Informing Low-Dose Aspirin in Gestation and Reproduction Through Novel Methods in Causal Inference

by

Gabriel Conzuelo-Rodriguez

MD, Autonomous University of the State of Mexico, 2014

MPH, University of Pittsburgh, 2016

Submitted to the Graduate Faculty of the

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Gabriel Conzuelo-Rodriguez

It was defended on

July 12, 2021

and approved by

Ashley I. Naimi, PhD, Department of Epidemiology
Emory Pittsburgh

Maria M. Brooks, PhD, Department of Epidemiology
University of Pittsburgh

Abdus S. Wahed, PhD, Department of Statistics
University of Pittsburgh

Edward H. Kennedy, PhD, Department of Statistics & Data Science
Carnegie Mellon University

Dissertation Director: Lisa M. Bodnar, PhD, Department of Epidemiology
University of Pittsburgh

Informing Low-Dose Aspirin in Gestation and Reproduction Through Novel Methods in Causal Inference

Gabriel Conzuelo Rodriguez, MD PhD

University of Pittsburgh, 2021

Abstract

Pregnancy loss is the most common complication of human reproduction, occurring in up to 20% of all recognized pregnancies. Aspirin, a widely available anti-inflammatory drug is hypothesized to improve pregnancy outcomes in women with a previous pregnancy loss if administered early in gestation. Under this premise, the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial was devised to evaluate the benefits of assigning preconception low-dose aspirin on live birth. While the study findings suggest a moderate increase in live birth rate of 5.1% (95% CI -0.84 to 11.2), this is currently of limited use due to (1) potential effect modification of the aspirin effect among heterogenous subgroups in the EAGeR population; (2) low generalizability ensuing after demographic differences between the trial sample and the U.S. population; and (3) measurement error associated with time-varying treatments. Presently, there is a critical need to develop epidemiologic methods to overcome these limitations.

This dissertation will focus on evaluating and developing epidemiologic methods to address these limitations. In section 2, we will conduct a simulation study to evaluate the performance of nonparametric doubly robust estimators (i.e., Augmented Inverse Probability Weighting and Targeted Minimum Loss-Based Estimation) against correctly specified Generalized Linear Models to quantify effect modification. Then, we will apply these methods in 1,228 women enrolled in the EAGeR trial to quantify the extent to which the effect of low-dose

aspirin on live birth is modified by pre-pregnancy body mass index. In Section 3, we address generalizability concerns in EAGeR that result from its highly selective recruitment process. Specifically, we will adapt the parametric g-formula to generalize the intention-to-treat (ITT) and per-protocol (PP) effects of aspirin to a more representative U.S. sample of childbearing age women with a previous pregnancy loss (National Survey of Family Growth). Finally, in Section 4, we will develop an approach based on the parametric g-formula to correct for measurement error of time-varying exposures in complex longitudinal settings. The results from this work will improve our understanding on preconception aspirin role in pregnancy loss. Furthermore, our methods will help to overcome major limitations present in modern epidemiological studies.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Pregnancy loss is the most common complication of human reproduction, occurring in up to 20% of all recognized pregnancies (Wilcox et al., 1988) and recurring in up to 30% of women with a previous pregnancy loss (Ford & Schust, 2009). Consequences of pregnancy loss extend beyond the physical sphere, as women are likely to experience psychological distress that can lead to anxiety and/or depression (Klock, Chang, Hiley, & Hill, 1997). Furthermore, we also expect to observe an increase in pregnancy loss incidence as the average childbearing age rises in most populations (Rasmark Roepke, Matthiesen, Rylance, & Christiansen, 2017). Unfortunately, current treatments to prevent a recurrent pregnancy loss are both costly and invasive (El Hachem et al., 2017). As such, they are usually reserved for women who have experienced at least two consecutive losses. However, there is an urgent need to develop safe, efficient, and affordable interventions that improve pregnancy outcomes in women who experienced a recent pregnancy loss.

Aspirin is a widely available anti-inflammatory drug that is known to improve many adverse perinatal outcomes when administered early in gestation (Turner, Robertson, Hartel, & Kumar, 2020). Evidence suggests that the effect of aspirin on perinatal outcomes is triggered by the inhibition of the cyclooxygenase enzyme, which promotes platelet aggregation through conversion of arachidonic acid into prostaglandins (Vane & Botting, 2003). The net effect is an increase in blood flow to the uterus and reproductive organs (Rubinstein, Marazzi, & Polak de Fried, 1999) that is recognizable for even low-doses of aspirin (e.g., 81 mg/day) (Weksler, Kent, Rudolph, Scherer, & Levy, 1985). Similar to other adverse perinatal outcomes, pregnancy loss is associated with restricted blood flow impairing placenta implantation (Rubinstein et al., 1999).

Therefore, it is hypothesized that low dose aspirin can reduce pregnancy loss, especially if administered early in gestation (Rubinstein et al., 1999; Vane & Botting, 2003).

Notwithstanding its theoretical benefits, few studies have evaluated the effect of low dose aspirin on livebirth among women with a previous pregnancy loss (de Jong, Kaandorp, Di Nisio, Goddijn, & Middeldorp, 2014). While informative, most of these studies assigned treatment after clinical recognition of the pregnancy (Dolitzky et al., 2006; Kaandorp et al., 2010; Tulppala et al., 1997), which is likely not as beneficial as pre-conceptional initiation. Furthermore, most women fulfilled the criteria for repeated pregnancy loss (i.e. three or more losses), which is a disease with several underlying mechanisms probably not affected by aspirin (Dolitzky et al., 2006; Kaandorp et al., 2010). Additionally, there is large heterogeneity concerning doses and comparison groups (de Jong et al., 2014). Under this premise, the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial was devised to evaluate the benefits of assigning preconception low-dose aspirin on live birth.

The EAGeR trial evaluated the benefits of assigning preconception low-dose aspirin on live birth (Schisterman et al., 2014, 2013). Unlike most trials, EAGeR design comprises two populations of women actively trying to conceive (Mumford & Schisterman, 2019; Schisterman et al., 2013). The first, "biologically-based" population, included women with a single previous pregnancy loss of less than 20 weeks of gestation within the last 12-months. It represents women who are more likely to benefit from low-dose aspirin because their overall reproductive health, these women were considered to be less predisposed to have a recurrent pregnancy loss (Schisterman et al., 2014). The second, "expanded" population, anticipates that clinicians would likely apply findings from EAGeR to a broader range of women; and considers difficulties in enrollment given the strict inclusion criteria. This expansion allowed the inclusion of women with

one or two previous losses of any gestational age at any time in the past. While the biologically based population experienced an absolute increase in livebirth rate of 9.2% (95% CI 0.5 to 17.9), the effect was diluted to 5.1% (95% CI = -0.84, 11.02) when combining both stratums (Schisterman et al., 2014). Only the primary intention-to-treat (ITT) findings from EAGeR are considered by the American College of Obstetricians and Gynecologists (ACOG) to recommend against the use of preconception low dose aspirin for pregnancy loss ("ACOG Practice Bulletin No. 200 Summary: Early Pregnancy Loss.," 2018; Porter, Gyamfi-Bannerman, & Manuck, n.d.). However, ITT analysis possibly overlook major limitations present in the EAGeR trial. Indeed, a recently concluded per-protocol analysis (PP) of the EAGeR trial (Naimi et al., 2021) found that adhering to aspirin for at least 5 days per week was associated with 15 more live births (95%CI = 7.65, 21.15) per 100 women. This dissertation finds its motivation on evaluating and developing epidemiological methods to quantify the effects of low dose aspirin in EAGeR, beyond those obtained with ITT principle. Specifically, we will focus on (1) potential effect modification of the aspirin effect among heterogenous subgroups in the EAGeR population; (2) low generalizability ensuing after demographic differences between the trial sample and the U.S. population; and (3) measurement error associated with time-varying treatments.

In general, it is accepted that treatment effects tend to fluctuate between individuals with certain clinical and demographic characteristics. This effect measurement modification is often of interest to applied researchers and policy makers. Unfortunately, most analysis evaluating effect modification rely on strong parametric assumptions, which are difficult to meet in modern epidemiologic studies (Greenland, 1993). In section 2, we will use a simulation study to evaluate the performance of nonparametric doubly robust estimators to detect complex relationships between treatment and effect modifiers. Then, we will apply these estimators to data from 1,228

women in EAGeR to quantify the change in the effect of daily low dose aspirin on live birth as a function of continuous pre-pregnancy body mass index (BMI). Unlike previous studies based on categorization and parametric models (Nobles et al., 2019; Sjaarda et al., 2017), our approach via doubly robust estimators will enable us to avoid categorization while providing enough flexibility to relax strong parametric assumptions. As a result, we will understand how the effect of low dose aspirin on livebirth is modified across the range of BMI values. Moreover, we will provide a framework to evaluate effect measure modification of binary and continuous exposures using doubly robust estimators.

As with most clinical trials, results from EAGeR suffer from threats to external validity given its strict inclusion criteria and highly selective recruitment. This resulted in a sample population of mostly white (94.6%), high-income women (40% had income exceeding 100,000 USD per year) with at least high-school education (86.2%) (Schisterman et al., 2014). In contrast, a representative sample of women living in the U.S. (National Survey of Family Growth) found that 77.2% of reproductive age women with a previous pregnancy loss were non-Hispanic white; 26% had income exceeding 100,000 USD per year; and 68.7% completed high-school education. Ultimately, these differences can lead to effect estimates in the trial sample that deviate from what would be expected in an otherwise random sample of the target population (Stephen R. Cole & Stuart, 2010; Hernán & VanderWeele, 2011a; Lesko et al., 2017; Westreich, Edwards, Lesko, Cole, & Stuart, 2019). Therefore, current EAGeR estimates are not immediately generalizable to a broader target population of women living in the U.S. In Section 3, we adapt the parametric g-formula to generalize findings from EAGeR to a more representative sample of childbearing age women with a previous pregnancy loss living in the U.S.

Lastly, we address concerns with measurement error in complex longitudinal data. While methods for measurement error correction based on Simulation Extrapolation (SIMEX) (Cook & Stefanski, 1994) and regression calibration (ROSNER, SPIEGELMAN, & WILLETT, 1990; Rosner, Willett, & Spiegelman, 2006) are well established for a single timepoint exposures, accounting for measurement error in time-varying settings is not straightforward. In Section 4, we conduct a simulation study to evaluate the performance of a measurement error correction tool for time-varying exposures based on the parametric g-formula. The development of such tool is essential for epidemiologists, as measurement error continues to be an important concern in most studies, and the availability of complex longitudinal data increases.

In summary, each section of this dissertation will focus on addressing specific limitations that are present in, but are not exclusive of, the EAGeR trial. Therefore, the methods presented throughout this work have the potential to impact not only in our understanding of the prophylactic use of periconceptional aspirin, but also in the entire field of epidemiology.

2.0 Performance Evaluation of Parametric and Nonparametric methods when Assessing Effect Measure Modification

2.1 Introduction

Epidemiologists are often interested in evaluating whether exposure effects differ between individuals with certain clinical or demographic characteristics. Such questions about effect measure modification, while essential, are subject to several difficulties. For instance, a large proportion of studies in epidemiology are underpowered to detect effect modification simply because sample sizes are powered to detect main effects exclusively. (Greenland, 1993; Rencher & Schaalje, 2007) This holds true even when efficient parametric models (e.g. maximum likelihood estimation) are used. Additionally, to obtain unbiased (more technically, asymptotically consistent) estimators with appropriate confidence interval coverage, these parametric models must be correctly specified.(Greenland, 1983, 1993; Lubin, Samet, & Weinberg, 1990; VanderWeele & Knol, 2014) That is, researchers must accurately model the functional form of continuous modifiers and the interaction between covariates, as well as selecting appropriate family distributions and link functions. Unfortunately, scenarios in epidemiology are often too complex to confidently support strong parametric assumptions about the true underlying data generating mechanisms with any degree of certainty.

In addition to concerns with power and model misspecification, categorizing continuous variables is also customary in effect measure modification analysis. For example, data from a recent randomized trial (Schisterman et al., 2013) were used to evaluate whether the effect of pre-conception aspirin on live birth differed by body mass index (BMI). To evaluate effect

modification, women were categorized into normal weight ($\leq 25\text{kg/m}^2$) and overweight/obese ($> 25\text{kg/mt}^2$) BMI categories; and effects were estimated within strata of these categorized data. (Sjaarda et al., 2017) While this common approach is often implemented by reason of simplicity, it is well understood that such categorization results in loss of information, power and underestimation of variability within levels of the categorical variables. (D G Altman, Lausen, Sauerbrei, & Schumacher, 1994; Douglas G Altman & Royston, 2006; MacCallum, Zhang, Preacher, & Rucker, 2002; Royston, Altman, & Sauerbrei, 2006) Likewise, optimal threshold selection requires accurate background knowledge of the relationship between exposure, outcome and the effect modifier. The lack of this substantive knowledge may lead researchers to choose 'optimal' thresholds based on data dredging. (D G Altman et al., 1994; Royston & Sauerbrei, 2004) Furthermore, particularly when based on quantiles, these arbitrary cut points may not be relevant beyond the study's sample.(Douglas G Altman & Royston, 2006)

Use of nonparametric methods, such as doubly robust machine learning based estimators, is increasing in epidemiology. These methods avoid the need to rely on correct parametric model assumptions. (Bang & Robins, 2005; Funk et al., 2011; Kang & Schafer, 2007; Kennedy, 2015; Naimi & Kennedy, 2017; Robins & Rotnitzky, 1995; Schuler & Rose, 2017) However, while nonparametric methods make fewer assumptions about the true underlying data generating mechanisms (such as the specific distribution that the outcome follows, or the functional form of the relation between the covariates and the outcome), they typically require much larger sample sizes to obtain a given level of accuracy. (Riley et al., 2020; van der Ploeg, Austin, & Steyerberg, 2014) The extent to which these losses materialize when effect measure modification is of primary interest is unknown. Furthermore, there are currently no empirical studies evaluating the

performance of nonparametric methods to evaluate effect modification of a continuous effect modifier.

Here, we use simulated data to first evaluate the performance of nonparametric double-robust estimators to quantify effect measure modification across binary and continuous modifiers. Our aim is to evaluate the trade-offs that result when using correctly specified parametric versus nonparametric methods. In the binary modifier case, we compare several parametric and nonparametric approaches, and evaluate the impact of sample-splitting when nonparametric approaches are used. In the continuous modifier case, we compare flexible parametric approaches such as splines and fractional polynomials to its nonparametric counterpart, the DR-learner. We lastly build upon previous BMI analysis in the Effects of Aspirin on Gestation and Reproduction (EAGeR) Trial data to illustrate how these approaches can be used to evaluate effect modification. Specifically, we estimate the extent to which the effect of aspirin assignment on live birth is modified by BMI, using the continuous version of the BMI variable. We demonstrate how doubly robust machine learning based methods compare to flexible parametric approaches to estimate how the effect of low-dose aspirin on live birth changes as a function of BMI.

2.2 Methods

2.2.1 Simulated data

2.2.1.1 Binary Effect Measure Modifier

To evaluate the performance of various methods to quantify EMM, we devised two data generating mechanisms. The first consisted of a continuous outcome, a binary exposure, a binary effect measure modifier, and two continuous confounders. The confounders (C_1 and C_2) were drawn independently from a multivariate normal distribution with mean = 0 and standard deviation (SD) = 1. The exposure was then generated from a binomial distribution using a logistic regression model as:

$$P(X = 1|C) = expit\{-0.5 + \log(1.5)\, C_1 + \log(1.5)\, C_2\}$$

where $expit(\circ) = \frac{1}{1+\exp(-\circ)}$ is the inverse of the logit function. An identical model was used to generate the binary effect measure modifier (EMM), which we denote M. Finally, the continuous outcome was obtained as:

$$Y = 120 + 6X + 6M - 3XM + 3C_1 + 3C_2 + \epsilon$$

where ϵ was randomly drawn from a normal distribution with mean equal to 0, and SD equal to 3 or 6. This outcome model yielded a mean difference between exposed and unexposed observations of 6 among those with $M = 0$, and 3 among those with and $M = 1$. A total of 1,000 Monte Carlo simulations were run, each with a sample of $N = 200$ or $N = 500$.

2.2.1.2 Continuous Effect Measure Modifier

We used the same data generating mechanisms to create the continuous confounders and binary exposure as described above. The continuous effect modifier was drawn from a uniform distribution, conditional on C and bounded between 0.1 and 13. The outcome was generated such that the mean difference for the exposure-outcome relation varied as a function of the continuous effect modifier in three different scenarios. The first scenario (Figure 1, panel A) was defined as a quadratic function using the following equation:

$$Y = 120 + 6X + 2.5M + X(M-6)^2 + \epsilon$$

The second scenario (Figure 1, panel B) was defined as an increasing monotonic function:

$$Y = 120 + 6X + 2.5M + X\ln(M) + \epsilon$$

Lastly, the third scenario (Figure 1, panel C) was defined as a complex non-monotonic relationship:

$$Y = 120 + 6X + 2.5M + X\left[4\sqrt{9M}\mathbb{I}(M < 2)\right] + \mathbb{I}(M \geq 2) \times |M-6|^2 + \epsilon$$

where $\mathbb{I}(\cdot)$ is an indicator function that evaluates to 1 if the argument is true and 0 otherwise. (Naimi & Balzer, 2018) For all the scenarios ϵ is drawn from a normal distribution with mean = 0 and SD = 6. These models generate a mean difference for the relation between the binary exposure and continuous effect modifier that changes across the range of the effect modifier. The performance of methods for continuous effect modifiers was evaluated via 200 Monte Carlo simulations, each with a sample of $N = 500$.

2.2.2 Analysis

2.2.2.1 Binary Effect Measure Modifier

We used several techniques to analyze simulated data generated with a binary EMM. For each Monte Carlo sample, we fit a correctly specified generalized linear model (GLM) with a Gaussian distribution and identity link that included an interaction term between M and X, with the estimated effect in each M stratum obtained as a contrast of coefficients from the model (correct model). We also fit two GLMs stratified by M to obtain a single coefficient for the exposure effect in each group. In our setting, this stratified modeling approach should be less efficient as it also (unnecessarily) allows all other coefficients in the model to vary between strata of M.

Two doubly robust approaches, namely targeted minimum loss-based estimation (TMLE) and augmented inverse probability weighting (AIPW) were also used. TMLE is a maximum likelihood-based method that optimizes bias-variance tradeoffs on the parameter of interest by using an extra 'targeting' step. (Schuler & Rose, 2017; van der Laan & Rubin, n.d.) The AIPW can be seen as 'augmenting' the IPW estimator with an outcome model to fully utilize the information in the conditioning set. (Glynn & Quinn, 2010; Robins, Rotnitzky, & Zhao, 1994) These estimators (i.e. TMLE and AIPW) are asymptotically equivalent. Both TMLE and AIPW estimators were stratified by M. For TMLE, initial nonparametric machine learning (ML) based models for the outcome $E(Y|X,C)$ and the exposure $\Pr(X=1|C)$ were fit. Subsequently, a no-intercept logistic regression model for the outcome was generated, using the initial exposure model as weights (IPW), and the outcome model as an offset. Predictions from this 'updated' model are generated by setting every individual to $X=0$ and then to $X=1$. Finally, the effect in the given stratum M was obtained as follows:

$$\hat{\psi}_{tmle} = \frac{1}{N}\sum_{i=1}^{N}[E^*(Y|X=1,C) - E^*(Y|X=0,C)]$$

where $E^*(Y|X=x,C)$ are the predictions from the updated model. For AIPW, we used predictions from the outcome model $E(Y|X,C)$ as well as the propensity score from the exposure model $\Pr(X=1|C)$ to estimate the effect by M using the following equation:

$$\hat{\psi}_{aipw} = \frac{1}{N}\sum_{i=1}^{N}\frac{(2X_i-1)[Y_i - E(Y|X,C)]}{(2X_i-1)\Pr(X|C) + (1-X_i)} + E(Y|X=1,C) - E(Y|X=0,C)$$

Both estimators were fit with and without 10-fold sample splitting, (Kravitz, Carroll, & Ruppert, 2019) which proceeds by dividing the sample into roughly 10 equal folds. Models are then fit in nine of these folds, and the effect estimate is then computed in the remaining fold. This process is repeated 10 times, yielding 10 effect estimates for each stratum. The overall estimate is then obtained by averaging each of these 10 estimates.

Double-robust estimators require specification of both the exposure and outcome models, which can both be estimated nonparametrically. We used a stacking algorithm (SuperLearner) for both that included: (1) random forests via the ranger package (500 trees with a minimum of 30 observations per node and 2 or 3 predictors sampled for splitting at each node); (2) generalized linear model via penalized maximum likelihood glmnet package with elastic-net mixing parameter from 0 to 1 by 0.2; (3) support vector machine via svm package with v = 0.25, 0.5 or 0.75, cost parameter = 1 and degree of polynomial = 3 or 4; (4) multivariate adaptive regression splines via earth package with degree of 2 and 3; (5) generalized additive models via gam package with 2, 3 and 4 knots; (6) Bayesian GLM via bayesglm package with normally distributed coefficient priors, mean = 0; (7) Generalized linear models via glm package with identity link; (8) multinomial log-linear models via neural networks via nnet package; and (9) standard mean estimator via mean

package. Ten-fold cross validation was used to estimate the weights for the SuperLearner (a process that is distinct from 10-fold sample splitting).

For each analytic approach, we computed the absolute bias, the mean standard error and 95% confidence interval coverage. A measure of accuracy was defined as the ratio of the average of all standard errors divided by the standard deviation of all estimates. Relative efficiency was also calculated by taking the inverse of the ratio of the mean squared error (MSE) between the correctly specified GLM and each of the other estimators. These parameters were all calculated by Mstratum, considering a true value of $X = 6$ when $M = 0$; and $X = 3$ when $M = 1$. Power to detect EMM (i.e. interaction between X and M) was computed by testing whether the difference in the risk differences was different from zero using a Z-test. To do this, we estimated the pooled variance for the difference and calculated a 95% CI. Then, we computed the proportion of times whenever the 95% CI included the null.

In addition to the effect modification scenarios, we evaluated the performance of all our estimators for all the pre-specified simulation conditions under no effect modification (i.e., interaction between X and M was set to zero). The result from these simulations is presented in the Appendix 6. Additionally, we present type I error rate for each estimator when testing for effect modification (Appendix 7).

2.2.2.2 Continuous Effect Measure Modifier

Continuous modifier data were analyzed using two flexible parametric models, restricted cubic splines and second-degree fractional polynomials. These approaches have been used before to model interactions of treatment with continuous variables.(Royston & Sauerbrei, 2004, 2013, 2014) Additionally, we used a doubly robust influence function-based estimator similar to AIPW, the DR-learner. (Kennedy, 2020; van der Laan & Luedtke, n.d.) The DR-learner is a flexible,

oracle efficient estimator, capable of providing model-free error bounds. (Kennedy, 2020) This estimator was used to compute efficient influence function (EIF) values based on predictions from the outcome model $E(Y|X,C,M)$ and the propensity score model $\Pr(X=1|C,M)$. The EIF was then computed as:

$$EIF = \frac{(2X_i - 1)\big(Y - E(Y|X,C,M)\big)}{(2X_i - 1)\Pr(X|C,M) + (1 - X_i)} + E(Y|X=1,C,M) - E(Y|X=0,C,M)$$

Intuitively, the EIF can be roughly interpreted as individual risk differences in the outcome. To estimate effect measure modification for the continuous effect modifier, we then regressed these EIF values against M using the SuperLearner algorithm with the same libraries described for the binary case. This regression step returns the risk difference for the outcome across the range of the continuous modifier. Finally, predictions from this Super Learner are then plotted against the effect modifier values to obtain a risk difference values across the range of the continuous effect modifier. Using these predictions, we also computed integrated absolute bias (IAB) on a grid set across the range of EMM values with an increment of $\Delta = 0.1$, defined as:

$$IAB = \sum_{i=1}^{130} \big|\hat{\psi}_i - \psi\big|\Delta$$

And integrated squared bias (ISB), defined as:

$$IAB = \sum_{i=1}^{130} \big(\hat{\psi}_i - \psi\big)^2\Delta$$

In both cases, $\hat{\psi}_i$ represents the mean estimated distribution of the 200 datasets derived from the Monte Carlo simulations, and the sum ($\sum_{i=1}^{130}(\cdot)$) is taken across the grid set.

2.2.3 Empirical Data

To evaluate how nonparametric methods for continuous effect modification perform in a realistic data setting, we used data from 1,228 women in EAGeR to quantify the change in the effect of daily low-dose aspirin on livebirth as a function of continuous pre-pregnancy BMI. Details on the EAGeR dataset are provided elsewhere. (Schisterman et al., 2013) We sought to quantify the intent-to-treat effect of being assigned to receive 81mg of aspirin per day ($N = 615$) versus placebo ($N = 613$), adjusted for baseline level of hsCRP. Pre-pregnancy BMI was the effect modifier, measured on the continuous scale in $\mathrm{kg/m^2}$. The outcome was an indicator of live birth status at the end of follow-up, which accrued for at most 6-menstrual cycles, and throughout pregnancy in those who became pregnant. EAGeR data were analyzed using restricted cubic splines, second degree fractional polynomials and the DR-learner.

2.3 Results

The results from the binary case simulations are presented in Table 1 and 2. As expected, for all our simulations, the correctly specified GLM model that included an interaction term was unbiased in both EMM strata, with coverage ranging from 0.93 to 0.97. Similarly, the stratified GLM models were unbiased, with coverage of at least 0.95 in both EMM strata.

The accuracy of the estimators was calculated as the ratio of the average SE to the SD of the estimates from each Monte Carlo sample. This measure of accuracy captures how well the SE estimates the true sampling variation. In our experiment, the estimates ($\hat{\psi}$) from each Monte Carlo simulation should follow a normal distribution and the SEs associated with each simulation (i.e.,

SE$(\widehat{\psi_i})$) should correspond to the standard deviation (i.e., SD$(\hat{\psi})$) of the distribution of estimates from all Monte Carlo simulations. Therefore, $\frac{\frac{1}{N}\sum_{i=1}^{N} SE(\hat{\psi}_i)}{SD(\hat{\psi})} \approx 1$. An illustration of the relationship between the numerator (i.e. average $SE(\hat{\psi}_i)$) and the denominator (i.e. $SD(\hat{\psi})$) used for accuracy calculations can be found in the Appendix. Under this definition, both GLM and stratified GLM were fully accurate (i.e., $\frac{\frac{1}{N}\sum_{i=1}^{N} SE(\hat{\psi}_i)}{SD(\hat{\psi})} = 1$).

Results from doubly robust estimators for the binary modifier are presented in Table 1 and 2, and differed from the parametric case in several aspects. First, in all the scenarios we explored, these estimators showed slightly larger bias and MSE compared to its parametric counterparts. However, they showed better performance in terms of coverage. Accuracy of doubly robust estimators was close to parametric models, but only when sample splitting was used, which underscores the importance of implementing this, or alternative techniques to reduce overfitting. For instance, in the scenario of $N = 500$ and $\epsilon \sim N(\mu = 0, SD = 6)$, the accuracy of TMLE went from 0.85 to 1.01 and from 1.81 to 1.05 in EMM stratum 0 and 1, respectively. AIPW showed a similar improvement going from 1.58 to 0.98 and from 2.58 to 0.87 in EMM stratum 0 and 1, respectively. A similar pattern was seen in all the additional scenarios.

The power to detect effect measure modification in all the scenarios is presented in Table 3. As expected, the greatest power for all our estimators was observed in the scenario where $N = 500$ and $\epsilon \sim N(\mu = 0, SD = 3)$. Overall, the GLM with an interaction term achieved the greatest power compared to all other estimators, ranging from 42% to 100% for the worst and best scenario, respectively. For AIPW, we observed improvements in power when sample splitting was used, regardless of the explored scenario. However, we did not observed any improvements in TMLE when sample splitting was used.

Figure 2 shows the conditional mean difference: $E(Y^1 - Y^0|M)$, plotted against the range of the continuous modifier. The solid black line in panels A to I represents the true mean difference across the values of effect modifier as defined by the quadratic (panels A to C), increasing monotonic (panels D to F) and complex (panels G to I) functions.

As depicted, the GLM with restricted cubic splines, the fractional polynomial of second-degree and the DR-learner estimator were able to capture the true quadratic and increasing monotonic functions (Figure 2, panels A to F). In these scenarios, the flexible parametric approaches had lower integrated bias and squared bias compared to the DR-learner (Table 4). Conversely, the DR-learner estimator demonstrates a greater capacity to model the complex function compared to the GLM model with restricted cubic splines and the fractional polynomials of second-degree (Figure 2, panels G to I). Indeed, the integrated absolute bias was 141.3, 251.7 and 209.0 for the DR-learner, GLM with splines and fractional polynomial, respectively. Similarly, integrated square bias was 339.4, 786.8 and 742.8 for the DR-learner, GLM with splines and fractional polynomial, respectively (Table 4).

Lastly, a comparison between the DR-learner and our two flexible parametric approaches is presented in Figure 3. This figure shows the conditional risk difference for the relation between pre-conception daily low-dose aspirin and the probability of live birth across the range of pre-pregnancy BMI values. The original distribution of BMI in the EAGeR population can be found in the appendix. In general, the DR-learner and the fractional polynomial of second-degree behave relatively similar across the entire range of BMI values. As for the GLM with restricted cubic splines, we observed a similar behavior to the other estimators for BMI levels ranging between 20 and 40$\mathrm{kg/m^2}$, followed by a sharp decline after this point.

2.4 Discussion

In this paper, we outline and evaluate an alternative approach to quantify effect modification based on nonparametric, doubly robust estimators. These estimators offer some degree of protection against model misspecification, particularly that arising from an incorrect functional form. (Keil et al., 2018) Such functional form assumptions typically include no confounder-confounder interactions or linear dose-response relations between continuous covariates and the outcome. Nevertheless, they also tend to suffer from losses in efficiency when compared to correctly specified parametric models.

Mounting theoretical and empirical evidence is suggesting that causal effect estimation with machine learning methods should only be done via doubly robust estimators. (Naimi & Kennedy, 2017; Zivich & Breskin, 2020) However, little is known about how well these methods perform when used to address questions about effect measure modification. In the simulated scenarios we explored, we found that machine learning methods with doubly robust estimators can perform comparatively well, but only when sample splitting was used. The use of sample splitting can reduce problems that result from overfitting and increase the accuracy and robustness of inferences when machine learning methods are used. (Kreif & DiazOrdaz, 2019; Rinaldo, Wasserman, G'Sell, & Lei, 2016) Importantly, these results align with previous simulation studies that demonstrate the importance of sample splitting.(Naimi & Kennedy, 2017; Zivich & Breskin, 2020)

The extension to the continuous effect modifier demonstrated clearer benefits for adopting nonparametric doubly robust estimators when compared to using flexible parametric models, especially when evaluating non-linear non-monotonic functions. In this scenario, we found a much smaller integrated absolute bias and integrated squared bias when the DR-learner was compared

to restricted cubic splines and second-degree fractional polynomials. Again, these results reflect the degree of flexibility between parametric and nonparametric methods. While we could have increased the degree of flexibility of these parametric models, we opted to use specifications that are most commonly used in applied epidemiological analyses. (Greenland, 1995; Howe et al., 2011) Additionally, at certain level of increased flexibility the concerns commonly invoked regarding the use of nonparametric methods (e.g., curse of dimensionality) would be important to consider.

To illustrate the application of these methods in EAGeR data, we evaluated the extent to which the effect of low dose aspirin on live birth was modified by BMI. In contrast to previous studies, (Sjaarda et al., 2017) we describe a functional relationship between BMI and aspirin assignment, hence enabling us to observe the risk difference change in live birth compared to pregnancy loss for aspirin assignment across the entire range of BMI values. Overall, the flexible parametric models and the DR-learner produced similar results. Women with BMI in the range of 20 to 40 kg/m^2 had a beneficial effect of aspirin on live birth, followed by a steady decline after this point. This may be the result of a dilution of the aspirin effect as BMI increases, as was observed in other studies outside perinatal epidemiology.(Patrono & Rocca, 2017; Rothwell et al., 2018)

Generalized linear models are among the most commonly used regression models in the applied sciences. When correctly specified, few methods will outperform generalized linear models. Furthermore, in our simulations, optimal performance was obtained even under mild misspecification (i.e., when effect modification was estimated via stratified GLM). It is reassuring then, that the nonparametric estimators we explored performed as well as the GLM approaches, particularly when sample splitting was used. Furthermore, in certain settings (such as when the

effect modifier was continuous, and the function defining effect measure modification was complex), the nonparametric methods we explored outperformed GLMs rather markedly.

However, as appealing as these methods are, several considerations should be made before use. First, there is currently a wide array of libraries that can be included in the stacking algorithm. The decision to include a given library must consider the research question as well as previous knowledge, if any, about the relationship between exposure, outcome and effect modifier. It is advisable to have a good balance between traditional parametric and data adaptive models in the final pool of libraries. (Naimi & Balzer, 2018) Second, data adaptive methods should be carefully tuned to yield optimal performance. Tuning can be achieved by including a wide array of diverse algorithms in a stacking library (as we did in our study), but also by including screening algorithms that select important variables and/or variable transformations from the covariate adjustment set. We did not explore the impact of varying the algorithms in the stacking library, nor did we include screening algorithms. Third, our results demonstrate the importance of sample splitting to obtain correct standard errors. In this study, we split our samples into ten folds, however other research has relied on different numbers of sample splitting folds ranging from 2 to 10. The tradeoff between choosing a smaller versus larger number of sample splitting folds is, to our knowledge, unexplored. Finally, in our simulation study of the continuous effect modifiers, as well as our evaluation of the effect of low-dose aspirin on the probability of live birth as a function of continuous BMI, we did not consider standard error estimation. At present, there is no viable method to accurately estimate standard errors for continuous functionals when machine learning methods are used.

In summary, although it is generally accepted that treatment effects vary according to sociodemographic and clinical characteristics, studies specifically design to detect EMM are rarely

encountered in epidemiological literature. Our study shows the utility of nonparametric, doubly robust, machine learning based methods to approach the effect modification. These estimators perform relatively well compared to parametric methods under correctly specified conditions. Losses in performance should be mitigated by using sample splitting or similar techniques that avoid overfitting. Furthermore, its use will enable the analyst to avoid relying on parametric assumptions and are preferable in conditions of limited sample size, like most of those involving effect measure modification.

2.5 Tables

Table 1. Performance of Different Estimators to Detect Effect Measurement Modification (Stratum Zero of the Effect Measure Modifier)

Feature	Interaction GLM	Stratified GLM	AIPW	AIPW SS	TMLE	TMLE SS
Simulation specifications: N = 200; SD = 3						
Mean bias	0.05	0.05	0.10	0.94	0.26	0.07
Mean SE	0.58	0.59	1.09	1.73	0.94	0.61
Coverage	0.94	0.95	0.99	0.95	0.96	0.95
Accuracy	0.98	0.98	1.59	1.05	1.09	1.01
Simulation specifications: N = 200; SD = 6						
Mean bias	0.10	0.10	0.11	2.32	0.49	0.14
Mean SE	1.16	1.17	2.01	3.92	1.85	1.22
Coverage	0.94	0.95	0.99	0.95	0.96	0.94
Accuracy	0.98	0.98	1.61	1.04	1.19	1.01
Simulation specifications: N = 500; SD = 3						
Mean bias	0.02	0.02	0.02	0.01	0.25	0.03
Mean SE	0.37	0.37	0.69	0.41	0.6	0.38
Coverage	0.95	0.95	0.99	0.95	0.93	0.95
Accuracy	1.02	1.02	1.54	0.77	0.63	1.01
Simulation specifications: N = 500; SD = 6						
Mean bias	0.03	0.03	0.06	0.05	0.38	0.06

Mean SE	0.73	0.74	1.29	0.8	1.19	0.76
Coverage	0.95	0.95	0.99	0.95	0.95	0.95
Accuracy	1.00	1.02	1.58	0.98	0.85	1.01

NOTE: Results from 1,000 Monte Carlo simulations

Augmented Inverse Probability Weighting (AIPW); Targeted Minimum Loss-based Estimation (TMLE); sample splitting (SS).

Accuracy = Average of SE ($\hat{\psi}_i$) / SD ($\hat{\psi}$)

Table 2. Performance of Different Estimators to Detect Effect Measurement Modification (Stratum One of the Effect Measure Modifier)

Feature	Interaction GLM	Stratified GLM	AIPW	AIPW SS	TMLE	TMLE SS
Simulation specifications: N = 200; SD = 3						
Mean bias	0.01	0.01	0.03	0.04	0.12	0.02
Mean SE	0.64	0.73	2.04	1.06	1.83	0.76
Coverage	0.93	0.95	1.00	0.96	0.99	0.94
Accuracy	0.90	1.00	2.59	1.05	2.21	1.01
Simulation specifications: N = 200; SD = 6						
Mean bias	0.02	0.02	0.08	0.14	0.22	0.10
Mean SE	1.29	1.46	3.87	2.05	3.63	1.52
Coverage	0.93	0.95	1.00	0.96	0.99	0.96
Accuracy	0.9	1.00	2.51	1.07	2.23	1.07
Simulation specifications: N = 500; SD = 3						

Mean bias	0.02	0.01	0.01	0.01	0.09	0.01
Mean SE	0.41	0.46	1.32	0.51	1.16	0.47
Coverage	0.94	0.97	1.00	0.98	0.99	0.96
Accuracy	0.96	1.06	2.66	0.99	0.67	1.05
Simulation specifications: N = 500; SD = 6						
Mean bias	0.03	0.03	0.04	0.02	0.24	0.04
Mean SE	0.81	0.91	2.51	1.05	2.3	0.93
Coverage	0.94	0.97	1.00	0.97	0.99	0.97
Accuracy	0.96	1.06	2.58	0.87	1.81	1.05

NOTE: Results from 1,000 Monte Carlo simulations

Augmented Inverse Probability Weighting (AIPW); Targeted Minimum Loss-based Estimation (TMLE); sample splitting (SS).

Accuracy = Average of SE ($\hat{\psi}_i$) / SD ($\hat{\psi}$)

Table 3. Power to Detect Effect Measure Modification by Several Estimators

Estimator	**N = 200**		**N = 500**	
	SD = 3	**SD = 6**	**SD = 3**	**SD = 6**
GLM	0.93	0.42	1.00	0.77
AIPW	0.58	0.30	0.82	0.58
AIPW SS	0.79	0.34	0.99	0.70
TMLE	0.90	0.41	0.99	0.73
TMLE SS	0.90	0.38	0.99	0.76

NOTE: Power to detect EMM was computed by testing whether the difference in the risk

differences were different from zero using a Z-test.

Augmented Inverse Probability Weighting (AIPW); Targeted Minimum Loss-based Estimation (TMLE); sample splitting (SS).

Table 4. Performance of Different Estimators to Detect Effect Modification of Continuous Modifier

Quadratic Function			
Feature	Restricted Cubic Splines	Fractional Polynomial (2^{nd} degree)	DR-Learner
IAB	13.1	24.3	64.4
ISB	2.0	7.6	62.2
Natural Logarithm Function			
IAB	18.6	29.4	31.8
ISB	5.0	14.7	22.1
Complex Function			
IAB	251.7	209.0	141.3
ISB	786.8	742.8	339.4

NOTE: Integrated Absolute Bias (IAB); Integrated Squared Bias (ISB)

2.6 Figures

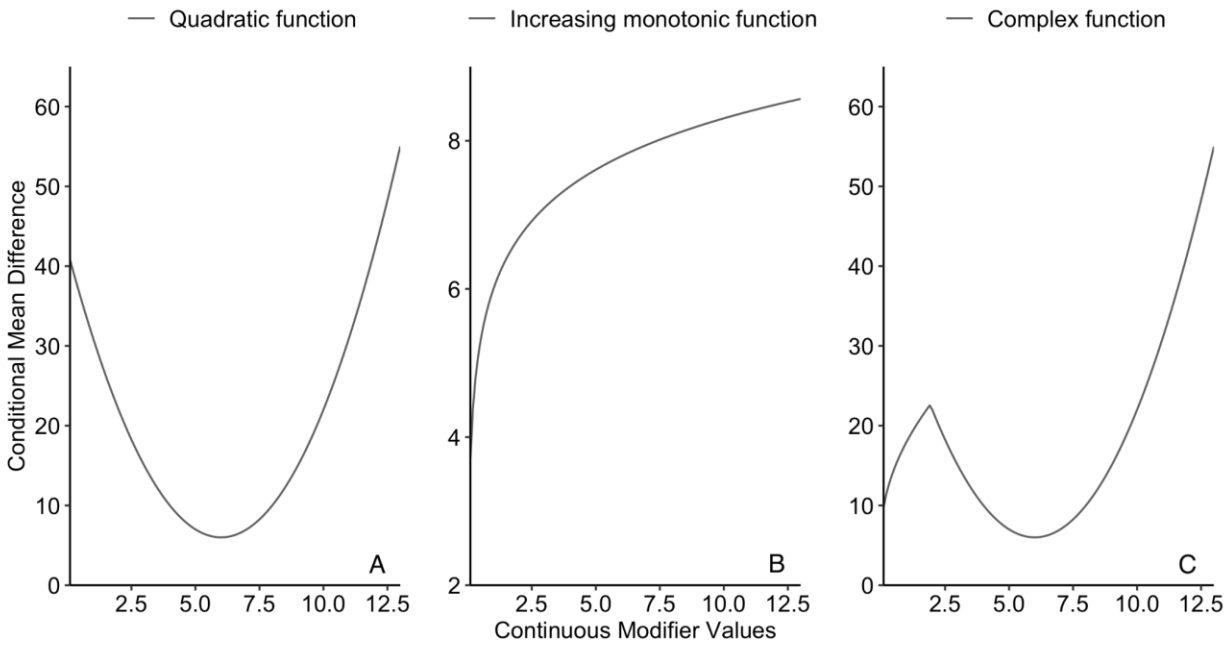


Figure 1. True mean difference for the relationship between a binary treatment and a continuous modifier using three different functions

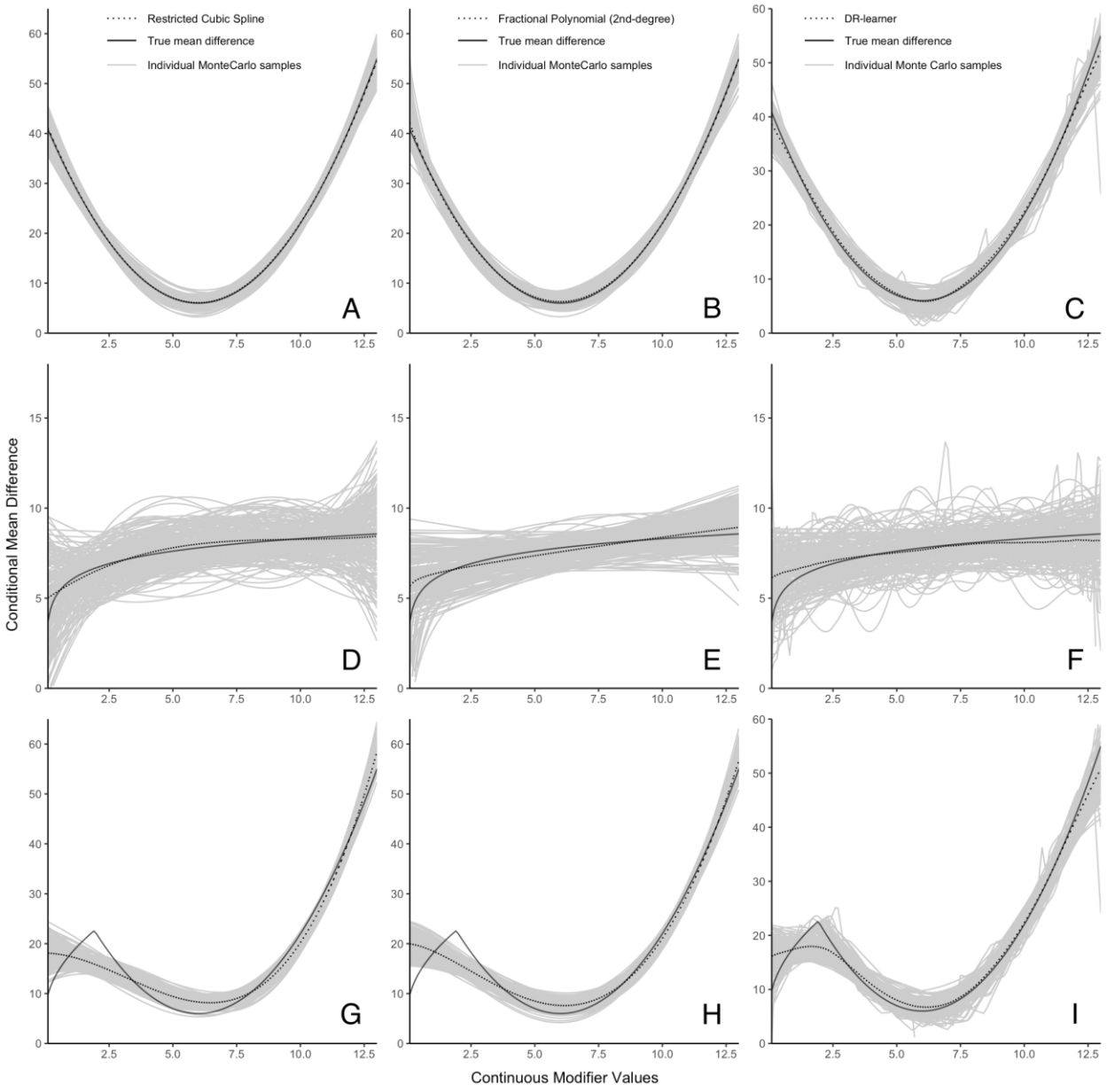


Figure 2. Estimation of continuous effect measurement modification with two flexible parametric models vs. nonparametric DR-learner estimator.

NOTE: Results from 200 Monte Carlo simulations of N = 500. True mean difference for a given function is presented across

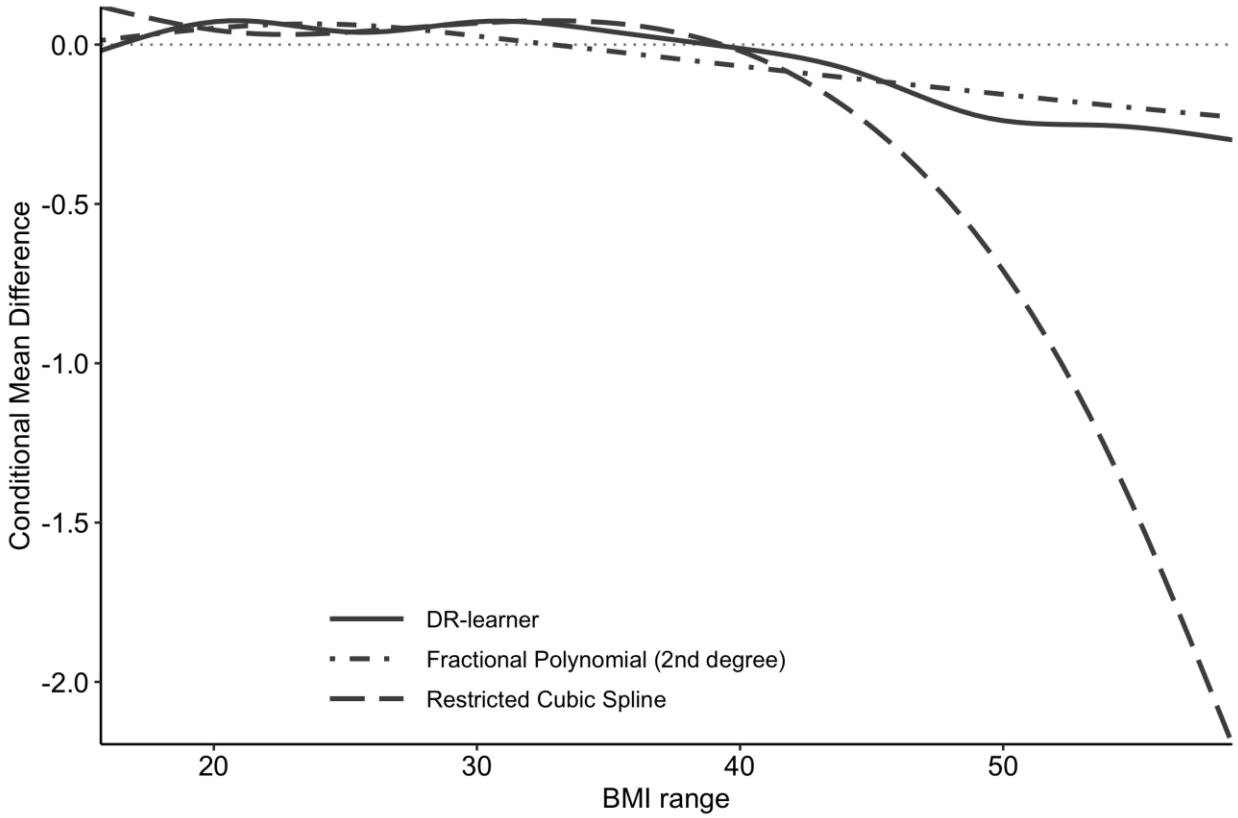


Figure 3. Conditional mean difference for the effect of aspirin on livebirth across BMI values in the EAGeR trial using flexible parametric vs. nonparametric DR-learner estimator.

NOTE: Results from 50 bootstrap resamples of N=1,228

3.0 Generalizing evidence from the Effects of Aspirin on Gestation and Reproduction trial

3.1 Introduction

For decades, aspirin has been recognized for its protective role against several complications of human reproduction, including pregnancy loss. (de Jong et al., 2014) Indeed, the beneficial effects of preconception low-dose aspirin on live birth are widely accepted in patients with anti-phospholipid syndrome.(Empson, 2002; Farquharson, 2002) Nonetheless, it is unclear whether other women at high risk of subsequent pregnancy loss would benefit from treatment in the same way.(de Jong et al., 2014) Under this premise, the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial was devised to evaluate the effects of assigning preconception low-dose aspirin on live birth among women trying to become pregnant after 1 or 2 prior pregnancy losses.(Schisterman et al., 2013) Unlike most previous trials, pre-conception initiation of aspirin in EAGeR allowed for an evaluation of its effects early in pregnancy, a potentially critical period. The primary intention-to-treat (ITT) analysis showed that women assigned to aspirin had about 5 more live births per 100 women (95%CI = -0.84, 11.02) when compared to placebo.(Schisterman et al., 2014) Furthermore, a recent per-protocol (PP) analysis reported that adhering to aspirin for at least 5 days per week was associated with 15 more live births (95%CI = 7.65, 21.15) per 100 women.(Naimi et al., 2021) Both analyses suggest some beneficial effects of pre-conception initiated low-dose aspirin on various pregnancy outcomes in women at high risk of pregnancy loss.

However, as with many clinical trials, generalizing EAGeR findings is difficult due to the set of strict eligibility criteria employed during recruitment, including the criterion that women must be willing to participate in a randomized clinical trial. For instance, women age 18-41 years

from any one of four fertility centers in the United States (Scranton, PA; Denver, CO; Buffalo, NY; Salt Lake City, UT) with a history of one or two prior pregnancy losses, and additional characteristics were eligible for inclusion in EAGeR.(Schisterman et al., 2013) Furthermore, While this strategy provided reliable enrollment, it also shifted the distribution of key potential effect measure modifiers of treatment and compliance. For example, in EAGeR, women were older and more educated compared to their counterparts from a representative sample of women living in the U.S. (e.g., National Survey of Family Growth [NSFG](National Center for Health Statistics (NCHS), 2017b)). Ultimately, differences in these key effect measure modifiers can lead to effect estimates in the trial sample that deviate from what would be expected in an otherwise random sample of the target population.(S. R. Cole & Stuart, 2010; Hernán & VanderWeele, 2011b; Stuart, Bradshaw, & Leaf, 2015) Therefore, current EAGeR estimates are not immediately generalizable to a broader target population of women living in the U.S.

Here, we adapt the parametric g-formula to generalize the ITT and PP analyses of EAGeR to a more representative sample of childbearing age women with a previous pregnancy loss living in the U.S.

3.2 Methods

3.2.1 The EAGeR data (Trial Sample)

We used data from the Effect of Aspirin in Gestation and Reproduction (EAGeR) trial.(Schisterman et al., 2013) Over the four years of recruitment, there were a total of 1,228 women enrolled across 7 sites in the U.S. Originally, women were eligible to enroll if they were

between 18 and 40 years old, had a single previous pregnancy loss at less than 20 weeks of gestation within the last 12-months. These criteria were later expanded to allow for one or two previous losses of any gestational age at any time in the past. Following enrollment, women were randomized to receive 81mg of aspirin per day (n = 615) or placebo (n = 613). Additionally, all participants received 400 mcg of daily folic acid supplementation. Women were followed for at most 6-menstrual cycles, if they did not become pregnant, and during pregnancy, if it occurred. Among those who conceived, treatment was continued until 36-weeks of gestation. The primary outcomes of interest were live birth and pregnancy loss. All participants provided written informed consent. Additional information on the study procedures and protocols can be found elsewhere.(Schisterman et al., 2014, 2013)

3.2.2 The National Survey of Family Growth data (Target Population)

We used data from the 2015-2017 period of the National Survey of Family Growth (NSFG), a continuous probability survey of the non-institutionalized U.S. population, aged 15 to 44 years.(Groves, Mosher, Lepkowski, & Kirgis, 2009; National Center for Health Statistics (NCHS), 2017b) NSFG used a multi-stage stratified clustered sampling frame of households in 65 primary sampling units chosen to represent the entire U.S. population.(National Center for Health Statistics (NCHS), 2017a) The interviews were conducted in person and in private settings by female interviewers trained specifically for the NSFG survey. One individual from each household was interviewed according to sampling and design criteria; once contact was established with an adult (18 years or older) member of a sampled household, the interviewer conducted the household screener to determine if any household member was age-eligible for the survey. If more than one age-eligible household member was identified, the pre-programmed survey selection algorithm

selected one person to be interviewed. If no one in the household was eligible, no further contact was made with the household. Because the retrospective nature of some questions in the NSFG, a recall tool called "Life History Calendar" was used to help the respondent record key personal events used as landmark events to cue memories of the dates of events measured in the survey. Additional details in design and methodology had been provided elsewhere.(National Center for Health Statistics (NCHS), 2017b, 2017a)

A total of 10,094 individuals (55% women) responded to the 2015-2017 NSFG survey.(National Center for Health Statistics (NCHS), 2017a) The response rate for women was 66.7%. This provided information on 9,553 pregnancies among 5,554 women.(National Center for Health Statistics (NCHS), 2017a) Similar to EAGeR, we restricted our analysis to women with two or three previous pregnancies, the most recent of which ended in miscarriage or stillbirth. We additionally excluded women that were pregnant at the time of the survey. However, unlike EAGeR, the NSFG sample was drawn from a nationally representative cohort of women in the United States. Our final analytic sample from the NSFG consisted of 806 childbearing age women with a previous pregnancy loss.

3.2.3 Statistical Analysis

3.2.3.1 Overview of the g-formula to generalize clinical trials findings

The parametric g-formula can be used to estimate the population average treatment effect (PATE) when the trial sample does not constitute a simple random sample of the target population.(Lesko et al., 2017) The PATE can be obtained for both the intention-to-treat (ITT) and the per-protocol (PP) effect:

$$PATE\ ITT = E(Y^{r=1}) - E(Y^{r=0})$$

$$PATE\ PP = E(Y^{r=1,\bar{c}=1}) - E(Y^{r=0,\bar{c}=1})$$

Where Y^r and $Y^{r,\bar{c}}$ represent the potential outcomes under randomized treatment r (i.e. aspirin = 1, placebo = 0), and under randomized treatment r and adherence to assigned treatment $\bar{c} = 1$ throughout the follow-up. To obtain valid estimates using this method, we must first define a set of covariates L, such that, conditional on L, we can reasonably assume exchangeability between individuals sampled into the study ($S = 1$) and those in the target population that were not included into the study ($S = 0$).

$$S \perp Y^r | L$$

Under this assumption of conditional exchangeability and in the presence of random treatment assignment (i.e., ITT), the PATE can be estimated using the following g-formula estimator (Lesko et al., 2017):

$$E[Y^r] = \sum_l E[Y|R = r, L = l, S = 1]P(L = l)$$

Where $E[Y|R = r, L = l, S = 1]$ is estimable from the trial sample and $P(L = l)$ from the external target population.

This approach can be extended to generalize effects of non-randomized interventions, which is analogous to the presence of non-adherence in randomized trials. In this scenario, one must assume that conditional exchangeability of treatment (i.e., adherence [C]) holds within a reasonable set of covariates L, that is: $C \perp Y^r | L, S = 1$.

3.2.3.2 Generalizability of the EAGeR trial using g-formula

To evaluate the effect of aspirin on our reproductive outcomes of interest (i.e., hCG detected pregnancy, pregnancy loss and live birth), we drew a directed acyclic graph (DAG) representing the causal relationships between adherence to aspirin, time-varying confounders and the outcomes of interest (Appendix). We first compared baseline demographic information available in EAGeR and the NSFG to identify potential modifiers of the aspirin and live birth relationship that also affected selection into the study. Specifically, we considered age, body mass index (BMI), income (<40,000 USD vs. ≥40,000 USD), race (white vs. non-white), education (high-school degree vs. no high-school degree), marital status (married vs. not married), employment (yes vs. no), alcohol use (ever vs. never), and tobacco use (ever vs. never). An absolute standardized mean difference (ASMD) greater than 0.2 was considered a substantial difference between populations. Information on adherence, an essential component to estimate PP effects, was available in EAGeR but not in the NSFG. We developed a logit Generalized Linear Model (GLM) to impute adherence in the NSFG based on baseline covariates that were considered predictors of adherence. A similar approach was considered to obtain initial estimates for nausea/vomiting, and bleeding.

Then, in the total EAGeR sample of 1,227 women (42,697 person-weeks of follow-up), we used 8 GLMs to model the relationship between the specified factors and the following outcomes: live birth, pregnancy loss, withdrawal from the study, hCG detected pregnancy, end of follow-up without pregnancy, adherence to treatment, bleeding, and nausea or vomiting. For each model, we also generated multiplicative interaction terms between three different types of variables: (1) adherence and assigned treatment; (2) adherence and baseline covariates that had different

distribution between EAGeR and the NSFG; and (3) time-varying confounders and baseline covariates that had different distribution between EAGeR and the NSFG.

The g-formula algorithm starts by fitting the above-mentioned models. Subsequently, we drew a Monte Carlo resample (M = 1,000) of the NSFG baseline data at a potential first week of treatment. The models that were fit in the EAGeR data were then used to predict a second week of follow-up, which in turn was used to predict the third week of follow-up. We repeated this process over a period of 60 weeks. To obtain generalized ITT estimates, we compared the probability of hCG detected pregnancy, pregnancy loss and live birth under a scenario where everyone was assigned to preconception low-dose aspirin vs. a scenario where everyone was assigned to placebo. For the generalized PP effects, we compared the probability of a given outcome if everybody were assigned to aspirin or placebo and adhered with treatment for at least 5 of 7 days a week. Full details on the g-computation algorithm are provided in the Appendix.

In addition to the analysis using the entire EAGeR sample (i.e. original and expanded inclusion criteria), we replicated the entire procedure in the subsample of EAGeR that met the original inclusion criteria (N = 548).

3.3 Results

There were substantial differences observed between participants in the EAGeR trial and the NSFG, particularly in sociodemographic characteristics. Overall, compared to the NSFG, women in EAGeR were older (28.7 vs. 25.9 years), mostly non-Hispanic white (94.6% vs. 77.2%), predominantly married (91.5% vs. 60.5%) and had at least a high school education (86.2% vs. 68.7%) (Table 1). Women in EAGeR were also less likely to report using alcohol (31.1% vs.

75.5%) and tobacco (12.3% vs. 36.7) during the previous 12 months. These patterns were similar after constraining EAGeR population to the original inclusion criteria. Only employment (0.02) and annual income under 40,000 USD (0.07) were below the 0.20 threshold for a meaningful absolute standardized mean difference (Figure 1).

The ITT effects generalized from EAGeR to the NSFG are presented in Table 2. In contrast to the estimated difference of 5.1 more live births per 100 women in the original EAGeR sample (95% CI: -0.8, 11.0), if we were to assign aspirin to women with the distribution of baseline covariates found in the NSFG, we would expect about 2 more live births per 100 women compared to a scenario where we assign all women to placebo (95%CI = -3.0, 6.9). As in the original EAGeR sample, the transported ITT effect increased to 4.9 (95% CI: -1.5, 11.3) when restricting EAGeR population to those meeting the original inclusion criteria, but using the distribution of baseline covariates found in the NSFG. However, these generalized effects were again lower in magnitude than what was observed in EAGeR. Similar patterns were seen for the ITT effects of aspirin on hCG detected pregnancies and pregnancy loss. The generalized ITT effect estimate (expressed per 100 women in the sample) was 4.0 (95%CI -1.6, 9.7) for hCG detected pregnancy, and 0.9 (95%CI -3.2, 5.0) for pregnancy loss (Table 2). These generalized ITT results aligned well with those in the original EAGeR trial.

For the per protocol effects, we found that after generalizing to a sample in which the distribution of baseline covariates were comparable to those found in the NSFG, the effect of adhering with aspirin for at least 5/7 days per week over the entire course of follow up (relative to adhering with placebo) resulted in an average of 3 more live births per 100 women (95% CI = -1.9, 7.3). Again, these findings contrast with per protocol analysis of the EAGeR data showing about 15 more live births (95%CI = 7.65, 21.15) per 100 women.(Naimi et al., 2021). Restricting

these generalized per protocol analyses to the sample of women in the original stratum yielded an estimated increase of 12 more live births per 100 women (95% CI: 6.5 to 18.1) (Table 2).

3.4 Discussion

We sought to generalize the estimated ITT and PP effects from the EAGeR trial to a more representative sample of US women with a previous pregnancy loss (NSFG 2015-2017). Overall, we found that assigning low-dose aspirin to a population with a distribution of treatment effect modifiers similar to the NSFG resulted in mild improvements in hCG detected pregnancy and live birth rates among women at high-risk of experiencing a subsequent pregnancy loss. Similarly, adhering to aspirin for at least 5/7 days a week results in mild benefits from treatment. These findings diverge from previous ITT and PP analysis of the EAGeR trial (Naimi et al., 2021; Schisterman et al., 2014).

The generalizability of a sample of study participants to their counterparts in a target population of interest has received more focused attention in recent years. Often, the trial sample is assumed to represent simple random sample of its corresponding population, and not a more representative population of interest, in which case generalizability would be met in expectation (i.e., trivially transportable).(Pearl & Bareinboim, 2014) That is, findings from the study sample are likely to be generalizable to its target population. Unfortunately, the relevance of the trial sample to a target population of interest is often complicated by factors including challenging logistical issues, costs or ethical challenges. This can be problematic when the prevalence of treatment effect modifiers affects selection into the trial. For example, in EAGeR, fertility centers were used for recruitment, which led to a non-representative demographic sample of the target

population with respect to potential modifiers of the aspirin effect (e.g., age, income). These differences may have led to non-exchangeability between women in EAGeR and those who were not included but were eligible to receive treatment. Consequently, the average treatment effect (ATE) obtained from the trial sample would not immediately align with the population average treatment effect (PATE) of interest.

Unlike many studies, the EAGeR trial had a unique enrollment based on two different eligibility criteria representing two different populations. The first consisted of women with a single previous pregnancy loss of less than 20 weeks of gestation within the last 12-months. Then, those criteria were expanded to accommodate women with one or two previous losses of any gestational age at any time in the past. The former criteria can be thought of as a biologically based target population (Mumford & Schisterman, 2019) (i.e., women who would benefit the most from treatment); the second anticipates that, in practice, EAGeR findings would be applicable to a wider range of women. The original analysis reported important differences in the aspirin effect by depending on eligibility group. Specifically, there was an increase in live birth rate of 9.2% (95% CI: 0.5, 17.9) in the biologically based target population. Our results followed a similar pattern, thus supporting aspirin use among this selected group of patients.

This work shows how the parametric g-formula can be implemented in data with time-varying exposures and confounders to generalize an average treatment effect estimate to a target population of interest. Under a set of assumptions including no measured or unmeasured confounding, no selection bias, no information bias, positivity, consistency, and no interference, the parametric g-formula is a consistent estimator of the average treatment effect of a time-varying exposure (Keil, Edwards, Richardson, Naimi, & Cole, 2014; Naimi, Cole, & Kennedy, 2017). In addition, to be a consistent estimator of the population average treatment effect, exchangeability

with respect to the modifiers of the average treatment effect that are associated with selection into the trial must be accounted for.(Lesko et al., 2017) To meet this additional exchangeability assumption, we relied on a set of baseline covariates that were common to both the NSFG and the EAGeR trial. However, as with the unmeasured confounding assumption, there are no guarantees that we appropriately adjusted for all relevant variables. In particular, we were unable to adjust for measures of physical activity and underlying cardiovascular conditions, which were unavailable in the NSFG data. Our results should thus be interpreted in light of this particular limitation.

Additional limitations influence the interpretation of this study. Primarily, the NSFG did not have any information on aspirin consumption practices, or any of the key time-varying confounders from EAGeR (i.e., bleeding and nausea/vomiting). To address this issue, we used the g-formula to generate measures of aspirin consumption, bleeding, and nausea/vomiting using the distribution of baseline data from NSFG, with models fit in the EAGeR data. In addition, as with all implementations of the parametric g-formula, our results rely on the assumption of correct model specification. In the generalizability setting, this requires correctly modeling all relevant interactions between the selected baseline covariates and the exposure. This creates challenges in any scenario, since including numerous interactions in a given model can lead to instability due to insufficient sample sizes. As a result, we had to judge which interactions were the most relevant to our analyses, and choose interactions selectively.

The need to carefully address challenges with generalizability has long been acknowledged. There is a growing body of literature demonstrating the use of formal approaches to generalize results from clinical trials. However, the insights generated from studies accounting for a lack of generalizability should be balanced against the additional challenges imposed by

pursuing such a task. Despite these challenges, our findings add to a growing body of evidence suggesting an important role of daily low-dose aspirin in improving pregnancy outcomes.

3.5 Tables

Table 5. Characteristics of women enrolled in EAGeR trial and women from the National Survey of Family Growth (2015-2017)

Potential Effect Measure Modifiers	NSFG Target Population [a] (*N* = 10,998,642)		EAGeR Full Population [b] (*N* = 1,227)		EAGeR Original Stratum [c] (*N* = 548)	
	Est.	**95% CI**	**Est.**	**95% CI**	**Est.**	**95% CI**
Age at conception, mean	25.9	25.3, 26.5	28.7	28.5, 29.0	28.0	27.8, 28.3
BMI, mean	28.7	27.8, 29.5	26.3	26.0, 26.7	25.8	25.4, 26.1
Race, %						
White	77.2	70.9, 82.0	94.6	93.2, 95.8	96.9	95.1, 98.1
Income, %						
<40K	36.5	32.2, 41.0	33.1	30.5, 35.8	35.0	31.1, 39.1
Education, %						
High school completed	68.7	64.1, 73.0	86.2	84.1, 88.0	89.8	87.0, 92.1
Marital status, %						
Married	60.5	55.4, 65.0	91.5	89.4, 93.0	94.9	92.3, 96.4
Employment, %						

Employed	74.1	68.7, 79.0	74.9	72.4, 77.3	79.8	76.2, 82.9
Alcohol use, %						
Never	75.5	68.8, 81.0	31.1	28.6, 33.8	29.2	25.5, 33.1
Tobacco use, %						
Never	36.7	29.9, 44.0	12.3	10.7, 33.8	9.7	7.4, 12.4

[a] Women with one or two previous pregnancies, the latest of which outcome was identified as miscarriage or stillbirth

[b] One previous pregnancy loss of less than 20 weeks of gestation within the last 12-months

[c] One or two previous losses of any gestational age at any time in the past

Table 6. Generalized ITT and PP effects from EAGeR population to the NSFG 2015-2017.

	Placebo	**Aspirin**	**Risk difference (95% CI)**
Based on the full (both strata) EAGeR population (*N* = 1,227)			
Intention to treat (ITT)			
Live birth	25.1	27.1	2.0 (-3.0, 6.9)
hCG detected pregnancy	45.4	49.4	4.0 (-1.6, 9.7)
Pregnancy loss	13.9	14.8	0.9 (-3.2, 5.0)
Per protocol			
Live birth	24.4	27.1	2.7 (-1.9, 7.3)
hCG detected pregnancy	45.4	49.4	4.1 (-1.1, 9.2)
Pregnancy loss	13.9	14.8	0.9 (-2.9, 4.7)
Based on the original inclusion criteria stratum (*N* = 548)			

Intention to treat (ITT)			
Live birth	35.3	40.3	4.9 (-1.5, 11.3)
hCG detected pregnancy	46.5	52.3	5.7 (-0.3, 11.7)
Pregnancy loss	15.0	16.2	1.1 (-3.1, 5.3)
Per protocol			
Live birth	31.5	43.8	12.3 (6.5, 18.1)
hCG detected pregnancy	59.0	63.7	4.7 (-1.6, 11.3)
Pregnancy loss	7.2	6.9	-0.3 (-4.6, 4.0)

NOTE: Risk difference presented as difference in the outcome per 100 women

3.6 Figures

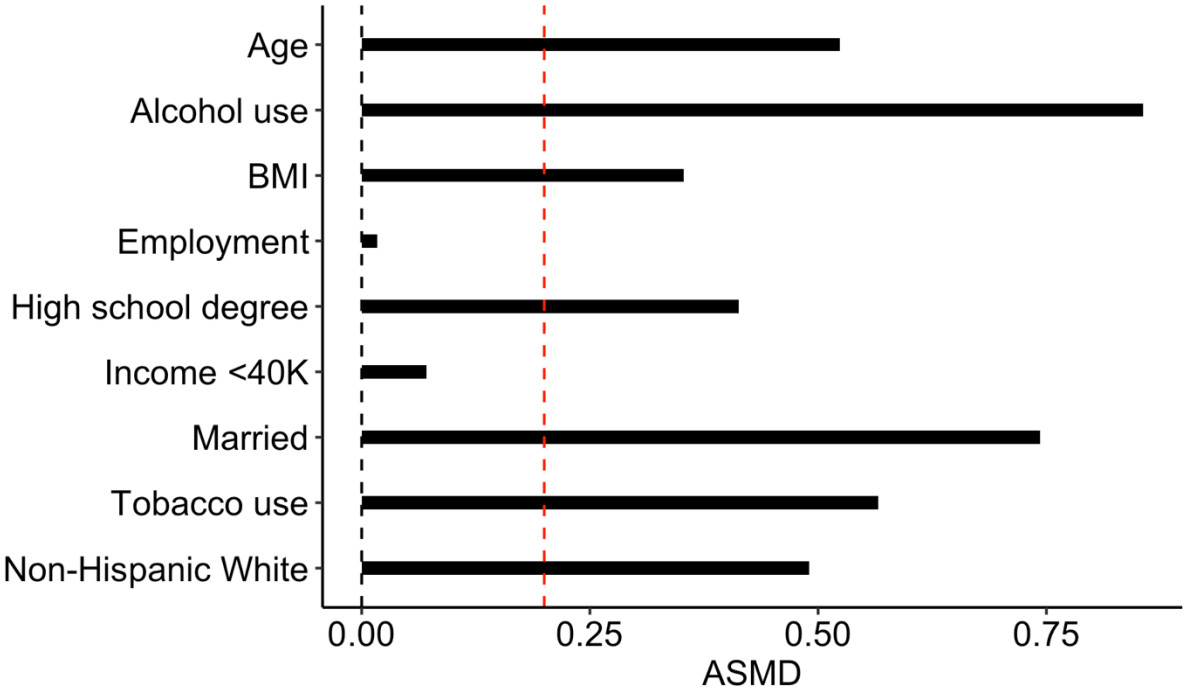


Figure 4. Absolute Standardized Mean Difference (ASMD) of potential treatment effect modifiers between the full EAGeR sample (N=1,227) and the NSFG (2015-2017).

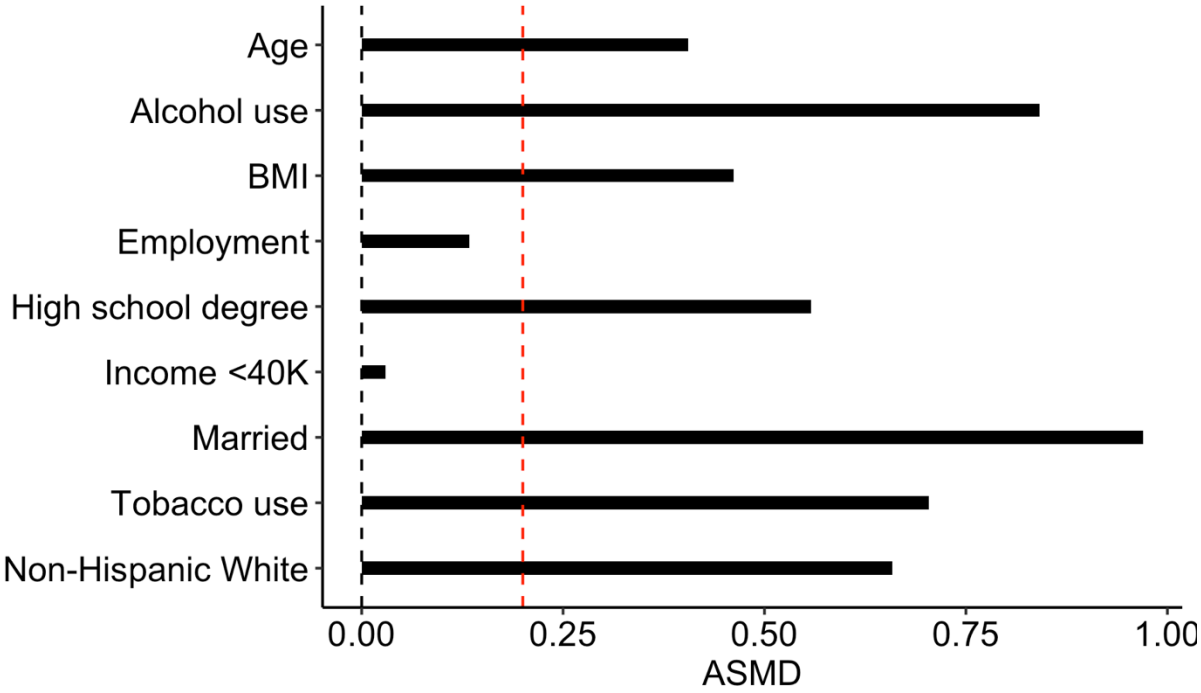


Figure 5. Absolute Standardized Mean Difference (ASMD) of potential treatment effect modifiers between the original EAGeR sample (N=548) and the NSFG (2015-2017).

4.0 Measurement error correction for time-varying exposures using the g-Formula

4.1 Introduction

Measurement error is a common concern in epidemiologic studies. Arguably, the majority of exposures routinely used by epidemiologists are ascertained with some degree of error. Researchers frequently claim that the measurement error present in a given study will only bias estimated effects towards the null.(Bross, 1954; Gullen, Bearman, & Johnson, 1968) However, the conditions needed for an exclusively null-directed bias to occur (i.e. including a truly binary exposure, exactly non-differential measurement error, independence in measurement error between covariates, and the absence of interactions with other systematic errors) (CHAVANCE, DELLATOLAS, & LELLOUCH, 1992; Jurek, Greenland, Maldonado, & Church, 2005; Sorahan & Gilthorpe, 1994) are often difficult to justify in modern epidemiologic analyses. When conventional approaches are used to analyze epidemiologic data, mis-measured exposures will lead to biases with both unpredictable directions and magnitudes, creating the need for some form of measurement error correction.(Jurek et al., 2005)

Appropriate measurement error correction most often requires incorporating information from a gold-standard exposure measurement in a subset of the study sample. The gold standard data from the subsample can then be used to correct the potential bias in the exposure effect estimate using data in the sample. For example, regression calibration (ROSNER et al., 1990; Rosner et al., 2006; Spiegelman, McDermott, & Rosner, 1997) and simulation extrapolation (SIMEX) (Cook & Stefanski, 1994) are two commonly used procedures to address measurement error. However, they are most often applied to scenarios where the exposure is measured once at

a fixed time point. Unfortunately, the extension of these methods to longitudinal settings where exposure is measured multiple times is not often feasible. There is a need to address the gap in methods for measurement error correction of time-varying exposures.

Correcting for measurement error in time-varying settings is complicated by several issues. First, obtaining "gold standard" measurements in a sizable subset of the study sample (i.e., the validation sample) can be challenging, and it is impractical to collect these additional measurements over the entire course of the follow-up, for in the validation subsample. Second, existing analytic procedures used to correct for measurement error are only generally applicable to time-fixed settings.

The parametric g-formula has been widely used to estimate the effect of a time-varying exposure subject to potential time-varying confounding, where the latter may also mediate the effect between the exposure and outcome of interest (Stephen R. Cole, Richardson, Chu, & Naimi, 2013; Keil et al., 2014; Westreich et al., 2012). Here, we build upon existing work by incorporating a measurement error correction step into the g-formula algorithm. Using simulated complex longitudinal data, we evaluate the performance of this measurement error correction step under a scenario where "gold standard" measurements are available at intermittent times during follow-up. We assess the impact of varying sensitivities and specificities of the "gold-standard" measurement tool, as well as varying sample sizes of the internal validation set.

4.2 Methods

4.2.1 Simulated data

4.2.1.1 Complex longitudinal data simulation

To simulate the longitudinal data, we used an approach proposed by Young et.al. (Young, Hernán, Picciotto, & Robins, 2010), which generates time-varying exposures and confounders from a structural nested model. Briefly, we generated 1,000 baseline (T_0) observations from an exponential distribution with constant hazard $\lambda_0 = 0.0375$. Each individual follow-up is then generated until the event occurs or the maximum time-point (T_{12}) is reached. At each time-point, we generated a time-varying confounder $Z(t)$ conditional on previous exposure and confounder values, $X(t-1)$ and $Z(t-1)$, respectively. Whether the outcome occurred at a given time-point ($Y(t)$) was determined by:

$$Y(t) = I\left\{\lambda_0 \times T_0 \leq \sum_{k=1}^{t-1} \exp[\log(\lambda_0) + \log(2.5)\, X(k)]\right\}$$

where $I()$ represents the indicator function which is set to 1 if its argument is true, and zero otherwise; and $\log(2.5)$ represents the true Hazard-Ratio associating the exposure $X(k)$ with the outcome. A directed acyclic graph summarizing the data generating mechanism can be found in Appendix 1.

The net result is a dataset with a time-to-event outcome (arranged in the Andersen-Gill data structure format (Therneau & Grambsch, 2000)), a binary time-varying exposure, and a binary time-varying outcome.

4.2.1.2 Measurement error generation

Using the complex longitudinal data generated above, we generated different scenarios for a mismeasured time-varying exposure [denoted $X'(k)$]. At each time-point, the error prone exposure was generated from a Bernoulli distribution as:

$$X' = \sim Be(\Pr[expit(\log(m) + \log(j)X(k) - t_m)])$$

where $expit(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$; t_m represents the follow-up time-point; and $\log(k)$, $\log(j)$ are user defined values yielding a specified sensitivity and specificity of mismeasured exposure (X') with respect to its true value (X). Additional details can be found in Appendix 2.

4.2.1.3 Validation set with gold standard information

For every Monte Carlo sample of $N = 1{,}000$ individuals, we randomly selected 50 (5%), 100 (10%) and 200 (20%) observations to be included in the validation set ($S = 1$). Then, at any given time k over follow-up, the probability of having gold standard information was given by: $\Pr(GS(k) = 1|S = 1) = 0.5$. This resulted in intermittent availability of gold standard measurements for those in the validation set. We set individuals with $GS(k) = 1$ to have their true exposure value $X(k)$, while the exposure value for those with $GS(k) = 0$ was set to $X'(k)$.

4.2.2 Analysis

4.2.2.1 Modified g-Formula estimator

In a typical setting with a time-to-event outcome $Y(k)$, time-dependent exposure $X(k)$, and time-dependent confounder $Z(k)$, generated from a data generating mechanism such as depicted in Appendix 1, the g formula estimator begins with the following factorization:

$$Y(k) = \sum_{m=0}^{k} \sum_{\bar{z}_k} \sum_{\bar{x}_k} \Bigg\{ P(Y_m = 1 | \bar{Z}_m = \bar{z}_m, \bar{X}_m = \bar{x}_m, \bar{Y}_m = 0, j \leq m) \times \prod_{j=0}^{m} \begin{bmatrix} P\left(Z_j = 1 \middle| \bar{Z}_{j-1} = \bar{z}_{j-1}, \bar{X}_j = \bar{x}_j\right) \times \\ P\left(X_j = 1 \middle| \bar{Z}_j = \bar{z}_j, \bar{X}_{j-1} = \bar{x}_{j-1}\right) \end{bmatrix} \Bigg\},$$

Under key identifiability assumptions (including exchangeability, counterfactual consistency, positivity, and no interference), setting the exposure history $\bar{X}_m$ to some specified value $\bar{d}_m$ can yield an estimate of the outcome that would be observed if (possibly contrary to the fact) X were set to d for all individuals in the sample (denoted $Y^{d(k)}$). To account for the fact that the gold standard measurement $X(k)$ is only available on a subset of the sample, we incorporate a model for the relationship between $X'(k)$ and $X(k)$ into the g-formula estimator above.

4.2.2.2 Implementation

To begin the modified g-formula procedure, using the data in the validation subset, we fit the following models:

(1) $\text{logit}(X_k) = \beta_o + \beta_1 X'_k + \beta_2 time$

(2) $\text{logit}(Z_k) = \beta_o + \beta_1 X_{k-1} + \beta_2 Z_{k-1} + \beta_3 time$

(3) $\text{logit}(X_k) = \beta_o + \beta_1 Z_k + \beta_2 X_{k-1} + \beta_3 Z_{k-1} + \beta_4 time$

(4) $\text{logit}(Y_k) = \beta_o + \beta_1 X_k + \beta_2 Z_k + \beta_3 X_{k-1} + \beta_4 Z_{k-1} + \beta_5 time$

We then take a Monte Carlo resample (with replacement) of the first time-point for all observations in the original data. For the first selected observation from the Monte Carlo dataset (i.e., ID = 1 at time-point $k = 1$) observation in the Monte Carlo resample, we carry out the following procedure:

1. If the observation contains a gold standard exposure measurement at $k = 1$, proceed to step 2. Otherwise, use model 1 to simulate the gold standard measure $X(k = 1)$.
2. With the true (if available) or simulated gold standard measure and other relevant variables, use model 2 to simulate $Z(k = 2)$.
3. With the true (if available) or simulated gold standard measure at $k = 1$, the simulated $Z(k = 2)$, and the measured $Z(k = 1)$, use model 3 to simulate $X(k = 2)$.
4. With the true (if available) or simulated gold standard measure at $k = 1$, the simulated $Z(k = 2)$, and the measured $Z(k = 1)$, and the simulated $X(k = 2)$, use model 4 to simulate $Y(k = 2)$.

If model 4 returns a value of $Y(k = 2) = 1$, we terminate the procedure for ID = 1, and proceed with the same procedure for ID = 2. Otherwise, we return to step 2 of the procedure and continue simulating X, Z, and Y for each subsequent time point until either $Y = 1$, or the pre-determined administrative censoring time is reached.

Because models 1 to 4 require the gold-standard measurements to be used, these models are only fit in the subsample of observations and person-time entries where X (i.e., gold standard assessment of the exposure status) is available. In the initial run of the above algorithm, we did not set individuals to any specific exposure value. However, to estimate the average treatment effect, we set all X values to their relevant levels.

The performance of our measurement error correction procedure was evaluated using a total of 27 different scenarios, which include the combination of the following elements: (1) sensitivity of 60%, 70% and 80% of the mismeasured exposure with respect to gold standard; (2) specificity of 60%, 70% and 80%; and (3) validation set size of 5%, 10% and 20%. All these scenarios were assessed using 500 Monte Carlo samples of $N = 1{,}000$. Additionally, we evaluated

an alternative scenario where the mismeasured exposure (at the above-mentioned combinations) was used in the g-formula. For each Monte Carlo dataset, we computed bias, mean squared error, and 95% confidence interval coverage.

4.3 Results

The results obtained by applying the initial measurement error correction step to the simulated data are presented in Tables 7-9. As expected, performance improved as specificity and sensitivity increased. The magnitude of the improvement was larger with increases in specificity as compared to increases in sensitivity in all of the scenarios. Post-correction specificity and sensitivity did not depend on the validation size.

As expected, results from the g-formula algorithm using mismeasured exposures were subject to a substantial degree of bias (Table 10). Overall, a larger degree of misclassification in the exposure resulted in a larger degree of bias. For example, when mismeasured exposure had 60% sensitivity and specificity with respect to the true exposure, we observed bias of 1.53 (SE = 0.21). Conversely, when mismeasured exposure sensitivity and specificity were set to 80%, bias was 1.01 (SE = 0.19). These results are displayed schematically in Appendix 3.

The performance of the g formula after implementing our correction approach is presented in Table 3. We observed that a larger validation set size yielding lower bias. These results are illustrated in Figures 2 to 4, which display the bias distribution from each of our simulation specifications. In particular, results from a fixed specificity at 60% (Figure 2), 70% (Figure 3) and 80% (Figure 4) are combined with sensitivity values of 60%, 70% and 80%, as well as validation set sizes of 5%, 10% and 20%. These results demonstrate that lower degrees of initial exposure

misclassification result in lower bias. For instance, initial sensitivity and specificity values of 80% had a bias of -0.29, -0.15 and -0.03 for a validation set size of 5%, 10% and 20%, respectively. In contrast initial sensitivity and specificity values of 60% had a bias of -0.49, -0.48 and- 0.16 for validation set sizes of 5%, 10% and 20%, respectively.

In addition to differences in bias, we observed less variation in the estimated exposure effect as the validation set size increased (Figure 2-4). Moreover, confidence interval coverage was closer to nominal with larger validation sizes, with values ranging from 59% to 95%, as shown in Table 11

4.4 Discussion

In this study, we evaluated the performance of a measurement error correction tool for complex longitudinal data within the framework of the parametric g-formula. Overall, our method shows promising results in reducing bias from mismeasured time-varying exposures.

Measurement error correction in complex longitudinal settings poses additional challenges compared to time-fixed scenarios. Indeed, attaining gold standard assessments at multiple timepoints in a subset of the study population is subject to several difficulties, including elevated costs, logistical challenges, and ethical dilemmas associated with invasive procedures. In most applied settings, researchers will be limited to a small validation set with intermittent gold standard assessments.

In addition to challenges concerning study design, some of the methods developed to correct for measurement error are only applicable under in restricted situations such as specific regression framework (e.g., Cox regression model) (Liao, Zucker, Li, & Spiegelman, 2011;

Spiegelman et al., 1997). This is problematic in complex longitudinal settings (i.e., time-varying confounders that are intermediates between exposure and outcome), where traditional regression models can lead to biased estimates of the exposure effects (Moodie & Stephens, 2010; Robins, Hernán, & Brumback, 2000). While more general methods developed using joint modeling approaches are available, such methods tend to be computationally intense for most practical applications (Tsiatis, Degruttola, & Wulfsohn, 1995).

Recently, Kyle et.al. (Kyle, Moodie, Klein, & Abrahamowicz, 2016) evaluated an approach based on Simulation Extrapolation (SIMEX) to correct for measurement error in complex longitudinal settings using marginal structural models (MSM). The authors demonstrate a reduction in bias and better coverage of SIMEX based procedures compared to naïve approaches (i.e., where mismeasured exposure is used). However, SIMEX itself can be computationally burdensome, which may explain why Kyle et. al.(Kyle et al., 2016) explored longitudinal data with only two timepoints. In our study, we examined measurement error correction in a dataset with up to 12 timepoints.

Our approach to correct for measurement error in complex longitudinal settings is based on the parametric g-formula estimator, which is also one of Robins' generalized methods (g-methods) for estimation of causal effects under less restrictive set of assumptions than traditional regression methods.(Naimi et al., 2017) Recently, the g-formula has become more widely used in applied epidemiological literature.(Stephen R. Cole et al., 2013; Keil et al., 2014; Taubman, Robins, Mittleman, & Hernán, 2009) However, little work has been done to address bias arising from mismeasured time-varying exposures when applying these methods. Here, we demonstrate the utility of a measurement error correction approach implemented within the g-formula

algorithm. This approach can be implemented in situations where the investigator is able to collect gold standard information for a subset of the entire sample at intermittent points over follow-up.

The potential benefits from adopting this approach should be weighed against its current limitations. First, we only considered a scenario where the investigator is able to randomly assign observations to the validation subset; and to obtain gold standard assessments of the exposure status. Currently, our approach depends heavily on the availability of this this type of validation subset, as it is used to fit the models for the g-formula algorithm. Second, all of our simulations are restricted to one mismeasured exposure, which will be uncommon in practice. Indeed, most studies addressing measurement error will need to consider the case of multiple, possibly dependent, error prone covariates. Third, for simplicity, we assumed that the surrogate (i.e., mismeasured) exposure was completely observed over the course of follow-up. Furthermore, we assumed that there were no losses of follow-up during the study. The impact of violating such assumptions, while essential, is out of the scope of this study. Fourth, as with other applications of the parametric g-formula, we must acknowledge that it is particularly sensitive to model misspecification. This will be an important limitation to consider in most practical applications dealing with multiple covariates and interactions between them.

Notwithstanding its limitations, our approach shows promising results to correct bias from mismeasured time-varying exposures within the framework of the parametric g-formula. This study adds to the growing body of literature addressing measurement error correction and provides investigators with a viable alternative to address this problem under certain conditions. Future research should focus on expanding the scope of this work to incorporate some of the more common situations found in modern epidemiologic studies.

4.5 Tables

Table 7. Specificity and sensitivity of mismeasured time-varying exposure, after initial measurement error correction using a validation set size of 5%

Initial Specificity (%)	Initial Sensitivity (%)	Post correction specificity (SE)	Post correction sensitivity (SE)
60	**60**	0.82 (0.03)	0.74 (0.04)
	70	0.84 (0.03)	0.75 (0.04)
	80	0.86 (0.03)	0.78 (0.04)
70	**60**	0.84 (0.03)	0.75 (0.04)
	70	0.85 (0.02)	0.78 (0.04)
	80	0.88 (0.02)	0.82 (0.04)
80	**60**	0.86 (0.03)	0.78 (0.04)
	70	0.88 (0.02)	0.81 (0.04)
	80	0.90 (0.02)	0.84 (0.03)

NOTE: Sensitivity and Specificity with respect to the true exposure. Results from 500 Monte Carlo simulations of N=1,000.

Table 8. Specificity and sensitivity of mismeasured time-varying exposure, after initial measurement error correction using a validation set size of 10%

Initial Specificity (%)	Initial Sensitivity (%)	Post correction specificity (SE)	Post correction sensitivity (SE)
60	**60**	0.83 (0.02)	0.72 (0.03)
	70	0.85 (0.02)	0.74 (0.03)
	80	0.87 (0.02)	0.77 (0.03)
70	**60**	0.84 (0.02)	0.73 (0.04)
	70	0.86 (0.02)	0.77 (0.03)
	80	0.89 (0.02)	0.80 (0.03)
80	**60**	0.86 (0.02)	0.77 (0.04)
	70	0.88 (0.02)	0.80 (0.03)
	80	0.90 (0.02)	0.84 (0.03)

NOTE: Sensitivity and Specificity with respect to the true exposure. Results from 500 Monte Carlo simulations of N=1,000.

Table 9. Specificity and sensitivity of mismeasured time-varying exposure, after initial measurement error correction using a validation set size of 20%

Initial Specificity (%)	Initial Sensitivity (%)	Post correction specificity (SE)	Post correction sensitivity (SE)
60	**60**	0.83 (0.01)	0.72 (0.02)
	70	0.85 (0.01)	0.74 (0.02)
	80	0.88 (0.01)	0.78 (0.02)
70	**60**	0.85 (0.01)	0.74 (0.02)
	70	0.87 (0.01)	0.77 (0.02)
	80	0.89 (0.01)	0.81 (0.02)
80	**60**	0.86 (0.01)	0.77 (0.02)
	70	0.88 (0.01)	0.81 (0.02)
	80	0.90 (0.01)	0.84 (0.02)

NOTE: Sensitivity and Specificity with respect to the true exposure. Results from 500 Monte Carlo simulations of N=1,000.

Table 10. Performance of g-Formula using mismeasured exposure

Specificity		Sensitivity		
		60%	**70%**	**80%**
60	**Bias (SE)**	1.53 (0.24)	1.30 (0.20)	1.26 (0.17)
	RMSE	2.40	1.73	1.63
	Coverage	0	0	0
70	**Bias (SE)**	1.50 (0.21)	1.22 (0.18)	1.23 (0.18)
	RMSE	2.30	1.53	1.55
	Coverage	0	0	0
80	**Bias (SE)**	1.27 (0.20)	1.11 (0.17)	1.01 (0.19)
	RMSE	1.66	1.26	1.06
	Coverage	0	0	0

NOTE: Sensitivity and specificity with respect to the true exposure

Table 11. Measurement error correction of partially observed time-varying exposures using g-Formula

Error prone exposure specification		Feature	g-Formula correction Validation set size (%)		
Sensitivity (%)	**Specificity (%)**		**5**	**10**	**20**
60		**Bias (SE)**	-0.49 (0.46)	-0.48 (0.31)	-0.16 (0.22)
	60	**RMSE**	0.45	0.33	0.07
		Coverage	0.81	0.65	0.87
		Bias (SE)	-0.64 (0.43)	-0.23 (0.32)	-0.18 (0.22)
	70	**RMSE**	0.59	0.16	0.08
		Coverage	0.66	0.90	0.89
		Bias (SE)	-0.69 (0.41)	-0.35 (0.30)	0.06 (0.22)
	80	**RMSE**	0.65	0.21	0.05
		Coverage	0.59	0.78	0.94
70		**Bias (SE)**	-0.41 (0.42)	-0.35 (0.31)	-0.08 (0.24)
	60	**RMSE**	0.34	0.22	0.06
		Coverage	0.82	0.79	0.94
		Bias (SE)	-0.30 (0.45)	-0.14 (0.34)	-0.06 (0.22)
	70	**RMSE**	0.30	0.13	0.06
		Coverage	0.90	0.94	0.94
		Bias (SE)	-0.13 (0.43)	-0.40 (0.32)	0.08 (0.20)
	80	**RMSE**	0.20	0.26	0.05
		Coverage	0.94	0.74	0.92

		Bias (SE)	-0.24 (0.47)	-0.05 (0.35)	0.01 (0.23)
	60	**RMSE**	0.28	0.13	0.05
		Coverage	0.90	0.96	0.97
		Bias (SE)	-0.29 (0.41)	0.02 (0.34)	-0.02 (0.22)
80	**70**	**RMSE**	0.25	0.12	0.05
		Coverage	0.90	0.95	0.95
		Bias (SE)	-0.29 (0.40)	-0.15 (0.35)	0.03 (0.21)
	80	**RMSE**	0.39	0.08	0.12
		Coverage	0.89	0.93	0.94

NOTE: Bias with respect to the true log(Hazard Ratio)

4.6 Figures

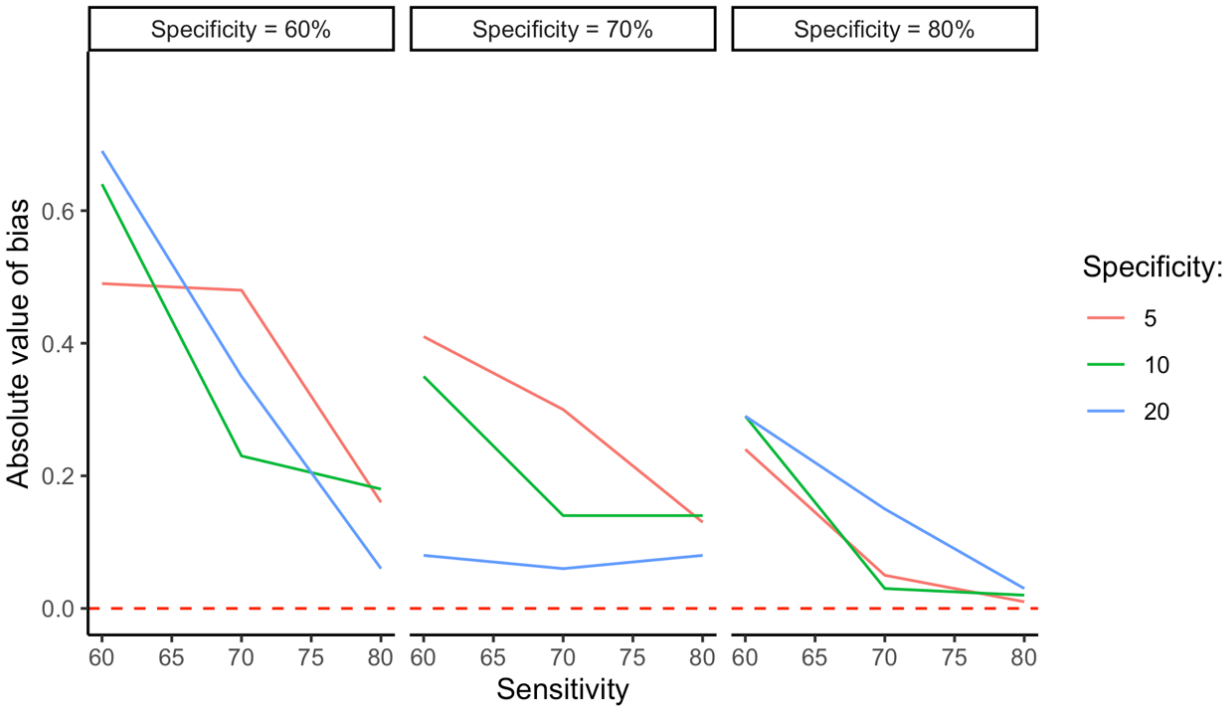


Figure 6. Bias from measurement error using the g-Formula

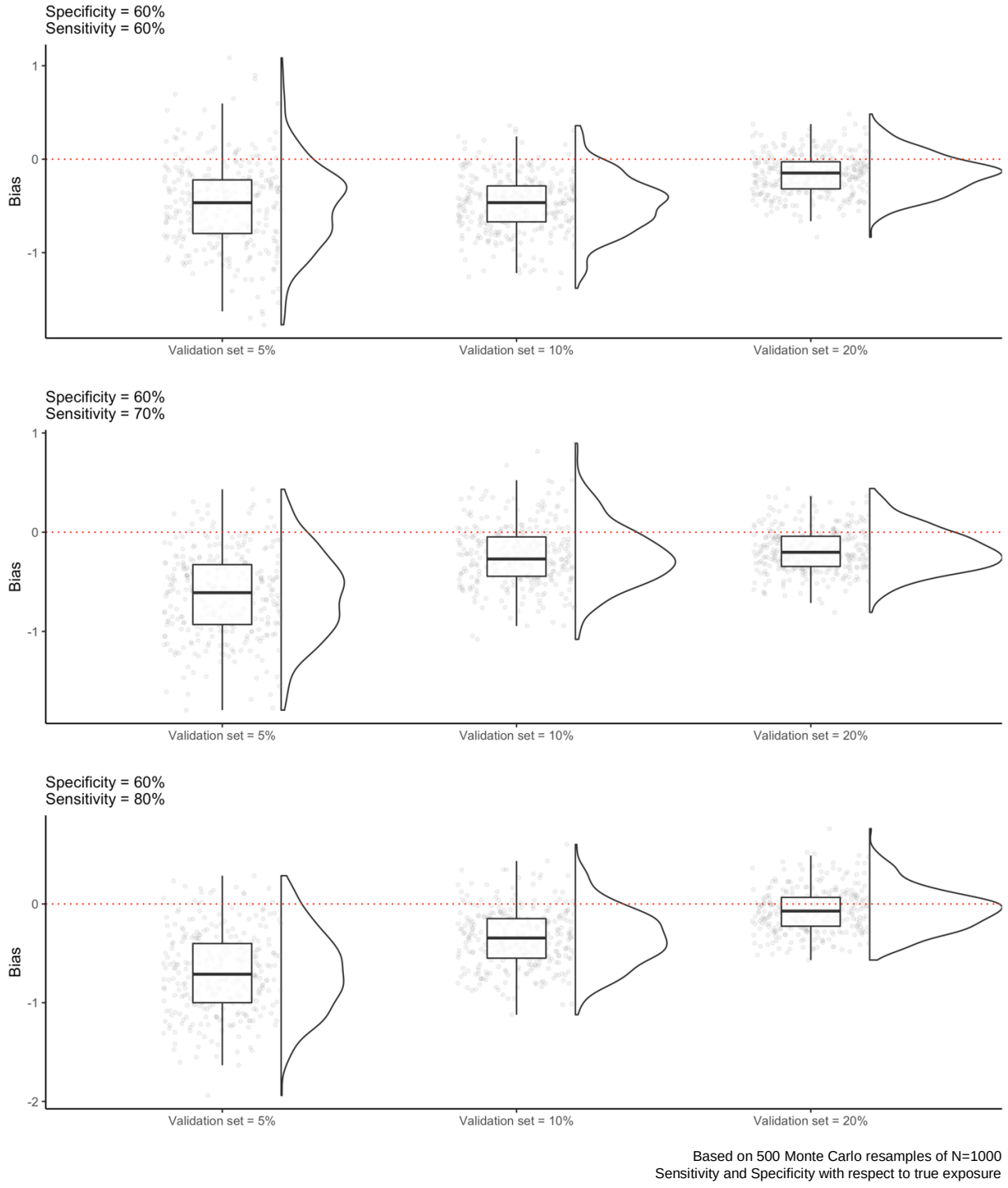


Figure 7. Distribution of bias from measurement error corretion using g-Formula, fixed specificity at 60%

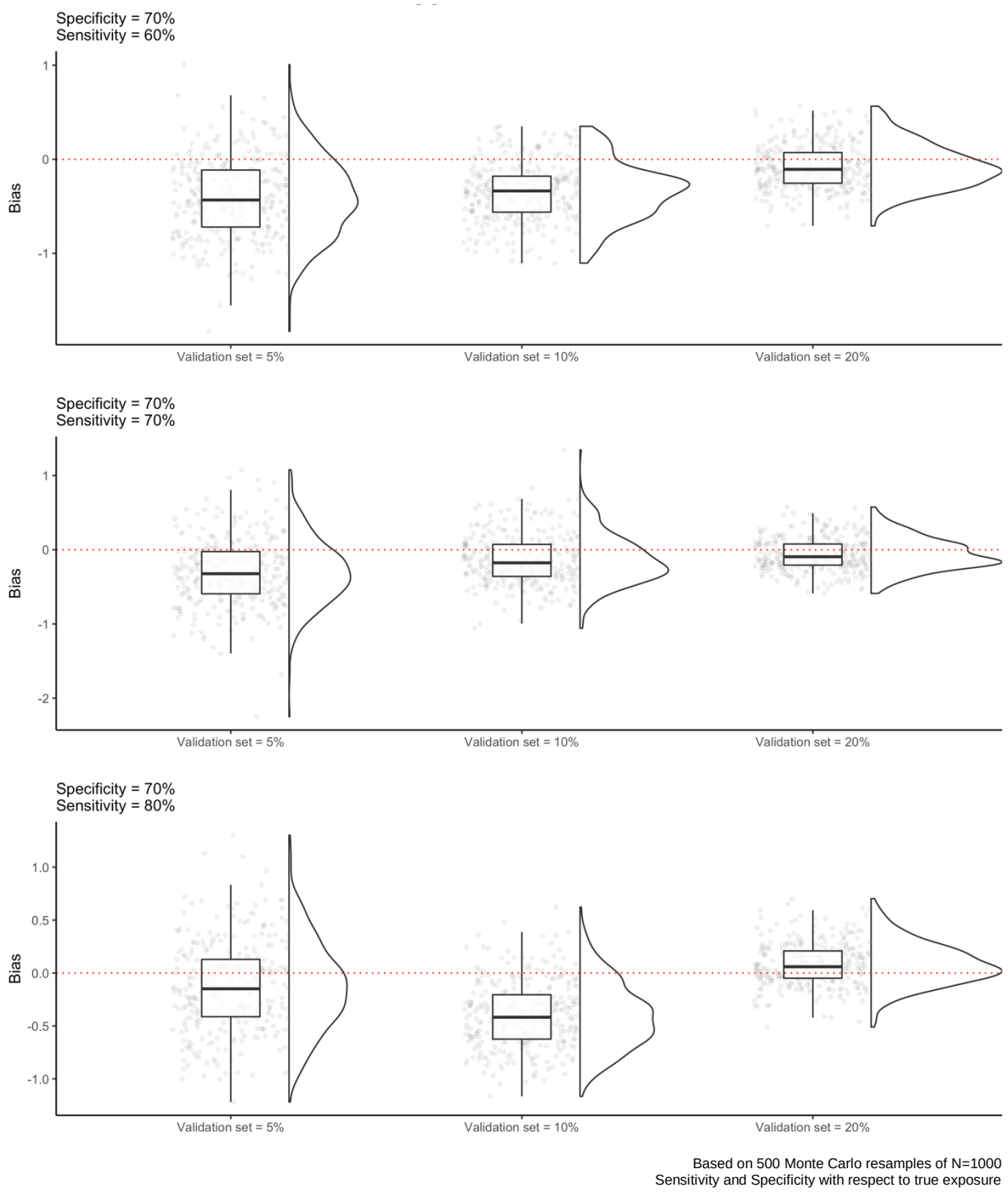


Figure 8. Distribution of bias from measurement error corretion using g-Formula, fixed specificity at 70%

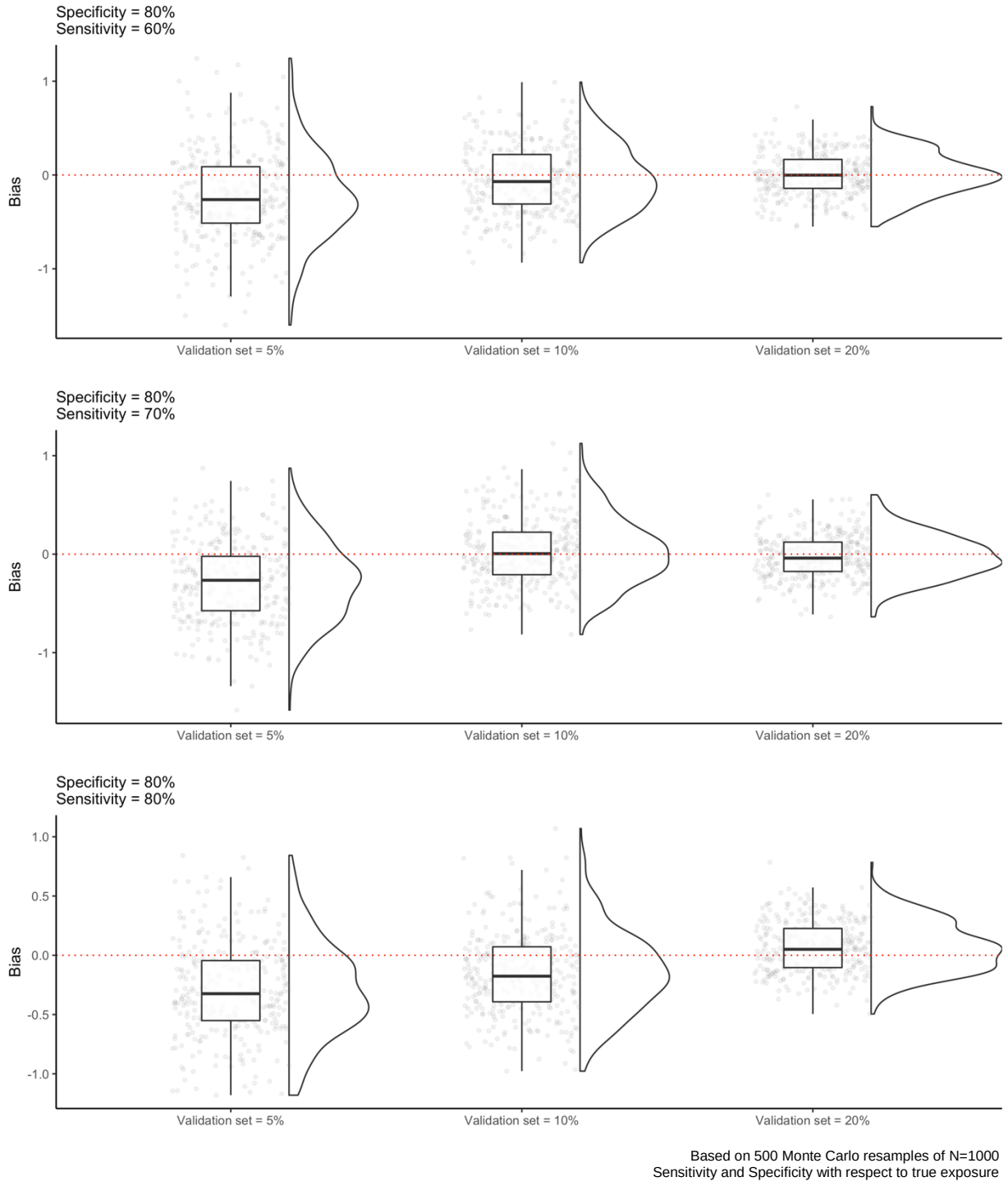


Figure 9. Distribution of bias from measurement error corretion using g-Formula, fixed specificity at 80%

5.0 Conclusions

This dissertation has focused on evaluating and developing epidemiologic methods that help to refine our understanding of the extent to which preconception low-dose aspirin may be used to prevent recurrent pregnancy loss. After highlighting the scarcity of studies addressing this question, we devoted our attention to the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial. Specifically, we build upon current limitations of the study that prevents clinicians and policy makers from fully utilizing these data to inform their decision-making process regarding the use preconception low-dose aspirin on pregnancy loss.

In Section 2, we presented doubly robust (DR) machine learning based estimators as a viable alternative to evaluate effect modification. While it is generally accepted that treatment effect tends to vary according to individuals' sociodemographic and clinical characteristics, the majority of studies in epidemiology are underpowered to detect such heterogeneity. Furthermore, most studies evaluating effect modification rely on parametric methods, and its corresponding strong assumptions. Here, we demonstrated how DR estimators perform relatively well compared to correctly specified parametric models. In doing so, we provided applied researchers with reliable alternatives to evaluate effect modification that does not rely on strong parametric assumptions and are preferable in situations of limited sample size.

Another limitation of the EAGeR trial, as with most clinical trials, is related to external validity. In Section 3, we highlighted differences between EAGeR and a more representative sample of childbearing age women with a previous pregnancy loss living in the U.S. (e.g., National Survey of Family Growth), especially with respect to key potential treatment effect modifiers. We discussed how these differences in distribution can lead to effect estimates in the trial sample that

deviate from what would be expected in an otherwise random sample of the target population. Furthermore, we demonstrated how to adapt the parametric g-formula to obtain generalized estimates of the intention-to-treat (ITT) and per-protocol (PP) analysis of the EAGeR trial to a potential target population of interest.

Finally, in Section 4, we discussed major logistical and methodological challenges related to measurement error in the context of complex longitudinal data. To address these concerns, we proposed and evaluated a method to correct for measurement error of a time-varying exposure in these settings. Our approach, based on the parametric g-formula, shows promising results to correct bias from mismeasured time-varying exposures. It also provides investigators with a feasible alternative to address this problem under certain conditions. Furthermore, it has great potential to be adopted by many investigators, as the number of analyses utilizing the g-formula becomes widely available in epidemiology.

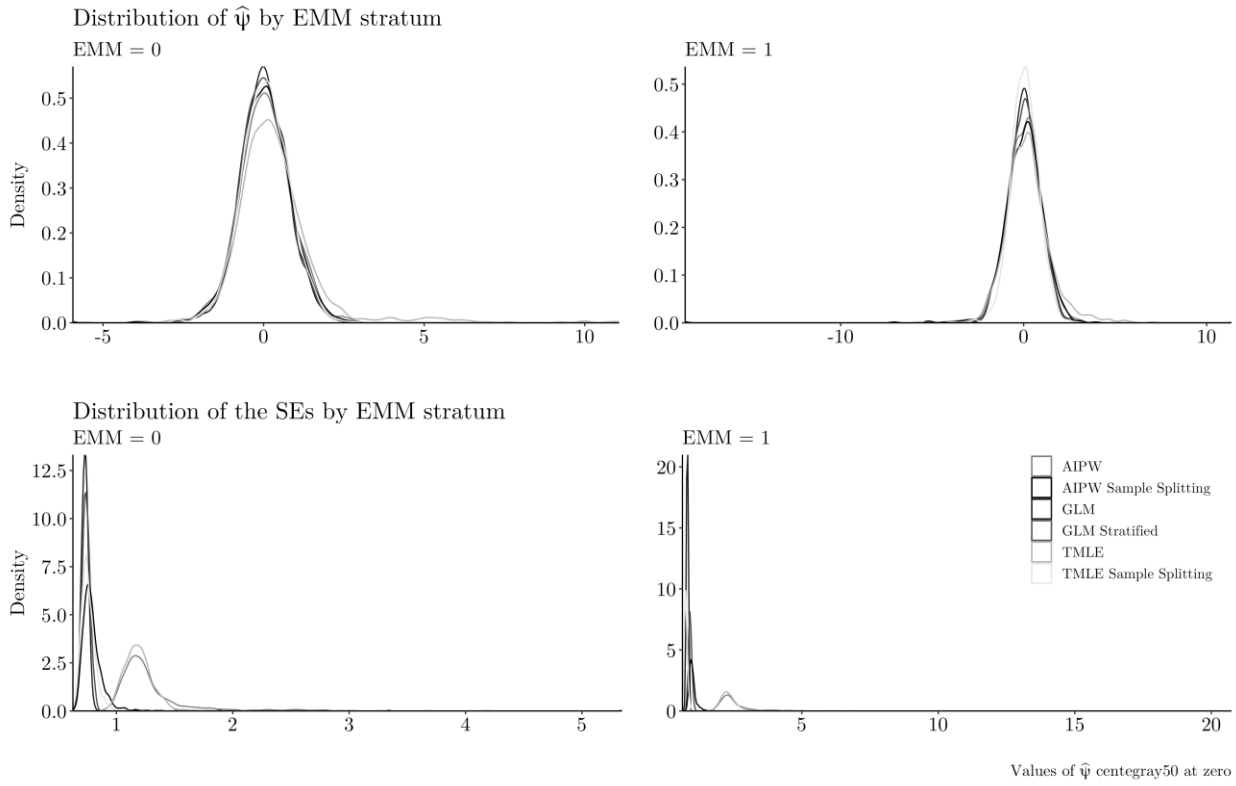
In the future, this dissertation can set the basis for new research. One promising opportunity relies on doubly robust estimation of heterogenous causal effects. Specifically, as mentioned in Section 2, developing error bounds for nonparametric DR estimators is a key question to address in the upcoming future. Equally essential questions remain on generalizability of clinical trials' findings, some of which are related to unobserved treatment effect modifiers in the external target population. In Section 3, we commented on this limitation when discussing the lack of information on aspirin consumption in the National Survey of Family Growth. This is expected to be a major limitation for numerous studies concerning generalizability, especially those relying on national surveys as their source of external population. Lastly, in Section 4, we emphasized the need to increase the extent of our measurement error correction tool to incorporate some of the more

common situations found in modern epidemiologic studies, including multiple mismeasured exposures, partially observed information on error prone covariates, and loss of follow-up.

In sum, this dissertation will add to the growing body of evidence suggesting an important role of daily low-dose aspirin in improving pregnancy outcomes. Furthermore, the methods evaluated and developed throughout this work will help other applied scientists, even outside reproductive and perinatal epidemiology, to tackle common limitations that permeate most clinical trials. Ultimately, we hope that this work contributes to advance epidemiology and generates a positive impact in public health and society.

Appendix A Performance Evaluation of Parametric and Nonparametric Methods when Assessing Effect Measure Modification

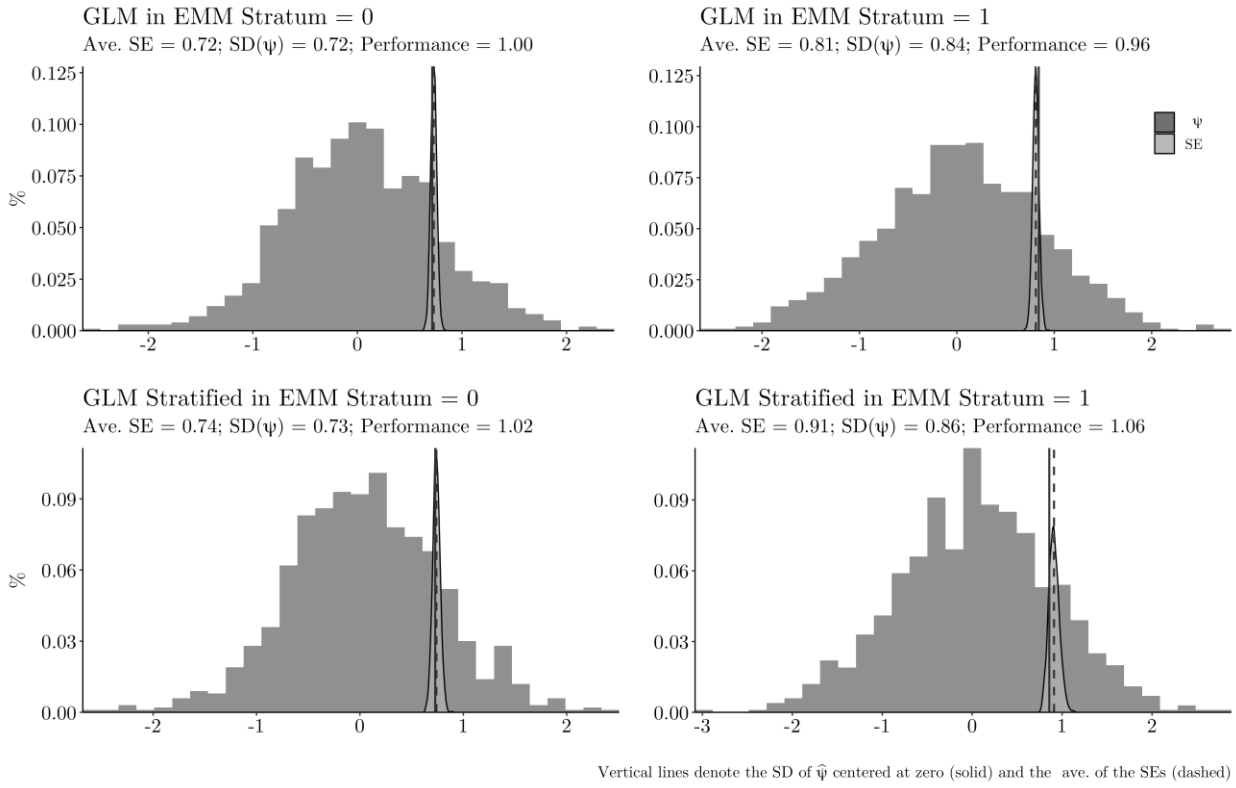
Appendix A.1 Accuracy of Different Estimators in Effect Measure Modification



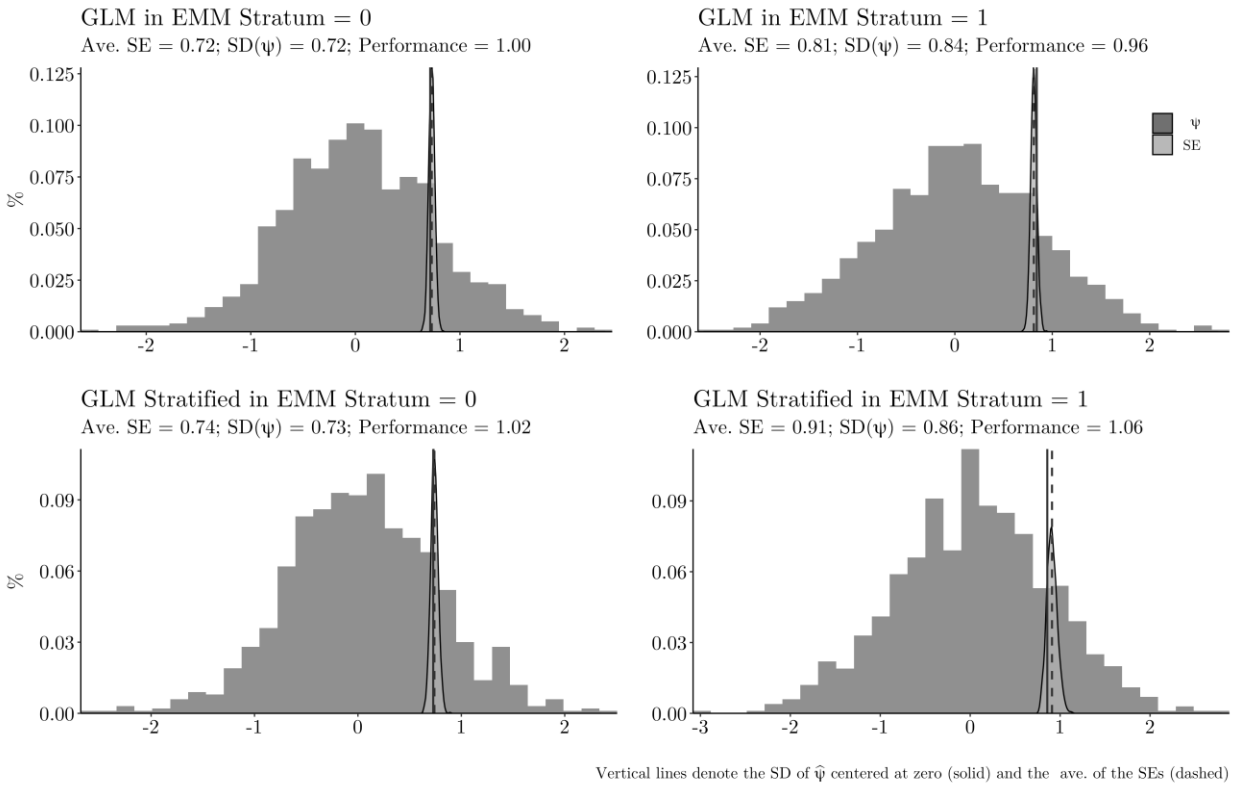
Appendix Figure 1. Distribution of Estimates and Standard Errors in 1,000 Monte Carlo Simulations (N=500)

Visual representation of the accuracy achieved by different estimators, namely: Generalized Linear Model (GLM) with interaction term; GLM stratified; Augmented Inverse Probability Weighting (AIPW); AIPW with Sample Splitting; Targeted Minimum Loss-Based Estimation (TMLE); and TMLE with Sample Splitting. Distribution of the estimates (ψ) is shown in the dark gray histogram, with its corresponding $SD(\psi)$ depicted as solid line. Similarly, the

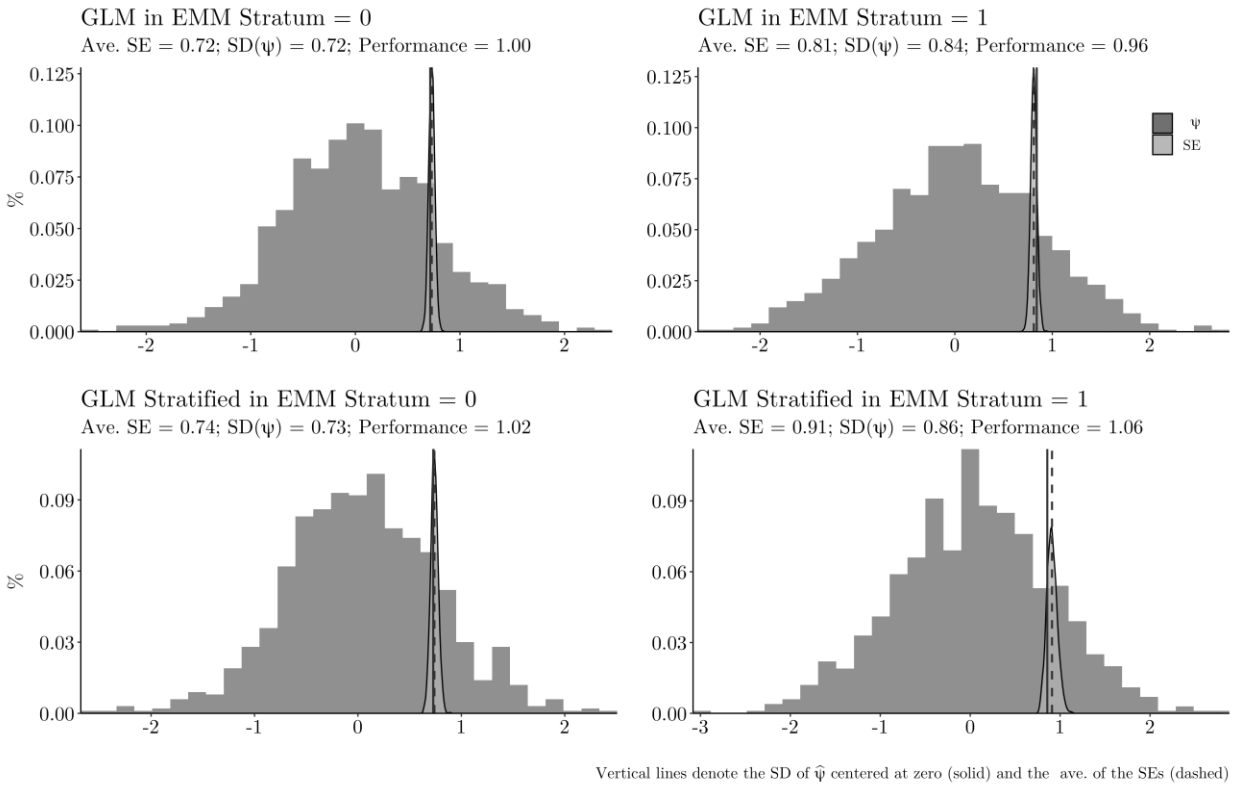
distribution of the SEs and its average is shown as light gray histogram and dashed line, respectively.



Appendix Figure 2. GLM with Interaction Term and Stratified GLM

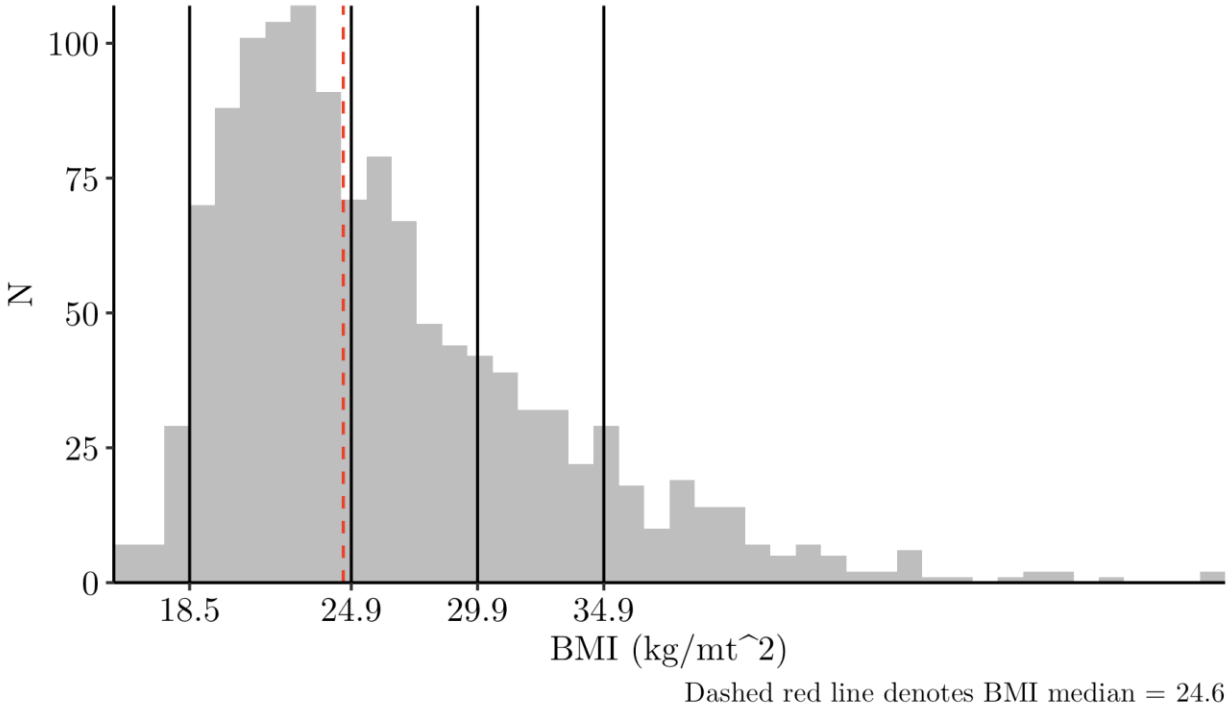


Appendix Figure 3. AIPW and TMLE without Sample Splitting



Appendix Figure 4. AIPW and TMLE with Sample Splitting

Appendix A.2 Pre-pregnancy Body Mass Index (BMI) in the EAGeR trial.



Appendix Figure 5. Distribution of Pre-pregnancy BMI in 1,228 women in the EAGeR Trial

Appendix A.3 Performance of different estimators under the scenario of no effect modification

Appendix Table 1. Performance of Different Estimators Under no Effect Modification (Stratum Zero of the Effect Measure Modifier)

Feature	Interaction GLM	Stratified GLM	AIPW	AIPW SS	TMLE	TMLE SS

Simulation specifications: N = 200; SD = 3						
Mean bias	0.05	0.06	0.10	0.07	0.27	0.07
Mean SE	0.36	0.35	0.49	0.51	0.72	0.39
Coverage	0.95	0.94	1.00	0.96	0.95	0.95
Accuracy	1.00	0.98	1.54	1.03	0.97	1.01
Simulation specifications: N = 200; SD = 6						
Mean bias	0.10	0.11	0.15	0.13	0.39	0.14
Mean SE	1.45	1.36	1.73	1.89	2.06	1.49
Coverage	0.95	0.94	1.00	0.96	0.97	0.95
Accuracy	1.00	0.99	1.51	1.04	1.16	1.00
Simulation specifications: N = 500; SD = 3						
Mean bias	0.02	0.02	0.04	0.03	0.21	0.03
Mean SE	0.13	0.13	0.19	0.16	0.42	0.14
Coverage	0.95	0.95	1.00	0.96	0.97	0.95
Accuracy	1.06	1.02	1.62	1.04	0.79	1.04
Simulation specifications: N = 500; SD = 6						
Mean bias	0.03	0.04	0.08	0.06	0.37	0.06
Mean SE	0.53	0.49	0.66	0.59	1.34	0.55
Coverage	0.95	0.95	1.00	0.95	0.94	0.95
Accuracy	1.04	0.99	1.62	1.04	0.88	1.04

NOTE: Results from 1,000 Monte Carlo simulations

Augmented Inverse Probability Weighting (AIPW); Targeted Minimum Loss-based Estimation (TMLE); sample splitting (SS).

Accuracy = Average of SE ($\widehat{\psi}_i$) / SD ($\widehat{\psi}$)

Appendix Table 2. Performance of Different Estimators Under no Effect Modification (Stratum One of the Effect Measure Modifier)

Feature	Interaction GLM	Stratified GLM	AIPW	AIPW SS	TMLE	TMLE SS
Simulation specifications: N = 200; SD = 3						
Mean bias	0.01	0.02	0.05	0.02	0.20	0.03
Mean SE	0.53	0.50	0.75	1.10	0.90	0.59
Coverage	0.95	0.93	1.00	0.96	0.99	0.94
Accuracy	0.92	1.00	2.39	0.96	1.73	1.01
Simulation specifications: N = 200; SD = 6						
Mean bias	0.02	0.02	0.10	0.01	0.36	0.05
Mean SE	2.04	2.11	2.50	4.47	3.73	2.33
Coverage	0.93	0.95	1.00	0.97	0.99	0.94
Accuracy	0.90	1.00	2.38	1.02	1.87	1.00
Simulation specifications: N = 500; SD = 3						
Mean bias	0.01	0.02	0.03	0.01	0.15	0.02
Mean SE	0.18	0.18	0.31	0.25	0.46	0.20
Coverage	0.97	0.94	1.00	0.97	0.99	0.96
Accuracy	0.98	1.06	2.56	1.07	1.63	1.07
Simulation specifications: N = 500; SD = 6						
Mean bias	0.03	0.03	0.08	0.03	0.24	0.04

Mean SE	0.78	0.74	0.99	0.92	0.96	0.78
Coverage	0.94	0.97	1.00	0.96	0.96	0.99
Accuracy	0.96	1.06	2.48	1.10	1.72	1.05

NOTE: Results from 1,000 Monte Carlo simulations

Augmented Inverse Probability Weighting (AIPW); Targeted Minimum Loss-based Estimation (TMLE); sample splitting (SS).

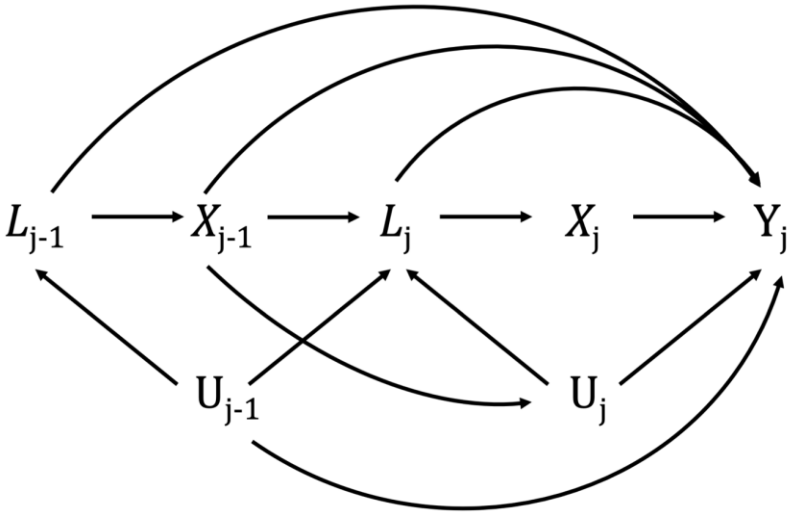
Accuracy = Average of SE ($\hat{\psi}_i$) / SD ($\hat{\psi}$)

Appendix Table 3. Type I Error Rate of Several Estimators Under no Effect Modification

Estimator	**N = 200**		**N = 500**	
	SD = 3	**SD = 6**	**SD = 3**	**SD = 6**
GLM	0.05	0.06	0.04	0.04
AIPW	0.01	0.03	0.01	0.03
AIPW SS	0.01	0.03	0.01	0.03
TMLE	0.07	0.06	0.07	0.07
TMLE SS	0.01	0.04	0.01	0.03

Appendix B Generalizing Evidence from the Effects of Aspirin on Gestation and Reproduction trial

Appendix B.1 Details on the parametric g-formula.



Appendix Figure 6. Directed Acyclic Graph (DAG) for the causal relationships between adherence to aspirin, time-varying confounders and the outcomes of interest.

In the above DAG, we represent a simplified version of the assumed causal relationships between our time-varying confounders $\boldsymbol{L}$ (i.e., bleeding, nausea/vomiting), adherence to aspirin $\boldsymbol{X}$, and the outcome of interest $\boldsymbol{Y}$ indexed at two arbitrary timepoints (j and $j-1$). Unknown or unmeasured factors are depicted as $\boldsymbol{U}$.

Appendix Table 4. Individual Specifications of each Logistic Regression Model Used to Generalize EAGeR findings using the g-Formula.

Dependent variable	Independent variables	Restrictions	Specification
Live birth	Randomization indicator (aspirin vs. placebo) Baseline covariates (high school education, marital status, employment, Non-Hispanic white race, alcohol use, smoking use, age, BMI) Time-varying confounders (adherence at week j and j – 1, bleeding at week j and j– 1, and nausea and/or vomiting at week j and j – 1)	At every week on study, fit among the sample of women who experienced an hCG pregnancy, did not experience pregnancy loss, and did not withdraw.	Age and BMI modeled using B-splines with three degrees of freedom. Week on study modeled using quadratic term. Interaction term between adherence and randomization indicator, and baseline covariates (except employment). Interaction term between time-varying covariates and alcohol, smoking use

	Week on study		
Pregnancy loss	Randomization indicator (aspirin vs. placebo) Baseline covariates (high school education, marital status, employment, Non-Hispanic white race, alcohol use, smoking use, age, BMI) Time-varying confounders (adherence at week j and j – 1, bleeding at week j and j– 1, and nausea and/or vomiting at week j and j – 1) Week on study	At every week on study, fit among the sample of women who experienced an hCG pregnancy and did not withdraw.	Age and BMI modeled using B-splines with three degrees of freedom. Week on study modeled using quadratic term. Interaction term between adherence and randomization indicator, baseline covariates (except employment) Interaction term between time-varying covariates and alcohol, smoking use

Withdrawal	Randomization indicator (aspirin vs. placebo) Baseline covariates (high school education, marital status, employment, Non-Hispanic white race, alcohol use, smoking use, age, BMI) Time-varying confounders (adherence at week j and j – 1, bleeding at week j and j– 1, and nausea and/or vomiting at week j and j – 1) Week on study	None	Age and BMI modeled using B-splines with three degrees of freedom. Week on study modeled using natural cubic splines. Interaction term between adherence and randomization indicator, baseline covariates (except employment). Interaction term between time-varying covariates and alcohol, smoking use

End of follow-up, no pregnancy	Randomization indicator (aspirin vs. placebo) Baseline covariates (high school education, marital status, employment, Non-Hispanic white race, alcohol use, smoking use, age, BMI) Time-varying confounders (adherence at week j and j – 1, bleeding at week j and j– 1, and nausea and/or vomiting at week j and j – 1) Week on study	At every week on study, fit among sample of women who had not yet experienced an hCG pregnancy and did not withdraw	Age, BMI and week on study modeled using natural cubic splines Interaction term between adherence and randomization indicator

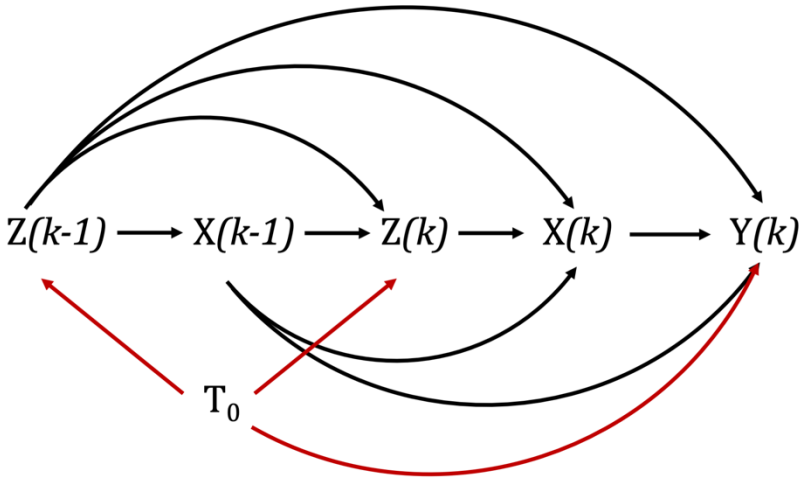
hCG pregnancy	Randomization indicator (aspirin vs. placebo) Baseline covariates (high school education, marital status, employment, Non-Hispanic white race, alcohol use, smoking use, age, BMI) Time-varying confounders (adherence at week j and j – 1, bleeding at week j and j– 1, and nausea and/or vomiting at week j and j – 1) Week on study	At every week on study, fit only among the sample of women whose previous week had an hCG value of 0.	Age, BMI and week on study modeled using B-splines with three degrees of freedom. Interaction term between adherence and randomization indicator, baseline covariates (except employment). Interaction term between time-varying covariates and alcohol, smoking use

Adherence to aspirin	Randomization indicator (aspirin vs. placebo) Baseline covariates (high school education, marital status, employment, Non-Hispanic white race, alcohol use, smoking use, age, BMI) Time-varying confounders (adherence at week j and j – 1, bleeding at week j and j– 1, and nausea and/or vomiting at week j and j – 1) Week on study	None	Age and BMI modeled using B-splines with three degrees of freedom. Week on study modeled using quadratic term. Interaction term between time-varying covariates and alcohol, smoking use

Bleeding	Randomization indicator (aspirin vs. placebo) Baseline covariates (high school education, marital status, employment, Non-Hispanic white race, alcohol use, smoking use, age, BMI) Time-varying confounders (adherence at week j – 1 and j-2, bleeding at week j– 1, and nausea and/or vomiting at week j and j – 1) Week on study	None	Age and BMI modeled using B-splines with three degrees of freedom. Week on study modeled using quadratic term. Interaction term between adherence at week j – 1 and randomization indicator, and baseline covariates (except employment). Interaction term between bleeding at week j – 1 and alcohol, smoking use. Interaction term between nausea and alcohol, smoking use.

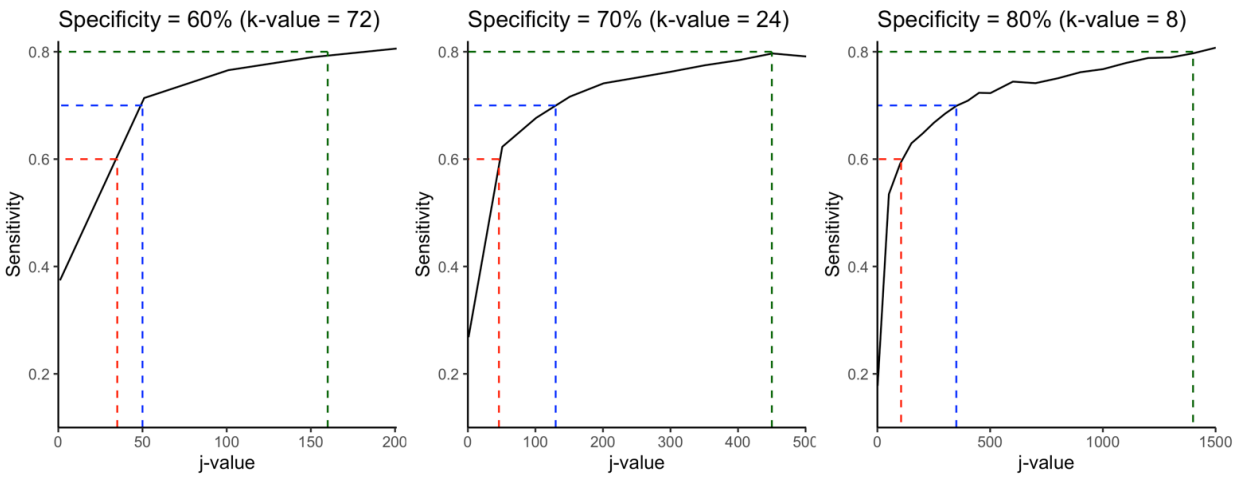
Nausea/Vomiting	Randomization indicator (aspirin vs. placebo) Baseline covariates (high school education, marital status, employment, Non-Hispanic white race, alcohol use, smoking use, age, BMI) Time-varying confounders (adherence at week j – 1 and j-2, bleeding at week j– 1, and nausea and/or vomiting at week j – 1) Week on study	None	Age and BMI modeled using B-splines with three degrees of freedom. Week on study modeled using quadratic term. Interaction term between adherence at week j – 1 and randomization indicator, and baseline covariates (except employment). Interaction term between nausea and bleeding at week j – 1 and alcohol, smoking use.

Appendix C Measurement Error Correction for Time-Varying Exposures Using the g-Formula

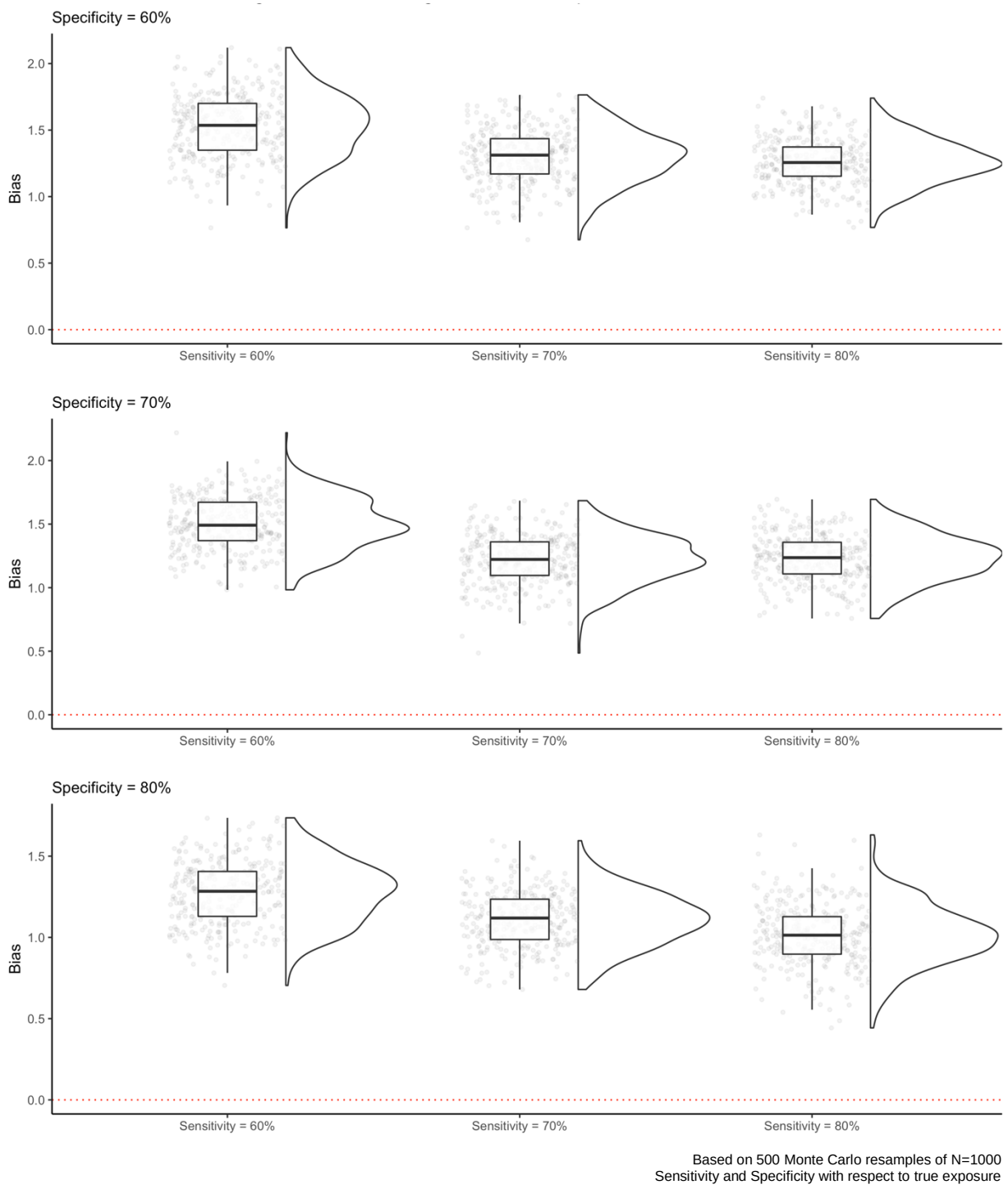


Appendix Figure 7. Directed acyclic graph pertaining to the data generating mechanism of the complex longitudinal data

NOTE: Z: time-varying confounding; X: time-varying exposure; Y: outcome



Appendix Figure 8. Selection of *k* and *j* values to produce a mismeasured exposure with a specified sensitivity and specificity with respect to the true exposure.



Appendix Figure 9. Distribution of bias from using mismeasured exposure in the g-Formula

Bibliography

ACOG Practice Bulletin No. 200 Summary: Early Pregnancy Loss. (2018). *Obstetrics and Gynecology*, *132*(5), 1311–1313. https://doi.org/10.1097/AOG.0000000000002900

Altman, D G, Lausen, B., Sauerbrei, W., & Schumacher, M. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst*, *86*(11), 829–835. https://doi.org/10.1093/jnci/86.11.829

Altman, Douglas G, & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ*, *332*(7549), 1080. https://doi.org/10.1136/bmj.332.7549.1080

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Bross, I. (1954). Misclassification in 2 X 2 Tables. *Biometrics*, *10*(4), 478. https://doi.org/10.2307/3001619

CHAVANCE, M., DELLATOLAS, G., & LELLOUCH, J. (1992). Correlated Nondifferential Misclassifications of Disease and Exposure: Application to a Cross-Sectional Study of the Relation between Handedness and Immune Disorders. *International Journal of Epidemiology*, *21*(3), 537–546. https://doi.org/10.1093/ije/21.3.537

Cole, S. R., & Stuart, E. A. (2010). Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology*, *172*(1), 107–115. https://doi.org/10.1093/aje/kwq084

Cole, Stephen R., Richardson, D. B., Chu, H., & Naimi, A. I. (2013). Analysis of occupational asbestos exposure and lung cancer mortality using the G formula. *American Journal of Epidemiology*, *177*(9), 989–996. https://doi.org/10.1093/aje/kws343

Cole, Stephen R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, *172*(1), 107–115. https://doi.org/10.1093/aje/kwq084

Cook, J. R., & Stefanski, L. A. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, *89*(428), 1314. https://doi.org/10.2307/2290994

de Jong, P. G., Kaandorp, S., Di Nisio, M., Goddijn, M., & Middeldorp, S. (2014). Aspirin and/or heparin for women with unexplained recurrent miscarriage with or without inherited thrombophilia. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD004734.pub4

Dolitzky, M., Inbal, A., Segal, Y., Weiss, A., Brenner, B., & Carp, H. (2006). A randomized study of thromboprophylaxis in women with unexplained consecutive recurrent miscarriages. *Fertility and Sterility*, *86*(2), 362–366. https://doi.org/10.1016/j.fertnstert.2005.12.068

El Hachem, H., Crepaux, V., May-Panloup, P., Descamps, P., Legendre, G., & Bouet, P.-E. (2017). Recurrent pregnancy loss: current perspectives. *International Journal of Women's Health*, *Volume 9*, 331–345. https://doi.org/10.2147/IJWH.S100817

Empson, M. (2002). Recurrent pregnancy loss with antiphospholipid antibody: a systematic review of therapeutic trials. *Obstetrics & Gynecology*, *99*(1), 135–144. https://doi.org/10.1016/S0029-7844(01)01646-5

Farquharson, R. (2002). Antiphospholipid syndrome in pregnancy: a randomized, controlled trial of treatment. *Obstetrics & Gynecology*, *100*(3), 408–413. https://doi.org/10.1016/S0029-7844(02)02165-8

Ford, H. B., & Schust, D. J. (2009). Recurrent pregnancy loss: etiology, diagnosis, and therapy. *Reviews in Obstetrics & Gynecology*, *2*(2), 76–83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19609401

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *Am J Epidemiol*, *173*(7), 761–767. https://doi.org/10.1093/aje/kwq439

Glynn, A. N., & Quinn, K. M. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis*, *18*(1), 36–56. Retrieved from http://www.jstor.org/stable/25791992

Greenland, S. (1983). Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med*, *2*(2), 243–251. https://doi.org/10.1002/sim.4780020219

Greenland, S. (1993). Basic problems in interaction assessment. *Environmental Health Perspectives*, *101*(suppl 4), 59–66. https://doi.org/10.1289/ehp.93101s459

Greenland, S. (1995). Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*, *6*(4), 356–365. https://doi.org/10.1097/00001648-199507000-00005

Groves, R. M., Mosher, W. D., Lepkowski, J. M., & Kirgis, N. G. (2009). Planning and development of the continuous National Survey of Family Growth. *Vital and Health Statistics. Ser. 1, Programs and Collection Procedures*, (48), 1–64. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20141029

Gullen, W. H., Bearman, J. E., & Johnson, E. A. (1968). Effects of misclassification in epidemiologic studies. *Public Health Reports (Washington, D.C. : 1896)*, *83*(11), 914–918. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4972198

Hernán, M. A., & VanderWeele, T. J. (2011a). Compound Treatments and Transportability of Causal Inference. *Epidemiology*, *22*(3), 368–377. https://doi.org/10.1097/EDE.0b013e3182109296

Hernán, M. A., & VanderWeele, T. J. (2011b). Compound Treatments and Transportability of Causal Inference. *Epidemiology*, *22*(3), 368–377. https://doi.org/10.1097/EDE.0b013e3182109296

Howe, C. J., Cole, S. R., Westreich, D. J., Greenland, S., Napravnik, S., & Eron Jr, J. J. (2011). Splines for trend analysis and continuous confounder control. *Epidemiology*, *22*(6), 874–875. https://doi.org/10.1097/EDE.0b013e31823029dd

Jurek, A. M., Greenland, S., Maldonado, G., & Church, T. R. (2005). Proper interpretation of non-differential misclassification effects: expectations vs observations. *International Journal of Epidemiology*, *34*(3), 680–687. https://doi.org/10.1093/ije/dyi060

Kaandorp, S. P., Goddijn, M., van der Post, J. A. M., Hutten, B. A., Verhoeve, H. R., Hamulyák, K., … Middeldorp, S. (2010). Aspirin plus Heparin or Aspirin Alone in Women with Recurrent Miscarriage. *New England Journal of Medicine*, *362*(17), 1586–1596. https://doi.org/10.1056/NEJMoa1000641

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, *22*(4), 523–539. https://doi.org/10.1214/07-STS227

Keil, A. P., Edwards, J. K., Richardson, D. B., Naimi, A. I., & Cole, S. R. (2014). The parametric g-formula for time-to-event data: Intuition and a worked example. *Epidemiology*, *25*(6), 889–897. https://doi.org/10.1097/EDE.0000000000000160

Keil, A. P., Mooney, S. J., Jonsson Funk, M., Cole, S. R., Edwards, J. K., & Westreich, D. (2018). RESOLVING AN APPARENT PARADOX IN DOUBLY ROBUST ESTIMATORS. *Am J Epidemiol*, *187*(4), 891–892. https://doi.org/10.1093/aje/kwx385

Kennedy, E. H. (2015). *Semiparametric theory and empirical processes in causal inference*. Retrieved from http://arxiv.org/abs/1510.04740

Kennedy, E. H. (2020). *Optimal doubly robust estimation of heterogeneous causal effects*.

Klock, S. C., Chang, G., Hiley, A., & Hill, J. (1997). Psychological Distress Among Women With Recurrent Spontaneous Abortion. *Psychosomatics*, *38*(5), 503–507. https://doi.org/10.1016/S0033-3182(97)71428-2

Kravitz, E. S., Carroll, R. J., & Ruppert, D. (2019). *Sample Splitting as an M-Estimator with Application to Physical Activity Scoring*.

Kreif, N., & DiazOrdaz, K. (2019). *Machine learning in policy evaluation: new tools for causal inference*.

Kyle, R. P., Moodie, E. E. M., Klein, M. B., & Abrahamowicz, M. (2016). Correcting for Measurement Error in Time-Varying Covariates in Marginal Structural Models. *American Journal of Epidemiology*, *184*(3), 249–258. https://doi.org/10.1093/aje/kww068

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., & Cole, S. R. (2017). Generalizing Study Results. *Epidemiology*, *28*(4), 553–561. https://doi.org/10.1097/EDE.0000000000000664

Liao, X., Zucker, D. M., Li, Y., & Spiegelman, D. (2011). Survival Analysis with Error-Prone Time-Varying Covariates: A Risk Set Calibration Approach. *Biometrics*, *67*(1), 50–58. https://doi.org/10.1111/j.1541-0420.2010.01423.x

Lubin, J. H., Samet, J. M., & Weinberg, C. (1990). Design issues in epidemiologic studies of indoor exposure to Rn and risk of lung cancer. *Health Physics*, *59*(6), 807–817. https://doi.org/10.1097/00004032-199012000-00004

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychol Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989x.7.1.19

Moodie, E. E. M., & Stephens, D. A. (2010). Using Directed Acyclic Graphs to detect limitations of traditional regression in longitudinal studies. *International Journal of Public Health*, *55*(6), 701–703. https://doi.org/10.1007/s00038-010-0184-x

Mumford, S. L., & Schisterman, E. F. (2019). New methods for generalizability and transportability: the new norm. *European Journal of Epidemiology*, *34*(8), 723–724. https://doi.org/10.1007/s10654-019-00532-3

Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: an introduction to super learning. *Eur J Epidemiol*, *33*(5), 459–464. https://doi.org/10.1007/s10654-018-0390-z

Naimi, A. I., Cole, S. R., & Kennedy, E. H. (2017). An Introduction to G Methods. *International Journal of Epidemiology*, *46*(2), 756–762. https://doi.org/10.1093/ije/dyw323

Naimi, A. I., & Kennedy, E. H. (2017). *Nonparametric Double Robustness*. Retrieved from http://arxiv.org/abs/1711.07137

Naimi, A. I., Perkins, N. J., Sjaarda, L. A., Mumford, S. L., Platt, R. W., Silver, R. M., & Schisterman, E. F. (2021). The Effect of Preconception-Initiated Low-Dose Aspirin on Human Chorionic Gonadotropin–Detected Pregnancy, Pregnancy Loss, and Live Birth. *Annals of Internal Medicine*, M20-0469. https://doi.org/10.7326/M20-0469

National Center for Health Statistics (NCHS). (2017a). 2015-2017 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods. Retrieved from Hyattsville, MD: CDC National Center for Health Statistics. Retrived from: https://www.cdc.gov/nchs/data/nsfg/PUF3-NSFG-2015-2017-Summary-of-DesignDataCollection_02Oct2019.pdf website: https://www.cdc.gov/nchs/data/nsfg/PUF3-NSFG-2015-2017-Summary-of-DesignDataCollection_02Oct2019.pdf

National Center for Health Statistics (NCHS). (2017b). 2015-2017 National Survey of Family Growth Public-Use Data and Documentation. In *Hyattsville, MD: CDC National Center for Health Statistics. Retrived from: http://www.cdc.gov/nchs/nsfg/nsfg_2*.

Nobles, C. J., Mendola, P., Mumford, S. L., Kim, K., Sjaarda, L., Hill, M., … Schisterman, E. F. (2019). Metabolic Syndrome and the Effectiveness of Low-dose Aspirin on Reproductive Outcomes. *Epidemiology*, *30*(4), 573–581. https://doi.org/10.1097/EDE.0000000000001019

Patrono, C., & Rocca, B. (2017). Type 2 Diabetes, Obesity, and Aspirin Responsiveness. *J Am Coll Cardiol*, *69*(6), 613–615. https://doi.org/10.1016/j.jacc.2016.11.049

Pearl, J., & Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, *29*(4). https://doi.org/10.1214/14-STS486

Porter, F. T., Gyamfi-Bannerman, C., & Manuck, T. (n.d.). Low-Dose Aspirin Use During Pregnancy. *ACOG COMMITTEE OPINION*, *743*.

Rasmark Roepke, E., Matthiesen, L., Rylance, R., & Christiansen, O. B. (2017). Is the incidence of recurrent pregnancy loss increasing? A retrospective register-based study in Sweden. *Acta Obstetricia et Gynecologica Scandinavica*, *96*(11), 1365–1372. https://doi.org/10.1111/aogs.13210

Rencher, A. C., & Schaalje, B. G. (2007). Multiple Regression: Estimation. In *Linear Models in Statistics* (pp. 137–184). https://doi.org/10.1002/9780470192610.ch7

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell Jr, F. E., Martin, G. P., Reitsma, J. B., … van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441

Rinaldo, A., Wasserman, L., G'Sell, M., & Lei, J. (2016). *Bootstrapping and Sample Splitting For High-Dimensional, Assumption-Free Inference*.

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, *90*(429), 122–129. Retrieved from http://www.jstor.org/stable/2291135

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818

ROSNER, B., SPIEGELMAN, D., & WILLETT, W. C. (1990). CORRECTION OF LOGISTIC REGRESSION RELATIVE RISK ESTIMATES AND CONFIDENCE INTERVALS FOR MEASUREMENT ERROR: THE CASE OF MULTIPLE COVARIATES MEASURED WITH ERROR. *American Journal of Epidemiology*, *132*(4), 734–745.

https://doi.org/10.1093/oxfordjournals.aje.a115715

Rosner, B., Willett, W. C., & Spiegelman, D. (2006). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, *8*(9), 1051–1069. https://doi.org/10.1002/sim.4780080905

Rothwell, P. M., Cook, N. R., Gaziano, J. M., Price, J. F., Belch, J. F. F., Roncaglioni, M. C., … Mehta, Z. (2018). Effects of aspirin on risks of vascular events and cancer according to bodyweight and dose: analysis of individual patient data from randomised trials. *Lancet*, *392*(10145), 387–399. https://doi.org/10.1016/S0140-6736(18)31133-4

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, *25*(1), 127–141. https://doi.org/10.1002/sim.2331

Royston, P., & Sauerbrei, W. (2004). A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Stat Med*, *23*(16), 2509–2525. https://doi.org/10.1002/sim.1815

Royston, P., & Sauerbrei, W. (2013). Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. *Stat Med*, *32*(22), 3788–3803. https://doi.org/10.1002/sim.5813

Royston, P., & Sauerbrei, W. (2014). Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis. *Stat Med*, *33*(27), 4695–4708. https://doi.org/10.1002/sim.6308

Rubinstein, M., Marazzi, A., & Polak de Fried, E. (1999). Low-dose aspirin treatment improves ovarian responsiveness, uterine and ovarian blood flow velocity, implantation, and pregnancy rates in patients undergoing in vitro fertilization: a prospective, randomized, double-blind placebo-controlled assay. *Fertility and Sterility*, *71*(5), 825–829. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10231040

Schisterman, E. F., Silver, R. M., Lesher, L. L., Faraggi, D., Wactawski-Wende, J., Townsend, J. M., … Galai, N. (2014). Preconception low-dose aspirin and pregnancy outcomes: Results from the EAGeR randomised trial. *The Lancet*, *384*(9937), 29–36. https://doi.org/10.1016/S0140-6736(14)60157-4

Schisterman, E. F., Silver, R. M., Perkins, N. J., Mumford, S. L., Whitcomb, B. W., Stanford, J. B., … Galai, N. (2013). A Randomised Trial to Evaluate the Effects of Low-dose Aspirin in Gestation and Reproduction: Design and Baseline Characteristics. *Paediatric and Perinatal Epidemiology*, *27*(6), 598–609. https://doi.org/10.1111/ppe.12088

Schuler, M. S., & Rose, S. (2017). Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *Am J Epidemiol*, *185*(1), 65–73. https://doi.org/10.1093/aje/kww165

Sjaarda, L. A., Radin, R. G., Silver, R. M., Mitchell, E., Mumford, S. L., Wilcox, B., …

Schisterman, E. F. (2017). Preconception Low-Dose Aspirin Restores Diminished Pregnancy and Live Birth Rates in Women With Low-Grade Inflammation: A Secondary Analysis of a Randomized Trial. *The Journal of Clinical Endocrinology and Metabolism*, *102*(5), 1495–1504. https://doi.org/10.1210/jc.2016-2917

Sorahan, T., & Gilthorpe, M. S. (1994). Non-differential misclassification of exposure always leads to an underestimate of risk: An incorrect conclusion. *Occupational and Environmental Medicine*. https://doi.org/10.1136/oem.51.12.839

Spiegelman, D., McDermott, A., & Rosner, B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *The American Journal of Clinical Nutrition*, *65*(4), 1179S-1186S. https://doi.org/10.1093/ajcn/65.4.1179S

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science*, *16*(3), 475–485. https://doi.org/10.1007/s11121-014-0513-z

Taubman, S. L., Robins, J. M., Mittleman, M. A., & Hernán, M. A. (2009). Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, *38*(6), 1599–1611. https://doi.org/10.1093/ije/dyp192

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. https://doi.org/10.1007/978-1-4757-3294-8

Tsiatis, A. A., Degruttola, V., & Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association*, *90*(429), 27–37. https://doi.org/10.1080/01621459.1995.10476485

Tulppala, M., Marttunen, M., Soderstrom-Anttila, V., Foudila, T., Ailus, K., Palosuo, T., & Ylikorkala, O. (1997). Low-dose aspirin in prevention of miscarriage in women with unexplained or autoimmune related recurrent miscarriage: effect on prostacyclin and thromboxane A2 production. *Human Reproduction*, *12*(7), 1567–1572. https://doi.org/10.1093/humrep/12.7.1567

Turner, J. M., Robertson, N. T., Hartel, G., & Kumar, S. (2020). Impact of low-dose aspirin on adverse perinatal outcome: meta-analysis and meta-regression. *Ultrasound in Obstetrics & Gynecology*, *55*(2), 157–169. https://doi.org/10.1002/uog.20859

van der Laan, M. J., & Luedtke, A. R. (n.d.). Targeted Learning of an Optimal Dynamic Treatment, and Statistical Inference for its Mean Outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series*, *Working pa*(September 2014).

van der Laan, M. J., & Rubin, D. (n.d.). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, *2*(1). https://doi.org/10.2202/1557-4679.1043

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res*

Methodol, *14*, 137. https://doi.org/10.1186/1471-2288-14-137

VanderWeele, T. J., & Knol, M. J. (2014). A Tutorial on Interaction. *Epidemiologic Methods*, *3*, 33–72.

Vane, J. ., & Botting, R. . (2003). The mechanism of action of aspirin. *Thrombosis Research*, *110*(5–6), 255–258. https://doi.org/10.1016/S0049-3848(03)00379-7

Weksler, B. B., Kent, J. L., Rudolph, D., Scherer, P. B., & Levy, D. E. (1985). Effects of low dose aspirin on platelet function in patients with recent cerebral ischemia. *Stroke*, *16*(1), 5–9. https://doi.org/10.1161/01.STR.16.1.5

Westreich, D., Cole, S. R., Young, J. G., Palella, F., Tien, P. C., Kingsley, L., … Hernán, M. A. (2012). The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine*, *31*(18), 2000–2009. https://doi.org/10.1002/sim.5316

Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R., & Stuart, E. A. (2019). Target Validity and the Hierarchy of Study Designs. *American Journal of Epidemiology*, *188*(2), 438–443. https://doi.org/10.1093/aje/kwy228

Wilcox, A. J., Weinberg, C. R., O'Connor, J. F., Baird, D. D., Schlatterer, J. P., Canfield, R. E., … Nisula, B. C. (1988). Incidence of early loss of pregnancy. *The New England Journal of Medicine*, *319*(4), 189–194. https://doi.org/10.1056/NEJM198807283190401

Young, J. G., Hernán, M. A., Picciotto, S., & Robins, J. M. (2010). Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis*, *16*(1), 71–84. https://doi.org/10.1007/s10985-009-9135-3

Zivich, P. N., & Breskin, A. (2020). *Machine learning for causal inference: on the use of cross-fit estimators*.