.1478
Prog	1.429	4.174	0.5698	2.508	0.01215

Likelihood ratio test=17.26 on 7 df, p=0.0158042 n= 5099, number of events= 16

Table 9: Fitting Proportional Hazards Regression Model: RCT Control

	coef	exp(coef)	se(coef)	z	p
sexMALE	0.7306	2.076	0.3891	1.878	0.0604
Age	0.0213	1.022	0.01408	1.512	0.1305
ECOGBfactor1	-0.1611	0.8512	0.3785	-0.4256	0.6704
ECOGBfactor2	-0.2442	0.7833	1.067	-0.229	0.8189
NPSTREGSfactor2	-0.6433	0.5255	0.5458	-1.179	0.2385
QoL	-0.01266	0.9874	0.009441	-1.34	0.1801
Prog	-0.4276	0.6521	0.4015	-1.065	0.2868

Likelihood ratio test=7.89 on 7 df, p=0.3426893 n= 1522, number of events= 30

F.1.2 Negative Control Analysis

Table 10: Fitting Proportional Hazards Regression Model: SAT Experimental

	coef	exp(coef)	se(coef)	z	p
sexMALE	-0.7271	0.4833	0.5901	-1.232	0.2179
Age	0.01839	1.019	0.0182	1.01	0.3123
ECOGBfactor1	-0.3376	0.7135	0.604	-0.5589	0.5762
ECOGBfactor2	0.7536	2.125	0.7105	1.061	0.2889
NPSTREGSfactor2	-0.6481	0.523	0.5126	-1.264	0.2061
QoL	-0.01885	0.9813	0.01302	-1.447	0.1478
Prog	1.429	4.174	0.5698	2.508	0.01215

Likelihood ratio test=17.26 on 7 df, p=0.0158042 n= 5099, number of events= 16

Table 11: Fitting Proportional Hazards Regression Model: RCT Experimental

	coef	exp(coef)	se(coef)	z	p
sexMALE	0.03874	1.04	0.3768	0.1028	0.9181
Age	0.03844	1.039	0.01448	2.655	0.007942
ECOGBfactor1	-0.3257	0.722	0.3985	-0.8175	0.4137
ECOGBfactor2	0.1583	1.171	0.702	0.2255	0.8216
NPSTREGSfactor2	-0.1102	0.8956	0.562	-0.1962	0.8445
QoL	-0.005568	0.9944	0.009365	-0.5945	0.5521
Prog	0.2844	1.329	0.4761	0.5974	0.5502

Likelihood ratio test=11.86 on 7 df, p=0.105344 n= 1542, number of events= 34

Table 12: Baseline demographics SAT vs Both RCT-arms. Complete study populations and remaining sample after aligning patient populations for positivity

Characteristic	Complete sample				Reduced for Positivity			
	Experimental - SAT, N = 879	Control - RCT, N = 158	Experimental - RCT, N = 162	Experimental - SAT, N = 467	Control - RCT, N = 147	Experimental - RCT, N = 152		
Sex								
FEMALE	503 (57%)	89 (56%)	88 (54%)	272 (58%)	82 (56%)	82 (54%)		
MALE	376 (43%)	69 (44%)	74 (46%)	195 (42%)	65 (44%)	70 (46%)		
Age	53 (44, 62)	49 (39, 59)	51 (41, 60)	53 (42, 63)	49 (41, 59)	51 (41, 59)		
Baseline ECOG Score								
0	249 (28%)	63 (40%)	72 (44%)	154 (33%)	57 (39%)	68 (45%)		
1	480 (55%)	84 (53%)	75 (46%)	248 (53%)	80 (54%)	70 (46%)		
2	119 (14%)	11 (7.0%)	15 (9.3%)	65 (14%)	10 (6.8%)	14 (9.2%)		
3	28 (3.2%)	0 (0%)	0 (0%)					
Unknown	3	0	0					
Number of Prior Drug Therapy Regimens								
1	181 (21%)	128 (81%)	142 (88%)	169 (36%)	121 (82%)	135 (89%)		
2	323 (37%)	29 (18%)	17 (10%)	298 (64%)	26 (18%)	17 (11%)		
3	189 (22%)	1 (0.6%)	3 (1.9%)					
≥4	183 (21%)	0 (0%)	0 (0%)					
Unknown	3	0	0					
baseline	50 (33, 75)	58 (42, 75)	58 (42, 75)	58 (33, 75)	58 (42, 75)	58 (42, 75)		
Unknown	41	10	7					
¹ n (%), Median (IQR)								