

# Collider Effect in the Analysis of Observational Data: Epidemiological Insight and Interactive Visualization

*Correlation between exposures and outcomes in the analysis of epidemiological data is common but correlation does not mean causation*

Miguel Angel Luque-Fernandez<sup>1</sup>, Daniel Rodondo-Sanchez<sup>2</sup>, Michael Schomaker<sup>3</sup>

## Abstract

## 1 Introduction

Causal quantities, such as the average treatment effect, causal odds ratio or causal risk ratio, can often be estimated from observational data if assumptions about the data generating process are made and met. In this context, Targeted Maximum Likelihood Estimation (TMLE) is one estimation option. Since the mid 2000's TMLE has been implemented successfully in numerous applications, initially for the effects of treatments or exposures at a single time point [1, 2, 3, 4]. In recent years, an increasing body of work has dealt with the application of TMLE in longitudinal settings [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16].

Longitudinal targeted maximum likelihood estimation (LTMLE) requires the fitting of models for the outcome, censoring, and treatment mechanisms, at each point in time. Alternative approaches include those based on estimating the treatment mechanism only (inverse probability of treatment weighted approaches)[17], the iterated outcome regressions only (one form of g-computation)[18] or an outcome regression combined with the conditional densities of the confounders ("traditional" parametric g computation) [19]. LTMLE is a doubly robust methodology which means that the causal quantity of interest is estimated consistently if either the iterated outcome regressions or the treatment and censoring mechanism are estimated consistently. Moreover, LTMLE allows analytic estimation of standard errors and confidence intervals, based on estimates of the influence curve, as opposed to the g-formula which makes use of bootstrapping [20]. To increase the chances of correct model specification, the use of machine learning algorithms are typically recommended. LTMLE, like other double robust methods, has the advantage over both inverse probability weighted and g-computation methods that it can more readily incorporate machine learning methods while retaining valid statistical inference.

The current literature is somewhat limited with respect to understanding the implications of using LTMLE for long follow-up times when dealing with small (or moderate) sample sizes and complex relationships among a multitude of confounders. Most studies to date were illustrative in nature and have dealt with a small number of follow-up time points, i.e. two to nine [5, 6, 7, 8, 9, 11, 12, 13, 14, 16], or large sample sizes [10, 14, 16], but to our knowledge never with small or moderate sample sizes and long follow-up at the same time. In spite of LTMLE's double robustness feature, consistent estimation of both treatment and censoring mechanisms and outcome regressions, with reasonable rates of convergence, is important to improve efficiency, reduce finite sample bias, and provide the basis for valid statistical inference. A small sample may limit the application of data-adaptive approaches, such as additive models, boosting, random forests, shrinkage estimators, support vector machines and others, because the selection of tuning parameters and convergence of fitting algorithms may not work well, or not at all, under these circumstances – particularly when confronted with multiple time-dependent confounders which influence outcome and treatment non-linearly and in interaction with each other. Interestingly, many studies using

---

<sup>1</sup>Biomedical Research Institute. Non-Communicable and Cancer Epidemiology Group (ibs.Granada), Andalusian School of Public Health, University of Granada, Granada, Spain, [miguel.luque.easp@juntadeandalucia.es](mailto:miguel.luque.easp@juntadeandalucia.es)

<sup>2</sup>Biomedical Research Institute. Non-Communicable and Cancer Epidemiology Group (ibs.Granada), Andalusian School of Public Health, University of Granada, Granada, Spain, [daniel.redondo.easp@juntadeandalucia.es](mailto:daniel.redondo.easp@juntadeandalucia.es)

<sup>3</sup>University of Cape Town, Centre for Infectious Disease Epidemiology and Research, Observatory, 7925; Cape Town, South Africa, [michael.schomaker@uct.ac.za](mailto:michael.schomaker@uct.ac.za)

LTMLE make restricted use of machine learning algorithms [6, 7, 10, 15, 16], posing the question whether this was done intentionally because of restricted sample size or computational limitations.

A common estimation approach for LTMLE requires model fitting on the subset of subjects who follow the intervention of interest (and remain alive and uncensored). This may pose a problem if follow-up time is long and therefore the sample size reduces gradually: under these circumstances it may no longer be possible to successfully fit a multitude of learners which makes correct model specification more challenging. There exist alternative estimation approaches [11], as well as other estimators, both of which utilize the full sample; it remains however to be investigated how the performance of these estimators differ in a finite sample.

For applying LTMLE in survival analyses, both the treatment and censoring mechanisms need to be modeled; in fact, for  $T$  time points, the cumulative inverse probabilities of remaining uncensored and following the intervention of interest, i.e. the product of  $T \times 2$  inverse probabilities, are part of the weighted fitting procedure. Unstable weights, possibly dealt with by truncation of the probabilities on which the weight calculation is based, could therefore emerge under long follow-up and non-ideal modeling strategies. It remains unclear to what degree this is a threat to the successful implementation of LTMLE.

The present article is based on our motivating question outlined in the next section. To answer it, one has to deal with long follow-up times, gradually reduced sample size (down to 10% of the original sample size of 2604) and multiple time-dependent confounders. Clinical knowledge suggests that the data-generating process is complex in the sense that modeling should go beyond standard regression approaches. We investigate challenges and solutions to apply LTMLE in this setting, both by means of simulations and data analyses. In particular, we try to answer the following questions: (i) is it feasible to apply LTMLE in such a setting; (ii) if yes, will the complex setup reduce the applicability of super learning and would this impact the success of LTMLE?; (iii) under such a challenging scenario, could a manual implementation which is informed by prior clinical knowledge and not data-adaptive be more successful than an automated LTMLE procedure?; (iv) how do long follow-up times, gradually reduced sample size, and truncated inverse probabilities impact on LTMLE for different intervention mechanisms? (v) what can be learned from these points that would benefit future analyses which attempt to apply LTMLE in complex data settings.

We introduce the motivating question in Section 2, followed by the theoretical framework in Section 3. After presenting our data analysis (Section 4) we discuss the implications of LTMLE in such a setting by means of both practical considerations and simulated studies in the following two sections. We conclude in Section 7.

## 2 Motivating Example

During the last decade the World Health Organization (WHO) updated their recommendations on the use of antiretroviral drugs for treating and preventing HIV infection several times. In the past, antiretroviral treatment (ART) was only given to a child if his/her measurements of CD4 lymphocytes fell below a critical value or if a clinically severe event (such as tuberculosis or persistent diarrhoea) occurred. Based on both increased knowledge from trials and causal modeling studies, as well as pragmatic and programmatic considerations, these criteria have been gradually expanded to allow earlier treatment initiation in children: ART has shown to be effective and to reduce mortality in infants and adults [21, 22, 23], but concerns remain due to a potentially increased risk of toxicities, early development of drug resistance, and limited future options for people who fail treatment. In late 2015, WHO decided to recommend immediate treatment initiation in all children and adults.

It remains important to investigate the effect of different treatment initiation rules on mortality, morbidity and child development outcomes; for example, because of possible unknown long-term treatment effects and the increase in drug related toxicities. However given the shift in ART guidelines towards earlier treatment initiation it is no longer ethically possible to conduct a trial which answers this question in detail. Moreover, trials are often restricted by testing intervention mechanisms in idealized, non-realistic, settings - for example by imposing regular medical visits and therefore attention which would not happen in most public health programmes.

Thus, observational data can be used to obtain effect estimates to compare early versus delayed treatment initiation rules. Methods such as inverse probability weighting of marginal structural models, the g-computation formula, and longitudinal targeted maximum likelihood estimation can be used to obtain results in complicated longitudinal settings where time-varying confounders affected by prior treatment are present — such as, for example, CD4 count which influences both the probability of ART initiation and outcome measures [24, 11].

Below we focus on children’s growth outcomes, measured by height-for-age z-scores, and explore the effect of different treatment initiation criteria on it. We are particularly interested in whether strategies that offer earlier treatment initiation are beneficial for children’s growth.

### 3 Methodological Framework

Consider  $n$  subjects studied at baseline ( $t = 0$ ) and during discrete follow-up times ( $t = 1, \dots, T$ ) where  $T$  defines the end of follow-up. Some subjects may be censored and not be followed until  $T$ . The data for each time point  $t$  consists of the outcome  $Y_t$ , an intervention variable  $A_t$ ,  $q$  time-dependent covariates  $\mathbf{L}_t = \{L_t^1, \dots, L_t^q\}$ , and a censoring indicator  $C_t$ . We assume this ordering to reflect the causal ordering for a given time point  $t$ . Note that  $\mathbf{L}_t$  may contain indicator variables which describe whether  $L_t^s$  has been measured at time  $t$  or not,  $s \in \{1, \dots, q\}$ . Moreover,  $\mathbf{L}_t$  may also include baseline variables  $\mathbf{V} = \{L_0^1, \dots, L_0^q\}$  and can potentially contain variables which refer to the outcome variable before time  $t$ , for instance  $Y_{t-1}$ . The treatment and covariate history of an individual  $i$  up to and including time  $t$  is represented as  $\bar{A}_{t,i} = (A_{0,i}, \dots, A_{t,i})$  and  $\bar{L}_{t,i}^s = (L_{0,i}^s, \dots, L_{t,i}^s)$  respectively.  $C_t$  equals 1 if a subject gets censored in the interval  $(t-1, t]$ , and 0 otherwise. Therefore,  $\bar{C}_t = 0$  is the event that an individual remains uncensored until time  $t$ . Further, let  $S_{t,i}$  be an indicator variable which is 1 if a patient  $i$  is still alive at time  $t$  and 0 otherwise.

The counterfactual outcome  $Y_{t,i}^{\bar{a}_t}$  refers to the hypothetical outcome that would have been observed at time  $t$  if subject  $i$  had received, possibly contrary to the fact, the treatment history  $\bar{A}_{t,i} = \bar{a}_t$ . Similarly,  $\mathbf{L}_{t,i}^{\bar{a}_t}$  are the counterfactual covariates related to the intervention  $\bar{A}_{t,i} = \bar{a}_t$ . The above notation refers to *static* treatment rules; a treatment rule may, however, depend on covariates, and in this case it is called *dynamic*. A dynamic rule  $d_t(\bar{\mathbf{L}}_t, S_t)$  assigns treatment  $A_{t,i} \in \{0, 1\}$  as a function of the covariate history  $\bar{\mathbf{L}}_{t,i}$ , conditional on survival. The vector of decisions  $d_t$ ,  $t = 0, \dots, T$ , is denoted as  $\bar{d}_T = \bar{d}$ . The notation  $\bar{A}_t = \bar{d}$  refers to the treatment history according to the rule  $\bar{d}$ . The counterfactual outcome related to a dynamic rule  $\bar{d}$  is  $Y_{t,i}^{\bar{d}}$ , and the counterfactual covariates are  $\mathbf{L}_{t,i}^{\bar{d}}$ . In the remainder of this paper we consider rules  $d_t$  that assign  $A_t$  as a function of  $\bar{\mathbf{L}}_t$  and set  $C_t$  deterministically to 0, conditional on survival.

Without loss of generalization we consider the expected value of  $Y$  at time  $T$ , under no censoring, for a given treatment rule  $\bar{d}$ , conditional on survival (under this rule) until at least the beginning of the time interval  $(t-1, t]$ , to be the main quantity of interest, that is  $\psi_T = \mathbb{E}(Y_T^{\bar{d}} | S_{T-1}^{\bar{d}} = 1)$ . Secondary quantities of interest are the expected values at different time points, i.e.  $\psi_t = \mathbb{E}(Y_t^{\bar{d}} | S_{t-1}^{\bar{d}} = 1)$ .

Below, in Sections 3.1, 3.2, and 3.3, we are going to introduce 3 estimators which are generally suitable to estimate  $\psi_T$ . These estimators are further discussed in Section 3.4, and are implemented and compared in both the data analysis (Section 4) and the simulation study (Section 6).

#### 3.1 The (parametric) g-formula

One possibility to identify the quantity of interest is by using the (parametric) g-formula [19]:

$$\begin{aligned} \psi_T = \mathbb{E}(Y_T^{\bar{d}} | S_{T-1}^{\bar{d}} = 1) = \\ \int_{\mathbf{I} \in \bar{\mathbf{L}}_T} \left\{ \mathbb{E}(Y_T | \bar{A}_{T-1} = \bar{d}_T, \bar{\mathbf{L}}_T = \bar{\mathbf{I}}_T, S_{T-1} = 1) \times \right. \\ \left. \prod_{t=1}^T \left[ \frac{f(\mathbf{I}_t | \bar{A}_{t-1} = \bar{d}_{t-1}, \bar{\mathbf{L}}_{t-1} = \bar{\mathbf{I}}_{t-1}, S_{t-1} = 1) \times \mathbb{P}(S_{t-1} = 1 | \bar{A}_{t-1} = \bar{d}_{t-1}, \bar{\mathbf{L}}_{t-1} = \bar{\mathbf{I}}_{t-1}, S_{t-2} = 1)}{\mathbb{P}(S_{t-1} = 1 | \bar{A}_{t-1} = \bar{d}_{t-1}, \bar{\mathbf{L}}_{t-1} = \bar{\mathbf{I}}_{t-1}, S_{t-2} = 1)} \right] \right\} d\bar{\mathbf{I}} \end{aligned} \quad (1)$$

For ordered  $\mathbf{L}_t = \{L_t^1, \dots, L_t^q\}$  we can write  $f(\mathbf{l}_t | \bar{A}_{t-1} = \bar{d}_{t-1}, \bar{\mathbf{L}}_{t-1} = \bar{\mathbf{l}}_{t-1}, S_{t-1} = 1)$  in (1) as

$$\prod_{s=1}^q f(l_t^s | \bar{A}_{t-1} = \bar{d}_{t-1}, \bar{\mathbf{L}}_{t-1} = \bar{\mathbf{l}}_{t-1}, L_t^1 = l_t^1, \dots, L_t^{s-1} = l_t^{s-1}, S_{t-1} = 1). \quad (2)$$

The above representation of the g-formula makes an implicit assumption about the time ordering of the relevant variables. For example, for  $s = 3$  we assume that the ordering is  $S_{t-1} \rightarrow L_t^1 \rightarrow L_t^2 \rightarrow L_t^3 \rightarrow Y \rightarrow A \rightarrow C \rightarrow S_t$  for  $\forall t$ , see Section 5.2 for more details on the chosen time ordering. We assume time points to represent intervals and thus both covariates and a fatal event like death can potentially be measured at the same time  $t$ .

The equality holds under several assumptions. One is *consistency* which says that the counterfactuals equal the observed data under assignment to the treatment actually taken; more technically,  $Y^{\bar{d}_t} = Y$  if  $\bar{A}_t = \bar{d}_t$  and  $\bar{\mathbf{L}}_t^{\bar{d}_t} = \bar{\mathbf{L}}_t$  if  $\bar{A}_{t-1} = \bar{d}_{t-1}$ . Another assumption is (sequential) conditional exchangeability (or *no unmeasured confounding*) which requires the counterfactual outcome under the assigned treatment rule to be independent of the observed treatment assignment, given the observed past:  $Y^{\bar{d}_t} \perp\!\!\!\perp A_t | \bar{\mathbf{L}}_t, \bar{A}_{t-1}$  for  $\forall \bar{A}_t = \bar{d}_t, \bar{\mathbf{L}}_t = \bar{\mathbf{l}}_t, t \in \{0, \dots, T\}$ . *Positivity* is the requirement that individuals have a positive probability of continuing to receive treatment according to the assigned treatment rule, given that they have done so thus far and irrespective of the covariate history:  $P(A_t = \bar{d}_t | \bar{\mathbf{L}}_t = \bar{\mathbf{l}}_t, \bar{A}_{t-1} = \bar{d}_{t-1}) > 0$  for  $\forall t, \bar{d}_t, \bar{\mathbf{l}}_t$  with  $P(\bar{\mathbf{L}}_t = \bar{\mathbf{l}}_t, \bar{A}_{t-1} = \bar{d}_{t-1}) \neq 0$ . In our setting, all above definitions apply to the subset of survivors at time  $t$ .

If we intervene upon censoring, e.g. we are interested in a treatment rule where all individuals remain uncensored, the above assumptions (consistency, sequential conditional exchangeability, positivity) also apply to  $C_t$ . An important consequence of this is that the censoring mechanism has to be non-informative (i.e. censoring may not depend on unmeasured variables).

More comprehensive interpretations of the above assumptions can be found in [24, 25, 26, 27, 28] and are discussed in Section 5 as well.

### 3.1.1 Practical implementation of the (parametric) g-formula:

the integral (1) can be approximated by means of simulation:

1. Estimate the conditional densities in (1) by means of suitable, possibly semi-parametric, regression models. One requires models for the  $s$  time-varying confounders  $L_t^s$ , the outcome  $Y_t$ , as well as the survival indicator  $S_t$ , for  $t = 1, \dots, T$ . Note that for this step some parametric assumptions are required, e.g. committing to a normal distribution of the residuals with homoscedastic variance when using a linear (additive) model.
2. Choose an intervention rule  $\bar{d}$ .
3. Simulate data for the chosen intervention rule *forward in time* based on the estimated conditional distributions from Step 1. For example, for  $t = 0$ , the data relates to the observed data. Then, for each  $t > 0$ , set  $A_t$  according to the treatment rule  $\bar{d}$  and make predictions based on the models fitted an step 1 ( $\tilde{L}_t^s, \tilde{Y}_t$ , and  $\tilde{S}_t$ ). Then, generate stochastic draws given the conditional distributions. For instance, for a linear regression model  $\mathcal{M}$  with outcome  $L_t^s$ , draw from a  $N(\tilde{L}_t^s, \hat{\sigma}_{\mathcal{M}}^2)$  distribution. This draw relates to the estimated counterfactual  $\hat{L}_t^{s, \bar{d}}$ .
4. Steps 1 to 3 produce a counterfactual data set for  $\bar{d}: \mathcal{D} = \{A^{\bar{d}}, \hat{\mathbf{L}}^{\bar{d}}, \hat{Y}^{\bar{d}}, \hat{S}^{\bar{d}}\}$ . Thus,  $\hat{\psi}_T = \frac{1}{n} \sum_{i \in \mathcal{S}} \hat{Y}_{T,i}^{\bar{d}}$ ,  $\mathcal{S} = \{i \text{ with } \hat{S}_i^{\bar{d}} = 1\}$ , can be easily calculated from this data set.
5. Confidence intervals are then created by means of bootstrapping.

## 3.2 The sequential g-formula

The g-formula integrates out time-dependent confounders, which are affected by prior treatment – with respect to their post intervention distribution. This can be interpreted as a standardization with respect

to them. The sequential g-formula, introduced below, shares the idea of standardization in the sense that one sequentially marginalizes the distribution with respect to  $\mathbf{L}$  given the intervention rule of interest. It is just a re-expression of the traditional g-formula where integration with respect  $\mathbf{L}$  is not needed [8, 12]. In general, using the iterative conditional expectation rule (together with the above stated assumptions of sequential conditional exchangeability, consistency, positivity and time ordering  $\mathbf{L}_t \rightarrow A_t \rightarrow Y_t$ ) it holds that

$$\mathbb{E}(Y_T^{\bar{d}}) = \mathbb{E}(\mathbb{E}(\dots \mathbb{E}(\mathbb{E}(Y_T | \bar{A}_T = \bar{d}_T, \bar{\mathbf{L}}_T) | \bar{A}_{T-1} = \bar{d}_{T-1}, \bar{\mathbf{L}}_{T-1}) \dots | \bar{A}_0 = d_0, \mathbf{L}_0) | \mathbf{L}_0). \quad (3)$$

However, our target quantity is  $\psi_T$ , which is defined as the expected value of  $Y$  at time  $T$ , for a given treatment rule  $\bar{d}$  and no censoring, conditional on all survivors. As noted above we further assume a time-ordering of  $S_{t-1} \rightarrow L_t^1 \rightarrow L_t^2 \rightarrow L_t^3 \rightarrow Y \rightarrow A \rightarrow C \rightarrow S_t$ . We show in Appendix A that the sequential g-formula in this case equates to

$$\begin{aligned} \psi_T &= \mathbb{E}(Y_T^{\bar{d}} | S_{T-1}^{\bar{d}} = 1) \\ &= \mathbb{E}(\mathbb{E}(\mathbb{E}(\dots \mathbb{E}(\mathbb{E}(Y_T | \mathbf{A}_{T-1} = \bar{d}_{T-1}, \bar{\mathbf{L}}_T, S_{T-1} = 1) | \mathbf{A}_{T-2} = \bar{d}_{T-2}, \bar{\mathbf{L}}_{T-1}, S_{T-1} = 1) \dots | \\ &\quad S_{T-1} = 1, \mathbf{L}_0, \mathbf{A}_0 = d_0) | S_{T-1} = 1, \mathbf{L}_0) | S_{T-1} = 1). \end{aligned} \quad (4)$$

### 3.2.1 Practical implementation of the sequential g-formula:

using equation (4) the target parameter can be estimated as follows:

For  $t = T, \dots, 0$ :

1. Use an appropriate model to estimate  $\mathbb{E}(Y_T | \bar{\mathbf{A}}_{T-1} = \bar{d}_{T-1}, \bar{\mathbf{L}}_T, S_{T-1} = 1)$ . The model is fitted on all subjects that are uncensored and alive (until  $T-1$ ). Note that the outcome refers to the measured outcome for  $t = T$  and to the prediction (of the conditional outcome) from step 2 if  $t < T$ .
2. Now, plug in  $\bar{A}_t = \bar{d}_t$  based on rule  $\bar{d}$  and use the regression model from step 1 to predict  $Y_t$  at time  $t$ .
3. For  $t = 0$  the estimate  $\hat{\psi}_T$  is obtained by calculating the arithmetic mean of the predicted outcome (of all survivors) from step 2.
4. Confidence intervals can be obtained using bootstrapping (or equation (6) below, which however may not yield nominal coverage).

Appendix C, which illustrates implementation details for the LTMLE estimator from Section 3.3, can also be seen as a guideline to implement the sequential g-formula: omitting steps 3 and 5 from the code yields the desired solution.

## 3.3 The LTMLE estimator

The LTMLE estimator is based on (3) and goes back to Bang and Robins [18] and Scharfstein et al. [29], and has been incorporated into a more general doubly robust framework by van der Laan and Rubin [30] and others. For our target parameter  $\psi_T$  the LTMLE framework can be applied as follows: estimate  $\psi_T$  with the sequential g-formula, but with an additional *targeted* step for each  $t$  which enables doubly robust inference with respect to  $\psi_T$ , the quantity of interest.

### 3.3.1 Practical implementation of the LTMLE estimator:

For  $t = T, \dots, 0$ :

1. Use an appropriate regression model to estimate  $\mathbb{E}(Y_T | \bar{\mathbf{A}}_{T-1} = \bar{d}_{T-1}, \bar{\mathbf{L}}_T, S_{T-1} = 1)$ . The model is fitted on all subjects that are uncensored and alive (until  $T-1$ ). Note that the outcome refers to the measured outcome for  $t = T$  and to the prediction (of the conditional outcome) from step 3d if  $t < T$ . (“ $Q^0$  model”)

2. Now, plug in  $\bar{\mathbf{A}}_{t-1} = \bar{d}_{t-1}$  based on rule  $\bar{d}$  and use the regression model from step 1 to predict the outcome at time  $t$ .
3. To improve inference with respect to the target parameter  $\psi_t$  update the initial estimate of step 2 by means of the following regression:
  - a) the outcome refers again to the measured outcome for  $t = T$  and to the prediction from item 3d if  $t < T$ .
  - b) the offset is the original predicted outcome from step 2.
  - c) the estimated “clever covariate” refers to the cumulative inverse treatment  $[j = 1]$  and censoring probabilities  $[j = 2]$  (estimated with the “g-model”):

$$\hat{H}(\bar{A}, \bar{C}, \bar{\mathbf{L}})_s = \prod_{s=0}^t \frac{I(\bar{A}_s^j = \bar{d}_s)}{\hat{\mathbb{P}}(A_s^j = \bar{d}_s | \bar{\mathbf{L}}_s = \bar{\mathbf{I}}_s, A_{s-1}^j = \bar{d}_{s-1})} \quad ; j \in \{1, 2\} \quad (5)$$

- d) Predict the (updated) outcome,  $\tilde{Y}_t^{\bar{d}}$ , based on the model defined through 3a, 3b, and 3c.

This model is fitted and evaluated for all subjects that are uncensored, alive (until  $T-1$ ), *and follow treatment rule*  $\bar{\mathbf{A}}_t = \bar{d}_t$  (“ $Q^1$  model”). Alternatively, the same model can be fitted with  $\hat{H}(\bar{A}, \bar{C}, \bar{\mathbf{L}})_s$  as a weight rather than a covariate [8].

4. For  $t = 0$  the estimate  $\hat{\psi}_T$  is obtained by calculating the mean of the predicted outcome (of all survivors) from step 3d.
5. Confidence intervals can, for example, be obtained using the vector of the estimated influence curve [31] of  $\psi_T$ , which can be written as

$$\widehat{\text{IC}}(\psi_T) = \left\{ \sum_{s=1}^T \hat{H}(\bar{A}, \bar{C}, \bar{\mathbf{L}})_{s-1} \left[ \tilde{Y}_s^{\bar{d}} - \tilde{Y}_{s-1}^{\bar{d}} \right] \right\} + \tilde{Y}_0^{\bar{d}} - \hat{\psi}_{T, \text{TMLE}} \quad (6)$$

An asymptotically normal 95% confidence interval is then given by

$$\left[ \hat{\psi}_{\text{TMLE}} \pm 1.959964 \sqrt{\widehat{\text{Var}}(\widehat{\text{IC}})/n} \right].$$

It has been shown that using the five steps above yields a doubly robust estimator of  $\psi_T$  [18]. Doubly robust means that the estimator is consistent as long as either the outcome model[s] (step 1) or the treatment model[s] (step 3c) are estimated consistently. If both are estimated consistently at reasonable rates, the estimator is asymptotically efficient. This is essentially because step 3, in particular the construction of the covariate in step 3c, guarantees that the estimating equation corresponding to the efficient influence curve is solved, which in turn yields desirable (asymptotic) inferential properties. The interested reader is referred to the appendix of van der Laan and Rose (2011) [20], or Schnitzer and Cefalu [32], for details on how the influence curve relates to inference in targeted maximum likelihood estimation.

### 3.4 Inferential considerations for the three presented estimators

Each of the above introduced estimators requires the estimation of conditional expectations which can be utilized with (parametric) regression models. Under the assumption that they are correctly specified this approach is valid. In a complex longitudinal setting such as ours, model mis-specification is a concern, and may yield a biased estimate of  $\psi_T$ .

An alternative to using regression models would be (supervised) machine learning, which is a loose term in the statistical literature referring to both traditional and modern estimation techniques, such as semi-parametric regression models, shrinkage estimators, boosting, random forests, support vector machines, and others. For a given data set, and an a priori specified set of estimation techniques, the method minimizing the expected prediction error (as estimated by  $k$ -fold cross validation) can be



chosen. As it is not clear beforehand which method this is, and whether a combination of learners might perform even better than a single one, one may use super learning. Super learning, which is also known as stacking [33], means considering a set of prediction algorithms (“learners”); Instead of choosing the algorithm with the smallest cross validation error, super learning chooses a weighted linear combination of different algorithms, that is the weighted combination which minimizes the  $k$ -fold cross validation error (for a given loss function, e.g. quadratic loss). This combination is linear, the weights are non-negative and sum up to one. This can be easily utilized with either quadratic programming or a non-negative least squares estimation approach, and is implemented in the *R*-package **SuperLearner** [34]. It can be shown that this weighted combination will perform at least as well as the best algorithm, if not better [35]. If none of the specified algorithms is a correctly specified parameteric model, then the super learner will perform asymptotically as well as the oracle selector (which chooses the best algorithm in the library), otherwise it will achieve an almost parameteric rate of convergence (and will therefore perform reasonably but will not be as good as the best candidate algorithm). For all methods a major advantage of using super learning (rather than a single regression model) is that the probability of model mis-specification is reduced because distributional assumptions are not needed and complex multivariate relationships can potentially be represented by the weighted combination of (potentially simple) learners. However, i) good prediction algorithms may still yield biased model estimates of the conditional expectations (in exchange for reduced variance) and ii) cross validation evaluates the expected prediction error with respect to the global fit of the conditional expectation, and not the scalar parameter  $\psi_T$ . TMLE incorporates the targeting step towards  $\psi_T$  and can thus use super learning, whereas the (sequential) g-formula may not necessarily benefit from it.

The successful use of super learning requires a good and broad selection of prediction algorithms, and the ability to estimate these on the data. This may be challenging for longitudinal data with many time points and small sample size, as in our data. Section 5.4 gives a thorough discussion on the challenges and appropriate use of machine learning for data setups like ours.

## 4 Data Analysis

We study 2604 children aged 1-5, with complete baseline data, for  $t = 0, 1, 3, 6, 9, \dots, 30$  months. The follow-up time points refer to the intervals  $(0, 1.5)$ ,  $[1.5, 4.5)$ ,  $[4.5, 7.5)$ ,  $[7.5, 10.5)$ ,  $[10.5, 13.5)$ ,  $[13.5, 16.5)$ ,  $[16.5, 19.5)$ ,  $[19.5, 22.5)$ ,  $[22.5, 25.5)$ ,  $[25.5, 28.5)$ ,  $[28.5, 31.5)$  months respectively. After 30 months of follow-up 1421 children were still alive and in care. Follow-up measurements, if available, refer to measurements closest to the middle of the interval. In our data  $Y_t$  refers to height for age z-score (HAZ) at time  $t$  (i.e. measured during the interval  $(t - 1, t]$ ). This growth outcome is of potential interest as the PREDICT trial, conducted in children aged 1-12 years, showed better height gain for children who start ART immediately [36, 37]. The binary treatment variable  $A_t$  refers to whether ART was taken at time  $t$  or not. Similarly,  $C_t$  and  $S_t$  describe whether a patient was censored (administratively or due to loss to follow-up), or died, at time  $t$  respectively. Children were defined as being lost to follow-up if at the time of database closure they had no contact with their health care facility for at least 9 months since their last recorded visit. Time-dependent confounders affected by prior treatment refer to  $\mathbf{L}_t = (L_t^1, L_t^2, L_t^3)$  which are CD4 count, CD4%, and weight for age z-score (WAZ) which is a proxy for WHO stage [38]. Similarly, we write  $L_t^{1m}, L_t^{2m}, L_t^{3m}$  for indicator variables which specify if  $L^s$  has been measured at time  $t$  or not. They are needed to facilitate estimation step 3 for the g-formula.  $\mathbf{V} = \mathbf{L}_0^V$  refers to baseline values of CD4 count, CD4%, WAZ, HAZ as well as sex, age, year of treatment initiation and region. The primary outcome is  $Y_{30}$  and therefore  $Y_t$ ,  $t < 30$ , could be regarded as additional confounder  $L_t^4$ . We consider the following four interventions, two of them static, two dynamic:

$$\bar{d}_t^1 = \{c_t = 0; \quad \text{and} \quad a_t = 1 \quad \text{for} \quad \forall t\}$$

$$\bar{d}_t^2(\text{CD4 count}_t, \text{CD4}\%_t) = \begin{cases} c_t = 0; \quad \text{and} \quad a_t = 1 & \text{if } \text{CD4 count}_t^{\bar{d}} < 750 \quad \text{or} \quad \text{CD4}\%_t^{\bar{d}} < 25\% \\ c_t = 0; \quad \text{and} \quad a_t = 0 & \text{otherwise} \end{cases}$$

$$\bar{d}_t^3(\text{CD4 count}_t, \text{CD4}\%_t) = \begin{cases} c_t = 0; & \text{and } a_t = 1 & \text{if } \text{CD4 count}_t^{\bar{d}} < 350 \text{ or } \text{CD4}\%_t^{\bar{d}} < 15\% \\ c_t = 0; & \text{and } a_t = 0 & \text{otherwise} \end{cases}$$

$$\bar{d}_t^4 = \{c_t = 0; \text{ and } a_t = 0 \text{ for } \forall t$$

The first intervention ( $\bar{d}_t^1$ ) means immediate treatment initiation, the last one ( $\bar{d}_t^4$ ) means no treatment initiation, and interventions two and three ( $\bar{d}_t^2, \bar{d}_t^3$ ) mean deferring treatment until a particular CD4 threshold is reached.

We consider the mean HAZ at 30 months, under no censoring, for a given treatment rule  $\bar{d}^j$ ,  $j = 1, \dots, 4$ , and for all survivors, to be the main quantity of interest, that is  $\psi_{30} = \mathbb{E}(Y_{30}^{\bar{d}^j} | S_{30}^{\bar{d}^j} = 1)$ . We estimate  $\psi_{30}$  based on four approaches:

- (i) using the g-formula as described in (1) with a manual implementation which includes prior clinical prior knowledge (i.e. using only additive regression models)
- (ii) using the sequential g-formula as described in (3), automated using the `ltmle` package in *R* [9, 39] with a data adaptive estimation approach.
- (iii) using TMLE as described in Section 3.3 with a manual implementation which includes prior clinical knowledge (i.e. using only additive regression models)
- (iv) using TMLE as described in Section 3.3 automated using the `ltmle` package in *R* [9, 39] with a data adaptive estimation approach.

This choice essentially compares using standard software approaches (`ltmle`, existing and applicable super learning wrappers) with a manual implementation which uses only one learner (additive models), however one which is complex, not implemented in a wrapper, and incorporates prior clinical knowledge. Section 5.5 provides more details on the choice of these estimators.

## 4.1 Model choice

The implementation details for both LTMLE estimators are given in appendices B and C, and are discussed further in Section 5. Briefly, for the automated implementations, i.e. estimators (ii) and (iv), we use the following specifications: first, we do not specify the regression models for outcome, treatment, and censoring parametrically, but use super learning instead. The candidate learners were the arithmetic mean, the full (generalized linear) regression model [GLM], GLM's based on an EM-algorithm-Bayesian model fitting, GLMs chosen by model selection with Akaike's Information Criterion [AIC], GLMs containing interaction terms, as well as additive models – i.e. those learners which could be successfully fit on our data.

Our manual implementation of the g-formula followed Schomaker et al. [37, Appendix]. Briefly, we used semi-parametric additive linear and logistic regression models to model the conditional densities in (1) and (2); To allow for flexible disease progression depending on how sick children are when they present at their first visit, interactions of baseline characteristics (represented in categories) with all other variables were added. Depending on the functional form of the covariates these interactions were either linear or non-linear (using penalized splines). Model selection by means of generalized cross validation was used to improve the bias-variance tradeoff in these models. All models implicitly assume that time-dependent risk factors measured before or on time  $t - 2$  do not predict the respective outcome. Note that the same approach of model building was used for our manual LTMLE estimator (iii). Sections 5.3, 5.4, and 5.5 give a thorough discussion of the issue of model choice.

## 4.2 LTMLE implementation

Both the automated and the manual implementation of the LTMLE algorithm do not use linear regression models to model the outcome, HAZ; rather, the data are transformed to lie between  $[0, 1]$  and a Quasi-binomial model ([40]) is used. This is the way the `ltmle` package deals with continuous outcomes with the aim to improve stability, particularly in the context of (near) positivity violations [41].



The clever covariate as defined in step 3 in Section 3.3 is included in the regression model not as a covariate, but as a weight [8]. This is another valid targeted maximum likelihood estimator, based on a refined loss function (and parametric submodel), and can improve stability [42]. This approach is again part of the `ltmle` implementation.

Moreover, the denominator of the clever covariate (5) is bounded in the sense that values  $< g$  are truncated at  $g$ . The automated implementation uses the default of  $g = 0.01$ , the manual implementation a stricter version of  $g = 0.05$  to deal with variable inverse probabilities; see Section 5.7 for a thorough discussion on truncation of the clever covariate.

### 4.3 Other details

Our automated implementation of the sequential g-formula, i.e. estimator (ii), follows the automated implementation of LTMLE, except that step 3 in Section 3.3 is not part of the inference procedure.

All four methods use the square root of CD4 count to work with a more symmetric variable.

### 4.4 Results

Figure 1 summarizes the results of the analyses, i.e the point and interval estimates for  $\psi_6$ ,  $\psi_{12}$ ,  $\psi_{24}$ ,  $\psi_{30}$ . For immediate ART initiation, point estimates of our manual implementation of the g-formula deviated from the other estimates by approximately -0.2 after 30 months (Figure 1, top left). The dynamic intervention of starting ART if either CD4 count  $< 750$  or CD4%  $< 25\%$  produced similar estimates for all methods, except for the manual implementation of LTMLE which produced slightly more optimistic estimates than the other methods after 30 months of follow-up.

Figure 1: Mean HAZ for different interventions, time points, and methods.

For the second dynamic intervention (Figure 1, bottom left) estimates from the manual implementations (i.e. estimators (i) and (iii)) were generally lower than the estimates from the automated implementations. One of the most interesting results can be seen for the intervention “no ART initiation”: here, automated estimation via LTMLE deviated by approximately -0.2 from all other methods (Figure 1, bottom right).

For all results confidence intervals were widest for our manual implementation of LTMLE, followed by our manual implementation of the g-formula. There is a tendency for all estimators to suggest better outcomes with earlier treatment initiation; however, the automated implementations are not so clear about this for the intervention that uses the 350/15% threshold. Also, the automated TMLE procedure suggests more drastic consequences of withholding antiretroviral treatment (“never ART”) than the other estimators.

Our manual implementations of the g-formula and TMLE suggest broadly similar growth differences between the different intervention strategies as Schomaker et al. [37], who implemented a similar analysis on a much bigger dataset using the g-formula with good performance under the natural course scenario (see also Section 5.6), but with an emphasis on mortality estimates. Our automated LTMLE procedure deviates from these results by estimating the 350/15% intervention more positive and the never ART rule more negative.

For all interventions and time points, super learning made heavy use of additive models in the Q-models (weights, averaged over time points, between 0.46 and 0.91), and moderate use in the g-models (weights, averaged over time points, between 0.16 and 0.41). It is worth pointing out that for the g-models a simple learner such as the arithmetic mean was important, with average weights between 0.19 and 0.57.

## 5 Challenges in Identifying the Causal Parameter of Interest in the Data Analysis

### 5.1 Long follow-up and sample size

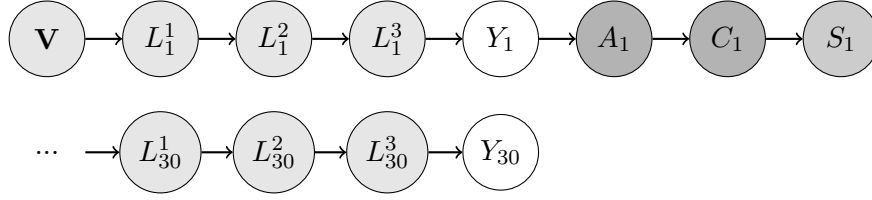
All 4 estimators under consideration require regression models (or data adaptive approaches) to be fitted on all subjects who are alive and uncensored at time  $t$ . It is evident that the longer the follow-up, the fewer subjects can be used to fit the respective models, e.g. after 30 months of follow-up only 1421 patients remain.

As we discuss below, this can be a problem as all estimators rely on correct model specifications and with smaller sample sizes it may no longer be possible to use more complicated methods, such as additive models with non-linear interactions or advanced machine learning algorithms. Moreover, this problem can potentially become even more severe in step 3 of the presented LTMLE algorithm: here, the models are estimated on the subset of patients alive, uncensored *and following treatment*  $\bar{A}_t = \bar{d}_t$ .

In our data, after 30 months of follow-up, we have 341 (13.1%) patients alive and uncensored following rule  $\bar{d}_{30}^1$ , 360 (13.8%) patients alive and uncensored following rule  $\bar{d}_{30}^2$ , 464 (17.8%) patients alive and uncensored following rule  $\bar{d}_{30}^3$ , and 259 (9.9%) patients alive and uncensored following rule  $\bar{d}_{30}^4$ . Note that after 30 months of follow-up we have evaluated 3 confounders, the outcome, as well as treatment at 11 different time points. Together with the baseline confounders this makes about 60 covariates which are potentially relevant to the respective regression models at time 30. With adding non-linear relationships, for example via penalized splines, the number of parameters that need to be estimated can therefore be close to  $n$ . The consequence is that complex estimation procedures, which may be needed for data adaptive approaches, may either fail or yield unstable solutions.

### 5.2 The implicit structural causal assumptions

For both the g-formula and the LTMLE estimators we make implicit assumptions about the time-ordering of the data. For our data analysis the assumptions can be visualized as follows:



Note that  $\mathbf{V}$  includes  $Y_0$ , and that we do not explicitly specify  $A_0$ ,  $C_0$ , and  $S_0$  here, as by definition everyone is uncensored, untreated and alive at  $t = 0$ . Equation (2) implies that the ordering of the time-dependent confounders  $\mathbf{L}_t$  can possibly affect the results of the g-formula: the confounder distribution which comes last, i.e the one referring to  $L_t^3$ , can be informed by the other confounders measured at the same time point,  $L_t^{1,2}$ , but not vice versa. While all factorizations of  $\mathbf{L}_t$  are valid, they may yield somewhat different answers in a finite sample. In our data, the obvious choice for  $L^3$  is WAZ, as it is a strong proxy for disease severity, which in turn predicts mortality [38]. It is clear that g-formula analyses benefit from sensitivity analyses with respect to the ordering of the time-dependent confounders. However, this may be very time-consuming or even unfeasible given the additional burden of bootstrapping. For LTMLE estimators the ordering of the  $L^s$  does *not* play a role, and neither is bootstrapping needed – which shows the method’s benefits in this regard.

Each time point refers to an interval and different choices of the ordering of treatment, censoring, confounder, and outcome distribution can be justified. Treatment decisions can potentially be informed by lab measurements ( $L^1, L^2$ ) and clinic measurements ( $L^3, Y$ ) which we have acknowledged in our time ordering. Since death and censoring can happen until the end of an interval, and a patient may still experience measurements before, we assume  $S$  and  $C$  to be measured at the end.

If we assume time points referring to intervals, as in our data set and most observational data analyses, time-ordering assumptions may not always be met exactly. While this point may not be of major practical relevance, it has hardly ever been acknowledged in the literature and deserves more research. Being aware of it can help making appropriate decisions in terms of data management.

### 5.3 Model selection in a structural model

The implicit ordering of variables as outlined above can be regarded as a structural model. This model is guided by prior knowledge. In our data analysis it is clear that decisions of treatment assignment are made based on CD4 count, CD4% and WHO stage, approximated by WAZ [38]. Therefore, the structure in terms of inclusion and exclusion of relevant variables is clear. The conditional expectations which need to be estimated for all three estimators should therefore, by definition, contain all specified variables from a philosophical point of view. However, as discussed in Section 5.1, this may be practically unfeasible because of reduced sample sizes for long follow-up times and many potential confounders. In this case, there is no way around model selection.

It is well-known that model selection in regression models yields typically biased estimates, over-confident confidence intervals, but possibly improved bias-variance tradeoff with respect to regression parameters and prediction [43, 44]. However, this is not of major relevance here because we are interested in making inference about  $\psi$  and not the regression parameters.

The traditional way to reduce the expected prediction error is using cross-validation [45]. LTMLE estimators make (valid) use of the extended concept of super learning, as explained above. It is worth highlighting that even after a prior decision on a pool of learners, model selection is needed, typically for tuning parameters. For example, a shrinkage estimator, say LASSO, is based on decisions around the number of possible tuning parameters and a criterion to choose it (e.g. generalized cross validation). The uncertainty around tuning parameter selection is typically ignored [46] and is one reason why the combination of the g-formula with data adaptive estimation can lead to invalid inference, i.e. confidence intervals which don’t adhere to the nominal coverage level [47].

Note that the **SuperLearner** package makes assumptions about the choice of tuning parameters, which – for a given data set – may not be ideal and would require modification.

## 5.4 Machine learning

The successful use of super learning for LTMLE estimators has been shown repeatedly in the literature [10, 20]. However, the problem for our motivating example is that particularly for longer follow-up times most of the available machine learning algorithms failed. In fact, we were only successful in using the learners described in Section 4. All other algorithms provided by the **SuperLearner** package failed. One could argue that under such circumstances the use of super learning is limited, unless one is able to and willing to write custom-made own wrappers.

Reasons for failure were manifold: (i) memory problems because of the use of complex fitting algorithms in the context of large covariate data (ii) non-convergence of fitting algorithms because of  $p \approx n$  (iii) sparse covariate data (iv) un-identified problems of the fitting algorithm (v) inability to use existing wrappers for Quasi-binomial model ([40], which are used in **ltmle** for bounded and transformed continuous outcomes) and (vi) others. Note that we need predictions from 3 prediction models/algorithms (one in step 1 and two in step 3c) for 11 time points, and for 4 interventions. Each time we work with a different subset of data with different covariate combinations. If problems occur for one model, potentially because of the specific sub-sample of data, super learning can fail as a whole.

## 5.5 Non-linear interactions

A clinical hypothesis in our data is that patients who are sicker at presentation will have a different disease trajectory from patients who are healthier at presentation. Children who arrive with severe disease will not be able to recover as fast as healthy children and will also not be able to reach the same CD4 count levels [37]. A model describing this relationship could be of the form

$$g(E(Y_t)) = \mathbf{X}\beta + f_1(L_\tau^1, L_\tau^2) + f_2(L_\tau^1, L_\tau^3) + f_3(L_\tau^2, L_\tau^3) \quad \tau \in \{1, \dots, t\}$$

where  $g(\cdot)$  represents a link function (i.e. the identity link if  $Y$  refers to HAZ),  $\mathbf{X}\beta$  refers to the parametric part of the model (e.g. binary baseline variables), and  $f_j$  are unspecified general smooth functions. Similar models can be specified if  $L_t^s$  is the outcome or when  $t$  refers to lagged variables. The major point here is that the above model is a generalized additive model in its wider sense [48]. No matter whether using the g-formula or an LTMLE estimator it is important to deal with many multiple non-linear interactions at the same time. The *R* library **mgcv** can fit models as specified above, and this is the reason why we use them in our manual implementations, i.e. estimators (i) and (iii). Using additive models in this context requires stepwise manual inclusion of non-linear interactions, if they improve the prediction error in terms of generalized cross validation, because inclusion of all possible interactions will make the fitting process too complex to converge and also results in models with too high variance.

Note that the automated estimation procedures (ii) and (iv) include the **gam** wrapper provided by the **SuperLearner** package. However, this wrapper doesn't use the **mgcv** implementation described above but the one from package **gam**. While the latter has advantages in that it uses the simple backfitting algorithm [49] for estimation, and could therefore be regarded as non-parametric rather than semi-parametric, it lacks a flexible approach of dealing with interactions. Moreover, it does not use automated smoothing parameter selection using cross validation but rather needs pre-specified knowledge for the degrees of freedom of the model. Also, the wrapper implicitly assumes no interactions. This fact can be irrelevant for super learning if enough learners are available, such that the combined weighted super learner contains the predictive information related to the interactions; however, if the number of (pre-implemented) learners that can be fitted is limited, as in our data analysis, the question arises whether it may be preferable to proceed with a manual implementation rather than a restricted automated one. This question will be answered later in this paper.

## 5.6 The natural course scenario (for the parametric g-formula)

Applying the g-formula assumes that the densities in (1) are estimated consistently. A necessary, but not sufficient, condition for correct specification is that the g-formula re-produces the raw data (at least approximately) if the treatment rule  $\bar{d}$  relates to the real treatment assignment process in the data. More specifically, the ‘‘natural course scenario’’ evaluates the stochastic treatment rule where treatment

assignment probabilities are estimated from  $\mathbb{P}(A_t = a_t | \bar{A}_{t-1} = \bar{a}_{t-1}, Y_t = y_t, \bar{\mathbf{L}}_t = \bar{\mathbf{l}}_t, \bar{C}_{t-1} = 0, S_{t-1} = 1)$  and censoring is not intervened upon but modeled via  $\mathbb{P}(C_t = 0 | \bar{A}_t = \bar{a}_t, Y_t = y_t, \bar{\mathbf{L}}_t = \bar{\mathbf{l}}_t, \bar{C}_{t-1} = 0, S_{t-1} = 1)$ . Then,  $\hat{\psi}_t$  (and also  $\hat{\bar{L}}_t^{s, \bar{d}^j}$ ) produced by the g-formula can be compared to the real data and should approximately match. Reporting the natural course scenario is a standard procedure in g-formula analyses [21, 27, 37].

Figure 2 shows the natural course scenario for our data. The mean HAZ for both the raw data (“observed data”) and the data produced by the g-formula (“no intervention”) are somewhat different. The graph suggests that there might be model mis-specification as the estimates of the natural intervention are lower than they should be. This may partly explain the results in Figure 1 (bottom left, top left).

Figure 2: Natural course scenario

## 5.7 Inverse probabilities, truncation and positivity

As opposed to the g-formula, LTMLE estimators require models for the treatment and censoring mechanism. This mechanism is reflected in the clever covariate used in step 3c of the LTMLE algorithm. While the LTMLE estimator is less sensitive than traditional inverse probability of treatment estimators [47] it nevertheless includes inverse probability weights in its fitting process which may lead to increased variance estimates when the weights become more variable [32].

For our motivating data example we have 22 treatment models relating to both the treatment and censoring mechanism of 11 time points. Ideally the fitted *cumulative* inverse probabilities are not too small, and values of the clever covariates are therefore not too large. Small estimated probabilities may indicate near-positivity violations; i.e. if there are covariate strata where the (estimated) probability of (not) receiving treatment is close to zero, for example when an extremely sick patient has not been treated although this would have been expected according to the data of similar patients. Practically positivity violations occur easily when dealing with many continuous confounders as in our data; but if the data is still “rich” enough both the g-formula and the LTMLE estimator may be able to deal with it and extrapolate based on the available data and estimated models [47].

One way to deal with small probabilities is to truncate them at a lower bound  $g$ . This reduces variability and improves stability, however at the cost of bias if the truncation is too strong [41].

Table 1 summarizes mean clever covariates (i.e. cumulative inverse weights at time 30) for different model specifications and interventions. Their mean should be about 1. More importantly, the percentage of truncated probabilities (here, [cumulative] probabilities  $< 1\%$  are truncated), should be as small as possible.

Table 1: Mean clever covariate for LTMLE estimators (at  $t = 30$ ) corresponding to different interventions and model specifications. The percentage of truncated values refers to the proportion of probabilities  $< 1\%$ .

Intervention	N	mean clever covariate				%truncated			
		SL	SL*	GAM	GLM	SL	SL*	GAM	GLM
no ART	259	3.32	5.74	8.64	10.33	5.3	13.5	73.4	100
350/15%	464	0.91	0.88	8.97	8.66	0	0	35.6	44.6
750/25%	360	0.87	0.56	4.91	2.00	0.3	0.3	19.7	10.6
immediate ART	341	0.82	0.42	4.28	0.53	0	0	16.7	0

Normal GLM’s which do not use any transformations or interactions and include covariate history up



to time  $t - 2$  yield highly variable results, i.e. the mean clever covariate ranges from 0.53 to 10.33 and between 0% and 100% of truncated observations. We know that these models are likely mis-specified, but they yield clever covariates that shouldn't cause a problem for the fitting process of step 3. Our manual implementation of TMLE [estimator (iii)], which uses additive models with non-linear interactions ("GAM"), leads to unstable weights with means that are far too high ( $> 4$ ) and many small probabilities which have been truncated. Using super learning ("SL") yields good mean clever covariates for 3 interventions, but somewhat too high values for the intervention "no ART", which is important to note given the volatile results according to Figure 1. Adding an additional learner ("SL\*"), i.e. random forests, does not improve the super learner as seen by the higher levels of truncation. This may be because this learner has fitting problems for later follow-up points. This highlights that poor learners can decrease the quality of the super learner in finite samples.

Table 2 lists point and interval estimates of LTMLE estimators ( $t = 30$ ), as well as the corresponding mean clever covariate and truncation percentages, for different model specifications for intervention  $\bar{d}^4$ .

Table 2: Point and interval estimates of LTMLE estimators ( $t = 30$ ), as well as the corresponding mean clever covariate, for different model specifications and truncation definitions. Results are provided for the intervention "no ART".

LTMLE	Estimates		mean CC		% truncated	
	g=0.01	g=0.05	g=0.01	g=0.05	g=0.01	g=0.05
GLM	-2.00 (-3.52; -0.47)	-2.00 (-2.31; -1.68)	10.33	2.06	100	100
GAM	-2.09 (-3.13; -1.06)	-2.07 (-2.32; -1.83)	8.64	1.99	73.4	92.3
SL	-2.25 (-2.42; -2.09)	-2.20 (-2.31; -2.09)	3.32	1.96	5.3	38.1
SL*	-2.12 (-2.21; -2.04)	-2.12 (-2.20; -2.04)	5.74	2.42	13.5	69.2

It can be seen that model specification has a considerable impact on the results. While truncation doesn't effect the point estimates much, it is worth noting that an improved mean clever covariate comes at the cost of truncation levels way above 35%, even with super learning. This raises the question whether positivity violations are so severe that one wouldn't trust any of the provided estimates for the fourth intervention of interest. Interestingly, our manual implementation of LTMLE (using additive models) yields estimates similar to the g-formula – in spite of the poor weights. This can be explained as follows: the cumulative probabilities are so small that most of them get truncated and the clever covariate can therefore hardly update the initial g-formula estimate in step 3 of the TMLE algorithm.

## 6 Simulation Study

The following simulation is meant to investigate under which conditions effect identification with LTMLE may be difficult, i.e. to what degree small samples, limited sets of learners, overuse of truncation and intervention definitions may affect its performance. Based on our data analysis and the discussions in Section 5 we generate longitudinal data ( $t = 0, 1, \dots, 12$ ) for 3 time-dependent confounders, as well as baseline data for 7 variables, using structural equations and the R-package `simcausal` [50]. Baseline data resembles region, sex, age, CD4 count, CD4%, WAZ and HAZ respectively ( $V^1, V^2, V^3, L_0^1, L_0^2, L_0^3, Y_0$ ). Follow-up data refers to CD4 count, CD4%, WAZ and HAZ ( $L_t^1, L_t^2, L_t^3, Y_t$ ), as well as antiretroviral treatment ( $A_t$ ) and censoring ( $C_t$ ). For simplicity, no deaths are assumed. The data generating mechanism is described in detail in Appendix D. It leads to the following baseline values: region A = 75.5%; male sex = 51.2%; mean age = 3.0 years; mean CD4 count = 672.5; mean CD4% = 15.5%; mean WAZ = -1.5; mean HAZ = -2.5. At  $t = 12$  the mean of CD4 count, CD4%, WAZ and HAZ are 1092, 27.2%, -0.8, -1.5 respectively. The target quantities  $\psi_{j,t}$  are defined as the expected value of  $Y$  at time  $t$ ,  $t \in \{6, 12\}$ , under no censoring, for a given treatment rule  $\bar{d}^j$ , where

$$\bar{d}_t^1 = \{c_t = 0; \text{ and } a_t = 1 \text{ for } \forall t$$

$$\bar{d}_t^2 = \begin{cases} c_t = 0; & \text{and } a_t = 1 & \text{if } \text{CD4 count}_t^{\bar{d}^2} < 750 \text{ or } \text{CD4}\%_t^{\bar{d}^2} < 25\% \text{ or } \text{WAZ}_t^{\bar{d}^2} < -2 \\ c_t = 0; & \text{and } a_t = 0 & \text{otherwise} \end{cases}$$

$$\bar{d}_t^3 = \begin{cases} c_t = 0; & \text{and } a_t = 1 & \text{if } \text{CD4 count}_t^{\bar{d}^3} < 350 \text{ or } \text{CD4}\%_t^{\bar{d}^3} < 15\% \text{ or } \text{WAZ}_t^{\bar{d}^3} < -2 \\ c_t = 0; & \text{and } a_t = 0 & \text{otherwise} \end{cases}$$

$$\bar{d}_t^4 = \{c_t = 0; \text{ and } a_t = 0 \text{ for } \forall t\}$$

The target parameters are  $\psi_{1,6} = -1.804$ ,  $\psi_{2,6} = -1.850$ ,  $\psi_{3,6} = -2.028$ ,  $\psi_{4,6} = -2.452$ ,  $\psi_{1,12} = -1.029$ ,  $\psi_{2,12} = -1.119$ ,  $\psi_{3,12} = -1.466$ , and  $\psi_{4,12} = -2.447$ .

For  $\mathcal{R} = 1000$  simulation runs we estimate the above parameters using `ltmle` under the following specifications:  $n \in \{200, 600, 1000\}$ ; truncation levels  $g \in \{0.01, 0.025, 0.04\}$  and three different sets of learners. Learner 1 consists of simple GLM's, learner 2 adds both the arithmetic mean and GLM's with interactions to the set of learners, and learner 3 contains additionally GLM's selected by AIC, standard additive models (no interactions, package `gam`), as well as GLM's based on an EM-algorithm-Bayesian model fitting. For computational feasibility, all fitted models included only variables measured at time  $t$ ,  $t - 1$ , and 0 – in line with the data generating process.

The average useable size of the sample at the end of follow-up (i.e. excluding censored patients and only including those who followed the respective treatment rule) varied between 39 (intervention  $\bar{d}^4$ ) and 79 (intervention  $\bar{d}^2$ ) for a baseline sample size of 200; between 116 (intervention  $\bar{d}^4$ ) and 238 (intervention  $\bar{d}^2$ ) for a baseline sample size of 600; and between 190 (intervention  $\bar{d}^4$ ) and 396 (intervention  $\bar{d}^2$ ) for a baseline sample size of 1000.

After 12 months of follow-up, for all interventions super learning (for learner 3) made heavy use of additive models when estimating the Q-model: the respective super learner weight, averaged over time and simulation runs, was always greater than 0.73. The g-models made particular use of GLM's selected by AIC.

Figure 3: Estimated bias and coverage probability, at  $t = 12$ , for different truncation levels, interventions, learner sets, and sample sizes.

Figure 3 shows the estimated bias and coverage probability, for different truncation levels, interventions, learner sets, and sample sizes.

One can see that the bias varies substantially with respect to the chosen intervention. Interestingly the interventions with the biggest data support (usable sample size) and with least data support,  $\bar{d}^4$  and  $\bar{d}^2$ , have the highest bias, showing that neither sample size nor type of intervention (static/dynamic) offers a reasonable guideline to judge the quality of the estimated target quantity. Those interventions with high bias have, obviously, also a poorer coverage.

In this particular simulation setup the level of truncation hardly affects the results. There are gains with a greater sample size, as one would expect; it is however worth pointing out that even with small sample sizes results are relatively stable and good.

Of course, learner 3, with the biggest set of learners, typically performs best. Surprisingly the other two learners also yield good results – which is encouraging given that the data-generating process contains interactions and non-linear associations as well, and this isn't specifically modeled by learner set 1 and learner set 2.

Since in the simulation the data-generating process is known, we can estimate  $P(A_t = \bar{d}_t | \bar{\mathbf{L}}_t = \bar{\mathbf{L}}_t, \bar{A}_{t-1} = \bar{d}_{t-1})$ , i.e the probability of continuing to receive treatment according to the assigned treatment rule, given that a patient has received treatment so far and irrespective of the covariate history. That is,

we can look at the indicator variable whether a patient has been treated according to the rule of interest at each time point  $t$ , or not, given the past covariates and conditional on those which were treated according to the rule at  $t - 1$ . If a substantial proportion of subjects have very small cumulative probabilities, then this could be a sign of positivity violations. Table 3 lists the summaries of the cumulative probabilities.

Table 3: Proportion of cumulative probabilities of continuing to receive treatment according to the assigned treatment rule (given that this has been done so far and irrespective of the covariate history) which are smaller than 0.01. The numbers are calculated based on the data-generating process in the simulation, and are estimated from the data set as well, and for both at the end of follow-up.

	$\bar{d}_t^1$	$\bar{d}_t^2$	$\bar{d}_t^3$	$\bar{d}_t^4$
simulation	0.1%	0.3%	0.2%	0.8%
data	0%	0.8%	0.4%	1.9%

The table clearly shows that the highest proportion of small cumulative probabilities is found for the second and fourth intervention, i.e. those interventions which performed worst in the simulations. This highlights the fact that interventions need data support to meet the positivity assumption. In general this can be also estimated in a data analysis, for a given estimation technique. When using a standard logistic regression model we obtain the highest proportion of small cumulative probabilities for the fourth intervention, i.e. the intervention that was most volatile in the data analysis. We suggest that in addition to summaries of the clever covariates and percentage of truncated values, the data support for each intervention, as presented in Table 3 should also serve as a diagnostic tool in LTMLE analyses.

## 7 Conclusions

To our knowledge this is the first successful implementation of LTMLE for a data set with  $> 10$  follow-up times *and* at least 3 time-dependent confounders (which are affected by prior treatment). Moreover, our data analysis was based on a realistic data set relevant to public health policy making. Large longitudinal data sets require specific care when using LTMLE: we have shown that long follow-up may come with sample sizes that are small compared to the number of parameters that need to be fitted for machine learning algorithms, including semi-parametric regression models. This can yield instabilities, memory problems and identifiability problems that need to be addressed.

We have argued that there is practically no way around model selection in causal inference with both the g-formula and LTMLE, despite the implicit assumption that some of the included confounders from the structural causal model are irrelevant. This is because the causal parameter of interest can be best identified with good predictions coming from an optimal bias-variance tradeoff. This model selection process could incorporate clinical knowledge. We have included non-linear interactions using additive model to address such hypotheses. While this may have potential benefits for the g-formula, our LTMLE estimator could not be improved by such a strategy. In fact, a small super learner, with few algorithms, led to more reliable and precise results.

Ideally large longitudinal causal analyses are best analyzed by several methods. If the results match, then this is perfect. If not, diagnostics such as clever covariate and truncation summaries for LTMLE, or the natural course scenario for the g-formula are indispensable. Our analyses have shown that neither a good natural course scenario for the g-formula nor considerations of sample size and type of intervention (static/dynamic) offers a reasonable guideline to judge the quality of the estimated target quantity; what matters is the support in the data for the intervention rule of interest: high levels of truncation and small cumulative probabilities to continue to follow the rule of interest are always a sign of problems, likely due to positivity violations or poor modeling strategies. They can be addressed by an increase of appropriate learners, sometimes even if they are simple, or – less preferably – a change in the target quantity. Whatever the methodological choices are, it remains important to consider the substantive question at hand as well; for example, in our analysis, the most conservative estimates are provided by

the g-formula, and a clinician may prefer to base her/his decisions on these estimates if the diagnostics for the method are acceptable.

We conclude the following with respect to the questions posed in the introduction:

- it is feasible to implement LTMLE in complex settings with long follow-up times, small sample size, multiple time-dependent confounders, and dynamic interventions.
- as seen in both the data analysis and the simulations, LTMLE can be successful even when confronted with limited modeling options, for example by a limited number of applicable learners when using super learning. Nevertheless, it would be beneficial to have more learners available that work with existing software packages and are computationally feasible in settings with long follow-up and therefore multiple variables. We have made first attempts to address this issue available [51]; and researchers with advanced programming capacities are advised to implement their own custom-made wrappers, tailored to their own analysis of interest, if possible.
- in our setting we couldn't find strong evidence that a flexible additive model, informed by prior clinical knowledge, possibly in conjunction with competing estimators such as the g-formula, may perform better than an automated LTMLE procedure with a limited set of learners.
- in the settings studied, a reduced sample size or truncation of inverse probabilities didn't have a major impact on the results; but it could be clearly seen that different interventions may have different support in the data and that the success of estimators such as LTMLE also varied with respect to the chosen intervention. This highlights the need to develop more diagnostic tools to diagnose practical positivity violations in the data. Reporting the percentage of truncated cumulative inverse probabilities, summaries of the clever covariates and the data support for each intervention (see Section 6) is highly recommended; but more research and guidance in this respect is needed.

## A Identification of our causal parameter of interest

We consider time-ordered data as discussed in Section 3. In general, we assume  $S_{t-1} \rightarrow L_t^1 \rightarrow L_t^2 \rightarrow L_t^3 \rightarrow Y \rightarrow A \rightarrow C \rightarrow S_t$ , and for  $t = 0$  we can write  $\underbrace{\mathbf{V} \rightarrow Y_0}_{\mathbf{L}_0} \rightarrow \underbrace{A_0 \rightarrow C_0}_{\mathbf{A}_0} \rightarrow S_0 \rightarrow \mathbf{L}_1 \rightarrow Y_1$ , taking

into account baseline variables. Note that  $A_0$  and  $C_0$  are 0 and ‘uncensored’ for everyone by definition and we don't intervene upon them, and thus  $\mathbf{L}_1$  could be considered as baseline as well. However, for the general case, we show the identification of the causal parameter for the two time points  $t = 0, 1$ , then  $t = t^*, t^* - 1$ , and then infer to the sequence  $t = 0, 1, \dots, T$ . The assumptions and definitions used are the rule of iterated nested expectations (“IE”), consistency (“cons.”), and exchangeability (“exch.”) as defined in Section 3.

(i) for  $t = 0, 1$  it holds that

$$\begin{aligned}
\mathbb{E}(Y_1^{d_0} | S_0 = 1) &\stackrel{\text{IE}}{=} \mathbb{E}(\mathbb{E}(Y_1^{d_0} | S_0 = 1, \mathbf{L}_0) | S_0 = 1) \\
&\stackrel{\text{IE}}{=} \mathbb{E}(\mathbb{E}(\mathbb{E}(Y_1^{d_0} | S_0 = 1, \mathbf{L}_0, \mathbf{A}_0) | S_0 = 1, \mathbf{L}_0) | S_0 = 1) \\
&\stackrel{\text{exch.}}{=} \mathbb{E}(\mathbb{E}(\mathbb{E}(Y_1^{d_0} | S_0 = 1, \mathbf{L}_0, \mathbf{A}_0 = d_0) | S_0 = 1, \mathbf{L}_0) | S_0 = 1) \\
&\stackrel{\text{IE}}{=} \mathbb{E}(\mathbb{E}(\mathbb{E}(\mathbb{E}(Y_1^{d_0} | S_0 = 1, \mathbf{L}_0, \mathbf{L}_1, \mathbf{A}_0 = d_0) | S_0 = 1, \mathbf{L}_0, \mathbf{A}_0 = d_0) | S_0 = 1, \mathbf{L}_0) | S_0 = 1)) \\
&\stackrel{\text{cons.}}{=} \mathbb{E}(\mathbb{E}(\mathbb{E}(\mathbb{E}(Y_1 | S_0 = 1, \bar{\mathbf{L}}_1, \mathbf{A}_0 = d_0) | S_0 = 1, \mathbf{L}_0, \mathbf{A}_0 = d_0) | S_0 = 1, \mathbf{L}_0) | S_0 = 1))
\end{aligned}$$

(ii) for  $t = t^*, t^* - 1$  it holds that

$$\begin{aligned}
\mathbb{E}(Y_{t^*}^{\bar{d}_{t^*-1}} | S_{t^*-1} = 1) &\stackrel{\text{IE}}{=} \mathbb{E}(\mathbb{E}(Y_{t^*}^{\bar{d}_{t^*-1}} | S_{t^*-1} = 1, \mathbf{L}_{t^*-1}) | S_{t^*-1} = 1) \\
&\stackrel{\text{IE}}{=} \mathbb{E}(\mathbb{E}(\mathbb{E}(Y_{t^*}^{\bar{d}_{t^*-1}} | S_{t^*-1} = 1, \mathbf{L}_{t^*-1}, \mathbf{A}_{t^*-1}) | S_{t^*-1} = 1, \mathbf{L}_{t^*-1}) | S_{t^*-1} = 1)
\end{aligned}$$

$$\begin{aligned}
&\stackrel{\text{exch.}}{=} \mathbb{E}(\mathbb{E}(\mathbb{E}(Y_{t^*}^{\bar{d}_{t^*-1}} | S_{t^*-1} = 1, \mathbf{L}_{t^*-1}, \mathbf{A}_{t^*-1} = \bar{d}_{t^*-1}) | S_{t^*-1} = 1, \mathbf{L}_{t^*-1}) | S_{t^*-1} = 1) \\
&\stackrel{\text{IE}}{=} \mathbb{E}(\mathbb{E}(\mathbb{E}(\mathbb{E}(Y_{t^*}^{\bar{d}_{t^*-1}} | S_{t^*-1} = 1, \mathbf{L}_{t^*-1}, \mathbf{L}_{t^*}, \mathbf{A}_{t^*-1} = \bar{d}_{t^*-1}) | S_{t^*-1} = 1, \mathbf{L}_{t^*-1}, \\
&\quad \mathbf{A}_{t^*-1} = \bar{d}_{t^*-1}) | S_{t^*-1} = 1, \mathbf{L}_{t^*-1}) | S_{t^*-1} = 1)) \\
&\stackrel{\text{cons.}}{=} \mathbb{E}(\mathbb{E}(\mathbb{E}(Y_{t^*} | S_{t^*-1} = 1, \bar{\mathbf{L}}_{t^*}, \mathbf{A}_{t^*-1} = \bar{d}_{t^*-1}) | \\
&\quad S_{t^*-1} = 1, \mathbf{L}_{t^*-1}, \mathbf{A}_{t^*-1} = \bar{d}_{t^*-1}) | S_{t^*-1} = 1, \mathbf{L}_{t^*-1}) | S_{t^*-1} = 1))
\end{aligned}$$

(iii) Therefore, by induction, a time-sequence  $t = 0, \dots, T$ , and for any general rule  $\bar{d}$ , it follows that

$$\begin{aligned}
\psi_T &= \mathbb{E}(Y_T^{\bar{d}} | S_{T-1}^{\bar{d}} = 1) \\
&= \mathbb{E}(\mathbb{E}(\mathbb{E}(\dots \mathbb{E}(\mathbb{E}(Y_T | \mathbf{A}_{T-1} = \bar{d}_{T-1}, \bar{\mathbf{L}}_T, S_{T-1} = 1) | \mathbf{A}_{T-2} = \bar{d}_{T-2}, \bar{\mathbf{L}}_{T-1}, S_{T-1} = 1) \dots | \\
&\quad S_{T-1} = 1, \mathbf{L}_0, \mathbf{A}_0 = d_0) | S_{T-1} = 1, \mathbf{L}_0) | S_{T-1} = 1)).
\end{aligned}$$

## B R code for automated estimation of $\Psi_6$ with ltmle

To use the `ltmle` function from the package `ltmle`, one needs to provide the data in wide format. Below we show 5 patients of our data set, where *b1*, *b2*, *b3*, *b4* refer to the baseline values of  $\sqrt{\text{CD4 count}}$ , CD4%, WAZ, and HAZ. Variable names ending with “.t” refer to measurements at time *t*. A.0 and C.0 are not needed and reported as by definition everyone is ART-naive and uncensored at baseline. The variable ‘dead.at’ indicates the time interval in which a person died (and is 33 if the person didn’t die within the first 30 months of follow-up). The choice of the data structure in terms of the ordering of the variables is discussed in Section 5.2.

	gender	age_fv	region2	b1	b2	b3	b4	cd4a_cf.1	cd4p_cf.1	wfa_cf.1	hfa_cf.1
1	1	4.752909	2	23.87467	15	-0.37	-0.30	23.87467	15	-0.37	-0.30
2	1	2.450377	2	26.98148	14	-0.81	-2.52	26.98148	14	-0.81	-2.52
3	1	2.318960	2	35.66511	14	-2.19	-3.13	35.66511	14	-2.19	-3.13
4	1	2.275154	2	34.46738	13	-2.13	-2.42	34.46738	13	-2.29	-2.38
5	1	2.091718	2	15.26434	6	-1.85	-1.62	15.26434	6	-2.02	-1.92
	ART.1	censored.1	cd4a_cf.3	cd4p_cf.3	wfa_cf.3	hfa_cf.3	ART.3	censored.3	cd4a_cf.6		
1	0	uncensored	23.87467	15	-0.50	-0.53	1	uncensored	23.87467		
2	0	uncensored	26.98148	14	-1.30	-2.20	1	uncensored	26.98148		
3	0	uncensored	35.66511	14	-2.54	-3.08	1	uncensored	35.66511		
4	1	uncensored	34.46738	13	-2.41	-2.28	1	uncensored	36.63332		
5	1	uncensored	15.26434	6	-2.14	-2.13	1	uncensored	22.40536		
	cd4p_cf.6	wfa_cf.6	hfa_cf.6	dead.at							
1	15	-0.50	-0.53	33							
2	14	-1.36	-1.99	33							
3	14	-2.72	-3.36	33							
4	17	-2.41	-2.28	33							
5	21	-1.88	-1.34	33							

Before applying the `ltmle` command we need to define the intervention matrix. For a static intervention, this could just be a vector such as  $c(0, 0, 0)$ , but needs to be a matrix for dynamic interventions as shown below. For illustrative purposes we use the intervention  $\bar{d}^2$ : “starting ART if either CD4 count  $< 750$  or CD4%  $< 25\%$ ” below. We also need to define the libraries used for the super learner, see Section 5.4 for more details.

```
# Define Intervention: start ART if CD4 count < 750 or CD4% < 25%
cd4_750.1 <- (mydata_wide$cd4a_cf.1 < sqrt(750) | mydata_wide$cd4p_cf.1 < 25)
cd4_750.3 <- (mydata_wide$cd4a_cf.3 < sqrt(750) | mydata_wide$cd4p_cf.3 < 25) | cd4_750.1
cd4_750.6 <- (mydata_wide$cd4a_cf.6 < sqrt(750) | mydata_wide$cd4p_cf.6 < 25) | cd4_750.3
abar750 <- as.matrix(cbind(cd4_750.1,cd4_750.3,cd4_750.6))
abar750[is.na(abar750)] <- 0

# Define collection of estimation methods to be used by SuperLearner
mylibrary <- list(Q=c("SL.mean","SL.glm","SL.stepAIC","SL.step.interaction","SL.gam","SL.bayesglm"),
                 g=c("SL.mean","SL.glm","SL.stepAIC","SL.gam","SL.bayesglm"))
```

Following (4),  $\Psi_6$  can be estimated easily on the data of survivors, as follows:

```
ltmle(mydata_wide[dead.at>=6,],
      Anodes=c("ART.1","ART.3","ART.6"),
      Cnodes=c("censored.1","censored.3","censored.6"),
      Lnodes=c("cd4a_cf.1","cd4p_cf.1","wfa_cf.1","cd4a_cf.3","cd4p_cf.3","wfa_cf.3",
               "cd4a_cf.6","cd4p_cf.6","wfa_cf.6"),
      Ynodes=c("hfa_cf.1","hfa_cf.3","hfa_cf.6"), Yrange=c(-15,15),
      Qform=NULL, gform=NULL, deterministic.g.function=MaintainTreatment, abar=abar750[dead.at>=6,],
      SL.library=mylibrary, estimate.time=T, variance.method="ic", gbounds = c(0.01, 1)))
```

Note that `Anodes`, `Lnodes`, `Cnodes`, `Ynodes` refer to the treatment, censoring, covariate, and outcome data respectively. By specifying a `SL.library`, and leaving the arguments for model formulae (`Qform`, `gform`) empty, model specification is done by super learning. `Yrange=c(-15,15)` ensures that the outcome is bounded (i.e. absolute z-scores greater than 15 are regarded to be impossible).

## C R code for manual LTMLE estimation of $\Psi_6$

The below implementation has been designed to be able to exactly reproduce the steps of the `ltmle` command; but to adhere to the framework and notation of Section 3 at the same time. As noted in the manuscript, we work with a transformed outcome and quasi-binomial models as `ltmle` does. We also add the clever covariate as weights, as `ltmle` does. Furthermore, we use additive models based on the package `mgcv` for model fitting. For comparing the manual and automated implementation, one could simply use simple GLMs (no super learning) and specify the model form in the `Qform` and `gform` arguments. The below code to truncate weights and to calculate confidence intervals is used later on in the main code.

```
gbounds <- function(x,bounds=c(0.05,1)){
  x[x<min(bounds)] <- min(bounds)
  x[x>max(bounds)] <- max(bounds)
  return(x)
}

CI <- function(est,infcurv){
  n <- length(infcurv)
  ciwidth <- 1.959964 * sd(infcurv)/sqrt(n)
  CI <- c(est-ciwidth, est+ciwidth)
  return(CI)
}

library(mgcv)
```

To estimate  $\psi_6$  we need to do the following:



```

# Data of survivors
mydata_wide6 <- mydata_wide[dead.at>=6,]
attach(mydata_wide6)

# Already prepare step 3c: fit treatment and censoring models
m1_c <- mgcv::gam(I(censored.1=="censored") ~ age_fv + s(b2) + s(b3) + s(b4), family=binomial)
m3_c <- mgcv::gam(I(censored.3=="censored") ~ gender + s(age_fv) + wfa_cf.1
+ s(I((hfa_cf.1 +15)/30)), family=quasibinomial)
m6_c <- mgcv::gam(I(censored.3=="censored") ~ s(b4) + s(wfa_cf.3) + I((hfa_cf.3 +15)/30),
family=quasibinomial)

m1_a <- mgcv::gam(ART.1 ~ s(cd4a_cf.1) + s(cd4p_cf.1) + s(wfa_cf.1) + s(b1) + s(b2) + s(b3),
data=mydata_wide6, family=binomial)
m3_a <- mgcv::gam(ART.3 ~ s(cd4p_cf.3) + s(wfa_cf.3) + s(cd4a_cf.1) + s(cd4p_cf.1) + s(b1) +
s(b2) + s(b3) + age_fv,
data=mydata_wide6[censored.1==1,], family=quasibinomial)
m6_a <- mgcv::gam(ART.6 ~ s(cd4a_cf.6) + s(cd4p_cf.6) + s(wfa_cf.6) + s(cd4p_cf.3) + s(cd4a_cf.3)
+ s(b3) + age_fv, data=mydata_wide6[censored.3==1,],family=quasibinomial)

# Already prepare step 3c: define subsets of patients uncensored and following the treatment rule
subset1 <- censored.1=="uncensored" & ART.1==abar[dead.at>=6,1]
subset3 <- censored.1=="uncensored" & ART.1==abar[dead.at>=6,1] & censored.3=="uncensored" &
ART.3==abar[dead.at>=6,2]
subset6 <- censored.1=="uncensored" & ART.1==abar[dead.at>=6,1] & censored.3=="uncensored" &
ART.3==abar[dead.at>=6,2] & censored.6=="uncensored" & ART.6==abar[dead.at>=6,3]

# Still prepare data for step 3c: calculate the cumulative treatment and censoring probabilities;
# also: apply cutoff at 0.05, i.e. probabilities <0.05 are defined as 0.05 -> clever covariate <= 20
p1a <- gbounds((ART.1 == 1)*predict(m1_a, type="response", newdata=mydata_wide6) +
(ART.1 == 0)*(1 - predict(m1_a, type="response", newdata=mydata_wide6)))
p3a <- gbounds((ART.3 == 1)*predict(m3_a, type="response", newdata=mydata_wide6) +
(ART.3 == 0)*(1 - predict(m3_a, type="response", newdata=mydata_wide6)))
p6a <- gbounds((ART.6 == 1)*predict(m6_a, type="response", newdata=mydata_wide6) +
(ART.6 == 0)*(1 - predict(m6_a, type="response", newdata=mydata_wide6)))

p1c <- gbounds((censored.1 == "censored")*predict(m1_c, type="response", newdata=mydata_wide6)+
(censored.1 == "uncensored")*(1 - predict(m1_c, type="response",newdata=mydata_wide6)))
p3c <- gbounds((censored.3 == "censored")*predict(m3_c, type="response", newdata=mydata_wide6)+
(censored.3 == "uncensored")*(1 - predict(m3_c, type="response",newdata=mydata_wide6)))
p6c <- gbounds((censored.6 == "censored")*predict(m6_c, type="response", newdata=mydata_wide6)+
(censored.6 == "uncensored")*(1 - predict(m6_c, type="response",newdata=mydata_wide6)))

# Still prepare data for step 3c: calculate clever covariate (=0 [weight 0] for those excluded)
mydata_wide6$cc1 <- 0
mydata_wide6$cc3 <- 0
mydata_wide6$cc6 <- 0
mydata_wide6$cc1[subset1] <- (1/(gbounds(p1a*p1c)))[subset1]
mydata_wide6$cc3[subset3] <- (1/(gbounds(p1a*p1c*p3a*p3c)))[subset3]
mydata_wide6$cc6[subset6] <- (1/(gbounds(p1a*p1c*p3a*p3c*p6a*p6c)))[subset6]

# Prepare for step 2: Define Intervention
newdata_wide6 <- mydata_wide6
newdata_wide6$ART.1 <- abar[dead.at>=6,1]
newdata_wide6$ART.3 <- abar[dead.at>=6,2]
newdata_wide6$ART.6 <- abar[dead.at>=6,3]

### LTMLE estimation ###

# time point 6: start to calculate innermost expectation
m6_y <- mgcv::gam(I((hfa_cf.6 +15)/30) ~ s(age_fv) + s(b2) + s(b3) + s(b4) + ART.1 + s(cd4p_cf.3)
+ s(wfa_cf.3, by=b3) + s(I((hfa_cf.3 +15)/30)), data=mydata_wide6[censored.3==
"uncensored" & censored.1=="uncensored",], family=quasibinomial) # STEP 1
mydata_wide6$ov6 <- predict(m6_y, newdata=newdata_wide6,type="response") # STEP 2
Qk6 <- glm("I((hfa_cf.6 +15)/30)~ 1 + offset(qlogis(ov6)) ", data=mydata_wide6, weights=cc6,

```

```

                                family = quasibinomial)                # STEP 3a-c
mydata_wide$y3.Q <- predict(Qk6, type="response",newdata=mydata_wide6)  # STEP 3d

# time point 3
m3_y <- mgcv::gam(y3.Q ~ (age_fv) + s(b3) + s(b4) + ART.1 + s(wfa_cf.1,by=b3)
               + s(I((hfa_cf.1 +15)/30),by=b3), data=mydata_wide6[censored.1=="uncensored"],
               family=quasibinomial)                                # STEP 1
mydata_wide$ov3 <- predict(m3_y, newdata=newdata_wide6,type="response") # STEP 2
Qk3 <- glm("y3.Q ~ 1 + offset(qlogis(ov3)) ", data=mydata_wide6, weights=cc3, # STEP 3a-c
          family = quasibinomial)
mydata_wide$y1.Q <- predict(Qk3, type="response",newdata=mydata_wide6)  # STEP 3d

# time point 1+0: calculate final TMLE estimate
# (as everything before ART.1/hfa_cf.1 is essentially baseline due to no prior intervention)
est_ltmle.6 <- mean(mydata_wide$y1.Q)*30-15                          # STEP 4

# Confidence Interval: calculate IC and set HAZ=0 if unmeasured at time t # Step 5
mydata_wide6$Y <- (mydata_wide6$hfa_cf.6+15)/30
mydata_wide6$Y[is.na(mydata_wide6$Y)]<-0
mydata_wide6$y3.Q[is.na(mydata_wide6$y3.Q)]<-0
IC <- mydata_wide6$cc3*(mydata_wide6$Y-mydata_wide6$y3.Q) + mydata_wide6$cc1*
      (mydata_wide6$y3.Q-mydata_wide6$y1.Q) + mydata_wide6$y1.Q-mean(mydata_wide6$y1.Q)
IC <- IC*30-15
CI(est_ltmle.6,IC) # Final Confidence Interval

```

## D Data generating process in the simulation study

Both baseline data ( $t = 0$ ) and follow-up data ( $t = 1, \dots, 12$ ) were created using structural equations using the *R*-package `simcausal` [50]. The below listed distributions, listed in temporal order, describe the data-generating process. Baseline data refers to region, sex, age, CD4 count, CD4%, WAZ and HAZ respectively ( $V^1, V^2, V^3, L_0^1, L_0^2, L_0^3, Y_0$ ). Follow-up data refers to CD4 count, CD4%, WAZ and HAZ ( $L_t^1, L_t^2, L_t^3, Y_t$ ), as well as a treatment ( $A_t$ ) and censoring ( $C_t$ ) indicator. In addition to Bernoulli ( $B$ ), uniform ( $U$ ) and normal ( $N$ ) distributions, we also use truncated normal distributions which are denoted by  $N_{[a,b]}$  where  $a$  and  $b$  are the truncation levels. Values which are smaller  $a$  are replaced by a random draw from a  $U(a_1, a_2)$  distribution and values greater than  $b$  are drawn from a  $U(b_1, b_2)$  distribution. Values for  $(a_1, a_2, b_1, b_2)$  are  $(0, 50, 5000, 10000)$  for  $L^1$ ,  $(0.03, 0.09, 0.7, 0.8)$  for  $L^2$ , and  $(-10, 3, 3, 10)$  for both  $L^3$  and  $Y$ . The notation  $\bar{\mathcal{D}}$  means “conditional on the data that has already been measured (generated) according the the time ordering”.

For  $t = 0$ :

$$\begin{aligned}
V^1 &\sim B(p = 4392/5826) \\
V^2|\bar{\mathcal{D}} &\sim \begin{cases} B(p = 2222/4392) & \text{if } V^1 = 1 \\ B(p = 758/1434) & \text{if } V^1 = 0 \end{cases} \\
V^3|\bar{\mathcal{D}} &\sim U(1, 5) \\
L_0^1|\bar{\mathcal{D}} &\sim \begin{cases} N_{[0,10000]}(650, 350) & \text{if } V^1 = 1 \\ N_{[0,10000]}(720, 400) & \text{if } V^1 = 0 \end{cases} \\
\tilde{L}_0^1|\bar{\mathcal{D}} &\sim N((L_0^1 - 671.7468)/(10 \cdot 352.2788) + 1, 0) \\
L_0^2|\bar{\mathcal{D}} &\sim N_{[0.06, 0.8]}(0.16 + 0.05 \cdot (L_0^1 - 650)/650, 0.07) \\
\tilde{L}_0^2|\bar{\mathcal{D}} &\sim N((L_0^2 - 0.1648594)/(10 \cdot 0.06980332) + 1, 0) \\
L_0^3|\bar{\mathcal{D}} &\sim \begin{cases} N_{[-5,5]}(-1.65 + 0.1 \cdot V^3 + 0.05 \cdot (L_0^1 - 650)/650 + 0.05 \cdot (L_0^2 - 16)/16, 1) & \text{if } V^1 = 1 \\ N_{[-5,5]}(-2.05 + 0.1 \cdot V^3 + 0.05 \cdot (L_0^1 - 650)/650 + 0.05 \cdot (L_0^2 - 16)/16, 1) & \text{if } V^1 = 0 \end{cases} \\
A_0|\bar{\mathcal{D}} &\sim B(p = 0) \\
C_0|\bar{\mathcal{D}} &\sim B(p = 0) \\
Y_0|\bar{\mathcal{D}} &\sim N_{[-5,5]}(-2.6 + 0.1 \cdot I(V^3 > 2) + 0.3 \cdot I(V^1 = 0) + (L_0^3 + 1.45), 1.1)
\end{aligned}$$

For  $t > 0$ :

$$\begin{aligned}
L_t^1 | \bar{\mathcal{D}} &\sim \begin{cases} N_{[0,10000]}(13 \cdot \log(t \cdot (1034 - 662)/8) + L_{t-1}^1 + 2 \cdot L_{t-1}^2 + 2 \cdot L_{t-1}^3 + 2.5 \cdot A_{t-1}, 50) & \text{if } t \in \{1, 2, 3, 4\} \\ N_{[0,10000]}(4 \cdot \log(t \cdot (1034 - 662)/8) + L_{t-1}^1 + 2 \cdot L_{t-1}^2 + 2 \cdot L_{t-1}^3 + 2.5 \cdot A_{t-1}, 50) & \text{if } t \in \{5, 6, 7, 8\} \\ N_{[0,10000]}(L_{t-1}^1 + 2 \cdot L_{t-1}^2 + 2 \cdot L_{t-1}^3 + 2.5 \cdot A_{t-1}, 50) & \text{if } t \in \{9, 10, 11, 12\} \end{cases} \\
L_t^2 | \bar{\mathcal{D}} &\sim N_{[0.06, 0.8]}(L_{t-1}^2 + 0.0003 \cdot (L_t^1 - L_{t-1}^1) + 0.0005 \cdot (L_{t-1}^3) + 0.0005 \cdot A_{t-1} \cdot \tilde{L}_0^1, 0.02) \\
L_t^3 | \bar{\mathcal{D}} &\sim N_{-5,5}(L_{t-1}^3 + 0.0017 \cdot (L_t^1 - L_{t-1}^1) + 0.2 \cdot (L_t^2 - L_{t-1}^2) + 0.005 \cdot A_{t-1} \cdot \tilde{L}_0^2, 0.5) \\
A_t | \bar{\mathcal{D}} &\sim \begin{cases} B(p=1) & \text{if } A_{t-1} = 1 \\ B(p = 1/(1 + \exp(-[-2.4 + 0.015 \cdot (750 - L_t^1) + 5 \cdot (0.2 - L_t^2) - 0.8 \cdot L_t^3 + 0.8 \cdot t]))) & \text{if } A_{t-1} = 0 \end{cases} \\
C_t | \bar{\mathcal{D}} &\sim B(p = 1/(1 + \exp(-[-6 + 0.01 \cdot (750 - L_t^1) + 1 \cdot (0.2 - L_t^2) - 0.65 \cdot L_t^3 - A_t]))) \\
Y_t | \bar{\mathcal{D}} &\sim N_{[-5,5]}(Y_{t-1} + 0.00005 \cdot (L_t^1 - L_{t-1}^1) - 0.000001 \cdot ((L_t^1 - L_{t-1}^1) \cdot \sqrt{\tilde{L}_0^1})^2 + 0.01 \cdot (L_t^2 - L_{t-1}^2) - \\
&\quad 0.0001 \cdot ((L_t^2 - L_{t-1}^2) \cdot \sqrt{\tilde{L}_0^2})^2 + 0.07 \cdot ((L_t^3 - L_{t-1}^3) \cdot (L_0^3 + 1.5135)) - 0.001 \cdot ((L_t^3 - L_{t-1}^3) \cdot (L_0^3 + 1.5135))^2 + \\
&\quad 0.005 \cdot A_t + 0.075 \cdot A_{t-1} + 0.05 \cdot A[t] \cdot A[t-1], 0.01)
\end{aligned}$$

## Acknowledgements

We are immensely grateful to Maya Petersen for her constructive and helpful feedback on earlier versions of this manuscript. We appreciate the continuous help of Joshua Schwab who helped solving multiple issues related to the implementation of LTMLE, particularly in conjunction with the *R* package `ltmle`. We further thank Matthias Aßenmacher for his dedicated help in constructing the simulation setup. We also thank Lorna Renner, Shobna Sawry, Sylvie N’Gbeche, Karl-Günter Technau, Francois Eboua, Frank Tanser, Haby Sygnate-Sy, Sam Phiri, Madeleine Amorissani-Folquet, Vivian Cox, Fla Koueta, Cleophas Chimbeta, Annette Lawson-Evi, Janet Giddy, Clarisse Amani-Bosse, and Robin Wood for sharing their data with us. We would also like to highlight the support of the International epidemiologic Databases to Evaluate AIDS Southern Africa (IeDEA-SA) and the West Africa Paediatric Working Group. The NIH has supported the above individuals, grant numbers U01AI069924 and U01AI069919. Miguel Angel Luque Fernandez is supported by the Spanish National Institute of Health, Carlos III Miguel Servet I Investigator Award (CP17/00206).

## References

- [1] S. Gruber and M. J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat*, 6(1):Article 18, 2010.
- [2] M. Pearl, L. Balzer, and J. Ahern. Targeted estimation of marginal absolute and relative associations in case-control data: An application in social epidemiology. *Epidemiology*, 27(4):512–517, 2016.
- [3] M. A. Luque-Fernandez, A. Belot, L. Valeri, G. Ceruli, C. Maringe, and B. Rachet. Data-adaptive estimation for double-robust methods in population-based cancer epidemiology: Risk differences for lung cancer mortality by emergency presentation. *American Journal of Epidemiology*, 2017.
- [4] M. A. Luque-Fernandez, M. Schomaker, B. Rachet, and M. Schnitzer. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 2018.
- [5] A. Decker, A. Hubbard, C. Crespi, E. Seto, and M Wang. Semiparametric estimation of the impacts of longitudinal interventions on adolescent obesity using targeted maximum-likelihood: Accessible estimation with the `ltmle` package. *Journal of Causal Inference*, 2(1):95–108, 2014.
- [6] M. A. Gianfrancesco, L. Balzer, K. E. Taylor, L. Trupin, J. Nititham, M. F. Seldin, A. W. Singer, L. A. Criswell, and L. F. Barcellos. Genetic risk and longitudinal disease activity in systemic lupus erythematosus using targeted maximum likelihood estimation. *Genes and Immunity*, 17(6):358–62, 2016.
- [7] A. Hubbard, F. Jamshidian, and N. Jewell. Adjusting for perception and unmasking effects in longitudinal clinical trials. *International Journal of Biostatistics*, 8(2):7, 2012.
- [8] N. Kreif, L. Tran, R. Grieve, B. deStavola, R. Tasker, and M. Petersen. Estimating the comparative effectiveness of feeding interventions in the paediatric intensive care unit: a demonstration of longitudinal targeted maximum likelihood estimation. *American Journal of Epidemiology*, in press, 2017.

- [9] S. Lendle, J. Schwab, M. Petersen, and M. Van der Laan. Itmle: An R package implementing targeted minimum loss-based estimation for longitudinal data. *Journal of Statistical Software*, 81:1–21, 2017.
- [10] R. Neugebauer, J. A. Schmittdiel, and M. J. van der Laan. Targeted learning in real-world comparative effectiveness research with time-varying interventions. *Statistics in Medicine*, 33(14):2480–520, 2014.
- [11] M. Petersen, J. Schwab, S. Gruber, N. Blaser, M. Schomaker, and M. van der Laan. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, 2:147–185, 2014.
- [12] M. L. Petersen. Commentary: Applying a causal road map in settings with time-dependent confounding. *Epidemiology*, 25(6):898–901, 2014.
- [13] M. E. Schnitzer, E. E. Moodie, M. J. van der Laan, R. W. Platt, and M. B. Klein. Modeling the impact of hepatitis C viral clearance on end-stage liver disease in an HIV co-infected cohort with targeted maximum likelihood estimation. *Biometrics*, 70(1):144–52, 2014.
- [14] M. E. Schnitzer, M. J. van der Laan, E. E. Moodie, and R. W. Platt. Effect of breastfeeding on gastrointestinal infection in infants: A targeted maximum likelihood approach for clustered longitudinal data. *Annals of Applied Statistics*, 8(2):703–725, 2014.
- [15] O. M. Stitelman, V. De Gruttola, and M. J. van der Laan. A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *International Journal of Biostatistics*, 8(1), 2012.
- [16] L. Tran, C. Yiannoutsos, B. Musick, K. Wools-Kaloustian, A. Siika, S. Kimaiyo, M. Van der Laan, and M. L. Petersen. Evaluating the impact of a HIV low-risk express care task-shifting program: A case study of the targeted learning roadmap. *Epidemiological Methods*, 5(1):69–91, 2016.
- [17] J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.
- [18] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 64(2):962–972, 2005.
- [19] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.
- [20] M. Van der Laan and S. Rose. *Targeted Learning*. Springer, 2011.
- [21] D. Westreich, S. R. Cole, J. G. Young, F. Palella, P. C. Tien, L. Kingsley, S. J. Gange, and M. A. Hernan. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine*, 31(18):2000–2009, 2012.
- [22] A. Edmonds, M. Yotebieng, J. Lusiana, Y. Matumona, F. Kitetele, S. Napravnik, S. R. Cole, A. Van Rie, and F. Behets. The effect of highly active antiretroviral therapy on the survival of HIV-infected children in a resource-deprived setting: a cohort study. *PLoS Medicine*, 8(6):e1001044, 2011.
- [23] A. Violari, M. F. Cotton, D. M. Gibb, A. G. Babiker, J. Steyn, S. A. Madhi, P. Jean-Philippe, and J. A. McIntyre. Early antiretroviral therapy and mortality among HIV-infected infants. *New England Journal of Medicine*, 359(21):2233–2244, 2008.
- [24] R. M. Daniel, S. N. Cousens, B. L. De Stavola, M. G. Kenward, and J. A. Sterne. Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32(9):1584–618, 2013.
- [25] J. Robins and M. A. Hernan. Estimation of the causal effects of time-varying exposures. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, editors, *Longitudinal Data Analysis*, pages 553–599. CRC Press, 2009.
- [26] R.M. Daniel, B.L. De Stavola, and S.N. Cousens. G- formula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal*, 11(4):479–517, 2011.
- [27] J. G. Young, L. E. Cain, J. M. Robins, E. J. O’Reilly, and M. A. Hernan. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in Biosciences*, 3(1):119–143, 2011.
- [28] J. Robins, M. A. Hernan, and U. Siebert. Effects of multiple interventions. In M. Ezzati, C. Murray, and A. Lopez, editors, *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*, pages 2191–2230. World Health Organization, 2004.
- [29] D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.
- [30] M. Van der Laan and D. Rubin. Targeted maximum likelihood learning. <http://biostats.bepress.com/ucbbiostat/paper213/>, 2006.
- [31] Edward H. Kennedy. *Semiparametric Theory and Empirical Processes in Causal Inference*, pages 141–167. Springer International Publishing, Cham, 2016.
- [32] M. E. Schnitzer and M. Cefalu. Collaborative targeted learning using regression shrinkage. *Statistics in Medicine*, in press, 2017.
- [33] L. Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

- [34] Eric Polley, Erin LeDell, Chris Kennedy, and Mark van der Laan. *SuperLearner: Super Learner Prediction*, 2017. R package version 2.0-22.
- [35] M. Van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6:Article 25, 2008.
- [36] T. Puthanakit, V. Saphonn, J. Ananworanich, P. Kosalaraksa, R. Hansudewechakul, U. Vibol, S. J. Kerr, S. Kanjanavanit, C. Ngampiyaskul, J. Wongsawat, W. Luesomboon, N. Ngo-Giang-Huong, K. Chettra, T. Cheunyam, T. Suwanlerk, S. Ubolyam, W. T. Shearer, R. Paul, L. M. Mofenson, L. Fox, M. G. Law, D. A. Cooper, P. Phanuphak, M. C. Vun, and K. Ruxrungtham. Early versus deferred antiretroviral therapy for children older than 1 year infected with HIV (PREDICT): a multicentre, randomised, open-label trial. *Lancet Infectious Diseases*, 12(12):933–41, 2012.
- [37] M. Schomaker, M. A. Davies, K. Malateste, L. Renner, S. Sawry, S. N’Gbeche, K. Technau, F. T. Eboua, F. Tanser, H. Sygnate-Sy, S. Phiri, M. Amorissani-Folquet, V. Cox, F. Koueta, C. Chimbete, A. Lawson-Evi, J. Giddy, C. Amani-Bosse, R. Wood, M. Egger, and V. Leroy. Growth and mortality outcomes for different antiretroviral therapy initiation criteria in children aged 1-5 years: A causal modelling analysis from West and Southern Africa. *Epidemiology*, 237-246, 2016.
- [38] M. Schomaker, M. Egger, J. Ndirangu, S. Phiri, H. Moultrie, K. Technau, V. Cox, J. Giddy, C. Chimbete, R. Wood, T. Gsponer, C. Bolton Moore, H. Rabie, B. Eley, L. Muhe, M. Penazzato, S. Essajee, O. Keiser, and M. A. Davies. When to start antiretroviral therapy in children aged 2-5 years: a collaborative causal modelling analysis of cohort studies from Southern Africa. *Plos Medicine*, 10(11):e1001555, 2013.
- [39] Joshua Schwab, Samuel Lendle, Maya Petersen, and Mark van der Laan. *ltmle: Longitudinal Targeted Maximum Likelihood Estimation*, 2016. R package version 0.9-9.
- [40] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 1989.
- [41] S. Gruber and M. J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *International Journal of Biostatistics*, 6(1):Article 26, 2010.
- [42] James Robins, Mariela Sued, Quanhong Lei-Gomez, and Mark Andrea Rotnitzky. Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science*, 22(4):544–559, 2007.
- [43] L. Hjort and G. Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98:879–945, 2003.
- [44] M. Schomaker and C. Heumann. Model selection and model averaging after multiple imputation. *Computational Statistics & Data Analysis*, 71:758–770, 2014.
- [45] M. Stone. Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36:111–147, 1974.
- [46] M. Schomaker. Shrinkage averaging estimation. *Statistical Papers*, 53(4):1015–1034, 2012.
- [47] M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012.
- [48] S. N. Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2006.
- [49] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple-regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.
- [50] Oleg Sofrygin, Mark J. van der Laan, and Romain Neugebauer. *simcausal: Simulating Longitudinal Data with Causal Inference Applications*, 2016. R package version 0.5.3.
- [51] Michael Schomaker. *MAMI: Model Averaging (and Model Selection) after Multiple Imputation*, 2017. R package version 0.9.10.