Methodological obstacles in causal inference: confounding, missing data, and measurement error

Penning de Vries, B.B.L.

7

Bias of time-varying exposure effects due to time-varying covariate measurement strategies

Bas B. L. Penning de Vries
Rolf H. H. Groenwold

Pharmacoepidemiology and Drug Safety 2021; 1–6

Abstract

Purpose. In studies of effects of time-varying drug exposures, adequate adjustment for time-varying covariates is often necessary to properly control for confounding. However, the granularity of the available covariate data may not be sufficiently fine, for example when covariates are measured for participants only when their exposure levels change. *Methods.* To illustrate the impact of choices regarding the frequency of measuring time-varying covariates, we simulated data for a large target trial and for large observational studies, varying in covariate measurement design. Covariates were measured never, on a fixed-interval basis, or each time the exposure level switched. For the analysis, it was assumed that covariates remain constant in periods of no measurement. Cumulative survival probabilities for continuous exposure and non-exposure were estimated using inverse probability weighting to adjust for time-varying confounding, with special emphasis on the difference between five-year event risks. *Results.* With monthly covariate measurements, estimates based on observational data coincided with trial-based estimates, with five-year risk differences being zero. Without measurement of baseline or post-baseline covariates, this risk difference was estimated to be 49% based on the available observational data. With measurements on a fixed-interval basis only, five-year risk differences deviated from the null, to 29% for six-monthly measurements, and with magnitude increasing up to 35% as the interval length increased. Risk difference estimates diverged from the null to as low as −18% when covariates were measured depending on exposure level switching. *Conclusion.* Our simulations highlight the need for careful consideration of time-varying covariates in designing studies on time-varying exposures. We caution against implementing designs with long intervals between measurements. The maximum length required will depend on the rates at which treatments and covariates change, with higher rates requiring shorter measurement intervals.

7.1 Introduction

In many pharmacoepidemiologic studies, the use of the drugs that are investigated may change over time. In case of such time-varying exposures, the exposure effect can be defined in different ways. For example, one could contrast initiating drug treatment at a particular point in time (irrespective of whether the use is continued) with not initiating, or continuous drug use with continuous non-use. While analyses of point interventions (e.g., a single-dose vaccination) require adjustment for confounding at baseline only, for analyses of a time-varying exposure, information on time-varying covariates might be required to mitigate bias due to time-varying confounding. However, the granularity of the available information about the time-varying covariates may not be sufficiently fine to adequately control for confounding.

One special case of where this issue may arise is where researchers choose to measure covariates for study subjects only when their exposure levels have changed since the last measurement. If exposure levels do not change, covariate levels are (implicitly) assumed to remain constant, which is an implementation of a method generally known as last-observation-carried-forward (LOCF). The accurateness of the observed covariate data may then depend on the observed exposure history. In studies of antidepressant use and the risk of hip fracture, for example, comorbidities and use of co-medication were assessed only at baseline and whenever patients switched exposure level or after every six months in the absence of switching (Ali et al., 2016; Souverein et al., 2016).

In this paper, we investigate the impact of various covariate measurement designs on the estimation of time-varying exposure effects in observational studies with time-varying confounding. We illustrate, by way of simulation, the potential for bias of inverse-probability-weighting (IPW) estimators under static designs of fixed-interval covariate measurement and under dynamic designs with covariates being measured depending on the observed exposure history. IPW estimators are considered as these are increasingly used for estimating causal effects of time-varying exposures, can accommodate exposure-covariate feedback (Hernán and Robins, 2020), and readily allow for 'adjusted' survival curves to be created (Cole and Hernán, 2004).

7.2 Methods

We first simulated data for a hypothetical study, the 'target trial', which if implemented on theoretical population of interest would readily allow us to identify the exposure effect of interest (Hernán and Robins, 2016). In practice, it

is not always possible to implement a target trial, but we use it here as a means to clarify the exposure effect of interest and we simulate from it to give a reference against which to compare results from analyses that are based on simulated data for observational studies. We considered multiple observational studies, each with the same data-generating mechanism but with different covariate measurement designs to evaluate their impact. Having simulated data, we then estimated the survival curves for the period of five years, using a weighting approach (described below) that was designed to keep treatment arms comparable throughout follow-up in terms of measured covariates. For each of the trial and observational studies, we first generated data on a single sample of $n = 150\,000$ individuals, which is sufficiently large to allow us to ignore sampling variability and regard differences between the survival curves as measures of the impact of the measurement designs on the large sample bias of the IPW estimators. The results corresponding to this single simulation run are described in detail below. In the online supplementary material, we summarise the results of 5000 independent simulation runs for sample sizes 150 000, 10 000, 1000, and 100. R code for the simulations is also provided as online supplementary material.

Set-up

The target trial has the following key design elements: (1) study participants (subjects who satisfy the eligibility criteria) are randomised at a well-defined baseline time point t_0 to either being issued a drug prescription ($A_0 = 1$)—say, a prescription for a daily dose of some antidepressant drug for the next one-month period—or to not being issued the prescription ($A_0 = 0$) at t_0; (2) participants are then followed over time until the occurrence of an event (e.g., the first hip fracture or death if the subject dies without having sustained a hip fracture during follow-up) or the administrative study end, whichever comes first; (3) provided event-free survival is long enough, study participants in the ($A_0 = 1$)-group are issued a further prescription after every month since t_0 and those in the ($A_0 = 0$)-group do not receive a prescription during follow-up. For a given subject, we define A_k to be the indicator variable that takes the value of 1 if the subject is on a one-month prescription on month k; $A_k = 0$ otherwise. We further define Y to be the amount of follow-up time between baseline and the subject's (first) event and let Y_k be that part of Y that relates to month k. We stipulate that study participants are event-free at the start of the study and that subjects do not get lost to follow-up before the administrative study end, which we stipulated to be five years (or $K = 60$ months) post-baseline.

The observational studies differ from the target trial in the following ways

only: (1) the decision to allocate a subject to $A_0 = 1$ versus $A_0 = 0$ is not made by randomisation; (2) the decisions to renew prescriptions for subjects in the $(A_0 = 1)$-group or to never issue a prescription throughout the follow-up period for those in the $(A_0 = 0)$-group are not determined by their baseline allocations A_0. Rather, for month $k = 0, 1, ...$, the decision to set exposure A_k to 0 or 1 is based only on past exposure history $(A_j : j < k)$ and certain binary covariates L_k. In this observational setting, subjects can switch at the start of each month between exposure levels 'being on prescription' (or 'exposed') versus 'not being on prescription' (or 'not exposed'). In variations on this setting, covariate data were measured according to one of the following measurement designs: (1) covariates were not measured at all, thus precluding any adjustment for confounding and effectively forcing us to implement a 'crude' estimator; (2) covariates were measured on a monthly basis, which is sufficient for identification of our target quantity; (3) covariates were measured on a six-monthly basis starting at baseline; (4) covariates were measured when the respective subject's exposure level switched; (5) covariates were measured with an exposure level switch and at a six-monthly basis in the absence of exposure level switching. We also considered variations on designs (3) and (5) where, instead of six months, the fixed measurement interval have a length of 2, 3, 9, 12, ..., or 60 months. Where design (3) means that measurement times are known before the start of follow-up, designs (4) and (5) are dynamic in the sense that whether or not a subject's covariate level is measured depends on the subject's time-varying variables.

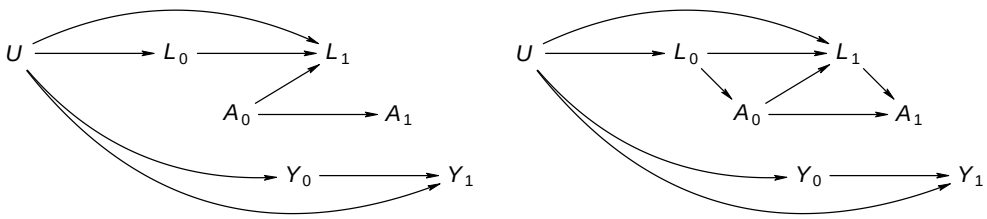

Figure 7.1: Directed acyclic graphs representing the data-generating mechanism for the first two months of the target trial (left) and observational study (right). Here, U represents a unmeasured common cause of the measured covariates L_0, L_1 and outcome variables Y_0, Y_1. The absence of directed paths from exposure variables to outcome variables reflects the absence of a causal exposure-outcome effect.

Data-generating mechanism

To simulate longitudinal data for a setting with time-varying confounding we used a variation on the approaches described by Havercroft and Didelez (2012) and Young and Tchetgen Tchetgen (2014). The data-generating mechanisms for the target trial and observational studies are described in the Appendix and produce data that are consistent with the directed acyclic graphs (DAGs) of Figure 7.1. In the trial setting (left panel of Figure 7.1), the absence of arrows going into the exposure variables reflects the absence of (time-varying) confounding. In the target trial, post-baseline exposures are fully determined by the baseline level of exposure, which takes the value of 1 for half of subjects (i.e., exposure status does not change over time). In the observational study, however, approximately 40% of subjects will have switched exposure level by the end of follow-up in each of the arms that are defined by baseline exposure level.

Defining and estimating the exposure effect

We define the exposure effect of interest as a contrast between continuous exposure ($A_j = 1$ for $j = 0, 1, ...$) versus continuous non-exposure ($A_j = 0$ for $j = 0, 1, ...$). In particular, we suppose that the interest lies with a contrast between the five-year event-free survival probabilities that we would observe had everyone received continuous exposure versus continuous non-exposure; i.e., a contrast that is identified in the target trial as

$$\Pr(Y \geq 60 | A_0 = 1) \quad \text{versus} \quad \Pr(Y \geq 60 | A_0 = 0).$$

As indicated by the absence of a directed path of arrows from the exposure variables to the outcome variables in the DAG for the target trial, the difference between these two survival probabilities is zero.

To account for time-varying confounding in the observational studies, we implemented IPW by applying a crude (Kaplan-Meier) estimator to an artificial data set where, given any time during follow-up, a subject received a weight of zero if the subject had experienced an exposure level switch by that time and otherwise a weight equal to the reciprocal of the product of the estimated probabilities of their observed exposure levels until that time given the respective measured exposure and covariate histories. That is, for $a = 0, 1$, a subject's weight for month k was

$$W_k = \prod_{j=0}^{k} \frac{1}{\Pr(A_j = a | Y \geq j, A_0 = ... = A_{j-1} = a, L_0, ..., L_j)}$$

if the subject received exposure level a in months 0 through k (i.e., $A_0 = ... = A_k = a$). Subjects were censored (i.e., received a weight of zero) from the time at which they switched to another exposure level. Apart from the covariate measurement design, the validity of the approach also rests on the correct specification of the model for the conditional treatment probabilities. To ensure correct specification for the reference measurement design (1), we assumed that the exposure A_k given survival and past exposure and covariate levels was Bernoulli distributed with mean equal to

$$\begin{aligned}&\Pr(A_k = 1 | Y \geq k, A_0, ..., A_{k-1}, L_0, ..., L_k) \\ &\quad = \frac{\exp[\alpha_0 + \alpha_1 I(k=0) + \alpha_2 A_{k-1} + \alpha_3 L_k]}{1 + \exp[\alpha_0 + \alpha_1 I(k=0) + \alpha_2 A_{k-1} + \alpha_3 L_k]}\end{aligned}$$

for some $\alpha_0, \alpha_1, \alpha_2, \alpha_3$, which were estimated by a pooled logistic regression under this model. Throughout, variables that were unobserved by measurement design were handled with LOCF.

7.3 Results

Figure 7.2 shows the estimated survival curves for the 'always treat' and 'never treat' protocols. Consistent with the absence of a directed path from the exposure variables to the outcome variables in the DAGs of Figure 7.1, the trial-based estimates of the survival curves overlap (Figure 7.2, panel A). Where we observed a five-year event risk of 31% in both arms of the target trial, in the observational setting, we observed a risk of 64% and 15% in those who do and those who do not receive a treatment prescription at baseline, respectively, giving a risk difference of 49% (panel B). With monthly covariate measurement, IPW resulted in survival curves that virtually coincide with those of the trial (panel C), for which we found a risk difference of zero. Six-monthly measurements (panel D), however, brought the curves closer to those of the no measurement setting (panel B), i.e., in the 'direction of confounding'. The five-year risks with six-monthly measurements were estimated to be to 50% and 21%, respectively, giving a risk-difference of 29%. In Figure 7.3, panel A, it is shown that the estimated risk differences at two and five years increase with the interval measurement length, until they reach a plateau of approximately 20% and 35%, respectively. When the interval length was set equal to the maximum follow-up duration (60 months), only baseline covariates were measured, which resulted in an estimated five-year risk difference that was approximately 15 percent points closer to the target than that of no covariate measurement at all (Figure 7.2, panel B). When we

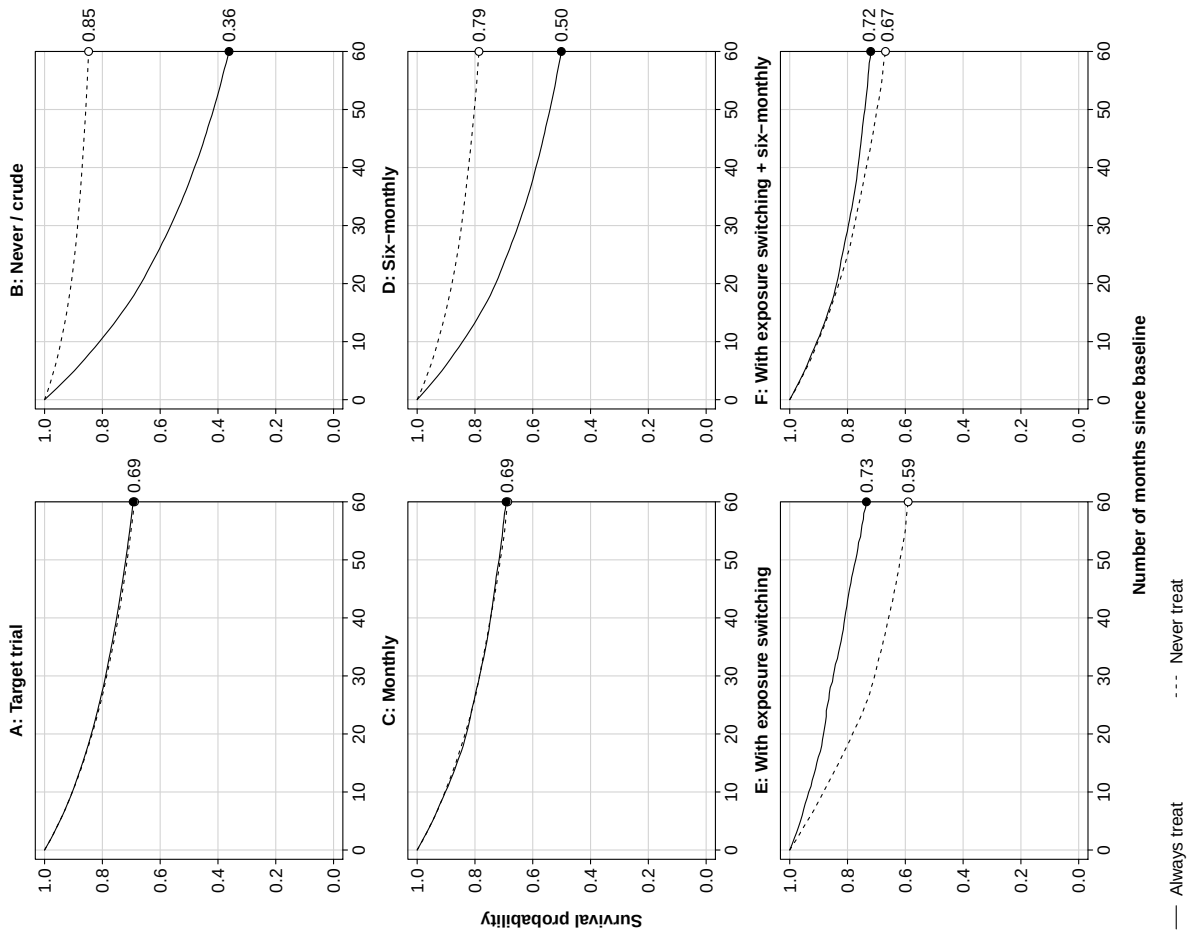

Figure 7.2: Estimated event-free survival curves for 'always treat' and 'never treat' protocols based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

implemented measurement design (4), the estimated 5-year risk difference flipped to the other side of the null, −14% (panel E), with five-year risks estimated to be 27% and 41% for the 'always treat' and 'never treat' protocols, respectively. For design (5), we observed a 5-year risk difference of −5%, somewhere between the results of design (3) and (4) (panel F). With increasingly large measurement intervals within periods of no switching, the estimated two-year risk difference steadily decreased to approximately −15% (Figure 7.3, panel B). The estimated five-year risk was also −15% with 60 months between measurements in periods of no switching, equal to the observed risk of design (4), as expected. However, it was lowest, approximately −18%, with an interval length of around 30 months.

The bias estimates of the survival curves and 5-year risk differences that were derived by averaging across 5000 independent samples of sizes 150 000, 10 000 and 1000 are nearly identical to the corresponding estimates described above and given in Figures 7.2 and 7.3 (cf. online supplementary material). For sample size 100, however, we observed substantial (small sample) bias for all measurement designs, even in the reference observational setting with full/monthly covariate measurement.

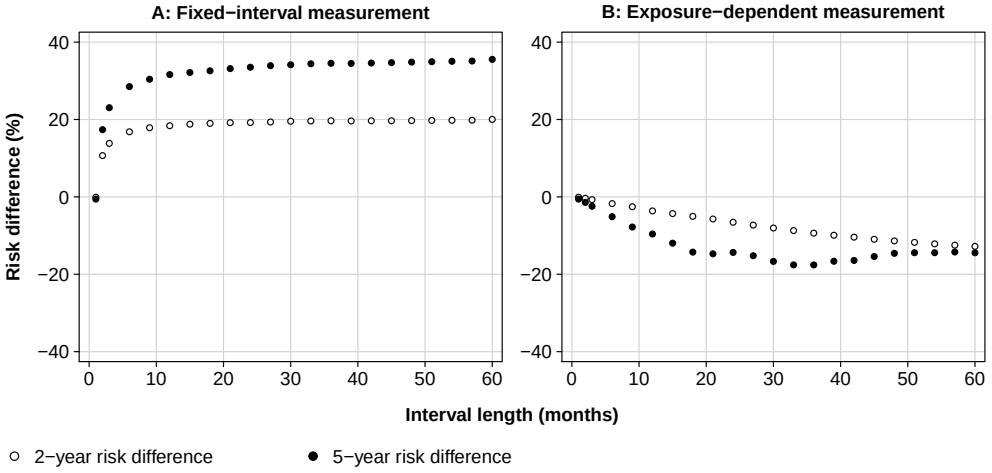

Figure 7.3: Estimated two- and five-year event risk differences comparing 'always treat' versus 'never treat' protocols. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

7.4 Discussion

We used simulation to study and illustrate the potential for bias due to measurement design choices in the estimation of the effects of time-varying exposures. The potential for bias in settings with static or fixed-interval covariate measurement designs has recently been illustrated already (Young et al., 2019). We additionally showed that bias might arise in settings where decisions to measure are driven by observed values of the time-varying exposure.

As expected, in our simulations, fixed-interval measurement resulted in bias in the direction of confounding, bias that is attributable to residual confounding. Interestingly, we found bias in the opposite direction when we implemented measurement designs where covariates were measured preferentially with exposure level switches. Together with LOCF, these measurement designs introduced a form of differential misclassification, which may result in bias even in the absence of confounding (Webster-Clark et al., 2020). Researchers familiar with DAGs might be alerted by the presence of colliders in the DAG that encodes part of the misclassification mechanism. For example, on the DAG of the right panel of Figure 7.1, the differential misclassification of L_1 can be represented by adding a measured version of L_1 with incoming arrows from L_0, L_1, A_0 and A_1. The measured variable can then be seen to be a collider on the path from A_1 to Y_1 via L_1 and U. By conditioning on the collider (and not the unmeasured variable L_1 or U), the path is opened, potentially leading to collider-stratification bias (Hernán and Robins, 2020).

In addition to adequate measurement of the time-varying covariates, the validity of IPW rests on the correct specification of the model for the distribution of the treatment variables given survival and past covariate and exposure levels. It is possible that the biases that we observed are partly due to model misspecification.

We considered a specific and relatively simple setting with a single, binary covariate, no censoring before the administrative study end and an interest in static rather than dynamic treatment strategies. These features are not required for valid inference with IPW (Hernán and Robins, 2020). However, the magnitude and direction of bias in other settings may differ from those observed in the current study. We stress that the bias that was observed in our simulation does not depend critically on the choice of IPW as a means to control for time-varying confounding. The choices regarding the frequency of covariate measurements will likely also affect the properties other methods, including the commonly applied Cox' regression analysis with time-varying covariates. The extent to which such choices impact a particular study are obviously context-specific. For example,

it will likely depend on the rate at which subjects cross over between treatment arms as well as on the extent to which covariates are subject to change over time.

In conclusion, our simulations highlight the need for adequate measurement of time-varying covariates in observational studies on the effects of time-varying exposures. Researchers should consider differential covariate misclassification as a possible source of bias when designing covariate measurement strategies (Webster-Clark et al., 2020). Whether or not covariates are measured with every exposure level switch, we caution against implementing measurement designs with long intervals between measurements, particularly when the impact of the design choices are poorly understood. The maximum interval length that is sufficient to yield negligible bias will depend on the rates at which treatments and covariates can change (Young et al., 2019), with higher rates requiring shorter measurement intervals.

References

Ali, M. S., R. H. Groenwold, S. V. Belitser, P. C. Souverein, E. Martín, N. M. Gatto, C. Huerta, H. Gardarsdottir, K. C. Roes, A. W. Hoes, et al. (2016): "Methodological comparison of marginal structural model, time-varying cox regression, and propensity score methods: the example of antidepressant use and the risk of hip fracture," *Pharmacoepidemiology and Drug Safety*, 25, 114–121.

Cole, S. R. and M. A. Hernán (2004): "Adjusted survival curves with inverse probability weights," *Computer methods and programs in biomedicine*, 75, 45–49.

Havercroft, W. and V. Didelez (2012): "Simulating from marginal structural models with time-dependent confounding," *Statistics in medicine*, 31, 4190–4206.

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

Hernán, M. A. and J. M. Robins (2016): "Using big data to emulate a target trial when a randomized trial is not available," *American journal of epidemiology*, 183, 758–764.

Souverein, P. C., V. Abbing-Karahagopian, E. Martin, C. Huerta, F. de Abajo, H. G. Leufkens, G. Candore, Y. Alvarez, J. Slattery, M. Miret, et al.

(2016): "Understanding inconsistency in the results from observational pharmacoepidemiological studies: the case of antidepressant use and risk of hip/femur fractures," *Pharmacoepidemiology and Drug Safety*, 25, 88–102.

Webster-Clark, M., M. Jonsson Funk, and T. Stürmer (2020): "Single-arm trials with external comparators and confounder misclassification: How adjustment can fail," *Medical Care*, 58, 1116–1121.

Young, J. G. and E. J. Tchetgen Tchetgen (2014): "Simulation from a known Cox MSM using standard parametric models for the g-formula," *Statistics in medicine*, 33, 1001–1014.

Young, J. G., R. Vatsa, E. J. Murray, and M. A. Hernan (2019): "Interval-cohort designs and bias in the estimation of per-protocol effects: a simulation study," *Trials*, 20, 552.

Supplementary Material

Table S7.1: Summary of estimated 5-year (always-versus-never-exposed) risk differences over 5000 simulation runs for sample sizes 150 000, 10 000, 1000, and 100. (Continued on next page.)

Study/measurement design[†]	Mean estimate (95% CI)[‡]	Empirical variance	Mean squared error
	Sample size: 150 000		
A: Target trial	-0.000 (-0.000, 0.000)	0.000	0.000
B: Observational study 1	0.485 (0.485, 0.485)	0.000	0.235
C: Observational study 2	-0.000 (-0.000, 0.000)	0.000	0.000
D: Observational study 3	0.286 (0.286, 0.286)	0.000	0.082
E: Observational study 4	-0.134 (-0.134, -0.134)	0.000	0.018
F: Observational study 5	-0.044 (-0.044, -0.043)	0.000	0.002
	Sample size: 10 000		
A: Target trial	0.000 (-0.000, 0.000)	0.000	0.000
B: Observational study 1	0.485 (0.485, 0.486)	0.000	0.236
C: Observational study 2	0.000 (-0.000, 0.001)	0.001	0.001
D: Observational study 3	0.286 (0.286, 0.287)	0.000	0.082
E: Observational study 4	-0.135 (-0.136, -0.134)	0.002	0.021
F: Observational study 5	-0.043 (-0.044, -0.042)	0.001	0.003

[†]The target trial and observational studies are described in the main text. Observational studies 1 through 5 differ in covariate measurement design: in observational study 1 (B), covariates were never measured; in study 2 (C), covariates were measured on a monthly basis; in study 3 (D), covariates were measured on a six-monthly basis starting at baseline; in study 4 (E), covariates were measured when the respective subject's exposure level switched; ...

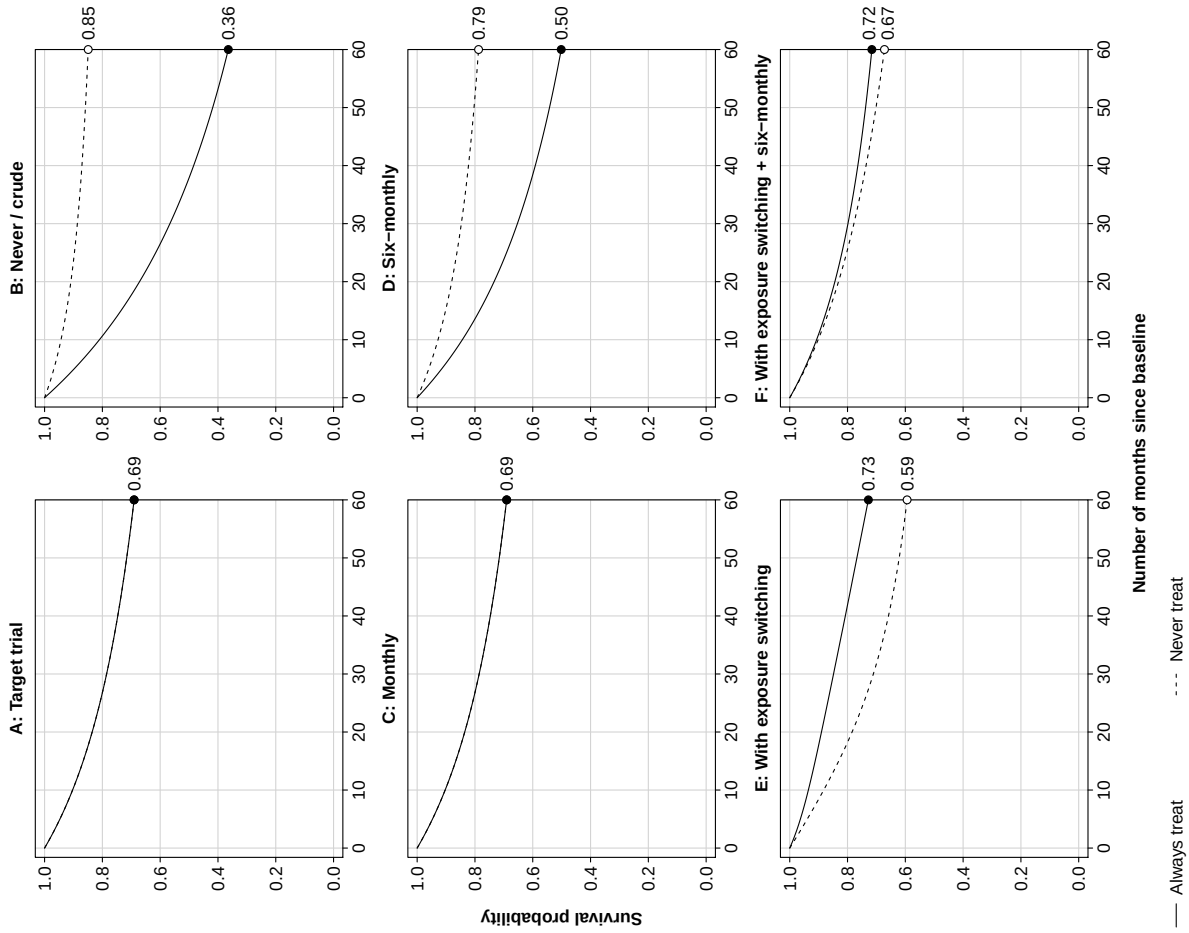

Figure S7.1: Mean estimated event-free survival probabilities across 5000 samples of size 150 000 based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

Table S7.1 continued.

Study/measurement design†	Mean estimate (95% CI)‡	Empirical variance	Mean squared error
	Sample size: 1000		
A: Target trial	0.000 (-0.001, 0.001)	0.001	0.001
B: Observational study 1	0.486 (0.485, 0.487)	0.002	0.238
C: Observational study 2	0.014 (0.012, 0.016)	0.007	0.007
D: Observational study 3	0.291 (0.289, 0.292)	0.003	0.087
E: Observational study 4	-0.132 (-0.136, -0.128)	0.026	0.044
F: Observational study 5	-0.028 (-0.031, -0.025)	0.009	0.009
	Sample size: 100		
A: Target trial	-0.000 (-0.003, 0.003)	0.009	0.009
B: Observational study 1	0.484 (0.481, 0.488)	0.016	0.251
C: Observational study 2	0.117 (0.110, 0.123)	0.060	0.073
D: Observational study 3	0.320 (0.315, 0.324)	0.026	0.129
E: Observational study 4	-0.004 (-0.014, 0.007)	0.154	0.154
F: Observational study 5	0.091 (0.084, 0.099)	0.072	0.080

... in study 5 (F), covariates were measured with an exposure level switch and at a six-monthly basis in the absence of exposure level switching. ‡95% CI refers to the pointwise 95% confidence interval $\hat{\mu} \pm 1.96\sqrt{\hat{\sigma}^2/5000}$, where $\hat{\mu}$ denotes the mean estimated risk difference and $\hat{\sigma}^2$ its empirical variance, i.e., the sample variance of the sample of 5000 estimates

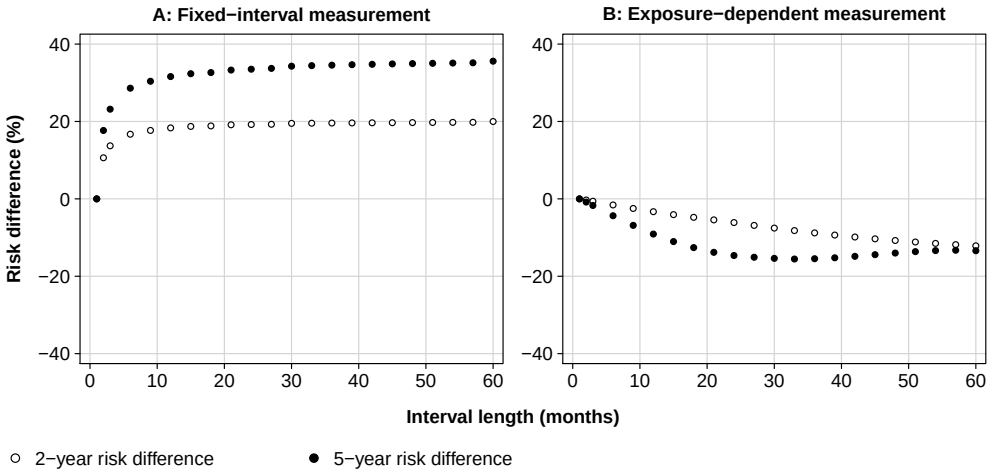

Figure S7.2: Mean estimated two- and five-year event risk differences across 5000 samples of size 150 000. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

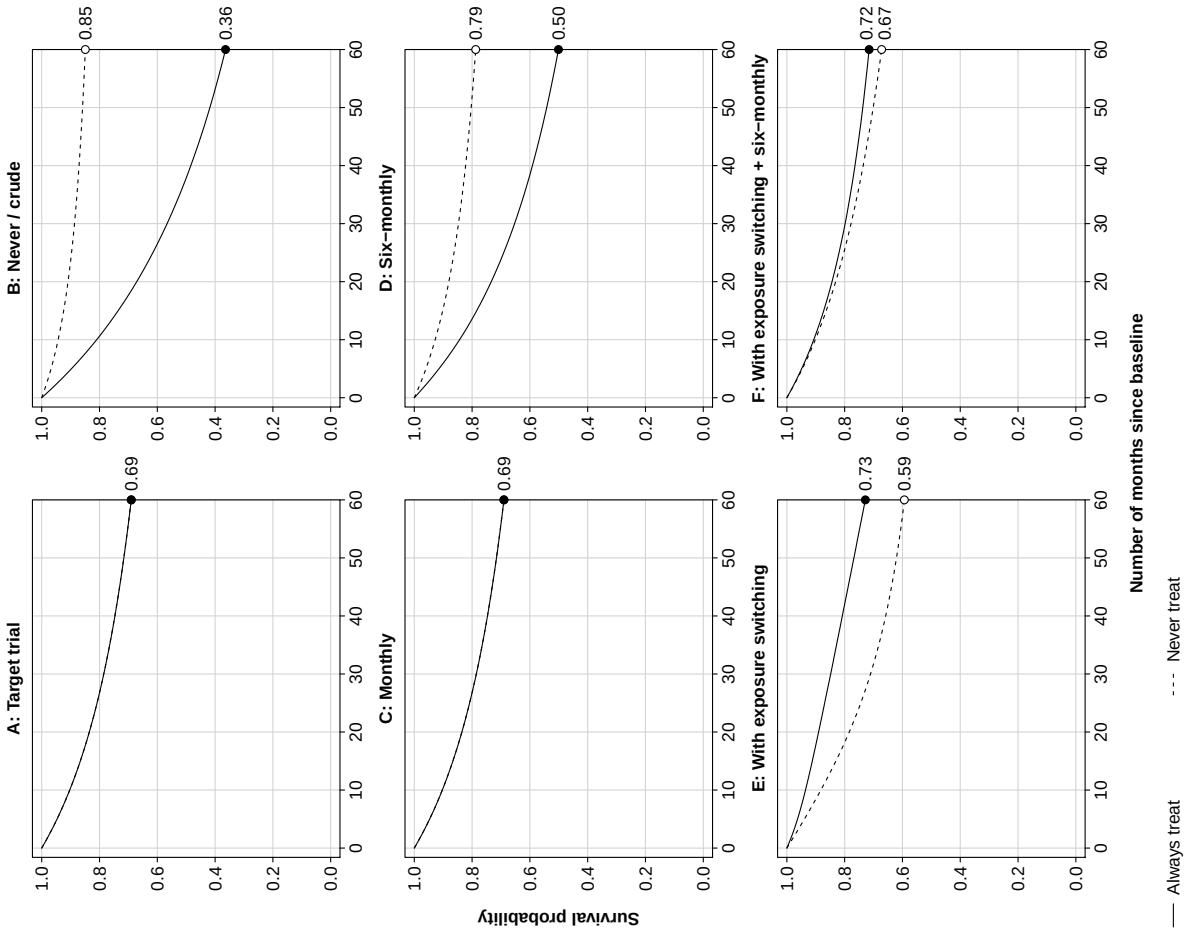

Figure S7.3: Mean estimated event-free survival probabilities across 5000 samples of size 10 000 based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

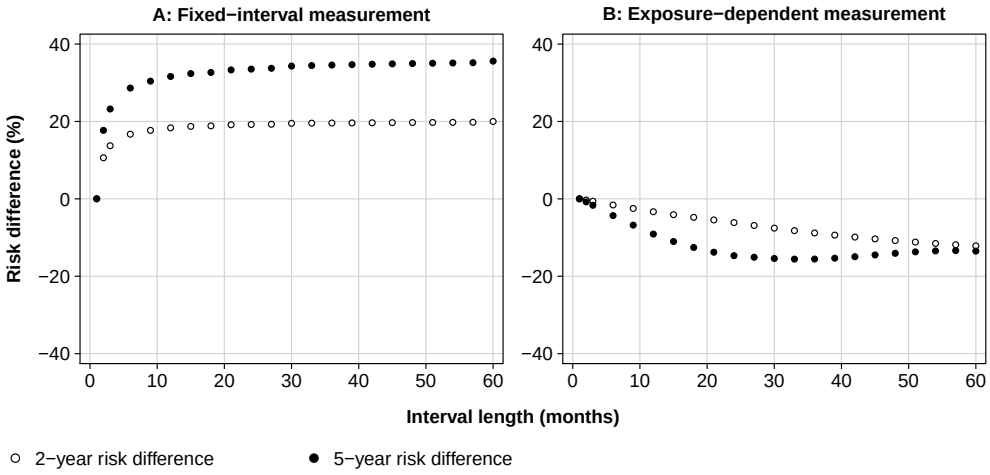

Figure S7.4: Mean estimated two- and five-year event risk differences across 5000 samples of size 10 000. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

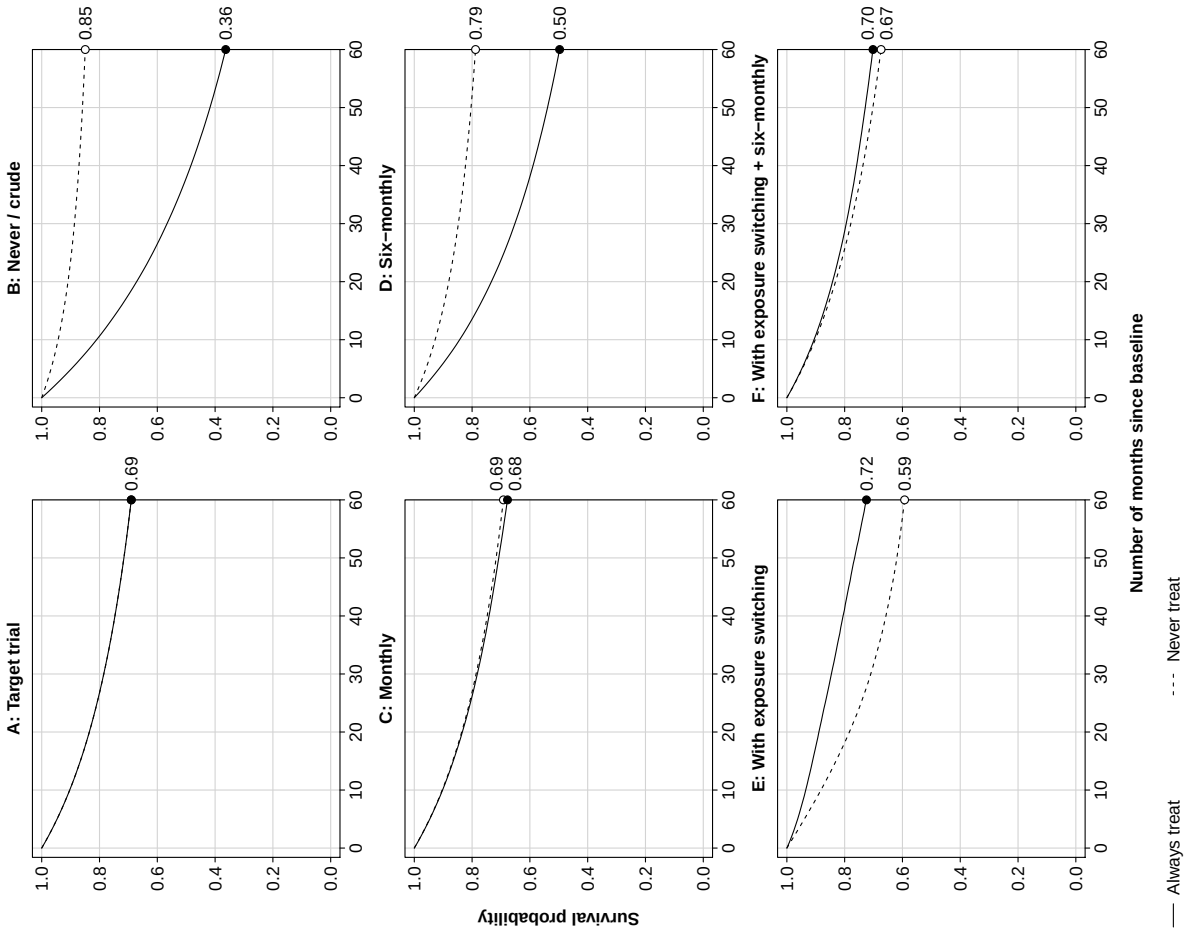

Figure S7.5: Mean estimated event-free survival probabilities across 5000 samples of size 1000 based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

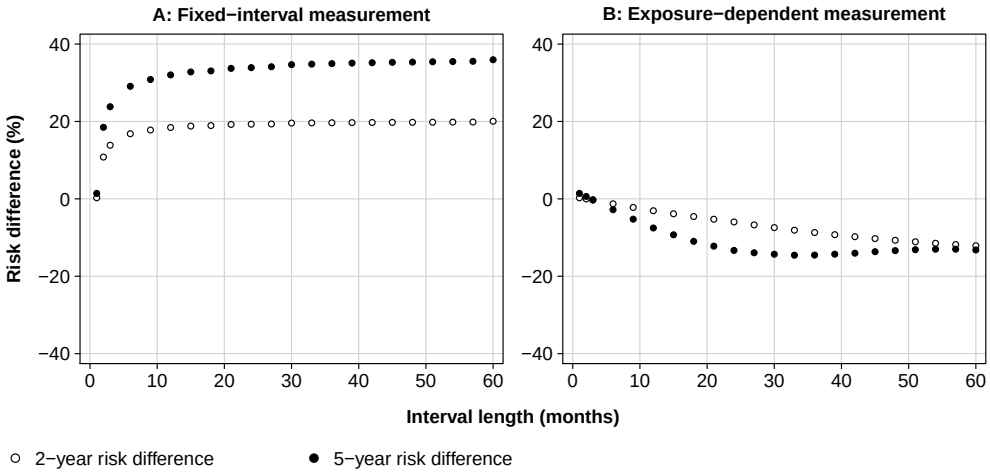

Figure S7.6: Mean estimated two- and five-year event risk differences across 5000 samples of size 1000. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

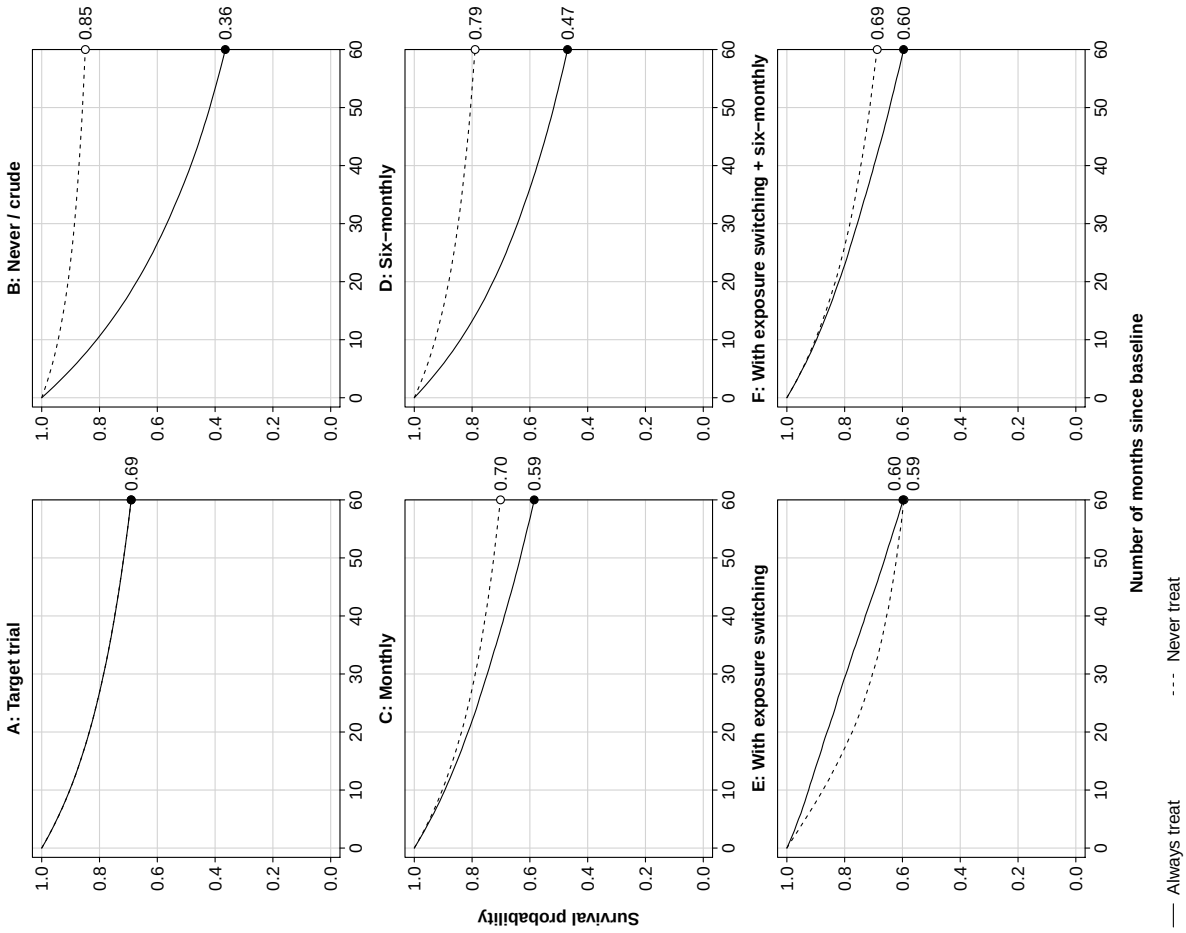

Figure S7.7: Mean estimated event-free survival probabilities across 5000 samples of size 100 based on target trial (panel A) and observational study (B through F) with varying covariate measurement designs: no covariate measurement (B), continuous to monthly covariate measurement (C), six-monthly covariate measurement (D), covariate measurement only with covariate level switching (E), and with exposure switching and six-monthly in periods without switching (F).

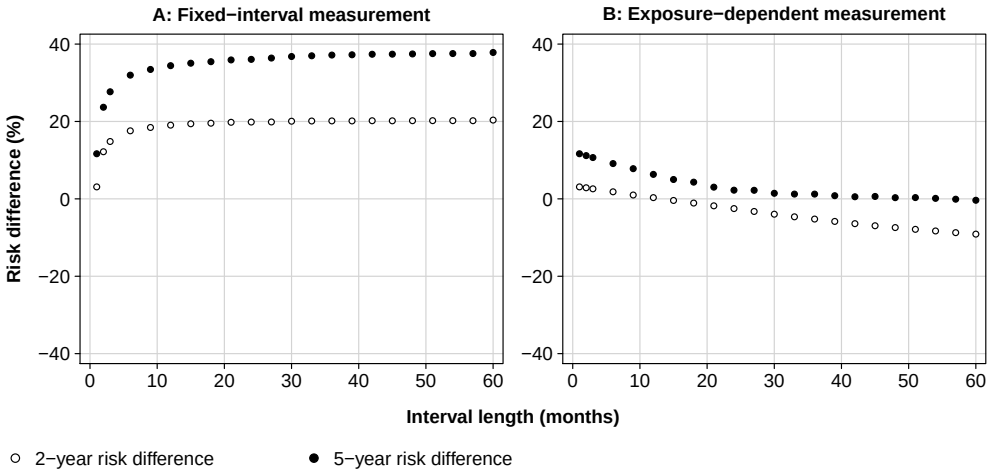

Figure S7.8: Mean estimated two- and five-year event risk differences across 5000 samples of size 100. Estimates derive from observational studies with varying covariate measurement designs. Panel A gives the estimates for fixed-interval measurement; panel B gives the estimates for covariate measurement with exposure switching and with fixed-length intervals in periods without switching.

Supplementary R code

```
# R code to supplement 'Bias of time-varying exposure effects due to
#   time-varying covariate measurement strategies'
# Compiled by Bas B.L. Penning de Vries (last updated: 7 Dec 2020)

# ====================================================================
# Preliminaries
# ====================================================================

# settings
K <- 60L # maximum number of months of follow-up
n <- 1.5e5L # sample size

# useful functions:
expit <- function(x) 1/(1+exp(-x))
locf <- function(x){
  # Last Observation Carried Forward
  isNA <- is.na(x)
  if(isNA[1L]) stop('the first element is NA.')
  y <- rep(x[!isNA],tabulate(cumsum(!isNA)))
  return(y)
}
qFirst <- function(x,last=FALSE){
  # Tests whether elements in x are the first occurrence of the
  #   corresponding values
  if(last) x <- rev(x)
  n <- length(x)
  w <- seq_len(n)
  o <- order(x)
  x <- x[o]
  y <- c(TRUE,x[-1L]!=x[-n])
  z <- y[match(w,o)]
  if(last) z <- rev(z)
  return(z)
}

# ====================================================================
# Data generating mechanism
# ====================================================================

drawSample <- function(n,trial=FALSE){
  sq <- seq(0,K-1e-6)
  A <- L <- matrix(nrow=n,ncol=length(sq))
  colnames(A) <- paste0("A",sq)
  colnames(L) <- paste0("L",sq)
  U <- runif(n)
  lagL <- lagA <- rep(0L,n)
  S <- rep(0,n)
  for(j in seq_along(sq)){
    Surv <- S>=(j-1L)
    Lk <- ifelse(Surv,runif(n)<expit(-.5+.25*(j==1L)+6*(U-.5)+
      .5*(lagA-.5)+1*(lagL-.5)),NA)
    g <- if(trial) {if(j==1L) rep(.5,n) else lagA} else
```

```
      expit(4*(j==1L)+10*(lagA-.5)+4*(Lk-.5))
    Ak <- ifelse(Surv,runif(n)<g,NA)
    L[,j] <- Lk
    A[,j] <- Ak
    eta <- 7*(U-.5)
    s <- suppressWarnings(rexp(n,rate=exp(-6+eta)))
    s[is.na(s)] <- 0
    S <- S+Surv*pmin(s,1L)
    Surv <- ifelse(Surv,s>1L,FALSE)
    lagL <- Lk; lagA <- Ak
  }
  status <- S<K
  S[!status] <- K
  return(data.frame(L*1L,A*1L,S=S,status=status))
}
coarsen <- function(data,design="I",after=6L){
  # assumes 'data' to be in long format
  out <- switch(design,
    I={ # exposure switch
      lagA <- ifelse(qFirst(data$unit),0L,c(0L,data$A[-nrow(data)]))
      M <- (data$start!=0L)&data$A==lagA&data$start>=0L
      data$L[M] <- NA
      return(data)
    },
    II={
    # set to missing if not multiple of 'after' months from baseline
      M <- data$start%%after!=0L&data$start>=0L
      data$L[M] <- NA
      return(data)
    },
    III={
    # set to missing if no switch of exposure AND not multiple of
    #   'after' months from since last exposure switch
      lagA <- ifelse(qFirst(data$unit),0L,c(0L,data$A[-nrow(data)]))
      M <- (data$start!=0L)&data$A==lagA&data$start>=0L
      cs <- cumsum(M)
      wh <- cumsum(!M)
      qf <- qFirst(wh)
      monthsSinceLastSwitch <- cs-cs[qf][match(wh,unique(wh))]
      M[!monthsSinceLastSwitch%%after] <- FALSE
      data$L[M] <- NA
      return(data)
    }
  )
}

# ========================================================================
# Data pre-processing functions
# ========================================================================

longFormat <- function(data){
  n <- nrow(data)
  w_L <- grep("^L-[:0-9:]+$",colnames(data))
  wL <- grep("^L[:0-9:]+$",colnames(data))
  wA <- grep("^A[:0-9:]+$",colnames(data))
```

```
  unit <- matrix(seq_len(n),nrow=n,ncol=length(wA),
    byrow=FALSE)[!is.na(data[,wA])]
  column <- matrix(seq_along(wA),nrow=n,ncol=length(wA),
    byrow=TRUE)[!is.na(data[,wA])]
  w <- cbind(unit,column)[order(unit),]
  out <- data.frame(unit=w[,1L],start=w[,2L]-1L)
  out$stop <- out$start+1L
  out$stop[qFirst(out$unit,TRUE)] <- data$S
  out$L <- data[,wL][w]
  out$A <- data[,wA][w]
  out$event <- FALSE
  out$event[qFirst(out$unit,TRUE)] <- data$status
  rownames(out) <- NULL
  return(out)
}
lagVariables <- function(data,m=1L){
  lagL <- matrix(nrow=nrow(data),ncol=m)
  colnames(lagL) <- paste0("lag",seq_len(m),"L")
  lagA <- matrix(nrow=nrow(data),ncol=m)
  colnames(lagA) <- paste0("lag",seq_len(m),"A")
  wh <- which(colnames(data)%in%c(colnames(lagL),colnames(lagA)))
  if(any(wh)) data <- data[,-wh,drop=FALSE]
  record <- data$start-min(data$start)+1L
  for(i in seq_len(m)){
    lagL[,i] <- ifelse(record>i,c(rep(0L,i),
      data$L[-(nrow(data)-0:(i-1))]),0L)
    lagA[,i] <- ifelse(record>i,c(rep(0L,i),
      data$A[-(nrow(data)-0:(i-1))]),0L)
  }
  data <- cbind(data,lagL,lagA)
  return(data)
}
LOCF <- function(data){
  data$L <- locf(data$L)
  return(data)
}
qAdhering <- function(data){
  switched <-
    ifelse(qFirst(data$unit),TRUE,c(TRUE,diff(data$A)!=0L))*1L
  cs <- cumsum(switched)
  mt <- with(data,match(unit,unique(unit)))
  return(cs==cs[qFirst(data$unit)][mt])
}

# ========================================================================
# Estimators
# ========================================================================

getPS <- function(data){
  fit <- glm(A~I(!start)+lag1A+L,data=data[data$start>=0L,,
    drop=FALSE],family=binomial)
  return(unname(predict(fit,newdata=data,type="response")))
}
estimateIPW <- function(data,ps){
  data$ps <- ps
```

```
  data <- data[data$start>=0L,]
  data <- data[qAdhering(data),]
  lp <- log(ifelse(data$A>0L,data$ps,1-data$ps))
  cs <- cumsum(lp)
  ql <- qFirst(data$unit,TRUE)
  data$W <- 1/exp(cs-c(0,cs[ql][-sum(ql)])[match(data$unit,
    unique(data$unit))])
  EW0 <- with(data[data$A==0L,],tapply(W,start,mean))
  EW1 <- with(data[data$A==1L,],tapply(W,start,mean))
  mt0 <- with(data,match(start,unique(start)))
  mt1 <- with(data,match(start,unique(start)))
  data$sW <- data$W/ifelse(data$A>0L,EW1[mt1],EW0[mt0])
  fit <- with(data,survival::survfit(survival::Surv(start,stop,
    event)~A,weights=sW,timefix=FALSE))
  smmry <- survival:::summary.survfit(fit,times=0:K)
  est <- split(smmry$surv,smmry$strata)
  names(est) <- c("surv0","surv1")
  est$surv0 <- c(est$surv0,rep(rev(est$surv0)[1L],
    K+1L-length(est$surv0)))
  est$surv1 <- c(est$surv1,rep(rev(est$surv1)[1L],
    K+1L-length(est$surv1)))
  return(est)
}
crudePP <- function(data){
  data <- data[data$start>=0L,]
  data <- data[qAdhering(data),,drop=FALSE]
  ql <- qFirst(data$unit,TRUE)
  time <- data$stop[ql]
  status <- data$event[ql]
  group <- data$A[ql]
  fit <- survival::survfit(survival::Surv(time,status)~group)
  smmry <- survival:::summary.survfit(fit,times=0:K)
  est <- split(smmry$surv,smmry$strata)
  names(est) <- c("surv0","surv1")
  est$surv0 <- c(est$surv0,rep(rev(est$surv0)[1L],
    K+1L-length(est$surv0)))
  est$surv1 <- c(est$surv1,rep(rev(est$surv1)[1L],
    K+1L-length(est$surv1)))
  return(est)
}

# ======================================================================
# Data generation & estimation
# ======================================================================

trial <- longFormat(drawSample(n,trial=TRUE))
wide <- drawSample(n)
long <- lagVariables(longFormat(wide))
longI <- lagVariables(LOCF(coarsen(long,"I")))
longII <- lagVariables(LOCF(coarsen(long,"II")))
longIII <- lagVariables(LOCF(coarsen(long,"III")))

estA <- crudePP(trial)
estB <- crudePP(long)
estC <- estimateIPW(long,getPS(long))
```

```
estD <- estimateIPW(longII,getPS(longII))
estE <- estimateIPW(longI,getPS(longI))
estF <- estimateIPW(longIII,getPS(longIII))

sq <- c(1L,2L,seq(3L,60L,3))
estVarD <- estVarF <- list()
for(i in seq_along(sq)){
  cat("\r",i,"/",length(sq),sep=""); flush.console()
  longIIi <- lagVariables(LOCF(coarsen(long,"II",after=sq[i])))
  estVarD[[i]] <- estimateIPW(longIIi,getPS(longIIi))
  longIIIi <- lagVariables(LOCF(coarsen(long,"III",after=sq[i])))
  estVarF[[i]] <- estimateIPW(longIIIi,getPS(longIIIi))
}
names(estVarD) <- names(estVarF) <- sq
```