American Journal of Epidemiology **Submitted Manuscript**[1]

Special Collection: N/A

Title: Immortal time bias from selection: a principal stratification perspective

Authors: Bronner P. Gonçalves (1), Etsuji Suzuki (2)

ORCiD IDs: 0000-0002-3329-6050 (BG), 0000-0002-1290-5793 (ES)

Correspondence Address: Bronner P. Gonçalves, Faculty of Health and Medical Sciences, University of Surrey, Daphne Jackson Rd, Guildford, United Kingdom (b.goncalves@surrey.ac.uk) Etsuji Suzuki, Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, 2-5-1 Shikata-cho, Kita-ku, Okayama 700-8558, Japan (etsuji-s@cc.okayama-u.ac.jp)

Joint Authorship: Bronner P. Gonçalves (1), Etsuji Suzuki (2)

Affiliations: (1) Faculty of Health and Medical Sciences, University of Surrey, Guildford, United Kingdom, (2) Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan

Clinical trial registration number and website: N/A

Key words: immortal time bias, principal stratification, potential outcomes, causal inference

Acknowledgments[2]**:** N/A

Funding: This work was supported by the Japan Society for the Promotion of Science (JSPS KAKENHI Grant Numbers JP19KK0418 and JP23K09740 to E.S.).

[1] N/A indicates not applicable.
[2] Study investigators, conference presentations, preprint publication information, thanks.

Conflict of Interest: None declared.

Disclaimer: N/A

Data Availability Statement: N/A

Abstract

Immortal time bias due to post-treatment definition of eligibility criteria can affect experimental and observational studies, and yet, in contrast to the extensive literature on the classical form of immortal time bias, it has seldom been the focus of methodological discussions. Here, we propose an account of eligibility-related immortal time bias that uses the principal stratification framework to explain the non-comparability of treatment arms (or exposure groups) conditional on selection. In particular, we show that the statistical estimand that conditions on observed eligibility after time zero of follow-up can be interpreted using partially overlapping principal strata. Further, we show that, under this perspective, as the timing of eligibility approaches time zero of follow-up, the probabilities of the outcome for eligible individuals monotonically approach the corresponding unconditional (in absence of selection) expected potential outcomes under different treatment levels. Our study provides a potential outcomes-based explanation of eligibility-related immortal time bias, and indicates that, in addition to the target trial emulation framework, principal effects might, for some studies, be useful causal estimands.

INTRODUCTION

Epidemiologists have long recognized immortal time as a source of bias. In short, it can be conceived of as a period of time during which study participants, by study or analytic design, could not have developed the outcome of interest [1-3]. Indeed, addressing the inferential problems posed by this bias is one of the advantages of the target trial emulation framework for causal analyses using observational data [4]. However, this bias does not correspond to a single type of process, but rather to multiple mechanisms [4, 5], and a structural description of two types of immortal time bias was proposed in a recent study [5]: one type relates to selection into analysis based on post-treatment eligibility, and the other type relates to outcome-dependent misclassification of follow-up time due to treatment assignment after eligibility criteria are met. The misclassification-related immortal time bias occurs, for example, when using observational data to define treatment strategies that are only distinguishable after a pre-determined number of doses or time period under a particular therapy; in such cases, occurrence of the outcome, and in particular its timing, can affect a study participant's exposure classification. On the other hand, in settings with selection-related immortal time bias, although treatment strategies are distinguishable at time zero, only individuals at risk of the outcome after a pre-defined period of time are included in the analysis; in this case, comparisons between subgroups of the treatment arms that are selected into the analysis might be biased. Note that this structural classification parallels a previously reported categorization of immortal time bias [4].

In this paper, we build upon the article by Hernán and colleagues [5] and aim to approach selection-related immortal time bias from another perspective to gain a further insight into this problem. In particular, we discuss how this type of bias can also be understood using potential outcomes and the principal stratification framework [6]. Our objective was twofold: (i) as with other inferential problems, an account of this bias using the principal stratification approach could improve conceptual clarity; (ii) a description of this bias using principal stratification might also help investigators to understand whether other estimands (e.g., principal effects) could be relevant for some studies, and could encourage methodologists to leverage the rich literature on principal stratification methods, possibly as secondary analyses, in combination with the target trial emulation framework.

The manuscript is organized as follows. First, we describe the notation and variables that represent the selection process and the outcome of interest, and define possible estimands. Next, we discuss an example and show that if post-treatment selection is considered necessary by investigators in the particular context of their study, the estimation of a principal effect is possible. Finally, we use this approach to explain the value of a key component of the target trial emulation framework, namely the temporal alignment of treatment, time zero of follow-up and eligibility, in terms of principal stratification parameters. It is worth noting that our description does not apply to the structurally different type of immortal time bias in which treatment assignment occurs after, or treatment strategies are not distinguishable at, time zero [5].

METHODS

Notation

Here, we will consider a hypothetical clinical trial where hospitalized patients with a severe infection are randomly assigned to an active treatment ($A = 1$) or to placebo ($A = 0$). We let T denote the time when outcome occurs, and T^a, the corresponding potential time had the treatment A, possibly contrary to the fact, been set to a. For each participant, treatment assignment coincides with time zero of follow-up and is the hospital admission. Further, we assume that investigators, in performing their analysis, only include study participants who remain alive, say, for at least three days after admission (for example, because they judge that individuals who are more likely to die soon after admission might not benefit, or might benefit less, from the treatment). As eligibility requires survival for at least three days, all individuals included in the analysis dataset were, by design, immortal during those days. We let t_s denote the time when eligibility is defined (here, $t_s = 3$ days).

Our approach to this problem involves defining a variable S that corresponds to survival during the three days after hospital admission (1 = survival during the first three days in hospital, 0 = death during the first three days in hospital); hence S can also be viewed as representing selection into the analysis. Moreover, we define the potential outcome S^a as the value that S would take had the treatment A been set to a. Patients can then be categorized into four principal strata based on the joint potential outcomes of S: $(S^1 = 1, S^0 = 1)$, $(S^1 = 1, S^0 = 0)$,

$(S^1 = 0, S^0 = 1)$, and $(S^1 = 0, S^0 = 0)$. We also let Y denote the outcome of interest, which could be, for instance, death during the period from day 4 to 7 after admission, Y_{4-7}, or death during the first week of hospitalization, Y_{1-7} (for both definitions, 1 = death during the relevant period, and 0 = survival during the same period). For each of these definitions, the potential outcome variables, Y^a_{4-7} and Y^a_{1-7}, correspond to the potential death outcome during the relevant period had the treatment A been set to a. Note that $S = 0$ implies Y_{4-7} is undefined and $Y_{1-7} = 1$. Note also that the following relations hold in our hypothetical study:

$\Pr(S^a = 1) = \Pr(T^a \geq 3)$
$\Pr(S^a = 0) = \Pr(T^a < 3)$
$E[Y^a_{1-7}] = \Pr(T^a < 7)$
$E[Y^a_{4-7} | S^a = 1] = \Pr(3 \leq T^a < 7) \,/\, \Pr(T^a \geq 3)$

In **Table 1**, we present possible response types based on the joint potential outcomes of S, Y_{4-7}, and Y_{1-7}. Using these potential outcome variables, the causal estimand, had the post-treatment selection been avoided, would thus be $E[Y^1_{1-7} - Y^0_{1-7}]$; note that as $E[Y^1_{4-7} - Y^0_{4-7}]$ is not defined in the presence of early mortality, this is not the target estimand of these studies.

When we consider re-defining eligibility using a new time $t_{s\prime}$, with $0 \leq t_{s\prime} < t_s$, the variable S also needs to be redefined; to reduce ambiguity, we use S' to refer to eligibility based on $t_{s\prime}$ and use S for the eligibility defined based on the original time t_s.

Numerical example

In **Table 1**, we also describe a numerical example that is used to illustrate selection-related immortal time bias; there, we assumed that treatment A was randomized, and thus that exchangeability holds (i.e., $(S^a, Y^a_{4-7}, Y^a_{1-7}) \perp\!\!\!\perp A \ (a = 0,1)$). In **Table 1**, the proportions of the population with a particular set of response types, that is, $(S^1, S^0, Y^1_{1-7}, Y^0_{1-7})$, are shown in the rightmost column. For instance, the proportion with combined response type $(S^1 = 1, S^0 = 1, Y^1_{1-7} = 1, Y^0_{1-7} = 0)$ is calculated using the proportion of the total population with $(S^1 = 1, S^0 = 1)$, which is 0.40, and the proportion of this subpopulation, with $(S^1 = 1, S^0 = 1)$, who have response type $(Y^1_{1-7} = 1, Y^0_{1-7} = 0)$, which is 0.10.

Note that the proportions used in the illustrative example in **Table 1** were chosen so that active treatment is protective on average with regard to survival during the first 3 days of follow-up, that is, $\{\Pr(S^1 = 1, S^0 = 1) + \Pr(S^1 = 1, S^0 = 0)\} - \{\Pr(S^1 = 1, S^0 = 1) + \Pr(S^1 = 0, S^0 = 1)\} = \{\Pr(S^1 = 1, S^0 = 0) - \Pr(S^1 = 0, S^0 = 1)\} > 0$. Similarly, the proportions of outcome (Y_{1-7} and Y_{4-7})-based principal strata for the stratum defined by $(S^1 = 1, S^0 = 1)$ were chosen so that active treatment is also protective within this stratum (that is, so that active treatment reduces risk of death among individuals who would survive the first 3 days under both treatment levels). As mentioned in the *Results* section, in **Tables S1** and **S2**, we present slightly different numerical examples, where proportions of the S-based principal strata were chosen so that active treatment reduces survival during the first 3 days of follow-up or that it has no effect on survival during the first follow-up days, respectively. Finally, we note that **Figure 1** and **Figure 2**, discussed in the next section, are based on the distribution of principal strata in **Table 1**, and that **Figure 2** shows how selection-related immortal time bias varies with timing of eligibility (precise definitions of the biases are provided below).

Table 1. Response types based on potential outcomes of variables corresponding to eligibility, S, and to the outcome of interest, Y_{4-7} and Y_{1-7}, and numerical example. As mentioned in the text, while $S = 1$ corresponds to survival during the first three days of follow-up, $Y_{4-7} = 1$ and $Y_{1-7} = 1$ correspond to death during days 4 – 7 and during the first week, respectively. We present proportions of different principal strata defined by S^a, Y^a_{4-7}, and Y^a_{1-7} ; the proportions of the strata defined by combining S^a response types and Y^a_{4-7}, or Y^a_{1-7}, response types are presented as products of the proportions of the corresponding S^a principal strata and the S^a principal stratum specific proportions of Y^a_{4-7}, or Y^a_{1-7}, response types; note that these proportions are the same for the two outcome definitions, and are shown in the rightmost column.

	S^a			Y^a_{4-7}		Y^a_{1-7}		
SY type	S^1	S^0	**Proportions**	Y^1_{4-7}	Y^0_{4-7}	Y^1_{1-7}	Y^0_{1-7}	**Proportions**
1	1	1	0.40	1	1	1	1	0.40 × 0.40 = 0.16
2	1	1		1	0	1	0	0.40 × 0.10 = 0.04
3	1	1		0	1	0	1	0.40 × 0.30 = 0.12
4	1	1		0	0	0	0	0.40 × 0.20 = 0.08
5	1	0	0.40	1	Undefined	1	1	0.40 × 0.30 = 0.12
6	1	0		0	Undefined	0	1	0.40 × 0.70 = 0.28
7	0	1	0.10	Undefined	1	1	1	0.10 × 0.40 = 0.04
8	0	1		Undefined	0	1	0	0.10 × 0.60 = 0.06
9	0	0	0.10	Undefined	Undefined	1	1	0.10 × 1.00 = 0.10

Brief descriptions of the target trial emulation and the principal stratification frameworks

The target trial emulation framework is widely used in epidemiologic practice and provides guidance on the use of observational data to answer causal questions; for example, it helps investigators to avoid design-related biases. In this framework, causal analyses of observational data involve two steps: description of the protocol of a target trial that would answer the causal question of interest; and emulation of this target trial using observational data (see Hernán et al. [7] for a more comprehensive, recent, discussion). In the next section, we use the principal stratification framework [6], which has been especially used to study problems that involve conditioning on a post-treatment variable; examples of such problems include non-compliance of treatment assignment, truncation by death or post-infection outcomes [8-12]. As explained below, the principal stratification approach typically uses causal effects in strata defined by joint potential outcomes. Although the target trial emulation framework has a much broader scope of application in epidemiology, when the immortal time relates to selection, investigators might also want to consider other effects, and for that, principal stratification could be useful.

RESULTS

Selection-related immortal time and principal stratification

Here, we argue that a comparison of Y_{4-7} or Y_{1-7} by treatment levels for individuals with (observed) $S = 1$ (i.e., $E[Y_{4-7}|S = 1, A = 1] - E[Y_{4-7}|S = 1, A = 0]$ or $E[Y_{1-7}|S = 1, A = 1] - E[Y_{1-7}|S = 1, A = 0]$), which is the analysis that is potentially affected by selection-related immortal time, does not correspond to the causal estimand of interest ($E[Y^1_{1-7} - Y^0_{1-7}]$). In fact, the comparison of Y_{4-7} or Y_{1-7} for individuals with $S = 1$ is not a causal estimand, as it is not a contrast of potential outcomes for a common set of units. Assuming consistency, individuals with $A = 1$ and $S = 1$ are members of either of the following principal strata ($S^1 = 1, S^0 = 1$) or ($S^1 = 1, S^0 = 0$); participants with $A = 0$ and $S = 1$ have, as joint potential outcomes, ($S^1 = 1, S^0 = 1$) or ($S^1 = 0, S^0 = 1$). Put in terms of the information in **Table 1**, the biased analysis compares individuals with *SY* types 1, 2, 3, 4, 5, and 6 with individuals with *SY* types 1, 2, 3, 4, 7, and 8. If the selection procedure is considered necessary by investigators in the context of their study (e.g. because investigators believe that participants who die during the days following treatment are less likely to benefit

from it or, relatedly, because the effect might only operate after an initial period following treatment [as is the case for vaccination]), an effect that could be defined is the principal effect, $E[Y^1_{4-7} - Y^0_{4-7}|(S^1 = 1, S^0 = 1)]$ or $E[Y^1_{1-7} - Y^0_{1-7}|(S^1 = 1, S^0 = 1)]$. Notice that these two effects coincide, as, for individuals with the potential outcomes $(S^1 = 1, S^0 = 1)$, all death events occur after day 3. In the literature on truncation by death [8, 11], the term survivor average causal effect is used to refer to this effect. Although we believe that the survivor average causal effect rarely corresponds to the inferential target of studies affected by immortal time, investigators could also consider this causal estimand in the presence of eligibility-related immortal time. Note that this estimand will not in general correspond to the comparison that conditions on observed eligibility; rather, it conditions on the joint potential outcomes for eligibility, and identification and estimation require additional assumptions, as with other problems where principal stratification is used.

For illustration, in **Table 2**, we show both the data that would be observed by investigators in the hypothetical study described in **Table 1** and the possible estimands and corresponding estimates. Without aligning eligibility, treatment and time zero, the different estimates are not compatible. In fact, as mentioned above, if temporal alignment is not possible, the survivor average causal effect would be a valid causal estimand, and partial and point identification approaches have been described for such settings [8, 13-15]. Crucially, the comparisons that condition on $S = 1$, and that, consequently, might be affected by selection-related immortal time, are biased (with reference to the target estimand $\mathrm{E}[Y^1_{1-7} - Y^0_{1-7}]$) and further do not correspond to causal estimands.

In Panel A of **Figure 1**, we present, for the same hypothetical study of **Table 1**, the survival in the different principal strata. Our objective in presenting **Figure 1** was to provide visual illustration of the contributions of the different strata (Panel A) to quantities that are observable (Panel B). Note that for simplicity, we assumed that T is uniformly distributed in the strata $(S^1 = 1, S^0 = 1)$, $(S^1 = 1, S^0 = 0)$ and $(S^1 = 0, S^0 = 1)$ in periods before or after t_s; for participants with $(S^1 = 0, S^0 = 0)$ and $A = 0$, however, the survival function was approximated by an exponential function. The following functions describe survival in each of the four principal strata:

$$P(T > t|\,(S^1, S^0) = (1,1), A = 1) = \begin{cases} 1 & (0 \le t < t_s) \\ 1 - \beta_1^D (t - t_s) & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\,(S^1, S^0) = (1,1), A = 0) = \begin{cases} 1 & (0 \le t < t_s) \\ 1 - \beta_0^D (t - t_s) & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\,(S^1, S^0) = (1,0), A = 1) = \begin{cases} 1 & (0 \le t < t_s) \\ 1 - \beta_1^C (t - t_s) & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\,(S^1, S^0) = (1,0), A = 0) = \begin{cases} 1 - t/t_s & (0 \le t < t_s) \\ 0 & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\,(S^1, S^0) = (0,1), A = 1) = \begin{cases} 1 - t/t_s & (0 \le t < t_s) \\ 0 & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\,(S^1, S^0) = (0,1), A = 0) = \begin{cases} 1 & (0 \le t < t_s) \\ 1 - \beta_0^P (t - t_s) & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\,(S^1, S^0) = (0,0), A = 1) = \begin{cases} 1 - t/t_s & (0 \le t < t_s) \\ 0 & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\,(S^1, S^0) = (0,0), A = 0) = \begin{cases} \exp(-\alpha\, t/t_s) & (0 \le t < t_s) \\ 0 & (t_s \le t \le t_y) \end{cases}$$

where β_1^D, β_0^D, β_1^C and β_0^P are parameters that describe survival after t_s, and t_y corresponds to the time when the outcome is assessed; α is the rate parameter of the exponential distribution (for the figure, we assumed $\alpha = 5$, so that $\exp(-\alpha) \approx 0$). In the *Appendix S1*, we present their definitions and expressions for survival in the treatment arms (Panel B of **Figure 1**), which is determined by strata-specific survival functions.

Figure 1. Principal stratification based interpretation of selection-related immortal time bias. We assume that the proportions of the principal strata are those in **Table 1**. In the subpanels, the y-axes represent survival (proportion), and the x-axes, time (in days); t_s, the time in reference to which the selection variable S is defined, is three days; and the outcome Y_{1-7} is measured on day 7 after admission. Note that, as mentioned in the text, the outcome Y_{1-7} represents death, rather than survival, and hence y-axes can be viewed as representing $\Pr(Y_{1-7} = 0)$. In Panel A, we present hypothetical survival curves for the different principal strata; survival under active treatment is represented by continuous lines and survival under placebo is represented by dashed-dotted lines. For the strata defined by $(S^1 = 1, S^0 = 0)$, $(S^1 = 0, S^0 = 1)$ and $(S^1 = 0, S^0 = 0)$, two curves are presented before t_s that correspond to survival during the first three days under different treatment levels; for the stratum defined by $(S^1 = 1, S^0 = 1)$, the survival curves under different treatment levels can only be differentiated after t_s. In the figure, we assume that for the stratum $(S^1 = 0, S^0 = 0)$, the active treatment $(A = 1)$ is, on average, beneficial before t_s and for the stratum $(S^1 = 1, S^0 = 1)$, the treatment is beneficial after t_s; these assumptions are only made for visualization, and are not essential to the approach we present. In Panel B, the subpanels present survival in the observed treatment groups; latent survival curves for the different principal strata as well as observed survival curves (thicker continuous line for $A = 1$, and thicker dashed-dotted line for $A = 0$) are shown. Notice that after t_s the slope of the survival curve of the treated group is a combination of the survival curves in the strata $(S^1 = 1, S^0 = 1)$ and $(S^1 = 1, S^0 = 0)$ under treatment; for the placebo group, the slope of the survival curve after t_s is a combination of the curves in the principal strata $(S^1 = 1, S^0 = 1)$ and $(S^1 = 0, S^0 = 1)$ under absence of treatment.

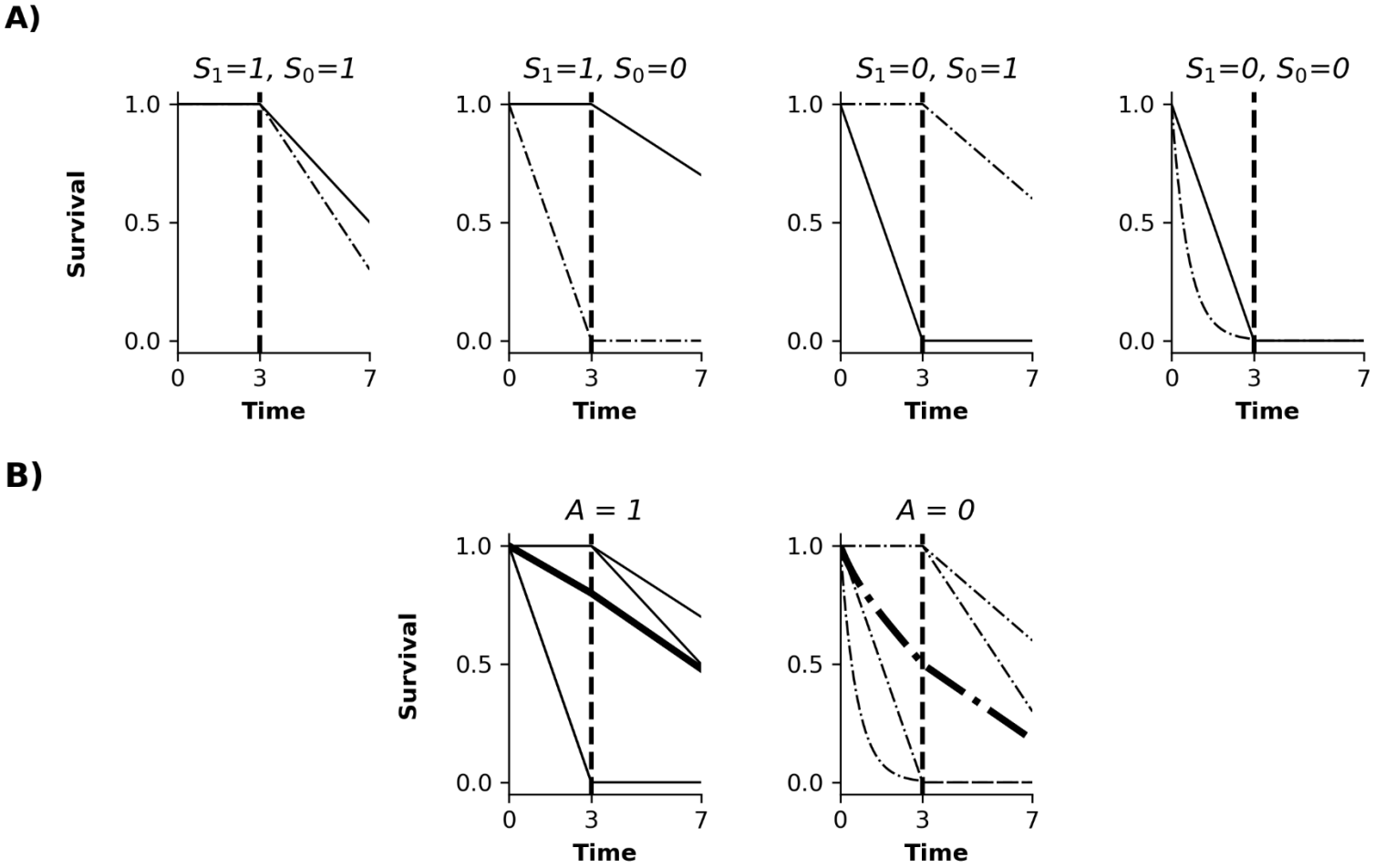

A)
S_1=1, S_0=1
S_1=1, S_0=0
S_1=0, S_0=1
S_1=0, S_0=0
Survival
1.0
0.5
0.0
0
3
7
Time
B)
A = 1
A = 0

Table 2. Observed data, estimands and estimates for the numerical example in **Table 1.** Here, we show quantities that would be observed by investigators as well as possible estimands and estimates. The first two estimands, which lead to biased effect estimates, could be more accurately described as measures of association; note that we refer to these estimands as biased with reference to the target estimand (if post-treatment selection can be avoided), $\mathrm{E}[Y^1_{1-7} - Y^0_{1-7}]$. Note also that the principal effects $\mathrm{E}[Y^1_{1-7} - Y^0_{1-7}|S^1 = 1, S^0 = 1]$ and $\mathrm{E}[Y^1_{4-7} - Y^0_{4-7}|S^1 = 1, S^0 = 1]$ are not identifiable without additional assumptions [8, 13], and here, as our focus is not on principal effect estimation, we used the unobserved proportions of the principal strata in **Table 1**.

Quantities	**Interpretation**	**Observed data**
$\Pr(S = 1\|A = 1)$	Proportion of survival in the first 3 days in treatment arm	0.80
$\Pr(S = 1\|A = 0)$	Proportion of survival in the first 3 days in control arm	0.50
$\Pr(Y_{1-7} = 1\|A = 1)$	Risk of death during week 1 in treatment arm	0.52
$\Pr(Y_{1-7} = 1\|A = 0)$	Risk of death during week 1 in control arm	0.82
Estimands		**Estimates**
$\Pr(Y_{1-7} = 1\|S = 1, A = 1) - \Pr(Y_{1-7} = 1\|S = 1, A = 0)$	Associational risk difference of death during week 1 among individuals who actually survive the first 3 days	-0.24
$\Pr(Y_{4-7} = 1\|S = 1, A = 1) - \Pr(Y_{4-7} = 1\|S = 1, A = 0)$	Associational risk difference of death during week 1 but after the first 3 days among individuals who actually survive the first 3 days	-0.24
$\mathrm{E}[Y^1_{1-7} - Y^0_{1-7}]$	Causal risk difference of death during week 1 for the entire study population	-0.30
$\mathrm{E}[Y^1_{4-7} - Y^0_{4-7}]$	Causal risk difference of death during week 1 but after the first 3 days for the entire study population	Undefined
$\mathrm{E}[Y^1_{1-7} - Y^0_{1-7}\|S^1 = 1, S^0 = 1]$	Causal risk difference of death during week 1 for individuals who would survive the first 3 days irrespective of treatment	-0.20
$\mathrm{E}[Y^1_{4-7} - Y^0_{4-7}\|S^1 = 1, S^0 = 1]$	Causal risk difference of death during week 1 but after the first 3 days for individuals who would survive the first 3 days irrespective of treatment	-0.20

Principal stratification based justification for the alignment component of target trial emulation

Alignment of eligibility, treatment, and start of follow-up is a key component of the target trial emulation framework [4, 16]. In this section, we show how changing the timing of eligibility, relative to hospital admission (that is, relative to treatment assignment and time zero), affects the distribution of principal strata in the analytical population, and, ultimately, the interpretation of the estimands.

Recall that S' denotes survival, and eligibility, based on a different time $t_{s\prime}$, with $t_{s\prime} < t_s$. Then, under the assumptions of consistency and exchangeability, $\mathrm{E}[Y^1_{1-7}]$ and $\mathrm{E}[Y_{1-7}|S'=1, A=1]$ can be written in terms of principal strata defined by S and of $t_{s\prime}$ as follows (see the *Appendix S2* for derivations):

$$\begin{aligned}&\mathrm{E}[Y^1_{1-7}]\\&= \Pr(Y^1_{1-7}=1|S^1=1,S^0=1)\Pr(S^1=1,S^0=1)\\&\qquad + \Pr(Y^1_{1-7}=1|S^1=1,S^0=0)\Pr(S^1=1,S^0=0) + \Pr(S^1=0,S^0=1)\\&\qquad + \Pr(S^1=0,S^0=0)\end{aligned}$$

(Expression 1)

$$\begin{aligned}&\mathrm{E}[Y_{1-7}|S'=1,A=1]\\&= \Pr(Y^1_{1-7}=1|S^1=1,S^0=1)\,\frac{\Pr(S^1=1,S^0=1)}{\Pr(S'^1=1)}\\&+ \Pr(Y^1_{1-7}=1|S^1=1,S^0=0)\frac{\Pr(S^1=1,S^0=0)}{\Pr(S'^1=1)}\\&+ \frac{\Pr(T^1\geq t_{s'}\,|S^1=0,S^0=1)\Pr(S^1=0,S^0=1) + \Pr(T^1\geq t_{s'}\,|S^1=0,S^0=0)\Pr(S^1=0,S^0=0)}{\Pr(S'^1=1)}\end{aligned}$$

(Expression 2)

In the *Appendix S2*, we present similar expressions for $\mathrm{E}[Y^0_{1-7}]$ and $\mathrm{E}[Y_{1-7}|S'=1,A=0]$. The rationale for deriving expressions for these quantities was the following: $\mathrm{E}[Y^1_{1-7}]$ and $\mathrm{E}[Y^0_{1-7}]$ represent the expected outcomes under the different treatment levels, and are important epidemiologic quantities in themselves and also determine the causal effect of interest;

$\mathrm{E}[Y_{1-7}|S'=1,A=1]$ and $\mathrm{E}[Y_{1-7}|S'=1,A=0]$ correspond to average outcomes among individuals alive at $t_{s'}$, and consequently are affected by immortal time, and allow quantification of the following comparison in the presence of selection, $\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y_{1-7}|S'=1,A=0]$. By expressing these latter observable quantities using principal strata defined by S, we can directly compare them with the expected outcomes, which are also definable in terms of proportions of the same strata in the population and stratum-specific outcome distributions.

In Expression 2, note that $\Pr(T^1 \geq t_{s'}|S^1=0,S^0=1)$ and $\Pr(T^1 \geq t_{s'}|S^1=0,S^0=0)$ correspond to the proportions of individuals in the principal strata $(S^1=0,S^0=1)$ and $(S^1=0,S^0=0)$, respectively (or equivalently, *SY* types 7 and 8, and *SY* type 9), who are still alive at (or equivalently, who have not developed the outcome by) $t_{s'}$. For instance, if, for a particular $t_{s'}$, $\Pr(T^1 \geq t_{s'}|S^1=0,S^0=1)=0.5$, then half of the participants in the stratum $(S^1=0,S^0=1)$ contributes to the corresponding numerator in the expression and to the denominator $\Pr(S'^1=1)$ and all of these individuals have the outcome by t_s, based on the definition of the principal stratum. As shown in the *Appendix S2*, the fractions in Expression 2 with denominator $\Pr(S'^1=1)$ correspond to relative weights for the S-defined stratum probabilities of the outcome (e.g., $\Pr(S^1=1,S^0=1|S'^1=1)$). In the *Appendix S2*, we also show that when $t_{s'}=0$, Expressions 1 and 2 become identical, and analogously, $\mathrm{E}[Y^0_{1-7}]$ and $\mathrm{E}[Y_{1-7}|S'=1,A=0]$ become identical.

In **Figure 2**, we show how the biases for expected outcomes and for the effect of interest change as $t_{s'}/t_s$ (which can be viewed as a standardised measure of time) varies between 0 and 1; the proportions of the principal strata are those in the numerical example in **Table 1**. The biases shown are:

$\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y^1_{1-7}]$,

$\mathrm{E}[Y_{1-7}|S'=1,A=0]-\mathrm{E}[Y^0_{1-7}]$,

and

$(\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y_{1-7}|S'=1,A=0])-(\mathrm{E}[Y^1_{1-7}]-\mathrm{E}[Y^0_{1-7}])$.

In **Figure 2**, we assume the same principal stratum specific survival functions as those in **Figure 1**. In this case, biases for both the expected outcomes ($\mathrm{E}[Y^1_{1-7}]$ and $\mathrm{E}[Y^0_{1-7}]$) and the target effect ($\mathrm{E}[Y^1_{1-7}] - \mathrm{E}[Y^0_{1-7}]$) decrease (in absolute terms) as $t_{s'}$ approaches 0. In fact, as can be seen in Expression 2, $\mathrm{E}[Y_{1-7}|S'=1, A=1]$ increases monotonically with decreasing $t_{s'}$, since Expression 2 is a weighted average, with monotonically increasing relative weights for the combined strata of individuals who would die before t_s (that is, for the combination of the strata defined by $(S^1=0, S^0=1)$ and $(S^1=0, S^0=0)$ [or equivalently, the strata of individuals with *SY* types 7–9], in which, under consistency, $Y_{1-7}=1$ whenever $A=1$). This point is also evident by noting that the last term in Expression 2 can be rewritten as follows:

$$\frac{\Pr(T^1 \geq t_{s'} \,|S^1=0, S^0=1)\Pr(S^1=0, S^0=1) + \Pr(T^1 \geq t_{s'} \,|S^1=0, S^0=0)\Pr(S^1=0, S^0=0)}{\Pr(S'^1=1)}$$

$$= \frac{\Pr(S'^1=1) - \{\Pr(S^1=1, S^0=1) + \Pr(S^1=1, S^0=0)\}}{\Pr(S'^1=1)}$$

$$= 1 - \frac{\Pr(S^1=1)}{\Pr(S'^1=1)},$$

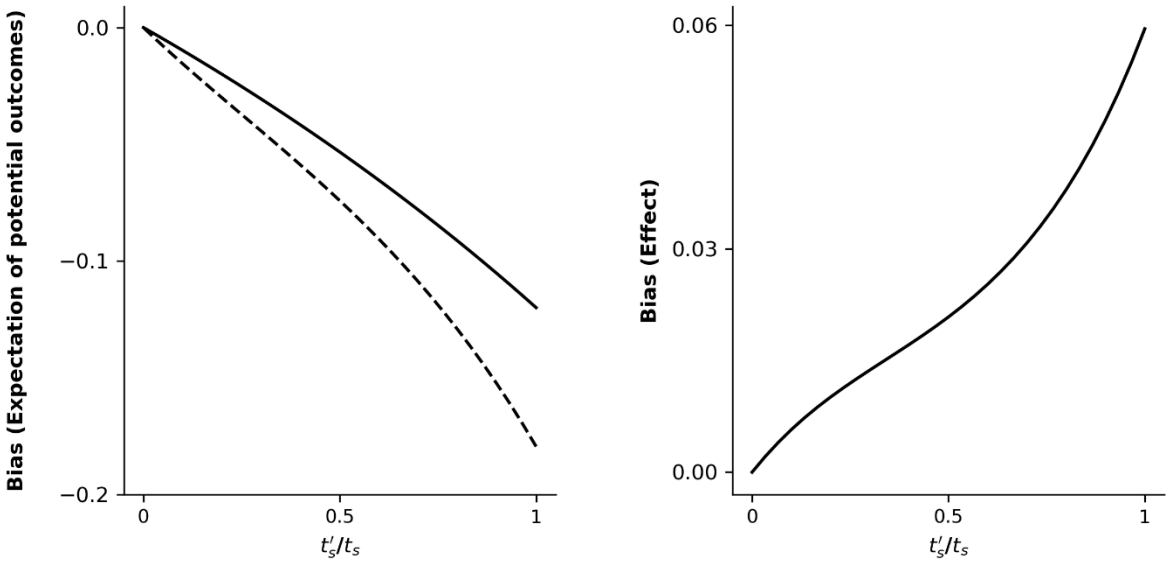
where the last expression monotonically increases with decreasing $t_{s'}$. For similar reasons, $\mathrm{E}[Y_{1-7}|S'=1, A=0]$ is monotonically increasing with decreasing values of $t_{s'}$.

In **Figure S1**, we show the corresponding biases when mortality is exponentially distributed before t_s in the stratum $(S^1=0, S^0=0)$ for both treatment levels; that is, $\Pr(T^a \geq t_{s'} \,|S^1=0, S^0=0) = \exp(-\alpha t_{s'}/t_s)$ for $a=0,1$, where we assumed a value for α so that $\exp(-\alpha) \approx 0$. We found monotonic patterns for the biases of $\mathrm{E}[Y^1_{1-7}]$ and $\mathrm{E}[Y^0_{1-7}]$, which is consistent with **Figure 2**. Although $\mathrm{E}[Y_{1-7}|S'=1, A=1] - \mathrm{E}[Y_{1-7}|S'=1, A=0]$ does not necessarily vary monotonically with $t_{s'}$ (as shown in **Figure S1**), when $t_{s'}=0$, each of these terms coincides with the corresponding expected outcome, and this expression and the target causal estimand become identical (see the *Appendix S2*).

It is important to emphasize that these figures are only numerical examples used for illustrative purposes, and that while the monotonic decrease in the magnitude of biases for the expected outcomes follows, as explained above, from the nature of Expression 2 for

$\mathrm{E}[Y_{1-7}|S'=1,A=1]$ (and the corresponding expression for $\mathrm{E}[Y_{1-7}|S'=1,A=0]$ in the *Appendix S2*), the precise shapes of these curves are dependent on the proportions of the S-based and Y-based principal strata as well as on the survival functions that describe $\Pr(T^a \geq t_{s'})$ for the relevant strata under each possible treatment level. Indeed, the visual differences between **Figure 2** and **Figure S1** in respect of aspects other than monotonicity confirm that, for example, assuming different survival functions for even only one of the S-based principal strata can lead to important changes in the evolution of biases with decreasing $t_{s'}$. To further illustrate this dependence on parametric assumptions, we show in **Figures S2** and **S3** biases under other assumptions for principal stratification parameters (**Table S1** and **Table S2**). For instance, in **Figure S3**, we observe that despite equal proportions of the strata defined by $(S^1=1,S^0=0)$ and $(S^1=0,S^0=1)$ and equal proportions of *SY* types 5 and 7, the biases in expected outcomes differ between treatment arms and is higher for the active treatment (versus placebo). For completeness, in the *Appendix*, we present formulas for the biases (*Appendix S3*) in terms of the principal strata, $t_{s'}/t_s$, and T^a, and present the derivatives of the bias expressions for the expected outcomes, with respect to $t_{s'}$, which confirm their monotonic nature (*Appendix S4*).

Figure 2. Selection-related immortal time bias and the timing of eligibility. For this figure, we used the proportions of the principal strata in **Table 1**. The x-axes represent the ratio of the (new) time when eligibility is defined over the original eligibility time, $t_{s'}/t_s$. The y-axis in the left panel corresponds to $\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y^1_{1-7}]$ (continuous line) and $\mathrm{E}[Y_{1-7}|S'=1,A=0]-\mathrm{E}[Y^0_{1-7}]$ (dashed line). The y-axis in the right panel shows $(\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y_{1-7}|S'=1,A=0])-(\mathrm{E}[Y^1_{1-7}]-\mathrm{E}[Y^0_{1-7}])$. Here, we assumed the same survival functions described for **Figure 1** (see text and the *Appendix S1*).



DISCUSSION

Although confounding is the most widely discussed threat for the internal validity of epidemiologic studies, other forms of biases are important and also need to be considered in the analysis and interpretation of observational studies [17]. The detailed structural description of immortal time bias by Hernán and colleagues [5] was an important step forward in the field and suggested that immortal time relates to either selection [18] or misclassification [19]. In its classical (misclassification-related) form, the immortal time bias occurs when treatment is assigned (or treatment strategies can only be discriminated) after the time zero of follow-up. Our focus was on selection-related immortal time bias, which, although less often discussed than the classical immortal time bias, might for instance affect vaccine studies (as suggested by Hernán and colleagues [5], see the discussion in the Appendix [section 4, Outcome Periods] of the study by Dagan and colleagues [20], where the authors explain that period-specific "risks", sometimes defined for periods starting several days after vaccination, should be interpreted as discrete-time hazards). Here, it is important to emphasize that our study does not to provide an alternative structural description of this type of bias; rather, in describing this bias using potential outcomes and principal strata, our aims were: (i) to improve conceptual clarity regarding the non-comparability of the exposure (treatment) groups using the principal stratification approach; (ii) to indicate that alternative estimands might also be useful in the presence of this bias, and to point to the literature on principal effects as a useful resource.

The interpretation of this bias that is advanced here complements, and is consistent with, previous work. In particular, comprehensive descriptions of this problem have been published [1, 21], and here we show that the statistical estimand that conditions on observed eligibility after time zero (here, $\mathrm{E}[Y_{1-7}|S=1,A=1]-\mathrm{E}[Y_{1-7}|S=1,A=0]$) can be interpreted using partially overlapping principal strata [6]. As with other problems interpreted using principal stratification, one benefit of using this framework is the coherent explanation of the non-comparability of exposure or treatment-defined groups in terms of response types. Furthermore, in other contexts, the principal stratification approach has led to novel analytical methods; examples abound in the non-compliance literature, as well as those of truncation by death and post-infection outcomes [8-11, 13-15]. It is our hope that this work will enrich the set of methods available to address this bias.

Moreover, under this interpretation, we can also understand how biases vary as the timing of eligibility approaches time zero of follow-up. For that, we can use proportions of response types for the eligibility variable (S), and the monotonic (in reference to the timing of eligibility) nature of Expression 2 above (and Expression 4 in the *Appendix S2*). Together, these analyses provide a formal justification for aligning different components of a study protocol, as routinely done in target trial emulation [4, 16]. A general problem with principal effects, however, is that they apply to subpopulations that are in general not discernible (those with particular combinations of potential outcomes for the post-treatment variable under different treatment values), as we can only observe at most one potential outcome for each individual.

Note that although the inferential problem here is not dissimilar to those operating in truncation by death [11, 22] and post-infection outcome [10] settings, the outcome of interest in studies with selection-related immortal time (if defined, in our example, by Y_{1-7}, and not by Y_{4-7}) can be defined for individuals with $S = 0$ and $S = 1$. In contrast, in the presence of truncation by death, a causal effect can only be defined for the stratum of the population who would survive under both treatment assignments, ($S^1 = 1, S^0 = 1$), and not for the entire population.

Finally, as a general advice, we believe that investigators should follow the target trial emulation framework as the primary approach for preventing selection-related immortal time bias; if the causal question of interest can be reasonably expressed as a hypothetical trial in which eligibility is aligned with treatment assignment, then following this framework would avoid immortal time. However, when investigators believe that it would also be valuable to estimate the effect of the exposure on the outcome after a particular time interval after assignment, in this case it could be useful to report principal effects (for example, as secondary analyses). The added value in doing so would be to gain insights into the underlying data generating mechanism.

References

1. Platt, R.W., J.A. Hutcheon, and S. Suissa, *Immortal Time Bias in Epidemiology.* Current Epidemiology Reports, 2019. **6**: p. 23-27.

2. Suissa, S., *Immortal time bias in observational studies of drug effects.* Pharmacoepidemiol Drug Saf, 2007. **16**(3): p. 241-9.
3. Suissa, S., *Immortal time bias in pharmaco-epidemiology.* Am J Epidemiol, 2008. **167**(4): p. 492-9.
4. Hernan, M.A., et al., *Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses.* J Clin Epidemiol, 2016. **79**: p. 70-75.
5. Hernan, M.A., et al., *A structural description of biases that generate immortal time.* Epidemiology, 2025. **36**(1): p. 107-114.
6. Frangakis, C.E. and D.B. Rubin, *Principal stratification in causal inference.* Biometrics, 2002. **58**(1): p. 21-9.
7. Hernan, M.A., et al., *The Target Trial Framework for Causal Inference From Observational Data: Why and When Is It Helpful?* Ann Intern Med, 2025. **178**(3): p. 402-407.
8. Chiba, Y. and T.J. VanderWeele, *A simple method for principal strata effects when the outcome has been truncated due to death.* Am J Epidemiol, 2011. **173**(7): p. 745-51.
9. Gruber, J.S., et al., *Estimation of treatment efficacy with complier average causal effects (CACE) in a randomized stepped wedge trial.* Am J Epidemiol, 2014. **179**(9): p. 1134-42.
10. Hudgens, M.G. and M.E. Halloran, *Causal Vaccine Effects on Binary Postinfection Outcomes.* J Am Stat Assoc, 2006. **101**(473): p. 51-64.
11. Rubin, D.B., *Causal Inference through Potential Outcomes and Principal Stratification: Application to Studies with 'Censoring' Due to Death.* Statistical Science, 2006. **21**(3): p. 299–309.
12. Goncalves, B.P. and E. Suzuki, *Effect Modification in Settings with "Truncation by Death".* Epidemiology, 2025. **36**(3): p. 374-380.
13. Chiba, Y., M. Taguri, and Y. Uemura, *On the Identification of the Survivor Average Causal Effect.* Journal of Biometrics & Biostatistics, 2011. **2**(5).
14. Wang, L., X.H. Zhou, and T.S. Richardson, *Identification and estimation of causal effects with outcomes truncated by death.* Biometrika, 2017. **104**(3): p. 597-612.
15. Tchetgen Tchetgen, E.J., *Identification and estimation of survivor average causal effects.* Stat Med, 2014. **33**(21): p. 3601-28.
16. Hernan, M.A., W. Wang, and D.E. Leaf, *Target Trial Emulation: A Framework for Causal Inference From Observational Data.* JAMA, 2022. **328**(24): p. 2446-2447.
17. Suzuki, E., et al., *Errors in causal inference: an organizational schema for systematic error and random error.* Ann Epidemiol, 2016. **26**(11): p. 788-793 e1.
18. Hernan, M.A., S. Hernandez-Diaz, and J.M. Robins, *A structural approach to selection bias.* Epidemiology, 2004. **15**(5): p. 615-25.
19. Hernan, M.A. and S.R. Cole, *Invited Commentary: Causal diagrams and measurement bias.* Am J Epidemiol, 2009. **170**(8): p. 959-62; discussion 963-4.
20. Dagan, N., et al., *BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting.* N Engl J Med, 2021. **384**(15): p. 1412-1423.
21. Hernan, M.A., et al., *A structural description of biases that generate immortal time.* Epidemiology, 2024.
22. Zhang, J.L. and D.B. Rubin, *Estimation of Causal Effects via Principal Stratification When Some Outcomes Are Truncated by "Death.".* Journal of Educational and Behavioral Statistics, 2003. **28**(4): p. 353–368.

Appendix

Table of contents

Appendix S1. *Expressions representing survival in Figure 1*

The following functions describe survival in the different principal strata in Panel A of **Figure 1** in the main text:

Principal stratum defined by $(S^1 = 1, S^0 = 1)$

$$P(T > t|\ (S^1, S^0) = (1,1), A = 1) = \begin{cases} 1 & (0 \le t < t_s) \\ 1 - \beta_1^D (t - t_s) & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\ (S^1, S^0) = (1,1), A = 0) = \begin{cases} 1 & (0 \le t < t_s) \\ 1 - \beta_0^D (t - t_s) & (t_s \le t \le t_y) \end{cases}$$

where $\beta_1^D = \frac{\Pr\left(Y_{1-7}^1 = 1 \middle| S^1 = 1, S^0 = 1\right)}{t_y - t_s}$ and $\beta_0^D = \frac{\Pr\left(Y_{1-7}^0 = 1 \middle| S^1 = 1, S^0 = 1\right)}{t_y - t_s}$.

Principal stratum defined by $(S^1 = 1, S^0 = 0)$

$$P(T > t|\ (S^1, S^0) = (1,0), A = 1) = \begin{cases} 1 & (0 \le t < t_s) \\ 1 - \beta_1^C (t - t_s) & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\ (S^1, S^0) = (1,0), A = 0) = \begin{cases} 1 - t/t_s & (0 \le t < t_s) \\ 0 & (t_s \le t \le t_y) \end{cases}$$

where $\beta_1^C = \frac{\Pr\left(Y_{1-7}^1 = 1 \middle| S^1 = 1, S^0 = 0\right)}{t_y - t_s}$.

Principal stratum defined by $(S^1 = 0, S^0 = 1)$

$$P(T > t|\,(S^1, S^0) = (0,1), A = 1) = \begin{cases} 1 - t/t_s & (0 \le t < t_s) \\ 0 & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\,(S^1, S^0) = (0,1), A = 0) = \begin{cases} 1 & (0 \le t < t_s) \\ 1 - \beta_0^P (t - t_s) & (t_s \le t \le t_y) \end{cases}$$

where $\beta_0^P = \frac{\Pr\left(Y_{1-7}^0 = 1 \middle| S^1 = 0, S^0 = 1\right)}{t_y - t_s}$.

Principal stratum defined by $(S^1 = 0, S^0 = 0)$

$$P(T > t|\,(S^1, S^0) = (0,0), A = 1) = \begin{cases} 1 - t/t_s & (0 \le t < t_s) \\ 0 & (t_s \le t \le t_y) \end{cases}$$

$$P(T > t|\,(S^1, S^0) = (0,0), A = 0) = \begin{cases} \exp(-\alpha\, t/t_s) & (0 \le t < t_s) \\ 0 & (t_s \le t \le t_y) \end{cases}$$

Note that $\exp(-\alpha\, t/t_s) = 1$ if $t = 0$, and that, in creating the figure, we assumed that $\alpha = 5$, so that $\exp(-\alpha\, t/t_s) \approx 0$ if $t = t_s$.

Below we present the expressions that represent survival in the treated and untreated arms (Panel B of **Figure 1**), which are combinations of the strata-specific functions above:

$P(T > t|A = 1)$

$$= 0.4 \times P(T > t|\,(S^1, S^0) = (1,1), A = 1) + 0.4 \times P(T > t|\,(S^1, S^0) = (1,0), A = 1) + 0.1 \times P(T > t|\,(S^1, S^0) = (0,1), A = 1)$$
$$+ 0.1 \times P(T > t|\,(S^1, S^0) = (0,0), A = 1)$$

$$= \begin{cases} 0.4 \times 1 + 0.4 \times 1 + 0.1 \times (1 - t/t_s) + 0.1 \times (1 - t/t_s) & (0 \le t < t_s) \\ 0.4 \times \left(1 - \beta_1^D (t - t_s)\right) + 0.4 \times \left(1 - \beta_1^C (t - t_s)\right) + 0.1 \times 0 + 0.1 \times 0 & \left(t_s \le t \le t_y\right) \end{cases}$$

$$= \begin{cases} 1 - t/5t_s & (0 \le t < t_s) \\ 0.8 - 0.4 \times (\beta_1^D + \beta_1^C)(t - t_s) & \left(t_s \le t \le t_y\right) \end{cases}$$

$P(T > t|A = 0)$

$$= 0.4 \times P(T > t|\,(S^1, S^0) = (1,1), A = 0) + 0.4 \times P(T > t|\,(S^1, S^0) = (1,0), A = 0) + 0.1 \times P(T > t|\,(S^1, S^0) = (0,1), A = 0)$$
$$+ 0.1 \times P(T > t|\,(S^1, S^0) = (0,0), A = 0)$$

$$= \begin{cases} 0.4 \times 1 + 0.4 \times (1 - t/t_s) + 0.1 \times 1 + 0.1 \times \exp(-\alpha\, t/t_s) & (0 \le t < t_s) \\ 0.4 \times (1 - \beta_0^D (t - t_s)) + 0.4 \times 0 + 0.1 \times (1 - \beta_0^P (t - t_s)) + 0.1 \times 0 & \left(t_s \le t \le t_y\right) \end{cases}$$

$$= \begin{cases} 0.9 - 2t/5t_s + 0.1 \times \exp(-\alpha\, t/t_s) & (0 \le t < t_s) \\ 0.5 - (0.4 \times \beta_0^D + 0.1 \times \beta_0^P)(t - t_s) & \left(t_s \le t \le t_y\right) \end{cases}$$

Appendix S2. *Principal stratification based logic for temporal alignment of protocol components*

Here, we first express the average potential outcomes in the total population, $\mathrm{E}[Y_{1-7}^{1}]$ and $\mathrm{E}[Y_{1-7}^{0}]$, in terms of proportions of principal strata defined by the variable S, and then express $\mathrm{E}[Y_{1-7}|S'=1,A=1]$ and $\mathrm{E}[Y_{1-7}|S'=1,A=0]$ in terms of the same principal stratification parameters. Next, we show that when $t_{s'}=0$, $\mathrm{E}[Y_{1-7}^{1}]=\mathrm{E}[Y_{1-7}|S'=1,A=1]$ and $\mathrm{E}[Y_{1-7}^{0}]=\mathrm{E}[Y_{1-7}|S'=1,A=0]$ hold. As our focus here is on selection-related immortal time, not confounding, in addition to consistency, we also assumed exchangeability (i.e., $(S^{a},Y_{4-7}^{a},Y_{1-7}^{a}) \perp\!\!\!\perp A$ $(a=0,1)$).

First, $\mathrm{E}[Y_{1-7}^{1}]$ can be expressed as follows:

$$
\begin{aligned}
&\mathrm{E}[Y_{1-7}^{1}]\\
&=\Pr(Y_{1-7}^{1}=1|S^{1}=1,S^{0}=1)\Pr(S^{1}=1,S^{0}=1)+\Pr(Y_{1-7}^{1}=1|S^{1}=1,S^{0}=0)\Pr(S^{1}=1,S^{0}=0)\\
&\qquad+\Pr(Y_{1-7}^{1}=1|S^{1}=0,S^{0}=1)\Pr(S^{1}=0,S^{0}=1)+\Pr(Y_{1-7}^{1}=1|S^{1}=0,S^{0}=0)\Pr(S^{1}=0,S^{0}=0)\\
&=\Pr(Y_{1-7}^{1}=1|S^{1}=1,S^{0}=1)\Pr(S^{1}=1,S^{0}=1)+\Pr(Y_{1-7}^{1}=1|S^{1}=1,S^{0}=0)\Pr(S^{1}=1,S^{0}=0)+\Pr(S^{1}=0,S^{0}=1)\\
&\qquad+\Pr(S^{1}=0,S^{0}=0)\quad(\because S^{1}=0\Rightarrow Y_{1-7}^{1}=1)
\end{aligned}
$$

(Expression 1)

Now, we derive an expression for $\mathrm{E}[Y_{1-7}|S'=1,A=1]$ in terms of the principal stratification parameters and the variables defined in the section *Principal stratification based justification for the alignment component of target trial emulation*, S' and T.

$$
\begin{aligned}
&\mathrm{E}[Y_{1-7}|S'=1,A=1]\\
&=\Pr\left(Y_{1-7}^{1}=1\middle|S'^{1}=1,A=1\right)\ (\because \text{consistency})
\end{aligned}
$$

$= \Pr\left(Y_{1-7}^{1} = 1 \middle| S'^{1} = 1\right)$ ($\because$ exchangeability)

$= \Pr(Y_{1-7}^{1} = 1|S^1 = 1, S^0 = 1, S'^1 = 1)\Pr(S^1 = 1, S^0 = 1|S'^1 = 1) + \Pr(Y_{1-7}^{1} = 1|S^1 = 1, S^0 = 0, S'^1 = 1)\Pr(S^1 = 1, S^0 = 0|S'^1 = 1)$
$+ \Pr(Y_{1-7}^{1} = 1|S^1 = 0, S^0 = 1, S'^1 = 1)\Pr(S^1 = 0, S^0 = 1|S'^1 = 1)$
$+ \Pr(Y_{1-7}^{1} = 1|S^1 = 0, S^0 = 0, S'^1 = 1)\Pr(S^1 = 0, S^0 = 0|S'^1 = 1)$

$= \Pr(Y_{1-7}^{1} = 1|S^1 = 1, S^0 = 1)\Pr\left(S^1 = 1, S^0 = 1 \middle| S'^{1} = 1\right) + \Pr(Y_{1-7}^{1} = 1|S^1 = 1, S^0 = 0)\Pr\left(S^1 = 1, S^0 = 0 \middle| S'^{1} = 1\right)$
$+ \Pr\left(S^1 = 0, S^0 = 1 \middle| S'^{1} = 1\right) + \Pr\left(S^1 = 0, S^0 = 0 \middle| S'^{1} = 1\right)$

$\left(\because \left(S^1 = 1 \Rightarrow S'^1 = 1\right) \wedge \left(S^1 = 0 \Rightarrow Y_{1-7}^{1} = 1\right)\right)$

$= \Pr(Y_{1-7}^{1} = 1|S^1 = 1, S^0 = 1)\frac{\Pr\left(S'^1=1|S^1=1,S^0=1\right)\Pr(S^1=1,S^0=1)}{\Pr\left(S'^1=1\right)} + \Pr(Y_{1-7}^{1} = 1|S^1 = 1, S^0 = 0)\frac{\Pr\left(S'^1=1|S^1=1,S^0=0\right)\Pr(S^1=1,S^0=0)}{\Pr\left(S'^1=1\right)} +$

$\frac{\Pr\left(S'^1=1|S^1=0,S^0=1\right)\Pr(S^1=0,S^0=1)}{\Pr\left(S'^1=1\right)} + \frac{\Pr\left(S'^1=1|S^1=0,S^0=0\right)\Pr(S^1=0,S^0=0)}{\Pr\left(S'^1=1\right)}$

($\because$ Bayes rule)

$= \Pr(Y_{1-7}^{1} = 1|S^1 = 1, S^0 = 1)\frac{\Pr(S^1=1,S^0=1)}{\Pr\left(S'^1=1\right)} + \Pr(Y_{1-7}^{1} = 1|S^1 = 1, S^0 = 0)\frac{\Pr(S^1=1,S^0=0)}{\Pr\left(S'^1=1\right)} +$

$\frac{\Pr(T^1\geq t_{s'}|S^1=0,S^0=1)\Pr(S^1=0,S^0=1)}{\Pr\left(S'^1=1\right)} + \frac{\Pr(T^1\geq t_{s'}|S^1=0,S^0=0)\Pr(S^1=0,S^0=0)}{\Pr\left(S'^1=1\right)}$

(Expression 2)

In the final equality, $\Pr(T^1 \geq t_{s'}\,|S^1 = 0, S^0 = 1)$ and $\Pr(T^1 \geq t_{s'}\,|S^1 = 0, S^0 = 0)$ represent the proportions of individuals in the corresponding strata who, under active treatment, remain alive by $t_{s'}$ and hence for whom $S'^1 = 1$. We also used $\Pr\left(S'^1 = 1|S^1 = 1, S^0 = 1\right) = 1$, which

follows from the ordering of $t_{s'}$ and t_s; similarly, we used that $\Pr(S'^1 = 1 | S^1 = 1, S^0 = 0) = 1$. Note that Expression 2 is a weighted average since the following equation holds.

$$\frac{\Pr(S^1 = 1, S^0 = 1)}{\Pr(S'^1 = 1)} + \frac{\Pr(S^1 = 1, S^0 = 0)}{\Pr(S'^1 = 1)}$$
$$+ \frac{\Pr(T^1 \geq t_{s'} | S^1 = 0, S^0 = 1) \Pr(S^1 = 0, S^0 = 1)}{\Pr(S'^1 = 1)} + \frac{\Pr(T^1 \geq t_{s'} | S^1 = 0, S^0 = 0) \Pr(S^1 = 0, S^0 = 0)}{\Pr(S'^1 = 1)}$$
$$= \Pr(S^1 = 1, S^0 = 1 | S'^1 = 1) + \Pr(S^1 = 1, S^0 = 0 | S'^1 = 1) + \Pr(S^1 = 0, S^0 = 1 | S'^1 = 1) + \Pr(S^1 = 0, S^0 = 0 | S'^1 = 1)$$
$$= 1$$

When $t_{s'} = 0$, $\Pr(S'^1 = 1)$ becomes, by definition, 1. Therefore, Expression 2 can be written as:

$$\Pr(Y^1_{1-7} = 1 | S^1 = 1, S^0 = 1) \frac{\Pr(S^1 = 1, S^0 = 1)}{1} + \Pr(Y^1_{1-7} = 1 | S^1 = 1, S^0 = 0) \frac{\Pr(S^1 = 1, S^0 = 0)}{1}$$
$$+ \frac{1 \times \Pr(S^1 = 0, S^0 = 1)}{1} + \frac{1 \times \Pr(S^1 = 0, S^0 = 0)}{1}$$
$$= \Pr(Y^1_{1-7} = 1 | S^1 = 1, S^0 = 1) \Pr(S^1 = 1, S^0 = 1) + \Pr(Y^1_{1-7} = 1 | S^1 = 1, S^0 = 0) \Pr(S^1 = 1, S^0 = 0) + \Pr(S^1 = 0, S^0 = 1)$$
$$+ \Pr(S^1 = 0, S^0 = 0)$$
$$= \mathrm{E}[Y^1_{1-7}] \quad \blacksquare$$

In a similar manner, $\mathrm{E}[Y^0_{1-7}]$ can be expressed as follows:

$$\mathrm{E}[Y_{1-7}^{0}]$$
$$= \Pr(Y_{1-7}^{0} = 1|S^1 = 1, S^0 = 1)\Pr(S^1 = 1, S^0 = 1) + \Pr(S^1 = 1, S^0 = 0) + \Pr(Y_{1-7}^{0} = 1|S^1 = 0, S^0 = 1)\Pr(S^1 = 0, S^0 = 1)$$
$$+ \Pr(S^1 = 0, S^0 = 0)$$

(Expression 3)

Then, $\mathrm{E}[Y_{1-7}|S' = 1, A = 0]$ can be written as follows:

$$\mathrm{E}[Y_{1-7}|S' = 1, A = 0]$$
$$= \Pr(Y_{1-7}^{0} = 1|S^1 = 1, S^0 = 1)\frac{\Pr(S^1 = 1, S^0 = 1)}{\Pr(S'^0 = 1)} + \frac{\Pr(T^0 \geq t_{s'}|S^1 = 1, S^0 = 0)\Pr(S^1 = 1, S^0 = 0)}{\Pr(S'^0 = 1)}$$
$$+ \Pr(Y_{1-7}^{0} = 1|S^1 = 0, S^0 = 1)\frac{\Pr(S^1 = 0, S^0 = 1)}{\Pr(S'^0 = 1)} + \frac{\Pr(T^0 \geq t_{s'}|S^1 = 0, S^0 = 0)\Pr(S^1 = 0, S^0 = 0)}{\Pr(S'^0 = 1)}$$

(Expression 4)

When $t_{s'} = 0$, Expression 4 can be written as:

$$\Pr(Y_{1-7}^{0} = 1|S^1 = 1, S^0 = 1)\frac{\Pr(S^1 = 1, S^0 = 1)}{1} + \frac{1 \times \Pr(S^1 = 1, S^0 = 0)}{1}$$
$$+ \Pr(Y_{1-7}^{0} = 1|S^1 = 0, S^0 = 1)\frac{\Pr(S^1 = 0, S^0 = 1)}{1} + \frac{1 \times \Pr(S^1 = 0, S^0 = 0)}{1}$$
$$= \mathrm{E}[Y_{1-7}^{0}] \ \blacksquare$$

Appendix S3. *Expression for biases*

We first present an expression for the bias defined in reference to the expected outcome under treatment. The expression is obtained by subtracting Expression 1 from Expression 2.

$$
\begin{aligned}
&\mathrm{E}[Y_{1-7}|S'=1, A=1] - \mathrm{E}[Y^1_{1-7}] \\
&\quad = \Pr(Y^1_{1-7}=1|S^1=1, S^0=1)\Pr(S^1=1, S^0=1)\,\frac{1-\Pr(S'^1=1)}{\Pr(S'^1=1)} \\
&\quad + \Pr(Y^1_{1-7}=1|S^1=1, S^0=0)\Pr(S^1=1, S^0=0)\frac{1-\Pr(S'^1=1)}{\Pr(S'^1=1)} \\
&\quad + \Pr(S^1=0, S^0=1)\frac{\Pr(T^1 \geq t_{s'}\,|S^1=0, S^0=1)-\Pr(S'^1=1)}{\Pr(S'^1=1)} \\
&\quad + \Pr(S^1=0, S^0=0)\frac{\Pr(T^1 \geq t_{s'}\,|S^1=0, S^0=0)-\Pr(S'^1=1)}{\Pr(S'^1=1)}
\end{aligned}
$$

(Expression 5)

Similarly, the expression for the bias in the expected outcome under placebo is given by subtracting Expression 3 from Expression 4:

$$\mathrm{E}[Y_{1-7}|S'=1,A=0]-\mathrm{E}[Y^0_{1-7}]$$

$$=\Pr(Y^0_{1-7}=1|S^1=1,S^0=1)\Pr(S^1=1,S^0=1)\frac{1-\Pr(S'^0=1)}{\Pr(S'^0=1)}$$

$$+\Pr(S^1=1,S^0=0)\frac{\Pr(T^0\geq t_{s'}\,|S^1=1,S^0=0)-\Pr(S'^0=1)}{\Pr(S'^0=1)}$$

$$+\Pr(Y^0_{1-7}=1|S^1=0,S^0=1)\Pr(S^1=0,S^0=1)\frac{1-\Pr(S'^0=1)}{\Pr(S'^0=1)}$$

$$+\Pr(S^1=0,S^0=0)\frac{\Pr(T^0\geq t_{s'}\,|S^1=0,S^0=0)-\Pr(S'^0=1)}{\Pr(S'^0=1)}$$

(Expression 6)

Note first that, when $t_{s'}=0$, Expressions 5 and 6 are zero, as the following quantities equal 1: $\Pr(S'^1=1)$, $\Pr(T^1\geq t_{s'}\,|S^1=0,S^0=1)$, $\Pr(T^1\geq t_{s'}\,|S^1=0,S^0=0)$, $\Pr(S'^0=1)$, $\Pr(T^0\geq t_{s'}\,|S^1=1,S^0=0)$, and $\Pr(T^0\geq t_{s'}\,|S^1=0,S^0=0)$. Further, note that Expressions 5 and 6 differ in all terms that relate to the specific principal strata. Indeed, except for the proportions of the principal strata defined by S^1 and S^0, all other quantities in the two expressions can differ. It is thus possible that different combinations of parameter values might allow for the magnitudes of the biases to be equal, as in **Figure S1**, where the curves cross. Moreover, when $t_{s'}=t_s$, the biases for the expected outcomes are:

$$\mathrm{E}[Y_{1-7}|S=1,A=1]-\mathrm{E}[Y^1_{1-7}]$$

$$= \{\Pr(Y_{1-7}^{1} = 1 | S^1 = 1, S^0 = 1) \Pr(S^1 = 1, S^0 = 1)$$
$$+ \Pr(Y_{1-7}^{1} = 1 | S^1 = 1, S^0 = 0) \Pr(S^1 = 1, S^0 = 0)\} \frac{1 - \Pr(S^1 = 1, S^0 = 1) - \Pr(S^1 = 1, S^0 = 0)}{\Pr(S^1 = 1, S^0 = 1) + \Pr(S^1 = 1, S^0 = 0)} - \Pr(S^1 = 0, S^0 = 1)$$
$$- \Pr(S^1 = 0, S^0 = 0)$$

$$\mathrm{E}[Y_{1-7} | S = 1, A = 0] - \mathrm{E}[Y_{1-7}^{0}]$$
$$= \{\Pr(Y_{1-7}^{0} = 1 | S^1 = 1, S^0 = 1) \Pr(S^1 = 1, S^0 = 1)$$
$$+ \Pr(Y_{1-7}^{0} = 1 | S^1 = 0, S^0 = 1) \Pr(S^1 = 0, S^0 = 1)\} \frac{1 - \Pr(S^1 = 1, S^0 = 1) - \Pr(S^1 = 0, S^0 = 1)}{\Pr(S^1 = 1, S^0 = 1) + \Pr(S^1 = 0, S^0 = 1)} - \Pr(S^1 = 1, S^0 = 0)$$
$$- \Pr(S^1 = 0, S^0 = 0)$$

Finally, we present below the expression for the magnitude of the bias of the effect by subtracting Expression 6 from Expression 5.

$$(\mathrm{E}[Y_{1-7} | S' = 1, A = 1] - \mathrm{E}[Y_{1-7} | S' = 1, A = 0]) - (\mathrm{E}[Y_{1-7}^{1}] - \mathrm{E}[Y_{1-7}^{0}])$$
$$= (\mathrm{E}[Y_{1-7} | S' = 1, A = 1] - \mathrm{E}[Y_{1-7}^{1}]) - (\mathrm{E}[Y_{1-7} | S' = 1, A = 0] - \mathrm{E}[Y_{1-7}^{0}])$$
$$= \left\{\frac{\Pr(Y_{1-7}^{1} = 1 | S^1 = 1, S^0 = 1)\left(1 - \Pr\left(S'^{1} = 1\right)\right)}{\Pr\left(S'^{1} = 1\right)} - \frac{\Pr(Y_{1-7}^{0} = 1 | S^1 = 1, S^0 = 1)\left(1 - \Pr\left(S'^{0} = 1\right)\right)}{\Pr\left(S'^{0} = 1\right)}\right\} \Pr(S^1 = 1, S^0 = 1)$$
$$+ \left\{\frac{\Pr(Y_{1-7}^{1} = 1 | S^1 = 1, S^0 = 0)\left(1 - \Pr\left(S'^{1} = 1\right)\right)}{\Pr\left(S'^{1} = 1\right)} - \frac{\Pr(T^0 \geq t_{s'} | S^1 = 1, S^0 = 0) - \Pr\left(S'^{0} = 1\right)}{\Pr\left(S'^{0} = 1\right)}\right\} \Pr(S^1 = 1, S^0 = 0)$$
$$+ \left\{\frac{\Pr(T^1 \geq t_{s'} | S^1 = 0, S^0 = 1) - \Pr\left(S'^{1} = 1\right)}{\Pr\left(S'^{1} = 1\right)} - \frac{\Pr(Y_{1-7}^{0} = 1 | S^1 = 0, S^0 = 1)\left(1 - \Pr\left(S'^{0} = 1\right)\right)}{\Pr\left(S'^{0} = 1\right)}\right\} \Pr(S^1 = 0, S^0 = 1)$$

$$+\left\{\frac{\Pr(T^1 \geq t_{s'} \,|S^1 = 0, S^0 = 0) - \Pr\left(S'^1 = 1\right)}{\Pr\left(S'^1 = 1\right)} - \frac{\Pr(T^0 \geq t_{s'} \,|S^1 = 0, S^0 = 0) - \Pr\left(S'^0 = 1\right)}{\Pr\left(S'^0 = 1\right)}\right\} \Pr(S^1 = 0, S^0 = 0)$$

(Expression 7)

When $t_{s'} = 0$, Expression 7 is, by definition, zero.

Appendix S4. *Proof of monotonicity of biases of expected outcomes*

In this section, we show that the biases $\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y_{1-7}^{1}]$ and $\mathrm{E}[Y_{1-7}|S'=1,A=0]-\mathrm{E}[Y_{1-7}^{0}]$ vary monotonically with $t_{s'}$. For that, we take the derivative of these expressions with respect to $t_{s'}$. We, first, present the proof for $\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y_{1-7}^{1}]$.

Below to facilitate visualization, we use:

$$Q1=\Pr(Y_{1-7}^{1}=1|S^{1}=1,S^{0}=1)\Pr(S^{1}=1,S^{0}=1)+\Pr(Y_{1-7}^{1}=1|S^{1}=1,S^{0}=0)\Pr(S^{1}=1,S^{0}=0)$$

We also use the following:

$$d/dt_{s'}\left(\Pr(S'^{1}=1)\right)=\Pr(S^{1}=0,S^{0}=1)\,d/dt_{s'}\,(\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=1)+\Pr(S^{1}=0,S^{0}=0)\,d/dt_{s'}\,(\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=0))$$

The derivative of the bias is:

$$
\begin{aligned}
&{}^{d}/_{dt_{s'}}\left(\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y^{1}_{1-7}]\right)\\
&=-Q1\,\frac{{}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{1}=1\right)\right)}{\Pr\left(S'^{1}=1\right)^{2}}\\
&+\frac{\{\Pr(S^{1}=0,S^{0}=1)\Pr(S'^{1}=1)\,{}^{d}/_{dt_{s'}}\left(\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=1)\right)\}-\{\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=1)\Pr(S^{1}=0,S^{0}=1)\,{}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{1}=1\right)\right)\}}{\Pr\left(S'^{1}=1\right)^{2}}+\\
&\frac{\{\Pr(S^{1}=0,S^{0}=0)\Pr(S'^{1}=1)\,{}^{d}/_{dt_{s'}}\left(\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\right)\}-\{\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\Pr(S^{1}=0,S^{0}=0)\,{}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{1}=1\right)\right)\}}{\Pr\left(S'^{1}=1\right)^{2}}.
\end{aligned}
$$

$$
\begin{aligned}
&=\frac{{}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{1}=1\right)\right)\,(-Q1-\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=1)\Pr(S^{1}=0,S^{0}=1)-\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\Pr(S^{1}=0,S^{0}=0))}{\Pr\left(S'^{1}=1\right)^{2}}\\
&+\frac{\{\Pr(S^{1}=0,S^{0}=1)\Pr(S'^{1}=1)\,{}^{d}/_{dt_{s'}}\left(\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=1)\right)\}}{\Pr\left(S'^{1}=1\right)^{2}}+\frac{\{\Pr(S^{1}=0,S^{0}=0)\Pr(S'^{1}=1)\,{}^{d}/_{dt_{s'}}\left(\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\right)\}}{\Pr\left(S'^{1}=1\right)^{2}}.
\end{aligned}
$$

$$
\begin{aligned}
&=\frac{{}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{1}=1\right)\right)\,(-Q1-\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=1)\Pr(S^{1}=0,S^{0}=1)-\Pr(T^{1}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\Pr(S^{1}=0,S^{0}=0))}{\Pr\left(S'^{1}=1\right)^{2}}\\
&+\frac{\{{}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{1}=1\right)\right)\,\Pr(S'^{1}=1)\}}{\Pr\left(S'^{1}=1\right)^{2}}.
\end{aligned}
$$

$$= \frac{{}^{d}/_{dt_{s'}}\left(\Pr(S'^{1}=1)\right)\left(\Pr(S'^{1}=1) - Q1 - \Pr(T^{1} \geq t_{s'} \,|S^{1}=0, S^{0}=1)\Pr(S^{1}=0, S^{0}=1) - \Pr(T^{1} \geq t_{s'} \,|S^{1}=0, S^{0}=0)\Pr(S^{1}=0, S^{0}=0)\right)}{\Pr(S'^{1}=1)^{2}}$$

$$= {}^{d}/_{dt_{s'}}\Big(\Pr(S'^{1} = 1)\Big)\frac{([\Pr(S^{1}=1, S^{0}=1) - \Pr(Y^{1}_{1-7}=1 \big| S^{1}=1, S^{0}=1)\Pr(S^{1}=1, S^{0}=1)] + [\Pr(S^{1}=1, S^{0}=0) - \Pr(Y^{1}_{1-7}=1 \big| S^{1}=1, S^{0}=0)\Pr(S^{1}=1, S^{0}=0)])}{\Pr(S'^{1}=1)^{2}}$$

$$= {}^{d}/_{dt_{s'}}\left(\Pr(S'^{1}=1)\right)\frac{[\Pr(S^{1}=1, S^{0}=1)\,(1 - \Pr(Y^{1}_{1-7}=1 | S^{1}=1, S^{0}=1))] + \left[\Pr(S^{1}=1, S^{0}=0)\left(1 - \Pr(Y^{1}_{1-7}=1 \big| S^{1}=1, S^{0}=0)\right)\right]}{\Pr(S'^{1}=1)^{2}}$$

which proves that the derivative is negative as ${}^{d}/_{dt_{s'}}\left(\Pr(S'^{1}=1)\right)$ is negative, and other term is positive (as each element within squared brackets are positive and the denominator is positive) or zero, when $\Pr(Y^{1}_{1-7}=1 | S^{1}=1, S^{0}=1) = \Pr(Y^{1}_{1-7}=1 | S^{1}=1, S^{0}=0) = 1$.

We now present the proof for $\mathrm{E}[Y_{1-7} | S'=1, A=0] - \mathrm{E}[Y^{0}_{1-7}]$.

We define:

$$Q0 = \Pr(Y^{0}_{1-7}=1 | S^{1}=1, S^{0}=1)\Pr(S^{1}=1, S^{0}=1) + \Pr(Y^{0}_{1-7}=1 | S^{1}=0, S^{0}=1)\Pr(S^{1}=0, S^{0}=1)$$

We also use the following:

$$d/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)=\Pr(S^{1}=1,S^{0}=0)\,d/_{dt_{s'}}\left(\Pr(T^{0}\geq t_{s'}\,|S^{1}=1,S^{0}=0\right)+\Pr(S^{1}=0,S^{0}=0)\,d/_{dt_{s'}}\left(\Pr(T^{0}\geq t_{s'}\,|S^{1}=0,S^{0}=0\right)$$

The derivative of the bias is:

$$d/_{dt_{s'}}\left(\mathrm{E}[Y_{1-7}|S'=1,A=0]-\mathrm{E}[Y^{0}_{1-7}]\right)$$

$$=-Q0\,\frac{d/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)}{\Pr\left(S'^{0}=1\right)^{2}}$$

$$+\frac{\{\Pr(S^{1}=1,S^{0}=0)\Pr(S'^{0}=1)\,d/_{dt_{s'}}\left(\Pr(T^{0}\geq t_{s'}\,|S^{1}=1,S^{0}=0)\right)\}-\{\Pr(T^{0}\geq t_{s'}\,|S^{1}=1,S^{0}=0)\Pr(S^{1}=1,S^{0}=0)\,d/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)\}}{\Pr\left(S'^{0}=1\right)^{2}}+$$

$$\frac{\{\Pr(S^{1}=0,S^{0}=0)\Pr(S'^{0}=1)\,d/_{dt_{s'}}\left(\Pr(T^{0}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\right)\}-\{\Pr(T^{0}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\Pr(S^{1}=0,S^{0}=0)\,d/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)\}}{\Pr\left(S'^{0}=1\right)^{2}}.$$

$$=\frac{d/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)(-Q0-\Pr(T^{0}\geq t_{s'}\,|S^{1}=1,S^{0}=0)\Pr(S^{1}=1,S^{0}=0)-\Pr(T^{0}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\Pr(S^{1}=0,S^{0}=0))}{\Pr\left(S'^{0}=1\right)^{2}}$$

$$+\frac{\{\Pr(S^{1}=1,S^{0}=0)\Pr(S'^{0}=1)\,d/_{dt_{s'}}\left(\Pr(T^{0}\geq t_{s'}\,|S^{1}=1,S^{0}=0)\right)\}}{\Pr\left(S'^{0}=1\right)^{2}}+\frac{\{\Pr(S^{1}=0,S^{0}=0)\Pr(S'^{0}=1)\,d/_{dt_{s'}}\left(\Pr(T^{0}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\right)\}}{\Pr\left(S'^{0}=1\right)^{2}}$$

$$= \frac{{}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)\left(-Q0-\Pr(T^{0}\geq t_{s'}\,|S^{1}=1,S^{0}=0)\Pr(S^{1}=1,S^{0}=0)-\Pr(T^{0}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\Pr(S^{1}=0,S^{0}=0)\right)}{\Pr\left(S'^{0}=1\right)^{2}}$$
$$+\frac{{}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)\Pr(S'^{0}=1)}{\Pr\left(S'^{0}=1\right)^{2}}.$$

$$= \frac{{}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)\left(\Pr(S'^{0}=1)-Q0-\Pr(T^{0}\geq t_{s'}\,\middle|S^{1}=1,S^{0}=0\right)\Pr(S^{1}=1,S^{0}=0)-\Pr(T^{0}\geq t_{s'}\,|S^{1}=0,S^{0}=0)\Pr(S^{1}=0,S^{0}=0)\right)}{\Pr\left(S'^{0}=1\right)^{2}}$$

$$= {}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)\frac{\left[\Pr(S^{1}=1,S^{0}=1)\left(1-\Pr(Y_{1-7}^{0}=1\middle|S^{1}=1,S^{0}=1)\right)\right]+\left[\Pr(S^{1}=0,S^{0}=1)\left(1-\Pr(Y_{1-7}^{0}=1\middle|S^{1}=0,S^{0}=1)\right)\right]}{\Pr\left(S'^{0}=1\right)^{2}}$$

which proves that the derivative is negative as ${}^{d}/_{dt_{s'}}\left(\Pr\left(S'^{0}=1\right)\right)$ is negative, and other term is positive (as each element within squared brackets are positive and the denominator is positive) or zero, when $\Pr(Y_{1-7}^{0}=1|S^{1}=1,S^{0}=1)=\Pr(Y_{1-7}^{0}=1|S^{1}=0,S^{0}=1)=1$.

Supplementary Tables

Table S1. Numerical example. Here, as in **Table 1** in the main text and in the corresponding **Figure S2**, below, we assumed that the treatment was randomized, and thus that exchangeability holds (i.e., $(S^a, Y^a_{4-7}, Y^a_{1-7}) \perp\!\!\!\perp A\ (a = 0,1)$). We present proportions of different principal strata defined by S^a, Y^a_{4-7}, and Y^a_{1-7} ; the proportions of the strata defined by combining S^a response types and Y^a_{4-7}, or Y^a_{1-7}, response types are presented as products of the proportions of the corresponding S^a principal strata and the S^a principal stratum specific proportions of Y^a_{4-7}, or Y^a_{1-7}, response types; note that these proportions are the same for the two outcome definitions, and are shown in the rightmost column. The strata in which parametric assumptions differ between this table and **Table 1** are indicated in red; in particular, the proportions of the $(S^1 = 1, S^0 = 0)$ and $(S^1 = 0, S^0 = 1)$ principal strata, as well as probability of the outcomes in these strata changed.

	S^a			Y^a_{4-7}		Y^a_{1-7}		
***SY* type**	S^1	S^0	**Proportions**	Y^1_{4-7}	Y^0_{4-7}	Y^1_{1-7}	Y^0_{1-7}	**Proportions**
1	1	1	0.40	1	1	1	1	0.40 × 0.40 = 0.16
2	1	1		1	0	1	0	0.40 × 0.10 = 0.04
3	1	1		0	1	0	1	0.40 × 0.30 = 0.12
4	1	1		0	0	0	0	0.40 × 0.20 = 0.08
5	1	0	0.10	1	Undefined	1	1	0.10 × 0.40 = 0.04
6	1	0		0	Undefined	0	1	0.10 × 0.60 = 0.06
7	0	1	0.40	Undefined	1	1	1	0.40 × 0.40 = 0.16
8	0	1		Undefined	0	1	0	0.40 × 0.60 = 0.24
9	0	0	0.10	Undefined	Undefined	1	1	0.10 × 1.00 = 0.10

Table S2. Numerical example. Here, as in **Table 1** in the main text and in the corresponding **Figure S3**, we assumed that the treatment was randomized, and thus that exchangeability holds (i.e., $(S^a, Y^a_{4-7}, Y^a_{1-7}) \perp\!\!\!\perp A\ (a = 0,1)$). We present proportions of different principal strata defined by S^a, Y^a_{4-7}, and Y^a_{1-7} ; the proportions of the strata defined by combining S^a response types and Y^a_{4-7}, or Y^a_{1-7}, response types are presented as products of the proportions of the corresponding S^a principal strata and the S^a principal stratum specific proportions of Y^a_{4-7}, or Y^a_{1-7}, response types; note that these proportions are the same for the two outcome definitions, and are shown in the rightmost column. The strata in which parametric assumptions differ between this table and **Table 1** are indicated in red; in particular here the proportions of the $(S^1 = 1, S^0 = 0)$ and $(S^1 = 0, S^0 = 1)$ principal strata are equal and the probability of the outcomes in two of these strata changed.

	S^a			Y^a_{4-7}		Y^a_{1-7}		
***SY* type**	S^1	S^0	**Proportions**	Y^1_{4-7}	Y^0_{4-7}	Y^1_{1-7}	Y^0_{1-7}	**Proportions**
1	1	1	0.40	1	1	1	1	0.40 × 0.40 = 0.16
2	1	1		1	0	1	0	0.40 × 0.10 = 0.04
3	1	1		0	1	0	1	0.40 × 0.30 = 0.12
4	1	1		0	0	0	0	0.40 × 0.20 = 0.08
5	1	0	0.25	1	Undefined	1	1	0.25 × 0.40 = 0.10
6	1	0		0	Undefined	0	1	0.25 × 0.60 = 0.15
7	0	1	0.25	Undefined	1	1	1	0.25 × 0.40 = 0.10
8	0	1		Undefined	0	1	0	0.25 × 0.60 = 0.15
9	0	0	0.10	Undefined	Undefined	1	1	0.10 × 1.00 = 0.10

Supplementary Figures

Figure S1. Selection-related immortal time bias and the timing of eligibility. Here, we used the proportions of principal strata in **Table 1**. The x-axes represent the ratio of the (new) time when eligibility is defined over the original eligibility time, $t_{s'}/t_s$. The y-axis in the left panel corresponds to $\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y^1_{1-7}]$ (continuous line) and $\mathrm{E}[Y_{1-7}|S'=1,A=0]-\mathrm{E}[Y^0_{1-7}]$ (dashed line). The y-axis in the right panel shows the bias for the target effect: $(\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y_{1-7}|S'=1,A=0])-(\mathrm{E}[Y^1_{1-7}]-\mathrm{E}[Y^0_{1-7}])$. The difference between this figure and **Figure 2** is that, for individuals with potential outcomes $(S^1=0,S^0=0)$, we assumed that mortality is exponentially distributed (which implies that death events occur independently and at a constant rate $\alpha>0$) under both treatment levels, $\Pr(T^a\geq t_{s'}\,|S^1=0,S^0=0)=\exp(-\alpha t_{s'}/t_s)$ for $a=0{,}1$. The latter expression is the complement of the cumulative distribution function of the exponential distribution; here, we assume a rate parameter α of 5, so that $\Pr(T^a\geq t_s\,|S^1=0,S^0=0)\approx 0$.

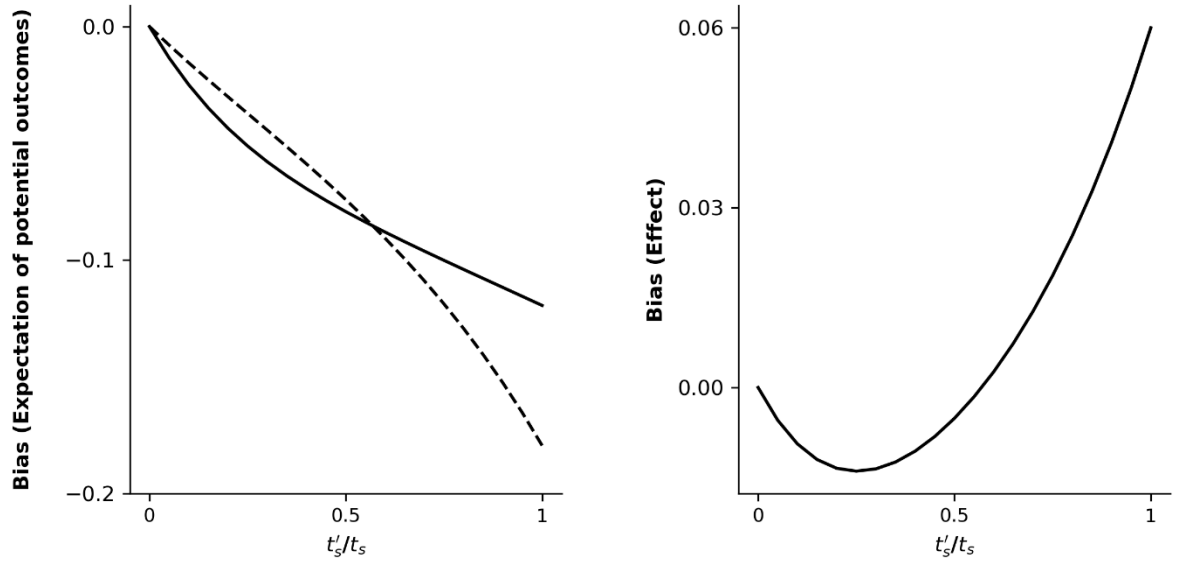


Figure S2. Selection-related immortal time bias and the timing of eligibility. For this figure, we used the proportions of the principal strata in **Table S1**. The x-axes represent the ratio of the (new) time when eligibility is defined over the original eligibility time, $t_{s'}/t_s$. The y-axis in the left panel corresponds to $\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y^1_{1-7}]$ (continuous line) and $\mathrm{E}[Y_{1-7}|S'=1,A=0]-\mathrm{E}[Y^0_{1-7}]$ (dashed line). The y-axis in the right panel shows $(\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y_{1-7}|S'=1,A=0])-(\mathrm{E}[Y^1_{1-7}]-\mathrm{E}[Y^0_{1-7}])$. Here, we assumed the same mortality functions as in **Figure 1**.

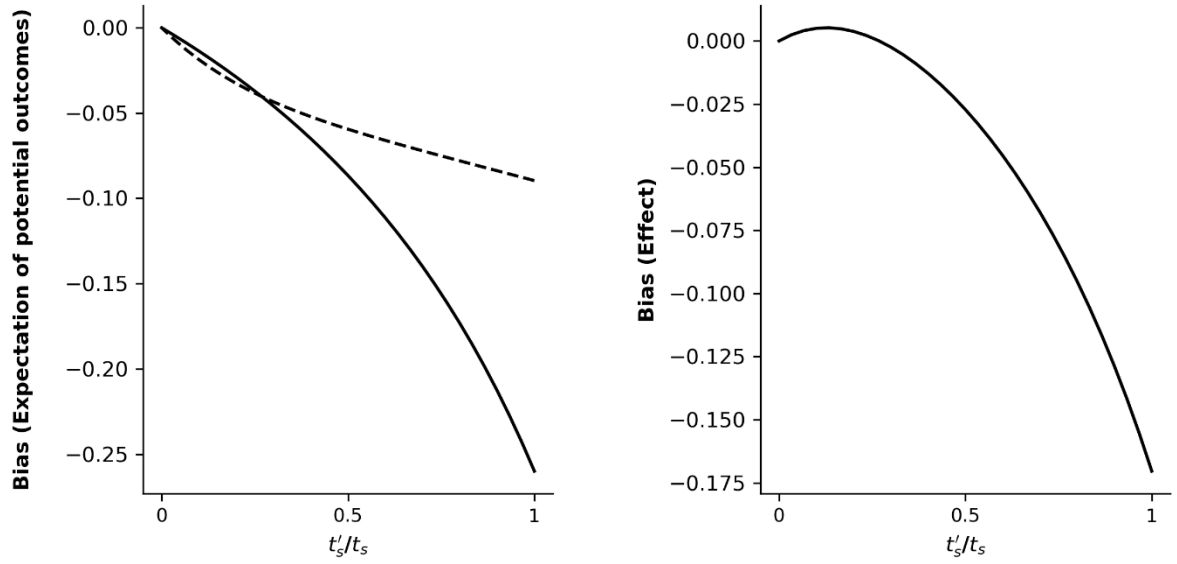


Figure S3. Selection-related immortal time bias and the timing of eligibility. For this figure, we used the proportions of the principal strata in **Table S2**. The x-axes represent the ratio of the (new) time when eligibility is defined over the original eligibility time, $t_{s'}/t_s$. The y-axis in the left panel corresponds to $\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y^1_{1-7}]$ (continuous line) and $\mathrm{E}[Y_{1-7}|S'=1,A=0]-\mathrm{E}[Y^0_{1-7}]$ (dashed line). The y-axis in the right panel shows $(\mathrm{E}[Y_{1-7}|S'=1,A=1]-\mathrm{E}[Y_{1-7}|S'=1,A=0])-(\mathrm{E}[Y^1_{1-7}]-\mathrm{E}[Y^0_{1-7}])$. Here, we assumed the same mortality functions as in **Figure 1.**