American Journal of Epidemiology
© The Author(s) 2022. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Vol. 192, No. 1
https://doi.org/10.1093/aje/kwac162
Advance Access publication:
September 16, 2022

Practice of Epidemiology

A Simulation Study Comparing the Performance of Time-Varying Inverse Probability Weighting and G-Computation in Survival Analysis

Jacqueline E. Rudolph*, Enrique F. Schisterman, and Ashley I. Naimi

* Correspondence to Dr. Jacqueline E. Rudolph, Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205 (e-mail: jacqueline.rudolph@jhu.edu).

Initially submitted August 16, 2021; accepted for publication September 13, 2022.

Inverse probability weighting (IPW) and g-computation are commonly used in time-varying analyses. To inform decisions on which to use, we compared these methods using a plasmode simulation based on data from the Effects of Aspirin in Gestation and Reproduction (EAGeR) Trial (June 15, 2007–July 15, 2011). In our main analysis, we simulated a cohort study of 1,226 individuals followed for up to 10 weeks. The exposure was weekly exercise, and the outcome was time to pregnancy. We controlled for 6 confounding factors: 4 baseline confounders (race, ever smoking, age, and body mass index) and 2 time-varying confounders (compliance with assigned treatment and nausea). We sought to estimate the average causal risk difference by 10 weeks, using IPW and g-computation implemented using a Monte Carlo estimator and iterated conditional expectations (ICE). Across 500 simulations, we compared the bias, empirical standard error (ESE), average standard error, standard error ratio, and 95% confidence interval coverage of each approach. IPW (bias = 0.02; ESE = 0.04; coverage = 92.6%) and Monte Carlo g-computation (bias = −0.01; ESE = 0.03; coverage = 94.2%) performed similarly. ICE g-computation was the least biased but least precise estimator (bias = 0.01; ESE = 0.06; coverage = 93.4%). When choosing an estimator, one should consider factors like the research question, the prevalences of the exposure and outcome, and the number of time points being analyzed.

bias; g-computation; inverse probability weighting; simulation; survival analysis; variance

Abbreviations: BMI, body mass index; EAGeR, Effects of Aspirin in Gestation and Reproduction; ESE, empirical standard error; ICE, iterated conditional expectations; IPW, inverse probability weighting; MC, Monte Carlo; RD, risk difference.

Epidemiologists routinely study cohorts followed over long time periods and are increasingly asking questions that leverage complex longitudinal data. We are often interested in assessing the causal effect of an exposure that changes over time and is subject to time-varying confounding (1, 2). Modern quantitative methods, including inverse probability weighting (IPW) and the parametric g-formula (g-computation), can account for time-varying confounders that are affected by past exposure (exposure-confounder feedback), in contrast to traditional regression (2, 3). These methods allow us to estimate causal effects in such cases, provided the relevant causal identifiability conditions have been met (1–5).

IPW in its simplest form involves modeling the probability of the exposure conditional on confounders. Each observation is weighted by the inverse of the estimated probability of receiving the observed exposure, and we conduct our analyses within the weighted population. In contrast, g-computation predicts the expected outcome under a specified exposure, standardized across the distribution of confounders. One then contrasts the average predicted outcomes from 2 exposure scenarios to estimate the effect of interest. IPW and g-computation can be used to estimate the same parameters, so how should we decide which approach to use in applied analyses?

Three important factors to consider include causal bias, statistical bias, and variance. By causal bias we mean the systematic errors that arise due to violations of the identifiability conditions, such as when there is uncontrolled confounding (violation of exchangeability), the intervention is poorly defined (violation of consistency), or the probability of being exposed conditional on adjustment variables is 1 or 0

Downloaded from https://academic.oup.com/aje/article/192/1/102/6702451 by Pitts Theology Library, Emory University user on 09 December 2025

(violation of positivity). These biases will need to be considered regardless of whether we use IPW, g-computation, or any other estimator. In most analyses, causal bias will be of greatest concern.

Statistical bias, or the average difference between the point estimate and the target statistical parameter, can differ between modeling approaches. The question here is how well the estimator captures the relationships in the observed data, when quantifying the effect of interest. Because IPW and g-computation estimate the same effect using different models (exposure vs. outcome models), one might see differences between these approaches in the amount of statistical bias in finite samples. Variance, capturing the precision of our estimates, is directly affected by the size of the sample, the distributions of key variables (e.g., whether there are random violations of the positivity assumption) (6), and the complexity and properties (e.g., efficiency) of the chosen statistical methods.

Ideally, we would choose the approach that minimizes all sources of bias and variance. Thus, we sought to compare the bias and variance of IPW and g-computation, specifically in the context of survival analysis in the presence of exposure-confounder feedback. Our primary interest was in how the estimators perform in finite sample sizes, which might differ from how they perform in theory under asymptotic conditions. To inform the use of these methods in applied analyses, we designed a plasmode simulation, using data from the Effects of Aspirin in Gestation and Reproduction (EAGeR) Trial (7, 8).

METHODS

Simulation

We generated a time-varying plasmode simulation from the EAGeR Trial data. The EAGeR Trial was a double-blind, placebo-controlled trial designed to evaluate the relationship between low-dose aspirin use and several pregnancy outcomes (7, 8). Participants were followed from randomization 1) until they dropped out of the study, 2) until they were administratively censored at 6 months (if they did not become pregnant), or 3) throughout pregnancy.

Here, we used the trial's longitudinal data (June 15, 2007–July 15, 2011) to ask: How did weekly exercise affect incidence of pregnancy by 10 weeks of follow-up? Our exposure ($X(t)$) was an indicator of whether a woman reported engaging in a moderate-to-high amount of exercise during week t of follow-up. In the EAGeR Trial, exercise was only measured at baseline; we simulated a time-varying version, as described below. We included 4 baseline confounders (**B**): a binary indicator for White race, a binary indicator for ever smoking cigarettes, continuous age (years), and continuous body mass index (BMI; weight (kg)/height $(m)^2$). We included 2 time-varying confounders: whether a woman took her assigned pill at least 5 out of 7 days in the week ($Z_1(t)$) and whether a woman experienced nausea during the week ($Z_2(t)$). Our outcome ($Y(t)$) was incidence of pregnancy, and women were allowed to drop out of the study ($D(t)$). To inform our simulation parameters, we modeled the relationships between these variables in the baseline EAGeR data (see Web Appendix 1, available at https://doi.org/10.1093/aje/kwac162).

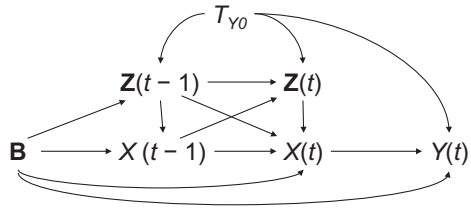

Figure 1. Causal diagram for a simulation comparing the performance of time-varying inverse probability weighting with that of g-computation in survival analysis, where $Y(t)$ is the outcome at time t, $X(t)$ is the exposure, $\mathbf{Z}(t)$ is the set of time-varying confounders, **B** is the set of baseline confounders, and T_{Y0} is baseline time to event (i.e., time to event under no exposure).

Baseline data for the simulation were generated by sampling with replacement from the observed EAGeR baseline data. We generated samples of various sizes (n = 500, n = 1,226, or n = 2,500) to examine the performance of the estimators in a sample of size equivalent to that of the EAGeR Trial and in samples of approximately half and double the original size.

We used the baseline values of all confounders to generate the baseline exposure ($X(1)$) from a Bernoulli distribution, with the following conditional probability:

$$P(X(1) = 1|\mathbf{B}, Z_1(1), Z_2(1)) = \text{expit}[1.45 + 1.29 \times \text{race} - 0.34 \times \text{smoking} - 0.05 \times \text{age} - 0.04 \times \text{BMI} - 0.58 \times Z_1(1) + 0.06 \times Z_2(1)].$$

We generated follow-up for the simulation by adapting the data-generating mechanism from Moodie et al. (9–11). This approach simulates longitudinal data with time-varying exposure-confounder feedback from a structural nested model that "satisfies" a Cox marginal structural model (10). Since g-computation can also be used to quantify the parameters of a marginal structural model, this data-generating mechanism can be used to compare the performance of IPW and g-computation (3). Figure 1 summarizes the causal structure for the simulation.

The algorithm first generated baseline time to event (T_{Y0}) and baseline time to censoring (T_{D0}) from exponential distributions with constant hazards $\lambda_Y = 0.05$ and $\lambda_D = 0.10$, respectively. By "baseline," we mean the time to event (or censoring) under no exposure. The outcome hazard was chosen because it was large enough to generate a reasonable number of outcomes, and it matched the intercept from the accelerated failure-time model that regressed the outcome against baseline exposure and baseline confounders (excluding $Z_1(t)$ and $Z_2(t)$). In the Discussion section, we discuss the limitations of using the exponential distribution to simulate time-to-event outcomes.

For all n individuals, the algorithm looped through each time point $t = 1, \ldots, 10$, until the event occurred or follow-up ended due to dropout. When $t > 1$, we generated the

Downloaded from https://academic.oup.com/aje/article/192/1/102/6702451 by Pitts Theology Library, Emory University user on 09 December 2025

time-varying confounders from Bernoulli distributions using the following conditional probabilities:

$$P(Z_1(t) = 1|\mathbf{B}, X(t-1), Z_1(t-1)) = \text{expit}[2.70 + \log(3) \times Z_1(t-1) - 0.29 \times X(t-1) + 0.94 \times \text{race} - 0.77 \times \text{smoking} + 0.04 \times \text{age} - 0.04 \times \text{BMI} + 2I(T_0 < c)];$$

$$P(Z_2(t) = 1|\mathbf{B}, X(t-1), Z_2(t-1)) = \text{expit}[-1.57 + \log(3) \times Z_2(t-1) + 0.05 \times X(t-1) + 1.02 \times \text{race} + 0.05 \times \text{smoking} - 0.05 \times \text{age} + 0.02 \times \text{BMI} + 2I(T_0 < c)],$$

where the term $I(T_{Y0} < c)$ with $c = 30$ was used to introduce confounding by creating a backdoor path between the time-varying confounders and $Y(t)$ through T_{Y0} (9, 11). We then generated $X(t)$ with the conditional probability

$$P(X(t) = 1|\mathbf{B}, X(t-1), Z_1(t), Z_1(t-1), Z_2(t), Z_2(t-1)) = \text{expit}[1.45 + \log(3) \times X(t-1) + 1.29 \times \text{race} - 0.34 \times \text{smoking} - 0.05 \times \text{age} - 0.04 \times \text{BMI} - 0.58 \times Z_1(t) - 0.58 \times Z_1(t-1) + 0.06 \times Z_2(t) + 0.06 \times Z_2(t-1)].$$

Whether an event occurred at time t was a function of $X(t)$, $\mathbf{B}$, T_{Y0}, and λ_Y:

$$Y(t) = I\Big\{\lambda_Y \times T_{Y0} \leq \sum_{k=1}^{t-1} \exp[\log(\lambda_Y) + \log(2)X(k) + 1.50 \times \text{race} - 0.71 \times \text{smoking} - 0.01 \times \text{age} - 0.04 \times \text{BMI}]\Big\}.$$

Time to pregnancy (T_Y,) was the time t at which the event occurred. Whether an individual dropped out at any given t was a function of $\mathbf{B}$, T_{D0}, and λ_D:

$$D(t) = I\Big\{\lambda_D \times T_{D0} \leq \sum_{k=1}^{t-1} \exp[\log(\lambda_D) + 0.63 \times \text{race} + 0.36 \times \text{smoking} - 0.09 \times \text{age} + 0.01 \times \text{BMI}]\Big\}.$$

Time to censoring (T_D) was the time t at which an individual dropped out. If an individual had the event and dropped out at t, we counted them as having had an event. We discretized time to match the structure of the EAGeR Trial, but we considered continuous T_Y in a supplementary analysis.

Estimators

Our goal was to estimate the causal risk difference (RD):

$$E[Y(10)^{\bar{x}=1}] - E[Y(10)^{\bar{x}=0}],$$

where $\bar{x}$ is assigned treatment history and $E[Y(10)^{\bar{x}}]$ is the counterfactual risk of the outcome by $t = 10$ (end of follow-up) under exposure $\bar{x}$. In all simulation settings where we controlled for all confounders, the identifiability assumptions held. (To assess whether there were random violations of the positivity assumption, we checked in select simulation iterations that the cumulative propensity scores were bounded away from 0) (6).) Therefore, we can link this counterfactual contrast to a contrast of the observed risks of the outcome, which we estimated using the approaches below.

When using the IPW estimator, we focused on modeling the exposure mechanism. Here, we modeled $X(t)$, conditional on historical variables, using pooled logistic regression (12):

$$\pi_d = P(X(t) = 1|\mathbf{B}, X(t-1), Z_1(t), Z_1(t-1), \overline{D}(t-1) = \overline{0}) = \text{logit}[\alpha_0 + \alpha_1 X(t-1) + \alpha_2 \text{race} + \alpha_3 \text{smoking} + \alpha_4 \text{age} + \alpha_5 \text{BMI} + \alpha_6 Z_1(t) + \alpha_7 Z_1(t-1) + \alpha_8 Z_2(t) + \alpha_9 Z_2(t-1) + f(T)],$$

where $f(T)$ is a function of time (specifically, indicator terms). While it was not necessary given our data-generating mechanism, we included a function of time in all pooled logistic regression models, as is commonly done in practice.

We then used this model to predict each individual's probability of being exposed, $\hat{\pi}_d$. To reduce the variability in our estimator, we stabilized the weights (as is recommended in practice) (13) by modeling the probability of being exposed not conditional on $\mathbf{B}$, $Z_1(t)$, or $Z_2(t)$:

$$\pi_n = P(X(t) = 1|X(t-1), \overline{D}(t-1) = \overline{0}) = \text{logit}[\alpha_0 + \alpha_1 X(t-1) + f(T)].$$

We again predicted each individual's probability of being exposed, $\hat{\pi}_n$. To control for dropout, we modeled the probability of not dropping out (η_d), conditional on $X(t)$, $Z_1(t)$, $Z_2(t)$, and $\mathbf{B}$, and the probability of not dropping out (η_n), conditional only on $X(t)$, using pooled logistic regression. Our stabilized weights (sw) for individual i at time point t then took the form

$$sw_{it} = \prod_{k=1}^{t} \frac{X_i(k) \times \hat{\pi}_n + (1 - X_i(k)) \times (1 - \hat{\pi}_n)}{X_i(k) \times \hat{\pi}_d + (1 - X_i(k)) \times (1 - \hat{\pi}_d)} \times \frac{(1 - D_i(k)) \times \hat{\eta}_n}{(1 - D_i(k)) \times \hat{\eta}_d}.$$

We estimated risk by fitting Kaplan-Meier survival curves, weighted by sw_{it} and stratified by exposure, and estimated the RD at $t = 10$ by contrasting the risks across exposure groups. We note that other IPW estimators exist, some of which rely on fewer parametric assumptions than our approach; however, the estimator described above is common in epidemiologic analyses (13, 14).

While IPW focuses on the exposure mechanism, g-computation estimates the target RD via model-based

Downloaded from https://academic.oup.com/aje/article/192/1/102/6702451 by Pitts Theology Library, Emory University user on 09 December 2025

standardization. The central idea is that we can express one of the counterfactual risks above, say $E[Y(t)^{\bar{x}=1}]$, as the product of a series of conditional probabilities, using the law of total probability (5, 15). For example, the g-formula for our example can be expressed as

$$E\left[Y(t)^{\bar{x}=1}\right]=\sum_{k=1}^{t}\sum_{\mathbf{z}}\left\{\begin{array}{l}P\left[Y(k)=1|\overline{X}(k)=1,\overline{\mathbf{Z}}(k)=\mathbf{z},\mathbf{B},\overline{Y}(k-1)=\overline{D}(k-1)=0\right]\times\\ \prod_{m=0}^{k}\left[\begin{array}{c}P\left[D(m)=0|\overline{\mathbf{Z}}(m)=\mathbf{z},\overline{X}(m)=1,\mathbf{B},Y(m-1)=\overline{D}(m-1)=0\right]\times\\ P\left[Z_1(m)|\overline{X}(m-1)=1,\overline{Z}_1(m-1)=z_1,\mathbf{B},Y(m-1)=\overline{D}(m-1)=0\right]\times\\ P\left[Z_2(m)|\overline{X}(m-1)=1,\overline{Z}_2(m-1)=z_2,\mathbf{B},Y(m-1)=\overline{D}(m-1)=0\right]\times\\ P\left[Y(m-1)=0|\overline{\mathbf{Z}}(m-1),\overline{X}(m-1)=1,\mathbf{B},Y(m-2)=\overline{D}(m-2)=0\right]\end{array}\right]\end{array}\right\},$$

where $\mathbf{Z}(k)$ indicates both $Z_1(t)$ and $Z_2(t)$. Since the way one carries out the g-computation algorithm can affect its performance, we examined 3 different implementations: 2 applications of the Monte Carlo (MC) estimator and the iterated conditional expectations (ICE) estimator (16, 17).

MC g-computation (or non-ICE g-computation) (16) estimates the counterfactual risk by modeling in the original data set the conditional probabilities in the above equation—for example, models for $Y(t)$ and $Z_1(t)$. Here, T_Y was modeled using an exponential accelerated failure-time model that controlled for $\mathbf{B}$, $X(t)$, $X(t-1)$, $Z_1(t)$, $Z_1(t-1)$, $Z_2(t)$, and $Z_2(t-1)$. We used pooled logistic regression to model $Z_1(t)$, conditional on $\mathbf{B}$, $Z_1(t-1)$, $X(t-1)$, and indicator variables for time (similarly for $Z_2(t)$).

We then took an MC resample of the original data, sampling with replacement from the baseline observations. We took this resample to ensure that even rare strata of covariates appeared with sufficient numbers to more reliably compute the g-computation integral via MC integration. Using the MC method to integrate the g-formula leads to a degree of simulation error resulting from the MC resample. Taking an MC resample that is as large as is computationally feasible minimizes this simulation error (18). We here examined the impact on estimator performance of using resample sizes $s=\{n,2n,4n\}$. When n equaled 2,500, we only examined $s=\{n,\ 2n\}$ for computational reasons. While it is not strictly necessary to resample the data when $s=n$, we did so here to be consistent across values of s.

Within the MC resample, we predicted follow-up under 2 scenarios: 1) the scenario where exposure was set to 1 at all t and dropout was eliminated and 2) the scenario where exposure was set to 0 at all t and dropout was eliminated. We approached this prediction in 2 ways. First, we marched through time, predicting T_Y based on the coefficients of the outcome model, set exposure, predicted time-varying confounders, and baseline confounders. If $T_Y<1$ at a given time point, the outcome occurred; otherwise, we moved to the next time point. We repeated this process until the unit either had the event or was administratively censored at $t=10$. Second, we followed all individuals across all time points, and instead of predicting T_Y, we stored their predicted hazard of the event ($h_{Yi}(t)$) at each time point.

We estimated the target RD by comparing the risks obtained under each exposure intervention. When we predicted T_Y, we estimated risk by taking the complement of the Kaplan-Meier survival curve. When we predicted $h_{Yi}(t)$, we first estimated the time-specific average of the individual hazards, $E[h_{Yi}(t)]$, and then estimated risk by taking

$$E[Y(t)^{\bar{x}}]=1-\prod_{k=1}^{t}\left\{1-E\left[h_{Yi}(t)\right]\right\}.$$

ICE g-computation targets the counterfactual risk by recognizing that the g-formula above can be rewritten as a series of nested expectations (15, 16):

$$\begin{aligned}E[Y(t)^{\bar{x}=1}]&=E\{E[\ldots E[E(Y(t)|\overline{X}(t)=1,\overline{\mathbf{Z}}(t),\mathbf{B},\overline{D}(t)\\&=0)\mid\overline{X}(t-1)=1,\overline{\mathbf{Z}}(t-1),\mathbf{B},\overline{D}(t-1)=0]\ldots|X(1)\\&=1,\mathbf{Z}(1),\mathbf{B}]\}.\end{aligned}$$

We estimated each of these expectations in turn, moving from the innermost (corresponding to the last time point) to the outermost (first time point). We began by modeling the outcome given observed exposure and time-varying and baseline covariates and then predicted the outcome under the exposure of interest. We regressed those predictions against the observed exposure and covariates from the previous time point and used that model to again predict outcomes under the exposure of interest. When we finished this process, we took the average of the predicted outcomes to estimate the risk and contrasted the risks obtained when we set everyone to be always exposed versus never exposed. Unlike MC g-computation, this approach did not require parametric models for the time-varying confounders. ICE g-computation was implemented using the ltmle package in R (19).

We compared the results from these 4 estimators with the crude RD, estimated by taking the difference between the exposure-specific Kaplan-Meier survival curves at $t=10$. To do this, we used a counting process approach and treated exposure as a time-varying variable. We also conducted 2 supplementary analyses, using the simulation where $n=s=$ 1,226. First, we implemented each approach controlling for baseline but not time-varying confounders. Second, to assess whether using discrete versus continuous time affected the results, we adapted the MC g-computation algorithm by modeling continuous time to event (i.e., the exact time at which the event occurred in the interval $(t-1,t]$ rather than just t) and using the algorithm-predicted T_Y when estimating RDs.

Downloaded from https://academic.oup.com/aje/article/192/1/102/6702451 by Pitts Theology Library, Emory University user on 09 December 2025

For all analyses, 95% confidence intervals were obtained using the standard error of the point estimates from 500 bootstrap resamples.

Simulation analyses

We iterated each simulation $M = 500$ times. We denote the true RD as ψ and estimates obtained in the simulation iterations as $\hat{\psi}$. To determine ψ, we generated the counterfactual follow-up for 1 million observations (separate from all simulations above) had they been exposed versus unexposed and then took the difference between the mean values of the counterfactual outcome indicators $Y(10)^{\bar{x}=1}$ and $Y(10)^{\bar{x}=0}$. We used this indirect approach for obtaining ψ because we could not determine ψ in a straightforward manner from the simulation parameters, which were specified as hazard ratios. We found that $\psi = 0.229$.

Our metrics for comparing the performance of the estimators were bias,

$$\text{Bias}_{\psi} = \left(\frac{1}{M}\sum\nolimits_{i=1}^{M}\hat{\psi}_i\right) - \psi;$$

the empirical standard error (ESE) of the $\hat{\psi}$ (the standard deviation of the $\hat{\psi}$),

$$\text{ESE} = \sqrt{\left[\frac{1}{M-1}\right]\sum_{i=1}^{M}\left(\hat{\psi}_i - \overline{\hat{\psi}}\right)};$$

the average standard error (ASE) for the $\hat{\psi}$,

$$\text{ASE} = \frac{1}{M}\sum_{i=1}^{M}\text{SE}\left(\hat{\psi}_i\right);$$

the standard error ratio (SER),

$$\text{SER} = \text{ASE}/\text{ESE};$$

and the 95% confidence interval coverage,

$$\text{Coverage} = \frac{1}{M}\sum_{i=1}^{M} I\left(\psi \in \hat{\psi}_i \pm 1.96 \times \text{SE}\left(\hat{\psi}_i\right)\right).$$

To determine what would be considered "acceptable" 95% confidence interval coverage, we used the formula published in Burton et al. (20). For 500 simulations, acceptable coverages fell within the range of 93.1%–96.9% (determined by $0.95 \pm 2 \times \sqrt{0.95(1-0.95)/500}$).

All analyses were carried out using R, version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria). Software code is provided in Web Appendix 2 and on GitHub.

RESULTS

When the sample size (n) was 1,226 (Table 1), the crude RD was 0.31 (bias = 0.08; ESE = 0.03; coverage = 14.6%). The RD estimated using ICE g-computation was 0.24 (bias = 0.01; ESE = 0.06; coverage = 93.4%). IPW estimated an RD of 0.25 (bias = 0.02; ESE = 0.04; coverage = 92.6%). Increasing s, the size of the MC resample, did not improve the performance of either MC g-computation approach; the small increase in bias as we increased s (from −0.01 to −0.02) was potentially the result of MC resampling error. We thus present the results for the case where $s = n$. When we predicted T_Y in the MC g-computation algorithm, the RD was 0.22 (bias = −0.01; ESE = 0.03; coverage = 94.2%). When we predicted $h_Y(t)$ in the MC g-computation algorithm, the RD was 0.25 (bias = 0.02; ESE = 0.03; coverage = 93.0%).

The results obtained when sample sizes were 500 and 2,500 are presented in Tables 2 and 3, respectively. We observed a consistent pattern of results across the 3 sample sizes. The results after applying IPW and g-computation were closer to the true RD than to the crude RD, indicating that these methods were at least partly controlling for the bias due to confounding and dropout. ICE g-computation was always the least biased but least precise estimator. MC g-computation (using either prediction approach) was marginally more precise than IPW. Whether the IPW-estimated RD or the MC g-computation-estimated RD was more biased depended on the sample size and the size of the MC resample. However, no approach was particularly biased, and differences in bias across methods was minimal. The largest absolute bias was 0.02 (for MC g-computation predicting T_Y when $n = 500$ and $s = 2{,}000$), which corresponds to a relative bias less than 10%.

All estimators had standard error ratios close to 1, which implies that the variance estimator was working as expected. Ninety-five percent confidence interval coverage was acceptable or close to acceptable, with one notable exception. When n equaled 2,500, both MC g-computation approaches had sufficient bias, with a sufficiently small ESE, to result in 95% confidence interval coverages that were considered less than acceptable (85.6%–89.6%).

In the supplementary analysis controlling for baseline but not time-varying confounders (with $n = s = 1{,}226$), results were generally more biased than when we controlled for time-varying confounders but less biased than the crude result. For example, the RDs for IPW and ICE g-computation were 0.25 (bias = 0.02; ESE = 0.04; coverage = 90.2%) and 0.26 (bias = 0.03; ESE = 0.06; coverage = 90.4%). The exception was MC g-computation predicting T_Y, which estimated an RD of 0.23 (bias < 0.01; ESE = 0.03; coverage = 95.8%). When we used continuous time to event in the g-computation algorithm, the estimated RD was 0.23 (bias < 0.01; ESE = 0.03; coverage = 95.8%), which was closer to the truth than when we used discrete time.

DISCUSSION

Here, we compared the performance of IPW and g-computation when estimating RDs in simulations with realistic sample sizes, time-varying exposure-confounder feedback, and a survival outcome. Under our data-generating mechanism, MC g-computation and IPW performed similarly in terms of bias and variance, while ICE g-computation was marginally the least biased but least precise estimator. In general, though, methods had similar performance, with

Downloaded from https://academic.oup.com/aje/article/192/1/102/6702451 by Pitts Theology Library, Emory University user on 09 December 2025

Table 1. Results From 500 Simulations Comparing the Performance of G-Computation With That of Inverse Probability Weighting (Sample Size: n = 1,226)[a]

Estimator	Average RD	Bias[b]	ESE	ASE	SER	Coverage, %
Crude	0.31	0.08	0.03	0.03	1.06	14.6
IPW	0.25	0.02	0.04	0.04	1.00	92.6
MC g-computation (T_Y)[c]						
MC resample: s = 1,226	0.22	−0.01	0.03	0.03	1.04	94.2
MC resample: s = 2,500	0.21	−0.02	0.03	0.03	1.04	91.4
MC resample: s = 5,000	0.21	−0.02	0.03	0.03	1.04	92.0
MC g-computation ($h_Y(t)$)[d]						
MC resample: s = 1,226	0.25	0.02	0.03	0.04	1.04	93.0
MC resample: s = 2,500	0.25	0.02	0.03	0.04	1.04	93.0
MC resample: s = 5,000	0.25	0.02	0.03	0.04	1.04	92.8
ICE g-computation	0.24	0.01	0.06	0.06	1.00	93.4

Abbreviations: ASE, average standard error; ESE, empirical standard error; ICE, iterated conditional expectations; IPW, inverse probability weighting; MC, Monte Carlo; RD, risk difference; SER, standard error ratio.

[a] Data were obtained from the Effects of Aspirin in Gestation and Reproduction (EAGeR) Trial (June 15, 2007–July 15, 2011).

[b] The true RD was 0.23.

[c] Time to event (T_Y) predicted.

[d] Hazard of event at time t ($h_Y(t)$) predicted.

relatively low bias (particularly compared with the crude RD), and standard error ratios were close to 1.

Our results for IPW agree with past work. Westreich et al. (21) showed that, in simulations with 1,500 individuals and 1 binary, time-varying confounder, Cox models with stabilized weights were unbiased, had efficiency greater than but comparable to that of a covariate-adjusted Cox model, and had nominal 95% confidence interval coverage. In our simulations, we observed little difference between MC g-computation and IPW in bias or variance. Theory states that MC g-computation will be asymptotically more efficient than IPW (5, 22). In our simulations, while MC g-computation was always marginally more precise than IPW, the difference between the 2 approaches was small.

Table 2. Results From 500[a] Simulations Comparing the Performance of G-Computation With That of Inverse Probability Weighting (Sample Size: n = 500)[b]

Estimator	Average RD	Bias[c]	ESE	ASE	SER	Coverage, %
Crude	0.31	0.08	0.04	0.04	1.06	56.5
IPW	0.25	0.02	0.06	0.06	0.95	91.9
MC g-computation (T_Y)[d]						
MC resample: s = 500	0.22	−0.01	0.05	0.05	1.03	93.8
MC resample: s = 1,000	0.21	−0.02	0.05	0.05	1.03	93.3
MC resample: s = 2,000	0.21	−0.02	0.05	0.05	1.03	92.7
MC g-computation ($h_Y(t)$)[e]						
MC resample: s = 500	0.24	0.01	0.06	0.06	1.03	94.0
MC resample: s = 1,000	0.24	0.01	0.06	0.06	1.03	93.8
MC resample: s = 2,000	0.23	<0.01	0.10	0.12	1.16	97.2
ICE g-computation	0.24	0.01	0.06	0.06	1.03	94.0

Abbreviations: ASE, average standard error; ESE, empirical standard error; ICE, iterated conditional expectations; IPW, inverse probability weighting; MC, Monte Carlo; RD, risk difference; SER, standard error ratio.

[a] Data were obtained from the Effects of Aspirin in Gestation and Reproduction (EAGeR) Trial (June 15, 2007–July 15, 2011).

[b] Results exclude 4 simulations where the models did not converge.

[c] The true RD was 0.23.

[d] Time to event (T_Y) predicted.

[e] Hazard of event at time t ($h_Y(t)$) predicted.

Downloaded from https://academic.oup.com/aje/article/192/1/102/6702451 by Pitts Theology Library, Emory University user on 09 December 2025

Table 3. Results From 500 Simulations Comparing the Performance of G-Computation With That of Inverse Probability Weighting (Sample Size: n = 2,500)[a]

Estimator	Average RD	Bias[b]	ESE	ASE	SER	Coverage, %
Crude	0.31	0.08	0.02	0.02	0.98	1.4
IPW	0.24	0.02	0.03	0.03	1.02	93.2
MC g-computation (T_Y)[c]						
MC resample: s = 2,500	0.21	−0.02	0.02	0.02	1.00	85.6
MC resample: s = 5,000	0.21	−0.02	0.02	0.02	1.00	85.8
MC g-computation ($h_Y(t)$)[d]						
MC resample: s = 2,500	0.25	0.02	0.03	0.03	1.00	89.6
MC resample: s = 5,000	0.25	0.02	0.03	0.03	1.00	89.4
ICE g-computation	0.23	<0.01	0.04	0.04	1.01	94.6

Abbreviations: ASE, average standard error; ESE, empirical standard error; ICE, iterated conditional expectations; IPW, inverse probability weighting; MC, Monte Carlo; RD, risk difference; SER, standard error ratio.

[a] Data were obtained from the Effects of Aspirin in Gestation and Reproduction (EAGeR) Trial (June 15, 2007–July 15, 2011).

[b] The true RD was 0.23.

[c] Time to event (T_Y) predicted.

[d] Hazard of event at time t ($h_Y(t)$) predicted.

The greater difference in bias and precision was observed between these 2 approaches and ICE g-computation.

One reason we may have seen little difference between MC g-computation and IPW could be that our results were specific to the simulation parameters used in the data-generating mechanism. For example, we specified an outcome generated from an exponential distribution, in small sample sizes, with no interaction terms in any of our structural models. The exponential distribution is a relatively simple distribution, with only 1 time-independent parameter; as such, it may not reflect the complexity inherent in many empirical data sets. We used this distribution as a simplifying component of our complex time-varying simulation and for the easy interpretation of its single parameter; future work could examine the impact of more complex time-to-event distributions.

Furthermore, the statistical bias and precision of these approaches in finite samples depends on the way they are implemented. When we modeled the exposure for IPW and the confounders and outcome for MC g-computation, our approach was highly parametric. We used logistic regression models that pooled data across time (i.e., we included no interaction terms or stratification by time) and across exposure trajectories (i.e., we did not subset the data to those who were continuously exposed or unexposed). In contrast, while we used logistic regression models to carry out ICE g-computation, this approach did stratify by time and did subset the sample to those who were continuously exposed (17, 19). This meant that ICE g-computation made fewer assumptions about the form of the data and was consequently less susceptible to bias due to statistical model misspecification than the implementations of MC g-computation and IPW used here. The difference in modeling approaches also explains why ICE g-computation was less precise than MC g-computation and IPW. Stratifying by time and dropping from the analysis those who deviated from the specified exposure trajectory reduced the amount of data available for model-fitting, especially at later time points.

Model misspecification bias could in fact be one explanation for the residual bias seen for MC g-computation and IPW. We knew the correct data-generating mechanism and could control for all relevant confounders. However, the time-varying data structure was relatively complex, and while our implementations of MC g-computation and IPW followed standard practice, the parametric assumptions may not have fully captured the true nature of the data. Different models might have been a better fit to the data.

To this point, we have largely focused on the statistical bias and variance of our estimators. In applied analyses, though, causal bias is as much as if not more of a threat to the validity of one's results. We observed that the RDs estimated using IPW and the g-computation approaches were much less biased than the crude RDs. This is because these methods appropriately controlled for confounding and right-censoring, even in the presence of time-varying exposure-confounder feedback. However, in nonsimulated data, we rarely know the true data-generating mechanism and must make educated guesses based on substantive knowledge. In time-varying data, identifying the adjustment set can be difficult because of the need to control for exposure and confounder histories. Violations of the other identifiability conditions (positivity and consistency) or of the assumption of no measurement error can also lead to causal bias. Data with long follow-up periods and time-varying exposures are fundamentally susceptible to random violations of positivity, due to the difficulty in observing individuals who remain continuously exposed or unexposed across time (within strata of confounders). Random violations can result in both bias and loss of precision; this is why it is critical to investigate the potential for nonpositivity in one's analysis—for example, by examining the distribution of cumulative weights or cumulative propensity scores (6).

Downloaded from https://academic.oup.com/aje/article/192/1/102/6702451 by Pitts Theology Library, Emory University user on 09 December 2025

Ultimately, how do the findings from our simulation experiment inform choice of analytical approach? The researcher needs to consider their research question, data structure, and specified causal model. Our simulation speaks to the case where one has longitudinal data with time-varying exposure-confounder feedback and a survival outcome. We observed similar performance for IPW and MC g-computation. If deciding between these 2 approaches, a researcher might consider the distribution of the exposure and outcome in their data. If the exposure is common, while the outcome is not, IPW might be preferred because it will be easier to estimate the probability of exposure. Conversely, if the researcher believes they can better model the outcome, they might choose g-computation.

We observed that ICE g-computation was less biased but less precise compared with IPW and MC g-computation. This was partly due to the modeling choices, but there are additional factors to consider. ICE g-computation only requires fitting models for the outcome, while MC g-computation also requires specifying a model for each time-varying confounder. This again means that ICE g-computation is less susceptible to model misspecification bias. However, ICE g-computation is designed to estimate the expectation of the outcome at each time point, conditional on the past, which in practice means the approach works best in data with a few, discrete time points (17). MC g-computation, on the other hand, can handle data structures where time to event is continuous (or practically continuous) (18). Consequently, if the researcher's data had many time points or continuous time to event, the researcher would need to pool the data into fewer time points to use ICE g-computation or use an alternative approach.

One should finally keep in mind that the estimators explored here are not the only options available. In particular, one could instead use a double-robust approach like augmented inverse probability weighting or longitudinal targeted minimum loss–based estimation, which have the advantage of relaxing the correct model specification assumptions in part because they can be implemented with machine learning algorithms (15, 17, 19, 23–25).

It bears repeating that the choice of analytical method should always be based on a wide array of considerations, most especially the research question and hypothesized causal model. We further emphasize that our results pertain to a few simulation settings, and we encourage researchers to set up their own simulation study to inform estimator selection.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States (Jacqueline E. Rudolph); Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States (Enrique F. Schisterman); and Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia, United States (Ashley I. Naimi).

This work was supported in part by the National Institutes of Health (grants R01-HD093602, R01-CA250851, and U01-DA036297) and by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (contracts HHSN267200603423, HHSN267200603424, and HHSN267200603426).

We conducted a plasmode simulation based on observed data from the Effects of Aspirin in Gestation and Reproduction (EAGeR) Trial. Researchers can apply to access EAGeR data from the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (https://dash.nichd.nih.gov/). Software code is provided in Web Appendix 2 and on GitHub (https://github.com/jerudolph13/IPW-g-comp-sim).

We thank Dr. Laura Balzer for her thoughtful, detailed comments on earlier drafts of this work. We also thank Dr. Erica Moodie for providing the software code used in the time-varying data-generating mechanism and for answering questions related to that code.

The views expressed in this article are those of the authors and do not reflect those of the National Institutes of Health.

Conflict of interest: none declared.

REFERENCES

1. Daniel RM, Cousens SN, De Stavola BL, et al. Methods for dealing with time-dependent confounding. *Stat Med.* 2013; 32(9):1584–1618.
2. Robins JM, Hernan MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, et al., eds. *Advances in Longitudinal Data Analysis*. New York, NY: Chapman and Hall/CRC Press; 2009.
3. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol.* 2017;46(2):756–762.
4. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560.
5. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model.* 1986;7(9):1393–1512.
6. Petersen ML, Porter KE, Gruber S, et al. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res.* 2012;21(1):31–54.
7. Schisterman EF, Silver RM, Perkins NJ, et al. A randomised trial to evaluate the effects of low-dose aspirin in gestation and reproduction: design and baseline characteristics. *Paediatr Perinat Epidemiol.* 2013;27(6):598–609.
8. Schisterman EF, Silver RM, Lesher LL, et al. Preconception low-dose aspirin and pregnancy outcomes: results from the EAGeR randomised trial. *Lancet.* 2014;384(9937):29–36.
9. Young JG, Hernan MA, Picciotto S, et al. Simulation from structural survival models under complex time-varying data structures. Presented at the 2008 Joint Statistical Meetings, Denver, Colorado, August 3–7, 2008.
10. Young JG, Hernan MA, Picciotto S, et al. Relation between three classes of structural models for the effect of a

Downloaded from https://academic.oup.com/aje/article/192/1/102/6702451 by Pitts Theology Library, Emory University user on 09 December 2025

time-varying exposure on survival. *Lifetime Data Anal.* 2010; 16(1):71–84.
11. Moodie EE, Stephens DA, Klein MB. A marginal structural model for multiple-outcome survival data: assessing the impact of injection drug use on several causes of death in the Canadian Co-infection Cohort. *Stat Med.* 2014;33(8): 1409–1425.
12. D'Agostino RB, Lee ML, Belanger AJ, et al. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Stat Med.* 1990;9(12): 1501–1515.
13. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11(5): 561–570.
14. Buchanan AL, Hudgens MG, Cole SR, et al. Worth the weight: using inverse probability weighted Cox models in AIDS research. *AIDS Res Hum Retroviruses.* 2014;30(12): 1170–1177.
15. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics.* 2005;61(4): 962–973.
16. Wen L, Young JG, Robins JM, et al. Parametric g-formula implementations for causal survival analyses. *Biometrics.* 2021;77(2):740–753.
17. Schomaker M, Luque-Fernandez MA, Leroy V, et al. Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions. *Stat Med.* 2019; 38(24):4888–4911.
18. Keil AP, Edwards JK, Richardson DB, et al. The parametric g-formula for time-to-event data: intuition and a worked example. *Epidemiology.* 2014;25(6):889–897.
19. Lendle SD, Schwab J, Petersen ML, et al. ltmle: an R package implementing targeted minimum loss-based estimation for longitudinal data. *J Stat Softw.* 2017;81(1):1–21.
20. Burton A, Altman DG, Royston P, et al. The design of simulation studies in medical statistics. *Stat Med.* 2006; 25(24):4279–4292.
21. Westreich D, Cole SR, Schisterman EF, et al. A simulation study of finite-sample properties of marginal structural Cox proportional hazards models. *Stat Med.* 2012;31(19):2098–2109.
22. van der Vaart AW. *Asymptotic Statistics*. New York, NY: Cambridge University Press; 1998.
23. Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms [published online ahead of print July 15, 2021]. *Am J Epidemiol.* 2021. (https://doi.org/10.1093/aje/kwab201).
24. Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. *Polit Anal.* 2010; 18(1):36–56.
25. Zhong Y, Kennedy EH, Bodnar LM, et al. AIPW: an R package for augmented inverse probability weighted estimation of average causal effects. *Am J Epidemiol.* 2021; 190(12):2690–2699.

Downloaded from https://academic.oup.com/aje/article/192/1/102/6702451 by Pitts Theology Library, Emory University user on 09 December 2025