Published in final edited form as:
Stat Methods Med Res. 2018 October ; 27(10): 3183–3204. doi:10.1177/0962280217694665.

A Bayesian approach to the g-formula

Alexander P. Keil[1], **Eric J. Daza**[2], **Stephanie M. Engel**[1], **Jessie P. Buckley**[1], and **Jessie K. Edwards**[1]

[1]Department of Epidemiology, University of North Carolina, Chapel Hill, USA

[2]Stanford Prevention Research Center, Stanford University School of Medicine, Palo Alto, USA

Abstract

Epidemiologists often wish to estimate quantities that are easy to communicate and correspond to the results of realistic public health interventions. Methods from causal inference can answer these questions. We adopt the language of potential outcomes under Rubin's original Bayesian framework and show that the parametric g-formula is easily amenable to a Bayesian approach. We show that the frequentist properties of the Bayesian g-formula suggest it improves the accuracy of estimates of causal effects in small samples or when data are sparse. We demonstrate an approach to estimate the effect of environmental tobacco smoke on body mass index among children aged 4–9 years who were enrolled in a longitudinal birth cohort in New York, USA. We provide an algorithm and supply SAS and Stan code that can be adopted to implement this computational approach more generally.

Keywords

Bayesian; causal inference; g-computation; semi-parametric

1 Introduction

Epidemiologists often wish to estimate quantities that are easy to communicate and correspond to the results of realistic public health interventions, such as changes to treatment guidelines or policy. Recent developments in the field of causal inference have shown that, under stated assumptions, observational data can be used to estimate such quantities.[14] Robins' generalized computation algorithm formula (the g-formula) is one example of an approach from that field.[30] Using a parametric version of the g-formula, we can easily compare risks or rates of a health outcome in a population of interest under different exposure distributions.[21]

One assumption of the parametric g-formula is that we have accurately modeled the association between the outcome of interest and both the exposures of interest and potential confounders. For example, if the rate for a particular outcome varies in a linear-quadratic

Reprints and permission: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Alexander P. Keil, Department of Epidemiology, University of North Carolina, Chapel Hill, 27514, USA, akeil@unc.edu.

function with exposure, then our model could capture that by including both a linear and a quadratic term for exposure. This assumption can be approximately satisfied by modeling the association using flexible models that incorporate non-linear functions and interaction terms. In particular settings, such as the study of complex exposure mixtures, complex longitudinal data, or small data sets, this may result in the model smoothing over strata in which we have little or no data, resulting in highly unstable estimates.

Numerous statistical approaches have arisen for stabilizing model estimates in such scenarios. Of note, Bayesian approaches have shown promise for dealing with sparsely-identified models through the use of parameter prior distributions to stabilize estimates that would be highly variable in an unpenalized maximum likelihood procedure.[23;42] Further, even a set of default shrinkage priors can improve parameter estimation in multivariable regression models. For example in Stein's paradox (see, for example, Efron and Morris's description[12]), the maximum likelihood estimator of a set of regression parameters will always be inferior to a well-chosen shrinkage estimator whenever the number of parameters exceeds two. The parametric g-formula is computationally intensive and requires modeling a large joint parameter space that includes the outcome of interest, confounders, and possibly exposure. When exposure and confounders are time-varying, the modeling requirements increase. When models are specified (approximately) correctly, the parametric g-formula is a powerful inferential tool with applications in many fields of knowledge generation such as public health, environmental health, and medicine. A significant drawback of the parametric g-formula in the frequentist construction is that it relies on inadmissible estimators (that is, we can always find a better estimator[40]) of a large number of model parameters, the maximum likelihood estimates.

The primary objective of the current manuscript is to outline a foundational framework for Bayesian causal inference in epidemiologic studies that can be adapted for situations in which frequentist estimators may not be ideal. Such situations arise in cases of sparse data due to small sample size, when the number of parameters approaches sample size, when meaningful prior information is available, when covariates are highly correlated,[23] or when accounting for measurement error.[20] In particular, we focus on the case of small samples when the nuisance parameters of the parametric g-formula may be highly unstable, leading to poor finite-sample characteristics of the parameters of interest. We hypothesize that, in the vein of Stein's paradox, a Bayesian approach with default shrinkage priors can result in estimates of marginal, population parameters that can improve on our predictions of the outcome of public health actions or interventions. This work work builds on previous applications of Bayesian, causal inference[2;7;18;39;45] by taking a more general approach to the Bayesian g-formula.

Our work also builds on non-Bayesian approaches to estimating parameters of the g-formula, such as those given by Taubman et al[41], Ahern et al [1], Westreich et al[46], and Keil et al[21]. In particular, the algorithm given in §3 is a Bayesian analogue of the algorithm given by Taubman et al[41], whereby we estimate the distribution of potential outcomes under exposure using the empirical distribution of predicted outcomes under hypothetical interventions on exposures. We show, using simulations, that finite-sample characteristics of the g-formula can be improved by adopting a Bayesian framework. Thus, provided that causal assumptions are met, our Bayesian approach may provide more accurate predictions of potential public health interventions than other approaches. For example, in a recent example of the non-Bayesian g-formula, Keil et al. estimated the potential impact of interventions on arsenic exposure in a cohort of copper smelters.[22] In that example, the authors addressed the effects of airborne arsenic exposure on lung cancer and heart disease mortality. The latter effect estimate was estimated imprecisely, in part due to the highly variable baseline risk of heart disease mortality. Our approach could be applied directly to improve such an analysis by providing shrinkage priors or by leveraging the numerous studies done of non-airborne arsenic exposure and heart disease, which could, for example, provide a reasonable prior range for the estimate of association between arsenic and heart disease mortality. We also show here that Bayesian Lasso is a viable approach to variance reduction. We are not aware of alternative approaches that can directly leverage prior information or shrinkage estimators while still providing impact estimates for hypothetical exposures. Such estimates are essential in deriving exposures standards in occupational and environmental settings, where precise estimates of risk are necessary to address the cost-benefit between exposure-control measures and population health.

The novelty of this work is based on 4 points: (1) we derive a Bayesian g-formula approach in the potential outcomes framework, (2) we show that the Bayesian g-formula is possible using off-the-shelf MCMC software, (3) we provide an example showing that the approach is naturally ‘modular’ and can be used to incorporate existing Bayesian solutions to problematic data, and (4) we demonstrate that the Bayesian g-formula has desirable frequentest properties under some choices of default shrinkage priors. This last point indicates that the Bayesian g-formula may have useful practical applications in scenarios when the frequentist g-formula may typically be used.

In Section 2 we define notation and demonstrate the Bayesian g-formula in time-fixed data, and we discuss some potential settings of interest. We outline an algorithm in Section 3 that can be used to develop software for this approach in a general Markov chain Monte Carlo estimation framework. This algorithm builds on non-Bayesian approaches to estimating parameters of the g-formula, such as those given by Taubman et al[41], Ahern et al[1], Westreich et al[46], and Keil et al.[21] In particular, we focus on the use of Bayesian methodology to improve on the finite-sample performance of the g-formula.

In Section 4 we assess the ability to improve parameter estimation using default shrinkage priors by simulation. In Section 5, we demonstrate our approach using a longitudinal study of environmental tobacco smoke exposure and childhood body mass index (BMI), and extend our approach to non-normal shrinkage priors using a version of the Bayesian LASSO (least absolute shrinkage and selection operator[27]). We close in Section 6 with a discussion of the assumptions underlying our approach and speculate about future extensions. In the Appendix, we derive our approach for both the time-fixed and longitudinal settings and we supply code in the Supporting Information to implement our approach for SAS, and Stan programming languages.

2 The Bayesian g-formula

2.1 The parametric g-formula

To simplify our presentation, in this section we nest our example in the context of an observed set of time-fixed variables, $O = (Y, X, L)$, where Y is some utility function, in our example a binary endpoint such as all-cause mortality, X is a binary exposure of interest, such as the level (high or low) of criteria air pollutants, (CO, NO_x, O_3, SO_2, Pb), and L is a vector of potential confounders such as residence in an urban (versus rural) setting and annual income (above or below $80,000, say). A more general setting is given in the Appendix.

In this setting, we are interested in the marginal distribution of the potential outcome $Y^g \in \{0, 1\}$, where g refers to some intervention value (or distribution) for the exposures of interest. For example, we may be interested in the risk (i.e. cumulative incidence) of lung cancer in the population in which g represents low concentrations of the pollutants. In other words, our intervention is to "set" each of the pollutants to "low." We assume that this hypothetical scenario corresponds to a well-defined intervention or set of interventions to reduce the pollutants to our specified level, such as mandating manufacturing changes to decrease vehicle exhaust emissions or restricting traffic on certain days. The risk under this intervention is given as $Pr(Y^g = 1)$[9]. For example, for binary L the g-formula estimate of the risk under intervention can be expressed using the law of total probability as

$$\Pr\left(Y^g = 1\right) = \sum_{\ell \in \{0,1\}} \Pr\left(Y^g = 1, L = \ell\right) = \sum_{\ell \in \{0,1\}} \Pr\left(Y^g = 1 \mid L = \ell\right) \Pr\left(L = \ell\right).$$

To facilitate the expression of these quantities in a Bayesian framework, we adopt a more general notation and instead express the distribution of the potential outcome as

$$p\left(y^g\right) = \int_{\ell} p\left(y^g \mid \ell\right) p\left(\ell\right) d\ell, \quad (1)$$

where the summation symbol has been replaced by an integral, and the discrete probability notation has been replaced by the generic probability function p. For a random variable A, if A is discrete then we let $p(a)$ denote $\Pr(A = a)$, the mass of A at realization a. Likewise, if A is continuous then we let $p(a)$ denote $f(a)$, the density of A at realization a. We also let $p(\cdot/a)$ denote $p(\cdot/A = a)$. To simplify notation we allow that the integral expression $\int g(a)da$ for a function $g(a)$ of a will be used to denote integration if A is continuous under Lebesgue measure, i.e., $\int g(a)da$, and summation if A is discrete under counting measure, i.e., $\Sigma_{\{a\}} g(a)$. The formula given in 1 yields the population distribution of the outcome under the intervention g, where the "population" referred to here is the set of individuals with the distribution of L given by $p(\ell)$.

The parametric g-formula extends the g-formula given in 1 by characterizing the conditional components $p(y^g|\ell)$ and $p(\ell)$ using parametric models. Thus, the parametric g-formula represention of the potential outcome distribution under intervention g is given as

$$p\left(y^g;\beta,\eta\right)=\int_\ell p\left(y^g\mid\ell;\beta\right)p\left(\ell,\eta\right)d\ell=\int_\ell p\left(y\mid g,\ell;\beta\right)p\left(\ell;\eta\right)d\ell \quad (2)$$

where the parameter vector β corresponds to the conditional change in the probability of Y for unit changes in X and L, and η corresponds to the parameters of the probability of L. The rightmost side of 2 is derived under counterfactual consistency[28] and relies on the assumptions of positivity, conditional exchangeability, and correct specification of models. Summary effect measures, such as the causal risk difference for a unit increase in X can be estimated using Monte Carlo methods.[21]

The utility of the g-formula becomes apparent when one considers that the functional form of the relationship between X and Y may be a complicated non-linear function in non-binary data, possibly with high order interaction terms. Note that, under such a scenario, the counterfactual distribution $p(\tilde{y}^g)$ is easily interpreted as the outcome distribution we would expect if we could have intervened to set X to g. Similarly, if L is high dimensional or X (and possibly L) is time-varying, simple interpretations are still possible and can correspond to realistic settings. Thus, the approach provides results that are easy to communicate to non-specialist stakeholders and have immediate public health utility. As a further consideration, the g-formula can be used to estimate unbiased net effects of exposure when X and L may affect each other over time, whereas regression model estimates will generally be biased for such parameters.[36]

The parameters β or η may not be stably estimated when L or X is of high dimension relative to the sample size, when estimating effects in small samples, or when there is high correlation between elements of the data (i.e., due to variance inflation or finite sample bias). In such settings, common approaches are to create a more parsimonious model, apply some penalization term/regularizing, or adopt a Bayesian framework to stabilize estimates using prior knowledge. We adopt an approach that recasts the parametric g-formula within a Bayesian framework as a way to embrace the interpretability of the g-formula approach while employing the variance reduction properties of Bayesian inference. We focus on shrinkage estimation, in which we allow the possible introduction of some bias in exchange for higher precision and an overall reduction in mean-squared error.

2.2 A Bayesian approach

Following Rubin [37] and Saarela et al. [39], we consider a Bayesian version of the potential outcome distribution, the posterior predictive distribution $p(\tilde{y}^g|o)$. The posterior predictive distribution of the potential outcomes given by the Bayesian approach differs from the estimated distribution of potential outcomes using the standard g-formula in that it marginalizes over the posterior distributions of the parameters (β and η).

Our quantity of interest is the posterior predictive distribution of the potential outcome under the intervention g, which is defined as $p(\tilde{y}^g|o) \equiv \int p(\tilde{y}^g|\theta)p(\theta|o)d\theta$, where $p(\theta|o)$ is the posterior distribution of the parameters θ (which describe the models of the relationships between Y, X, and L).

The posterior predictive distribution of the potential outcome under intervention g is given in our example as

$$p\left(\tilde{y}^g \mid o\right) = \int_{\tilde{\ell}} p\left(\tilde{y} \mid g, \tilde{\ell}, o\right) p\left(\tilde{\ell} \mid o\right) d\tilde{\ell} = \int_{\tilde{\ell}} \int_{\theta} p\left(\tilde{y} \mid g, \tilde{\ell}, \theta\right) p\left(\tilde{\ell} \mid \theta\right) p\left(\theta \mid o\right) d\theta d\tilde{\ell} \, .$$

Under counterfactual consistency and the assumptions of conditional exchangeability, positivity, and parameter variation independence (detailed in the Appendix), $p(\tilde{y}^g|o)$ is a function only of the observed data, the target population distribution of L, the intervention value set by the analyst, and prior parameter distributions. We make these assumptions for the remainder of this section.

Given a likelihood $\mathcal{L}(\theta|o) = \mathcal{L}(\beta, a, \eta|o)$ that factors into three distinct components $\mathcal{L}(\beta|y, x, \ell) \times \mathcal{L}(a|x, \ell) \times \mathcal{L}(\eta|\ell)$ we show in the Appendix that the posterior predictive distribution of Y^g is proportional to:

$$p\left(\tilde{y}^g \mid o\right) \propto \int_{\tilde{\ell}} \int_{\eta} \int_{\beta} p\left(\tilde{y} \mid g, \tilde{\ell}, \beta\right) p\left(\tilde{\ell} \mid \eta\right) \mathscr{L}\left(\beta \mid y, x, \ell\right) \mathscr{L}\left(\eta \mid \ell\right) \pi\left(\beta\right) \pi\left(\eta\right) d\beta d\eta d\tilde{\ell} \quad (3)$$

To conceptualize the computation, $\tilde{Y}^g$ acts like a missing variable that can be imputed after setting exposure to some value g.[11] In general, 3 and it's extension to longitudinal data given in the Appendix will not be given in closed-form, which necessitates MCMC approaches. MCMC software that can accommodate imputation of missing data may be used for the Bayesian g-formula.

3 Bayesian g-formula algorithm

Here we present an algorithm for estimation using the Bayesian g-formula to numerically approximate 3, as well as quantities derived from a comparisons of $\tilde{Y}^g$ under two interventions, such as the 30-year risk difference comparing the study population under two different exposures, one version of the average causal effect. To generalize the outline given in section 2 to longitudinal settings, we describe an algorithm for estimating parameters of the Bayesian g-formula when exposure and covariates may be time-varying (the Bayesian g-formula in the time-fixed setting is just a special case of this longitudinal approach). For simplicity, we limit this approach to static, deterministic interventions such as "always treat" or "never treat."

Generalizing our approach to the longitudinal setting requires choosing some time scale, such as time-on-study, calendar time, or age, and we denote specific points on that time scale as t, which we assume to be fine, yet discrete. We denote the value of the exposure at time t

as $X(t)$ and let the history of exposure up to and including time t be $\bar{X}(t) \equiv (X(0), \ldots, X(t))$. Time-varying covariate vectors $\bar{L}(t)$ and outcomes $\bar{Y}(t)$ are denoted similarly. In our notation $\bar{X}(t)$, $\bar{L}(t)$, and $\bar{Y}(t)$ are multivariate vectors with elements at times $0, \ldots, t$. In practice, these usually correspond to functional summaries over time that include only relevant aspects, such as the cumulative exposure through time t. The vector $(L(0), X(0), Y(0))$ contains the values of the covariates, exposure and outcome we observe at the origin of the time scale of interest. Generally, $L(0)$ includes time-fixed quantities such as race, sex at birth, or income at baseline.

We let $\tilde{Y}^g(t)$ and $\tilde{L}(t, g)$ represent values of the posterior predictive distributions of the potential outcome $Y^g(t)$ and the covariate vector $L(t)$ under intervention g. These variable vectors are denoted in different ways to emphasize that we may be able to estimate the posterior distribution of the potential outcomes (a causal interpretation) under intervention g without making the additional assumptions necessary to estimate the distribution of the other post-exposure covariates. We assume that $L(t)$ occurs temporally before $X(t)$, which both occur before $Y(t)$, implying that $\tilde{L}(0, g) = \tilde{L}(0)$. This allows us to non-parametrically sample $\tilde{L}(0)$ from the population empirical distribution $p_N(l(0))$. Because we let $p_N(l(0)) = p(l(0)/\eta)$ and we let $\tilde{\ell}(0) = \ell(0)$, our algorithm is more accurately described to estimate parameters of a semi-Bayesian, semi-parametric g-formula.

After we have selected an appropriate target population (defined by $p(\tilde{\ell}(0))$ and intervention of interest (defined by g), the Bayesian g-formula algorithm can be summarized in the following steps:

1. Specify the model for the joint likelihood

Specify a joint model in the source population for $\{L(t), X(t), Y(t)\}$. For static regimes, this corresponds to the following

1. Specify a model for each component of $L(t)$. One could conceivably fit a separate model at each t, which, for multivariate $L \equiv (L_1, \ldots, L_p)$ and $t = (0, \ldots, K)$ equals $p \times K$ models. In practice, the elements $L(t)$ are modeled using time-averaged (pooled) models that average coefficient values across time. If $L(t)$ consists of a single dichotomous variable this could be given as

$$\operatorname{logit}\left\{L(t) = 1 \mid \bar{x}(t-1), \bar{l}(t-1), \bar{Y}(t-1) = \bar{0}; \eta\right\} = \eta_0 + \eta_1 t + \eta_2 \sum_{j=0}^{t-1} x(j) + \eta_3 \sum_{j=0}^{t-1} \ell(j)$$

2. Specify a time-averaged model (or set of time-specific models) for $Y(t)$. Assuming that $Y(t)$ is an indicator of death by time t, this model could be

$$\operatorname{logit}\left\{Y_1(t) = 1 \mid \bar{x}(t), \bar{l}(t), \bar{Y}(t-1) = \bar{0}; \beta\right\} = \beta_0 + \beta_1 t + \beta_2 \sum_{j=0}^{t} x(j) + \beta_3 \sum_{j=0}^{t} \ell(j)$$

Both of these relatively simple models should be considered for illustrative purposes only because they are likely over-simplifications for realistic settings.

2. Specify the prior distributions

For a static intervention, we need only define prior distributions $\pi(\beta)$ and $\pi(\eta)$ because modeling exposure is unnecessary. In most scenarios, investigators may be able to derive meaningful and relatively precise prior information for β because this model often corresponds to regression models used in the literature. Parameters for η may be more challenging because covariate model parameters are rarely reported.

In many scenarios, the analyst will be unable to specify informative, non-null distributions for some (or possibly all) parameters. One may, instead, specify reasonable generic information, such that the conditional log-odds ratios are likely to be within some interval. In epidemiologic practice, it is rare to see (for example) odds ratios outside of the range 1/10 to 10.[19] Therefore, weakly informative default priors result in shrinkage estimators that may improve predictive accuracy, which is a common justification for using Bayesian inference.[17]

3. Sample from the target population

Define and sample from a target population defined by $p(\tilde{L}(0)$. This is often identical to the population by the observed sample, and the $L(0)$ can be sampled by taking the empirical distribution of the baseline covariates in the data. The target population will be considered to be under observation for some fixed time k, such that $t \equiv \{1, .., k\}$.

4. Set the intervention values

Define a static intervention $\tilde{g}(m)$ such as "always exposed", e.g., $g(t) = g = 1$for all t. Set the exposures at all time points to this static intervention; i.e., $\tilde{X}(t) = g$ for all t. A parameter of interest could be the posterior predictive causal risk difference comparing "always exposed" to "never exposed."

5. Draw from the posterior parameter distribution

From our simple models in step 1, draws from the posterior distribution of the model parameters given in step 1 ($\hat{\eta}_0, \ldots, \hat{\eta}_3, \hat{\beta}_0, \ldots, \hat{\beta}_3$) are interpreted as posterior, conditional log-odds-ratios. These quantities are not of particular interest for our purposes, but are used in the next step to generate posterior predictive values. This is the point where typical epidemiologic Bayesian analyses stop, because the posterior log-odds ratio is the parameter of interest in many settings.

6. Draw from the posterior predictive distribution

Draw new values of $\tilde{L}(t)$ and $\tilde{Y}(t)$ in the target population by iteratively sampling values from the models in step 1 under intervention $g(t)$. In our example, one could impute values of $\tilde{L}_i(1)$ by simulating random values from a Bernoulli distribution with a mean equal to $logit^{-1}(\hat{\eta}_0 + \hat{\eta}_1 t + \hat{\eta}_2 g + \hat{\eta}_3 \ell(0))$. Similarly, $\tilde{Y}(1)$ would be imputed as a draw from a Bernoulli distribution with a mean of $logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 g + \hat{\beta}_3(\tilde{\ell}(1, g) + \ell(0)))$. One would then similarly draw values of $\tilde{L}(2)$ using $logit^{-1}(\hat{\eta}_0 + \hat{\eta}_1 t + 2\hat{\eta}_2 g + \hat{\eta}_3(\ell(0) + \tilde{\ell}(1))$ among those with $\tilde{y}(1) = 0$. This process of sampling from conditional posterior distributions would continue iteratively for times $t = 2, \ldots, k$. Note that the posterior distributions of $\tilde{L}(t)$ and $\tilde{Y}(t)$ depend on the histories of the posterior draws up to time t.

7. Estimate the parameter of interest

For the survival outcome $Y(t)$, we are interested in causal contrasts of the posterior potential outcome distribution $p(\tilde{\bar{y}}^g \mid o)$. For example, we can estimate the posterior potential risk difference at the end of follow-up, time m, defined as: $rd^{(1,0)}(m) \equiv E(\tilde{Y}^1(m)) - E(\tilde{Y}^0(m))$, by taking the difference of the sample posterior means of the imputed $\tilde{Y}_i^g(m)$ values:

$$\widehat{rd}^{(1,0)}(m) = n^{-1}\left(\sum_{i=1}^{N} \tilde{Y}_i^1(m) - \sum_{i=1}^{N} \tilde{Y}_i^0(k)\right) = n^{-1}\sum_{i=1}^{n}\left(\tilde{Y}_i^1(m) - \tilde{Y}_i^0(m)\right).$$

The 95% posterior intervals could then be calculated by taking the 2.5th and 97.5th percentile of the posterior distribution of the risk difference or using an approximation such as taking ±1.96 times the standard deviation of the posterior distribution (a plug-in estimator of the standard error).

This step could be modified by considering only the expected values of $\{L(t, g), Y^g(t)\}$ at each time rather than simulating discrete values, as in Taubman et al's notable application of the frequentist g-formula[41].

4 Simulation study

We examine the impact of null-shrinkage priors in a Bayesian g-formula approach using 2 simple simulation studies in scenarios in which shrinkage parameters might typically be used. Our first simulation is in a time-fixed setting in which we have measured an exposure and a confounder that are highly correlated. Such settings are common in environmental epidemiology, such as in the air pollution literature where the effects of one pollutant might be difficult to estimate precisely due to high correlation with another pollutant from a similar source. Our second simulation deals with a simple, time-varying data scenario in small samples.

For both simulations, we are primarily interested in $rd^{(\bar{1},\bar{0})}$, the causal risk difference comparing "always exposed" and "never exposed" (referent). Thus, the hypothetical intervention g corresponds to exposing everyone or completely preventing exposure in a way that has no side-effects (i.e. we assume the intervention is an instrument for exposure, much in the way that randomization is an instrument for treatment in a randomized clinical trial). The target population in both scenarios is the population implied by the observed sample. We estimate $\widehat{rd}^{(\bar{1},\bar{0})}$ using both the frequentist g-formula ("standard g-formula") and the Bayesian g-formula. We calculate bias as $rd^{(\bar{1},\bar{0})} - \widehat{rd}^{(\bar{1},\bar{0})}$. For each simulation, we perform the analyses on $M = 1, 000$ simulated datasets and average the bias and mean-squared error over samples.

For the standard g-formula, the risk difference and its standard error, were estimated using a non-parametric bootstrap. For each of the M iterations, we created $S = 1, 000$ bootstrap samples of each of the M simulated datasets and calculated a risk difference using the Monte Carlo approach we described previously[21], which involves taking D draws from each bootstrap sample, where we set $D = 2000$. The risk difference for each of the M iterations was taken to be the mean risk difference across the S samples, and the standard error was calculated as the sample standard deviation for these S samples. Thus, for each analysis, the standard g-formula required simulation across $S \times M \times D = 2, 000, 000, 000$ units.

For the Bayesian g-formula, we analyzed each dataset using the MCMC procedure of SAS 9.4. Each analysis was performed using $C = 10, 000$ iterations after a burn-in of $B = 1, 000$ iterations. We estimated $\widehat{rd}^{(\bar{1}, \bar{0})}$ using the sample mean of the risk difference across all MCMC iterations, and we estimated the standard error of $\widehat{rd}^{(\bar{1}, \bar{0})}$ by taking the sample standard deviation of the posterior draws of the risk difference. For each analysis using the Bayesian g-formula, we made $(C + B) \times M = 11, 000, 000$ posterior draws.

We calculated 95% confidence/credible intervals for each analysis by as the 2.5th and 97.5th percentiles of the bootstrap/MCMC estimates. Confidence/credible interval coverage was calculated as the proportion of all datasets for which the 95% confidence/credible intervals contained the true risk difference. We also compared the standard deviation of the bias across the M samples for each approach as a measure of precision. Mean-squared error (MSE), our primary metric of comparison, was calculated as the average across the M iterations of the squared bias plus the variance. We also calculated the mean-squared (in-sample) prediction error (MSPE) as the mean of the squared difference between individual predictions of $rd^{(\bar{1},\bar{0})}$ and the true value. For most analyses MSE≈MSPE, so we only report MSE. We compared the MSE between the Bayesian g-formula and the standard g-formula using the ratio of the MSE for the Bayesian g-formula and the MSE for the standard g-formula. A value of the MSE ratio below one indicates that the Bayesian g-formula performs better than the standard g-formula in a bias-variance tradeoff (with MSE as the loss-function).

For computational considerations, all simulations are performed on samples of 100 or fewer units. Data generating mechanisms were designed to maximize precision of the parameter of interest in such small data situations. We expect that the choice between the Bayesian g-formula and the standard g-formula will be driven by the sparsity of data, which occurs even in large samples as the number of parameters grows. We expect that general statements about bias-variance tradeoff from our simulation results will apply to larger sample sizes in more realistic scenarios, though the (unknown) ideal amount of shrinkage that minimizes MSE will vary according to the data, the model, and the parameter of interest.

4.1 Time-fixed example

We simulate a simple dataset to represent a scenario in which two highly correlated, measured variables may be independently associated with an outcome, and we are interested in the independent effect of a single exposure X. The simulated data consist of a binary outcome Y, a binary exposure X and a binary covariate L, where X and L are correlated with coefficient ρ_{XL}.

The data are simulated as follows:

- $U \sim uniform(0, 1)$
- Select $\nu_1 - \nu_4$ that define the correlation coefficient $\rho_{XL} = \frac{(\nu_1 * \nu_4 - \nu_2 * \nu_3)}{\sqrt{(\nu_1 + \nu_3) * (\nu_2 + \nu_4) * (\nu_1 + \nu_2) * (\nu_3 + \nu_4)}}$ subject to the constraint $(\nu_1 + \nu_2) = (\nu_1 + \nu_3) = 0.5$.
- $L = 1$ if $U < \nu_1 + \nu_2$, 0 otherwise
- $X \sim Bernoulli(\pi)$, where $\pi = \begin{cases} \nu_1 + \nu_4 & \text{If } L = 1 \\ \nu_2 + \nu_3 & \text{If } L = 0 \end{cases}$
- $Y \sim Bernoulli(0.4 + U/10 + X * rd^{(1,0)})$

We repeated each of these simulations for values of the correlation coefficient equal to (0.4, 0.8, 0.9). Each analysis was repeated under a true risk difference of 0 and 0.2.

In a time-fixed setting, the standard g-formula requires only a single model to predict the outcome (if we take $p(\ell/\eta)$ to be the empirical distribution of L). For both the standard and the Bayesian g-formula, we predict Y using a logistic model with terms for X, and L, with coefficients β. For the model intercept we used a vague prior with a normal distribution with a mean of log(0.5) and a variance of 1000 ($\mathcal{N}(log(0.5), 1000)$). Other coefficients were given identical prior distributions of $\beta \sim \mathcal{N}(0, 3)$ (null, moderately informative priors). We also report MSE as a function of the prior variance on β, where we consider variances corresponding to upper 95% prior-odds ratios of 1.5, 3, 10, 30, 80, and 500.

As expected, the bias was larger for the Bayesian g-formula than for the standard g-formula (Table 1). The sample variance was higher for the standard g-formula estimator, however. With respect to MSE, the Bayesian g-formula under moderately informative priors outperformed the standard g-formula at correlation coefficients of 0.8 and 0.9, and 0.4. Credibile interval coverage was, in general, conservative, while confidence interval coverage in the standard g-formula was anti-conservative at higher values of the correlation coefficient. Overly precise, null-centered prior information can lead to over-shrinkage when the true risk difference is non-null, in which the MSE starts to increase as one posits increased precision (decreased variance) for the prior distributions of β for the Bayesian g-formula. For example, the minimum MSE for a risk difference of 0.2 occurred at a prior variance of 0.3. At prior variances larger than 0.3, the MSE increases due to larger posterior variance of the risk difference. At prior variances smaller than 0.3, the MSE increases due to larger bias in the posterior mean of the risk difference. When the true risk difference is null, decreasing prior variance leads to vanishing MSE because null values for individual coefficients in the Bayesian g-formula imply a null value for the posterior risk difference. (Figure 1.)

4.2 Time-varying example

We simulate a simple dataset with a time-varying exposure. The data, collected at two time points, $t = (0, 1)$, include a binary outcome measured at the end of follow up $Y(1)$, a binary exposure at two time points $\{X(0), X(1)\}$, and a binary confounder of the exposure-outcome relationship that is also affected by prior exposure $L(1)$. $L(1)$ is related to the outcome by the common cause U, which we consider to be unmeasured. This particular simulation setup has been used previously to illustrate other causal inference methods.[34] It addresses a central feature of the Bayesian g-formula: control of confounding by covariates that may also be affected by exposure.

The data are simulated as follows:

- $U \sim uniform(0.4, 0.5)$
- $X(0) \sim Bernoulli(0.5)$
- $L(1) \sim Bernoulli(logit^{-1}(-1 + X(0) + U))$
- $X(1) \sim Bernoulli(logit^{-1}(-1 + X(0) + L(1)))$
- $Y \sim Bernoulli(U + (X(0) + X(1))/2 * rd^{(\bar{1}, \bar{0})})$

In a time-varying setting, the parametric g-formula requires a model to predict $L(1)$ and $Y(1)$. For both the standard and the Bayesian g-formula, we use the correct model (i.e. the logistic model implied by the data generating mechanism) to predict $L(1)$ (with coefficients η) and we predict $Y(1)$ using a logistic model with terms for $X(0), X(1)$, and $L(1)$, with coefficients β. All model coefficients were given normal priors ($\mathcal{N}(ln(0.5), 1000)$ for intercepts and $\beta, \eta \sim \mathcal{N}(0, 3)$ for all other coefficients). Each analysis was repeated under a true risk difference of 0 and 0.2 We also report MSE as a function of the prior variance on (β, η), where we consider variances corresponding to upper 95% prior-odds ratios of 1.5, 3, 10, 30, 80, and 500.

As in our time-fixed simulations, under moderately informative, null priors, the bias was larger and the sample variance was smaller for the Bayesian g-formula than for the standard g-formula. In the samples we examined, the Bayesian g-formula outperformed the standard g-formula with respect to MSE (Table 2). As in the previous example, at increasingly higher precision of prior information, the MSE for the Bayesian g-formula can increase as a result of over-shrinkage (Figure 2.) Unlike the simulation shown in Figure 1, there is no inflection point for which a decrease or increase in prior variance will lead to an increase in the MSE. This is likely due to the small number of values we used for the prior variance, and we speculate that MSE reaches a minimum at some prior variance between 0.04 and 0.3.

5 Application

We estimated the impact of interventions on environmental tobacco smoke on childhood BMI z-scores in a prospective birth cohort. Here we propose to estimate the effects of an intervention that could affect smoking without otherwise influencing BMI (to reiterate, the intervention is assumed to be an instrument for environmental tobacco smoke exposure with respect to BMI). The Mount Sinai Children's Environmental Health and Disease Prevention

Research Center enrolled 479 primiparous women with singleton pregnancies in New York City between 1998 and 2002. The final cohort consists of 404 mother-infant pairs who met previously described inclusion criteria.[13] Children were invited to attend follow-up visits at approximately 4–5.5, 6–6.5, and 7–9 years of age (hereafter referred to as visits 1, 2, and 3, respectively) and 69 children attended all three visits. For simplicity, we assume that loss-to-follow-up is missing completely at random. Maternal baseline characteristics were ascertained via questionnaire at enrollment. At each follow-up visit, we calculated age- and sex-standardized BMI z-scores and classified children as physically active or inactive as described by Buckley et al.[5]

We estimated posterior distributions for the BMI z-score using the Bayesian g-formula approach. Our specific approach involved a pooled-logistic model for physical activity and a pooled-linear model for the BMI z-score. The predictors in the physical activity model included all baseline covariates (maternal age [quadratic polynomial], maternal pre-pregnancy BMI [quadratic polynomial], maternal height [linear], smoking during pregnancy [yes, no], race [white, black, or other], education [some high school, high school grad, some college, college grad]) as well as time-varying covariates for cumulative number of visits at which physical activity was reported [lagged one visit, range 0–2], cumulative visits at which ETS was reported [lagged one visit, range 0–2], child's age (months), and product terms for cumulative ETS and all other variables. The model for BMI z-score used identical predictors and additional terms for current ETS, cumulative years of ETS, and current PA.

For the physical activity model, we used moderately informative priors with $\mathcal{N}(0, 3)$ distributions for all covariates, which represent a moderate skepticism that we captured any strong predictor of PA. For the BMI z-score model, we used a Bayesian LASSO (least absolute shrinkage and selection operator) approach similar to that of Park and Casella [27]. Briefly, this approach utilizes a double exponential prior, assuming all regression coefficients are exchangeable. This approach differs from placing normally distributed priors on the coefficients in that it places much of the prior mass at the mean. We specified that the prior distribution for each parameter had a mean zero, again suggesting prior skepticism that any particular predictor of BMI z-score is strong. The motivation for using the Bayesian LASSO was both to demonstrate that hierarchical models may be used in the Bayesian g-formula approach as well as that the hierarchical LASSO approach allows the data to inform the amount of shrinkage that occurs (we allow that alternative prior structures may also have this property). The model for the BMI z-score at time t using the Bayesian LASSO has the following hierarchical representation (the justification of which is given in section 5 of Park and Casella (2008)[27]):

$$\mathrm{BMI}_n(t) \mid \mu, \bar{Z}(t), \beta, \sigma^2 \sim N_n(\mu 1_n + \bar{Z}(t)\beta, \sigma^2 I_n)$$

$$\beta \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim N_p(0_p \mid \sigma^2 \tau_p^2)$$

$$\sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim \pi(\sigma^2) d\sigma^2 \prod_j^p \frac{\lambda}{2} \exp\left(-\frac{\tau_j^2 \lambda^2}{2}\right) d\tau_j^2 \quad \sigma^2, \tau_1^2, \ldots, \tau_p^2 > 0$$

Where $\bar{Z}(t)$ denotes the $n \times p$ matrix of model covariates at time t and τ_p are the elements of a diagonal $p \times p$ matrix. We placed an inverse Gamma distribution prior $IG(1, 10)$ on σ^2, a Gamma distribution prior $G(1, 1)$ on λ^2, and μ was given an independent, flat prior. We note that the LASSO parameter (λ) is not fixed, which allows the data and model to inform the degree to which the β parameters are shrunk towards 0.

A linear model for the BMI z-score is standard in the literature. We assessed the fit of a linear model for the BMI z-score using the White test for heteroscedasticity,[47] which yielded an Lagrange-Multiplier statistic of 207 (p=0.45). A plot of standardized residuals versus the predicted BMI z-score is given in Appendix Figure A.1. For visits 1, 2, and 3, we estimated the median BMI z-score under the interventions “always exposed to ETS” and “never exposed to ETS”, as well as the difference in the mean BMI z-score under these interventions. We report the median of the posterior distribution of the population mean BMI z-score under each intervention (reported as “mean” or “mean z-score”). We report the 2.5th and 97.5th percentile of the posterior distributions as 95% credible intervals.

Approximately 25% of the children were exposed to ETS at visit 1, and this proportion increased to 35% at visits 2 and 3. The average BMI observed among the children in the study was slightly above the age-specific population norms, with mean z-scores between 0.5 and 0.6 at visits 1,2, and 3 (Table 3.) At all ages, we estimated that the BMI z-score was higher in the “always exposed” intervention group compared with the “never exposed” intervention group. Comparing the BMI z-score in the always exposed to never exposed (referent), the difference in the mean BMI z-score was largest at visit 2 (posterior median [95% CrI] = 0.42 [−0.22, 1.05]), though the credible intervals included the null at all ages. Our results suggest that environmental tobacco smoke is associated with higher BMI z-score throughout childhood, though the small study size precludes precise estimation in the study population.

Using a standard linear regression model adjusted for age and all baseline variables, we estimated that the BMI z-score at each visit was 0.15 higher among those actively exposed to ETS than among those not exposed (95% confidence intervals = −0.20, 0.49). After also adjusting for physical activity, this association increased to 0.22 (95% confidence intervals = −0.12, 0.57). The regression results do not correspond to potential interventions, as in the Bayesian g-formula and so the two results are difficult to compare directly. We note that the magnitude of association was smaller in the regression analysis. In the regression models, we can potentially capture an unbiased estimate of the association between BMI and the ETS behavior reported *at the current visit*. Thus, this regression model does not capture potential cumulative effects of ETS exposure at prior ages. We fit a model using cumulative ETS exposure, but this model may be subject to uncontrolled confounding by physical activity. After adjusting for physical activity, our estimate could be interpreted (under the assumptions outlined in the appendix) as a cumulative direct effect of ETS on BMI that does not operate through physical activity. Further, if there is effect measure modification by physical activity on the absolute scale, then the adjusted regression estimates will not have an easily defined interpretation with respect to informing interventions.

While our results should be seen as primarily illustrative, they show that the effects of interventions on childhood risk factors for obesity can be estimated using standard epidemiologic data, provided that the interventions act as instruments for exposure (a ‘no-side-effects’ or ‘treatment-version irrelevance’ assumption[8]). Such inference is valuable in placing the results of observational studies in real-world contexts to prioritize public health actions. We could easily compare the benefits of interventions on smoking to the benefits of interventions on physical activity or diet, for example.[41] Such comparisons are difficult with standard regression approaches, even under conditions in which estimators from regression models are unbiased.

One potential weakness of the Bayesian LASSO approach is that we assume that all model coefficients are exchangeable. It is likely that prior BMI is a much stronger predictor of current BMI than the other confounders considered, so future applications of the Bayesian g-formula could implement non-parametric Bayesian approaches that allow coefficients to shrink to different means. To assess whether our results might be sensitive to the shape of the LASSO prior, we estimated the risk difference using null-centered, normal variables ($\mathcal{N}(0, 3)$) in the model for the BMI z-score. The new priors resulted in changes to the risk difference at each visit of less than 0.03 (7%), and results were slightly more precise. Thus, even with the limitations of our illustrative approach, our results appear to be robust to a reasonable range of priors. We also examined the prior predictive distribution of the mean difference (which can be sampled from by removing the likelihood from the algorithm given in §3), which was diffuse and did not suggest a strong influence on the posterior.

6 Discussion

From a Bayesian perspective, the g-formula is a natural approach to causal inference. The frequentist version of the g-formula uses the basic rules of probability calculus, along with the language of counterfactuals and explicit identification assumptions, to make inference on the effects of treatment regimes or interventions. Other studies investigating postnatal exposure to environmental tobacco smoke in relation to child body size have also reported positive associations.[25;26;35;48] For example, in the Menorca subcohort of the Spanish INMA study, children living with one or two smoking parents at age 4 years had higher BMI z-scores at ages 4–14 years.[35] BMI z-scores were on average 0.21 (−0.05 to 0.48) or 0.27 (−0.07 to 0.61) units higher among children living with one or two smoking parents compared to nonsmoking parents, respectively.

The observed reduction of the MSE makes the Bayesian g-formula attractive for data analysis from a frequentist perspective. Our simulations suggest that, even in a limited longitudinal setting, default shrinkage priors on conditional model coefficients can result in improved frequentist results for marginal, population parameters (for whom the prior is a complex function of the model parameters and target population covariate distribution). Over-shrinkage can result in an increase in the MSE, as well. Our example in §5 shows that the Bayesian g-formula is useful in the epidemiologic setting that originally motivated the g-formula: longitudinal data in which exposures, covariates, and outcomes vary over time.[21] The approach is useful for making easily interpretable inference, even when data or the relationships between variables are highly complex. Thus, the results are easily communicated with stakeholders, public agencies, and non-statisticians.

Our approach has much in common with recent examples of Bayesian causal inference and Bayesian predictive inference. Namely, Arjas and co-authors developed an approach analogous to the Bayesian g-formula under a causal framework that does not utilize potential outcomes (described in a general setting in[2] and under a Bayesian framework in chapter 7 of Berzuini et al. [4], with a practical example given in[3 and38]). Robins refers to the approach that does not utilize the language of potential outcomes as an "agnostic" causal model.[33] In contrast, our approach is derived under the original causal framework of Rubin, which bases causal inference on the concept of potential outcomes.[37] The approach of Chib [7] is identical to the g-formula in the time-fixed setting. A previous example by Wang et al. used the Bayesian g-formula to estimate the effect of a change in ventilation practices on survival in patients with acute lung injury.[45] Notably, none of the examples mentioned focused on Bayesian inference as a way to improve the frequentist performance of causal inference methods. Using our general framework, we build on these applications by integrating the Bayesian Lasso, which demonstrates the potential utility of our approach to use other shrinkage or dimension reduction procedures to improve causal inference. The modularity of our approach allows for the straightforward incorporation of other Bayesian regression solutions to problems such as inference with highly correlated covariates,[23] or accounting for measurement error.[20]

As we show, Bayesian causal inference can be framed as a prediction problem. De Los Campos et al. [10] previously used a Bayesian LASSO approach to predict quantitative traits in wheat and mice based on genetic markers. In contrast to their approach, our work frames prediction in terms of causal inference, where the goal is not simply to predict observed outcomes, but to predict such outcomes *had we been able to change an exposure or treatment*. Interestingly, De Los Campos et al. [10] showed that the Bayesian LASSO can be used when the number of parameters exceeds the sample size, a canonical problem in high-dimensional data. Our approach could be used to extend these results to make causal contrasts in higher dimensional problems.

An characteristic feature of our approach is that we do not place priors directly on the quantity of interest. Thus, the bias-variance tradeoff of the Bayesian g-formula is not targeted towards the causal estimand. Similar claims can be made about frequentist parametric g-formula. Recent likelihood-based approaches, such as Targeted Maximum Likelihood (Minimum Loss) Estimation (TMLE), remedy this feature of standard causal inference approaches by 'targeting' the estimator at the parameter of interest, rather than the nuisance parameters that go into predicting the outcome or the treatment/exposure.[44] TMLE is typically combined with SuperLearner, which is an ensemble learning approach that averages predictions across many (possibly semi-parametric or non-parametric) models.[29] More generally, TMLE falls under the heading of "double-robust" methods, which are consistent if either the outcome model or the treatment/exposure model are correct. Such methods are typically based on restricted-moment models developed in the semi-parametric literature,[43] which do not imply a full likelihood and thus are not easily incorporated into a Bayesian framework. Graham et al. [15] uses an approach on Bootstrap methods, which has approximate Bayesian inference for the posterior predictive distribution of causal effects. Because such an approach is outside of a formal Bayesian framework, it is difficult to ascertain whether it could easily incorporate other aspects of a formal Bayesian analysis that accounts for problems such as measurement error or missing data. However, it is more robust to model misspecification than our approach. We address this shortcoming by ensuring that the posterior is flexibly parameterized, which allows for a variety of model forms. Thus, the Bayesian g-formula can be cast as one alternative to double-robust estimation as a way to reduce model specification assumptions. Bayesian model averaging[24] may be an additional way to incorporate robustness to model misspecification in our framework.

In more general settings with higher dimensional data, the need for flexibility grows even further as approximating correct model specification is more difficult. Other causal approaches, such as the use of inverse probability weighting to estimate the parameters of a marginal structural model, can marginalize over some of these parameters. However, such an approach necessitates modeling the (possibly longitudinal) exposure mechanism and a marginal model for the outcome, and thus does not completely absolve one from the modeling task. In some settings, such as studies of occupational exposures where time-varying confounders may have deterministic relationships with the exposure of interest, such alternative approaches may not be possible.[6] Using multiple causal approaches with different modeling assumptions to answer the same question can help triangulate the effects of interventions. Further work is needed regarding how to select model forms for $p(\ell(t)/\cdot)$ and $p(y(t)/\cdot)$ because the g-formula estimator is a function of both models. Our work suggests that some model selection choices can be avoided by introducing some flexibility and leveraging the shrinkage properties of Bayesian methods. This shows that previous knowledge about shrinkage estimation in standard regression models[16] may also apply to causal analysis where the estimator is a complex function of multiple parameters.

One potential concern in using the g-formula is the g-null paradox, in which model misspecification leads to hypothesis tests that inevitably reject the null hypothesis as sample size increases, even when the causal null hypothesis is true.[32] Part of this paradox is explained by considering that some choices of model form can rule out the null hypothesis *a priori* because no parameter values in the model parameter space are consistent with the causal null hypothesis - we say that such models are “incompatible.” Similar concerns may arise with prior specifications that rule out the null. Allowing for flexible models increases the parameter space but may necessitate the use of shrinkage priors, which may reduce concern over this paradox in the Bayesian setting, provided that the prior parameter space includes the “g-null” hypothesis. In the Bayesian setting, presence of the g-null paradox can be assessed by examining whether the prior predictive distribution of the potential outcomes (which is a function both of the priors, the intervention, the model, and the target population) rules out the g-null hypothesis. In general, we recommend examining the prior predictive distribution of the marginal parameter of interest to conceptualize how the joint function of the regression priors might influence the marginal parameter of interest in possibly unexpected ways.

We demonstrated that hierarchical modeling is feasible in the Bayesian g-formula, which suggests multiple future directions. Our example leveraged the naturally modular aspect of Bayesian modeling: once a probability model can be used to describe a process (e.g. measurement error, missing data), it is relatively straightforward to include that model as part of a Bayesian hierarchical model. This modular aspect of the Bayesian g-formula opens up many possibilities for improving causal inference in real-world data where some variables are not measured, and those that are measured may be recorded with error.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

National Institutes of Health Grant # 5T32ES007018-38; The Mount Sinai Children's Environmental Health Study was supported by NIEHS/EPA Children's Center grants ES09584 and R827039, The New York Community Trust, and ATSDR/CDC/ATPM

References

1. Ahern J, Hubbard A, Galea S. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. Am J Epidemiol. 2009; 169(9):1140–1147. [PubMed: 19270051]
2. Arjas E, Parner J. Causal reasoning from longitudinal data*. Scand J Stat. 2004; 31(2):171–187.
3. Arjas E, Saarela O. Optimal dynamic regimes: Presenting a case for predictive inference. Int J Biostat. 2010; 6(2):1–21.
4. Berzuini C, Dawid P, Bernardinell L. Causality: Statistical perspectives and applications. JohnWiley & Sons; 2012.
5. Buckley JP, Engel SM, Mendez MA, Richardson DB, Daniels JL, Calafat AM, Wolff MS, Herring AH. Prenatal phthalate exposures and childhood fat mass in a new york city cohort. Environ Health Perspect. 2015
6. Buckley JP, Keil AP, McGrath LJ, Edwards JK. Evolving methods for inference in the presence of healthy worker survivor bias. Epidemiology. 2015; 26(2):204–12. [PubMed: 25536456]
7. Chib S. Analysis of treatment response data without the joint distribution of potential outcomes. Journal of Econometrics. 2007; 140(2):401–412.
8. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? Epidemiology. 2009; 20(1):3–5. [PubMed: 19234395]
9. Cole SR, Hudgens MG, Brookhart MA, Westreich D. Risk. Am J Epidemiol. 2015; 181(4):246–50. [PubMed: 25660080]
10. De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes J. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 2009; 182(1):375–385. [PubMed: 19293140]
11. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. Int J Epidemiol. 2015
12. Efron B, Morris CN. Stein's paradox in statistics. Scientific American. 1977; 236(5):119–127.
13. Engel SM, Berkowitz GS, Barr DB, Teitelbaum SL, Siskind J, Meisel SJ, Wetmur JG, Wolff MS. Prenatal organophosphate metabolite and organochlorine levels and performance on the brazelton neonatal behavioral assessment scale in a multiethnic pregnancy cohort. Am J Epidemiol. 2007; 165(12):1397–404. [PubMed: 17406008]

14. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. Annu Rev Public Health. 2013; 34:61–75. [PubMed: 23297653]

15. Graham McCoyStephens. Approximate Bayesian Inference for Doubly Robust Estimation. Bayesian Analysis. 2016; 11(1):47–69.

16. Greenland S. Principles of multilevel modelling. Int J Epidemiol. 2000; 29(1):158–67. [PubMed: 10750618]

17. Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. Stat Med. 2015

18. Gustafson P. Discussion of "on bayesian estimation of marginal structural models". Biometrics. 2015; 71(2):291–293. [PubMed: 25773237]

19. Hamra GB, MacLehose RF, Cole SR. Sensitivity analyses for sparse-data problems-using weakly informative bayesian priors. Epidemiology. 2013; 24(2):233–9. [PubMed: 23337241]

20. Keil AP, Daniels JL, Hertz-Picciotto I. Autism spectrum disorder, flea and tick medication, and adjustments for exposure misclassification: the CHARGE (CHildhood Autism Risks from Genetics and Environment) case–control study. Environmental Health. 2014; 13(1):889–897.

21. Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: Intuition and a worked example. Epidemiology. 2014; 25(6):889–897. [PubMed: 25140837]

22. Keil AP, Richardson DB. Reassessing the link between airborne arsenic exposure among anaconda copper smelter workers and multiple causes of death using the parametric g-formula. Environ Health Perspect. 2016 In press.

23. MacLehose R, Dunson D, Herring A, Hoppin J. Bayesian methods for highly correlated exposure data. Epidemiology. 2007; 18(2):199–207. [PubMed: 17272963]

24. Madigan D, Raftery AE, Volinsky C, Hoeting J. Bayesian model averaging. Proceedings of the AAAI Workshop on Integrating Multiple Learned Models; Portland, OR. 1996. 77–83.

25. McConnell R, Shen E, Gilliland FD, Jerrett M, Wolch J, Chang CC, Lurmann F, Berhane K. A longitudinal cohort study of body mass index and childhood exposure to secondhand tobacco smoke and air pollution: the southern california children's health study. Environ Health Perspect. 2015; 123(4):360–6. DOI: 10.1289/ehp.1307031 [PubMed: 25389275]

26. Møller SE, Ajslev TA, Andersen CS, Dalgård C, Sørensen TIA. Risk of childhood overweight after exposure to tobacco smoking in prenatal and early postnatal life. PLoS One. 2014; 9(10):e109184.doi: 10.1371/journal.pone.0109184 [PubMed: 25310824]

27. Park T, Casella G. The bayesian lasso. J Am Stat Assoc. 2008; 103(482):681–686.

28. Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? Epidemiology. 2010; 21(6):872–5. [PubMed: 20864888]

29. Polley EC, van der Laan MJ. SuperLearner: Super Learner Prediction. 2012. URL http://CRAN.R-project.org/package=SuperLearner. R package version 2.0-9

30. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. Math Mod. 1986; 7(9):1393–1512.

31. Robins JM. Causal inference from complex longitudinal data. Latent variable modeling and applications to causality. 1997; 120:69–117.

32. Robins JM. General methodological considerations. J Econometrics. 2003; 112(1):89–106.

33. Robins JM, Greenland S. Causal inference without counterfactuals: comment. J Am Stat Assoc. 2000; 95(450):431–435.

34. Robins JM, Hernán MA. Longitudinal Data Analysis. Chapman & Hall/CRC; 2009. Estimation of the causal effects of time-varying exposures.

35. Robinson O, Martínez D, Aurrekoetxea JJ, Estarlich M, Somoano AF, Íñiguez C, Santa-Marina L, Tardón A, Torrent M, Sunyer J, Valvi D, Vrijheid M. The association between passive and active tobacco smoke exposure and child weight status among spanish children. Obesity (Silver Spring). 2016; 24(8):1767–77. DOI: 10.1002/oby.21558 [PubMed: 27367931]

36. Rosenbaum P. The consquences of adjustment for a concomitant variable that has been affected by the treatment. J R Stat Soc Ser A–G. 1984; 147:656–666.

37. Rubin D. Bayesian inference for causal effects: The role of randomization. Ann Stat. 1978; 6(1): 34–58.
38. Saarela O, Arjas E, Stephens DA, Moodie EEM. Predictive bayesian inference and dynamic treatment regimes. Biom J. 2015
39. Saarela O, Stephens DA, Moodie EE, Klein MB. On bayesian estimation of marginal structural models. Biometrics. 2015
40. Stein C, et al. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley symposium on mathematical statistics and probability. 1956; 1(399):197–206.
41. Taubman SL, Robins JM, Mittleman MA, Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. Int J Epidemiol. 2009; 38(6):1599–1611. [PubMed: 19389875]
42. Thomas DC, Witte JS, Greenland S. Dissecting effects of complex mixtures: who's afraid of informative priors? Epidemiology. 2007; 18(2):186–90. [PubMed: 17301703]
43. Tsiatis AA. Semiparametric Theory and Missing Data. Springer-Verlag; New York: 2006. Springer Series in Statistics
44. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. The International Journal of Biostatistics. 2006; 2(1):1.
45. Wang W, Scharfstein D, Wang C, Daniels M, Needham D, Brower R. the NHLBI ARDS Clinical Network. Estimating the causal effect of low tidal volume ventilation on survival in patients with acute lung injury. J R Stat Soc Ser C Appl Stat. 2011; 60(4):475–496.
46. Westreich D, Cole SR, Young JG, Palella F, Tien PC, Kingsley L, Gange SJ, Hernán MA. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident aids or death. Stat Med. 2012; 31(18):2000–2009. [PubMed: 22495733]
47. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: Journal of the Econometric Society. 1980:817–838.
48. Yang S, Decker A, Kramer MS. Exposure to parental smoking and child growth and development: a cohort study. BMC Pediatr. 2013; 13:104.doi: 10.1186/1471-2431-13-104 [PubMed: 23842036]
49. Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernán MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. Stat Biosci. 2011; 3(1):119–143. [PubMed: 24039638]

7 Appendix - Derivations

7.1 Notation

For any random variable A, if A is discrete then let $p(a)$ denote $\Pr(A = a)$, the mass of A at a. Likewise, if A is continuous then let $p(a)$ denote $f(a)$, the density of A at a. Let $p(\cdot|a)$ denote $p(\cdot|A = a)$, and likewise let $E(\cdot|a)$ denote $E(\cdot|A = a)$, where $E(\cdot)$ denotes the expectation function. In a slight abuse of notation, the integral expression $\int g(a)da$ for a function $g(a)$ of a will be used to denote integration if A is continuous, i.e., $\int g(a)da$, and summation if A is discrete, i.e., $\Sigma_{\{a\}}\ g(a)$. Suppose we have an observed sample of $i = 1, \dots, n$ measurements. Let Y, X, and L denote random variables or vectors that represent an outcome, a binary exposure or treatment such that $X = 0, 1$, and other covariates, respectively. Let $O = \{Y,X,L\}$ represent these observed quantities, where $A = \{A_1, \dots, A_n\}$ for a random variable $A = Y,X,L$. Let Y^g represent the potential outcome Y under exposure or treatment $X = g$.

7.2 Bayesian G-Formula

The mean potential outcome, the g-formula estimand of interest, may be estimated within the Bayesian framework as the posterior predictive mean potential outcome. The frequentist approach estimates the parametrized expected potential outcome $E(Y^g; \beta, \eta)$ by using the information in o to estimate β and η, e.g., as $\hat{\beta}$ and $\hat{\eta}$, respectively, in order to calculate $E(Y^g; \hat{\beta}, \hat{\eta})$ as an estimate of $E(Y^g; \beta, \eta)$. Contrast this with the following Bayesian quantity, the posterior predictive distribution of the potential outcome, which explicitly conditions on o and marginalizes over any parameters.

$$p(\widetilde{y}^g \mid o) = \int p(\widetilde{y}^g \mid \theta, o)\, p(\theta \mid o)\, d\theta. \quad \text{(A1)}$$

The Bayesian analogue of a potential outcome distribution under intervention g is

$$\begin{aligned} E\left(\widetilde{Y}^g \mid o\right) &= E_\theta\left\{E_{\widetilde{Y}^g}\left(\widetilde{Y}^g \mid \theta, o\right)\middle| o\right\} && \text{(A2)} \\ &= E_\theta\left[E_{\widetilde{L}}\left\{E_{\widetilde{Y}^g}\left(\widetilde{Y}^g \mid \widetilde{L}, \theta, o\right)\middle|\theta, o\right\}\middle| o\right] \\ &= \int E_{\widetilde{L}}\left\{E_{\widetilde{Y}^g}\left(\widetilde{Y}^g \mid \widetilde{L}, \theta, o\right)\middle|\theta, o\right\} p(\theta \mid o)\, d\theta \\ &= \int\left\{\int E\left(\widetilde{Y}^g \mid \widetilde{\ell}, \theta, o\right) p\left(\widetilde{\ell} \mid \theta, o\right) d\widetilde{\ell}\right\} p(\theta \mid o)\, d\theta \\ &= \int\left[\int\left\{\int \widetilde{y}^g p\left(\widetilde{y}^g \mid \widetilde{\ell}, \theta, o\right) d\widetilde{y}^g\right\} p\left(\widetilde{\ell} \mid \theta, o\right) d\widetilde{\ell}\right] p(\theta \mid o)\, d\theta, \end{aligned}$$

where θ denotes a set of parameters.

7.2.1 Assumptions

In all following subsections, we assume the following. We do not include the positivity assumption here, as it will first be motivated through the derivations, and then included in the summary in subsection 7.2.6. Let $\theta = \{\beta, \alpha, \eta\}$.

Conditional exchangeability— $Y^g \perp\!\!\!\perp X | L, \theta$ for $g = 0, 1$

Consistency— $Y = Y^1 X + Y^0(1 - X)$ where the support of Y^1 and Y^0 are identical.

We also make the following set of independence assumptions regarding the parameters

I1 $\{\tilde{A}\} \perp\!\!\!\perp O|\theta$ for any set $\{\tilde{A}\}$ that is a subset of $\{\tilde{Y}, g, \tilde{L}\}$. Conditional on θ, any subset of future outcomes $\{\tilde{A}\}$ of $\{\tilde{Y}, g, \tilde{L}\}$ is independent of all observed outcomes O.

I2 β, α, and η are mutually and jointly independent

I3 $Y \perp\!\!\!\perp \alpha, \eta | X, L, \beta \Leftrightarrow p(y|x, \ell, \theta) = p(y|x, \ell, \beta)$

I4 $X, L \perp\!\!\!\perp \beta | \alpha, \eta \Leftrightarrow p(x, \ell|\theta) = p(x, \ell|\alpha, \eta)$

I5 $X \perp\!\!\!\perp \eta | L, \alpha \Leftrightarrow p(x|\ell, \alpha, \eta) = p(x|\ell, \alpha)$

I6 $L \perp\!\!\!\perp \beta|\alpha, \eta \Leftrightarrow p(\ell|\theta) = p(\ell|\alpha, \eta)$ and $L \perp\!\!\!\perp \alpha|\eta \Leftrightarrow p(\ell|\alpha, \eta) = p(\ell|\eta)$

7.2.2 Conditional Distribution of $\tilde{y}$

The conditional distribution of $\tilde{y}^g$ in (A2) is

$$p\left(\tilde{y}^g \mid \tilde{\ell}, \theta, o\right) = p\left(\tilde{y}^g \mid \tilde{\ell}, y, x, \ell, \theta\right) = \frac{p\left(\tilde{y}^g, \tilde{\ell}, y, x, \ell, \theta\right)}{p\left(\tilde{\ell}, y, x, \ell, \theta\right)}.$$

The probability function in the numerator is

$$p\left(\tilde{y}^g, \tilde{\ell}, y, x, \ell, \theta\right) = \left\{p\left(\tilde{y} \mid g, \tilde{\ell}, \beta\right) p\left(y \mid x, \ell, \beta\right) \pi\left(\beta\right)\right\}\left\{p\left(\tilde{\ell} \mid \eta\right) p\left(x, \ell \mid \alpha, \eta\right) \pi\left(\alpha\right) \pi\left(\eta\right)\right\}$$

by Bayesian conditional exchangeability, consistency, I1–I4, and I6. Likewise, the probability function in the denominator is

$$p\left(\tilde{\ell}, y, x, \ell, \theta\right) = \left\{p\left(y \mid x, \ell, \beta\right) \pi\left(\beta\right)\right\}\left\{p\left(\tilde{\ell} \mid \eta\right) p\left(x, \ell \mid \alpha, \eta\right) \pi\left(\alpha\right) \pi\left(\eta\right)\right\} \quad \text{(A3)}$$

by I1–I4 and I6. Hence,

$$\begin{aligned} p\left(\tilde{y} \mid g, \tilde{\ell}, \theta, o\right) &= \frac{\left\{p\left(\tilde{y} \mid g, \tilde{\ell}, \beta\right) p\left(y \mid x, \ell, \beta\right) \pi\left(\beta\right)\right\}\left\{p\left(\tilde{\ell} \mid \eta\right) p\left(x, \ell \mid \alpha, \eta\right) \pi\left(\alpha\right) \pi\left(\eta\right)\right\}}{\left\{p\left(y \mid x, \ell, \beta\right) \pi\left(\beta\right)\right\}\left\{p\left(\tilde{\ell} \mid \eta\right) p\left(x, \ell \mid \alpha, \eta\right) \pi\left(\alpha\right) \pi\left(\eta\right)\right\}} \\ &= p\left(\tilde{y} \mid g, \tilde{\ell}, \beta\right). \end{aligned}$$

(A4)

7.2.3 Conditional Distribution of $\tilde{\ell}$

The conditional distribution of $\tilde{\ell}$ in (A2) is

$$p\left(\tilde{\ell} \mid \theta, o\right) = p\left(\tilde{\ell} \mid \eta\right) \quad \text{(A5)}$$

by I1, I2, and I6.

7.2.4 Posterior Distributions

The posterior distribution of θ can be expanded as

$$p(\theta \mid o) = \frac{p(o \mid \theta)\,\pi(\theta)}{\int p(o \mid \theta)\,\pi(\theta)\,d\theta}. \quad \text{(A6)}$$

The conditional probability function in the numerator and denominator is

$$p(o \mid \theta) = p(y \mid x, \ell, \beta)\,p(x \mid \ell, \alpha)\,p(\ell \mid \eta) \quad \text{(A7)}$$

by I3–I6. Hence,

$$p(\theta \mid o) = \left\{\frac{p(y \mid x, \ell, \beta)\,\pi(\beta)}{\int p(y \mid x, \ell, \beta)\,\pi(\beta)\,d\beta}\right\}\left\{\frac{p(x \mid \ell, \alpha)\,\pi(\alpha)}{\int p(x \mid \ell, \alpha)\,\pi(\alpha)\,d\alpha}\right\}\left\{\frac{p(\ell \mid \eta)\,\pi(\eta)}{\int p(\ell \mid \eta)\,\pi(\eta)\,d\eta}\right\}$$

by I2. Note that the posterior distribution of β is

$$p(\beta \mid y, x, \ell) = \frac{p(y \mid x, \ell, \beta)\,\pi(\beta)}{\int p(y \mid x, \ell, \beta)\,\pi(\beta)\,d\beta} \quad \text{(A8)}$$

by I2–I4. Also note that the posterior distribution of α is

$$p(\alpha \mid x, \ell) = \frac{p(x \mid \ell, \alpha)\,\pi(\alpha)}{\int p(x \mid \ell, \alpha)\,\pi(\alpha)\,d\alpha}$$

by I2, I5, and I6. We therefore have

$$p(\theta \mid o) = p(\beta \mid y, x, \ell)\,p(\alpha \mid x, \ell)\,p(\eta \mid \ell). \quad \text{(A9)}$$

7.2.5 Two Equalities

Consider the following equality, whereby the posterior predictive distribution of the potential outcomes is instead expanded as

$$E\left(\tilde{Y}^g \mid o\right) = \int E\left(\tilde{Y}^g \mid \tilde{\ell}, o\right) p\left(\tilde{\ell} \mid o\right) d\tilde{\ell}. \quad \text{(A10)}$$

The posterior predictive distribution of $\tilde{\ell}$ is

$$p\left(\tilde{\ell} \mid o\right) = \frac{\int p\left(\tilde{\ell}, \ell, y, x, \theta\right) d\theta}{\int p\left(\ell, y, x, \theta\right) d\theta}.$$

The probability function in the numerator is

$$p\left(\tilde{\ell}, y, x, \ell, \theta\right) = \{p(y \mid x, \ell, \beta)\pi(\beta)\}\{p(x \mid \ell, \alpha)\pi(\alpha)\}\left\{p\left(\tilde{\ell} \mid \eta\right)p(\ell \mid \eta)\pi(\eta)\right\}$$

by I1–I6. Likewise, the probability function in the denominator is

$$p(y, x, \ell, \theta) = \{p(y \mid x, \ell, \beta)\pi(\beta)\}\{p(x \mid \ell, \alpha)\pi(\alpha)\}\{p(\ell \mid \eta)\pi(\eta)\}$$

by I2–I6. Hence, the equality

$$p\left(\tilde{\ell} \mid o\right) = \int p\left(\tilde{\ell} \mid \eta\right)p(\eta \mid \ell)\, d\eta = p\left(\tilde{\ell} \mid \ell\right)$$

7.2.6 Summary of Calculation Requirements

The relevant parameter posterior distributions are

$$p(\beta \mid y, x, \ell) = \frac{\mathscr{L}(\beta \mid x, \ell, y)\pi(\beta)}{\int \mathscr{L}(\beta \mid x, \ell, y)\pi(\beta)\, d\beta}$$

by (A8), and

$$p(\eta \mid \ell) = \frac{\mathscr{L}(\eta \mid \ell)\pi(\eta)}{\int \mathscr{L}(\eta \mid \ell)\pi(\eta)\, d\eta},$$

where $\mathscr{L}(\cdot)$ denotes the likelihood function. Let $O_i = \{Y_i, X_i, L_i\}$ represent the observed data for measurement i. For example, for an observed sample of $i = 1, \ldots, n$ independent measurements, we have

$$\mathscr{L}(\beta \mid x, \ell, y) = \prod_{i=1}^{n} p\left(y_i \mid x_i, \ell_i, \beta\right),$$
$$\mathscr{L}(\eta \mid \ell) = \prod_{i=1}^{n} p\left(\ell_i \mid \eta\right)$$

Expressions (A4), (A5), and (A9) may now be substituted into expression (A2) for the posterior predictive mean, yielding

$$E\left(\tilde{Y}^{g} \mid o\right) = \int\left[\int\left\{\int_{\tilde{y}^{g}} \tilde{y}\, p\left(\tilde{y} \mid g, \tilde{\ell}, \theta, o\right) d\tilde{y}\right\} p\left(\tilde{\ell} \mid \theta, o\right) d\tilde{\ell}\right] p(\theta \mid o)\, d\theta$$
$$= \int\int\int\left[\int\left\{\int_{\tilde{y}^{g}} \tilde{y}\, p\left(\tilde{y} \mid g, \tilde{\ell}, \beta\right) d\tilde{y}\right\} p\left(\tilde{\ell} \mid \eta\right) d\tilde{\ell}\right] p(\beta \mid y, x, \ell)\, p(\alpha \mid x, \ell)\, p(\eta \mid \ell)$$
$$d\eta d\alpha d\beta\,.$$

(A11)

Note that (A11) is well defined only if all denominator quantities used in its definition are not equal to zero. That is, the following additional assumptions have been made implicitly up to this point.

- The expanded denominator in (A3) is greater than 0 if $p(\tilde{\ell}|\eta) > 0$, $p(y|x, \ell, \beta) > 0$, $p(x|\ell, \alpha) > 0$ by I5 and $p(\ell|\eta) > 0$ by I6, and $\pi(\theta) > 0$. Note that $p(\tilde{\ell}|\eta) > 0$ and $p(\ell|\eta) > 0$ are implied by assuming $p(\ell|\eta) > 0$ for all ℓ.
- The denominator used to define (A5) is greater than 0 if $\pi(\theta) > 0$.
- The denominator in (A6) is greater than 0 if $p(o) > 0$.

For the support (i.e., allowed values) of the random variable θ, $p(\theta) > 0$ is always true. Note that $p(o) > 0$ by definition because o represents observed quantities. Hence, the key assumption that must me made is:

Bayesian positivity—$p(\ell|\eta) > 0$ for all ℓ and η; $p(y|x, \ell, \beta) > 0$ for all y, x, ℓ, and β; and $p(x|\ell, \alpha) > 0$ for all x, ℓ, and α.

One reason to explicitly acknowledge the Bayesian positivity assumption is that it may help set realistic constraints on the possible values of $\theta = \{\beta, \alpha, \eta\}$. The following variable values and models must be observed, set, or specified.

- Observed values: y, x, ℓ
- Values set by the intervention: g
- Models specified: $p(y|x, \ell, \beta)$, $p(\ell|\eta)$, $\pi(\beta)$, $\pi(\eta)$

Importantly, note that $\tilde{\ell}$ need not be set because it is integrated out.

7.2.7 Some Binary Outcome Results

Suppose the outcome is binary such that $\tilde{Y}^g = 0, 1$ for $x = 0, 1$. Hence,

$$E\left(\tilde{Y}^{g} \mid z\right) = \Pr\left(\tilde{Y}^{g} = 1 \mid z\right)$$

by the definition of an expectation for any set of variable realizations z (including the empty set). Suppose assumptions Bayesian conditional exchangeability, consistency, and I1–I6 are true. From the equivalence of lines 1 and 2 of (A11), $E(\tilde{Y}^g|\theta, o) = E(\tilde{Y}^g|\theta)$. Hence, (A2) is therefore equal to

$$\Pr\left(\tilde{Y}^g = 1 \mid o\right) = \int \Pr\left(\tilde{Y}^g = 1 \mid \theta\right) p\left(\theta \mid o\right) d\theta .$$

Equivalently,

$$p\left(\tilde{y}^g \mid o\right) = \int p\left(\tilde{y}^g \mid \theta\right) p\left(\theta \mid o\right) d\theta$$

Similarly, (A10) is equal to

$$\Pr\left(\tilde{Y}^g = 1 \mid o\right) = \int \Pr\left(\tilde{Y}^g = 1 \mid \tilde{\ell}, o\right) p\left(\tilde{\ell} \mid o\right) d\tilde{\ell}$$

Equivalently,

$$p\left(\tilde{y}^g \mid o\right) = \int p\left(\tilde{y} \mid g, \tilde{\ell}, o\right) p\left(\tilde{\ell} \mid o\right) d\tilde{\ell}$$

by conditional exchangeability and consistency. Finally, from the equivalence of lines 1 and 2 of (A11) and the penultimate line of (A2),

$$\Pr\left(\tilde{Y}^g = 1 \mid o\right) = \int \left\{ \int \Pr\left(\tilde{Y} = 1 \mid g, \tilde{\ell}, \theta\right) p\left(\tilde{\ell} \mid \theta\right) d\tilde{\ell} \right\} p\left(\theta \mid o\right) d\theta$$

by conditional exchangeability and consistency. Equivalently,

$$p\left(\tilde{y}^g \mid o\right) = \int \int p\left(\tilde{y} \mid g, \tilde{\ell}, \theta\right) p\left(\tilde{\ell} \mid \theta\right) p\left(\theta \mid o\right) d\theta d\tilde{\ell}$$

and

$$p\left(\tilde{y}^g \mid o\right) \propto \int \int \int p\left(\tilde{y} \mid g, \tilde{\ell}, \beta\right) p\left(\tilde{\ell} \mid \eta\right) \mathscr{L}\left(\beta \mid y, x, \ell\right) \pi\left(\beta\right) \mathscr{L}\left(\eta \mid \ell\right) \pi\left(\eta\right) d\beta d\eta d\tilde{\ell}$$

by (A9), I3, and I6. Suppose Y^g (and therefore, Y by consistency), X, and L are binary such that $y^g = 0, 1$, $x = 0, 1$, and $\ell = 0, 1$. For an observed sample of $i = 1, \ldots, n$ independent measurements, we have

$$\mathscr{L}(\theta \mid o) = \prod_{i=1}^{n} \Pr\left(Y_i = 1 \mid X_i = x_i, L_i = \ell_i, \beta\right)^{y_i} \Pr\left(Y_i = 0 \mid X_i = x_i, L_i = \ell_i, \beta\right)^{1 - y_i}$$
$$\times \Pr\left(X_i = 1 \mid L_i = \ell_i, \alpha\right)^{x_i} \Pr\left(X_i = 0 \mid L_i = \ell_i, \alpha\right)^{1 - x_i} \times \Pr\left(L_i = 1 \mid \eta\right)^{\ell_i} \Pr\left(L_i = 0 \mid \eta\right)^{1 - \ell_i}$$

by (A7). Note that

$$\mathscr{L}(\theta \mid o) = \mathscr{L}(\beta \mid y, x, \ell)\,\mathscr{L}(\alpha \mid x, \ell)\,\mathscr{L}(\eta \mid \ell)$$

by (A7).

7.3 Time-varying quantities

Given a set of longitudinal data $O(t) \equiv (\bar{L}(t), \bar{X}(t), \bar{Y}(t))$, the Bayesian g-formula for the posterior predictive distribution of the potential outcomes under a static intervention g is given as proportional to:

$$p\left(\tilde{y}^g(t) \mid o\right) \propto \int_\beta \int_\eta \left[\int_{\breve{y}(t-1)} \int_{\breve{\ell}(t)} \prod_{j=0}^{t} \left[p\left(\tilde{y}(j) \mid g, \breve{\ell}(j), \breve{y}(j-1), o, \beta\right) \times p\right.\right. \tag{A12}$$
$$\left.\left.\left(\tilde{\ell}(j) \mid g, \breve{\ell}(j-1), \breve{y}(j-1), o, \eta\right)\right] d\breve{y}\, d\breve{\ell}\right] \mathscr{L}(\beta \mid o(t))\, \mathscr{L}(\eta \mid o(t))\, \pi(\beta)\, \pi(\eta)\, d\beta d\eta$$
$$\propto \int_\theta \int_{\breve{y}(t-1)} \int_{\breve{\ell}(t)} \prod_{j=0}^{t} \left[p\left(\tilde{y}(j) \mid g, \breve{\ell}(j), \breve{y}(j-1), o, \beta\right) \times p\right.$$
$$\left.\left(\tilde{\ell}(j) \mid g, \breve{\ell}(j-1), \breve{y}(j-1), o, \eta\right)\right] \mathscr{L}(\theta \mid o(t))\, \pi(\theta)\, d\theta d\breve{y}\, d\breve{\ell}$$

where the vector of posterior predictions is given as $\breve{A}(t) \equiv (\tilde{A}(0), \ldots, \tilde{A}(t))$ and $\int_{\breve{a}(k)} \equiv \int_{\tilde{a}(0)}, \ldots, \int_{\tilde{a}(k)}$. We now show that this definition comes from a the full likelihood of the parameters $\theta \equiv (\alpha, \eta, \beta)$, given the data O, the assumptions of positivity and conditional exchangeability, with consistency linking the observed outcomes with the potential outcomes, and the definition of a posterior predictive distribution.

We can express the likelihood under a set of correctly specified models for $L(t)$, $X(t)$ and $Y(t)$ as

$$\mathscr{L}(\theta \mid o(t)) = \left[p(y(t) \mid \bar{x}(t), \overline{\ell}(t-1), \bar{y}(t-1), \beta) \times \prod_{j=1}^{t} \left[p(x(j) \mid , \bar{x}(j-1), \overline{\ell}(j), \bar{y}(j-1), \alpha) \times p(\ell(j) \mid \bar{x}(j-1), \overline{\ell}(j\right.\right.$$
$$\left.\left. - 1), \bar{y}(j-1), \eta) \times p(y(j-1) \mid \bar{x}(j-1), \overline{\ell}(j-2), \bar{y}(j-2), \beta) \times\right] p(\ell(0), \eta)\right]$$

Where we define $(\bar{\ell}(-1), \bar{y}(-1)) = \varnothing$. We also note that, under a static intervention/regime in which we set X to g, $p(x_j |, \bar{x}(j-1), \bar{\ell}(j), \bar{y}(j-1), \alpha)$ is a degenerate distribution with probability 1 at $x = g$, and the posterior will be independent of α. Thus, we can reduce our expression for the likelihood to

$$\mathscr{L}(\theta \mid o(t)) = \mathscr{L}(\beta \mid o(t))\mathscr{L}(\eta \mid o(t)) = \prod_{j=0}^{t} \left[p(y(j) \mid \bar{x}(j), \overline{\ell}(j), \bar{y}(j-1), \beta) \times p(\ell(j) \mid \bar{x}(j-1), \overline{\ell}(j-1), \bar{y}(j-1), \eta)\right.$$

$$\left.\right]$$

We define a new assumption of sequential conditional exchangeability (no unmeasured confounding) as

$$Y^{g}(j) \perp\!\!\!\perp X(j) \mid \bar{L}(j), \bar{Y}(j-1), \bar{X}(j-1) = g, \theta \text{ for } g = 0, 1, j = 1, \ldots, t$$

where, for simplicity of notation, we redefine g in a longitudinal setting to mean that we set $X(j)$ to g for $j = 0, \ldots, t$.

The predictive distribution of interest is the joint posterior predictive distribution of the potential outcomes given by $p\,(\tilde{y}^g(0), \ldots, \tilde{y}^g(t)|o)$, which can be expressed as a complex function of the posterior predictive joint distribution of the Y and L at a fixed value for the exposure $X = g$, given as $p(\tilde{y}(0), \ldots, \tilde{y}(t), \tilde{\ell}(0), \ldots, \tilde{\ell}(t)|g, o)$.

Under the above assumptions, a draw from the posterior predictive distribution of the potential outcome at time t can be expressed as a recursive application of the law of total probability and the "one step ahead innovations" of Robins [31] (Theorem 3.2; This proof is also given more explicitly by Young et al. [49]). The posterior predictive generating function (which takes as inputs posterior draws of the coefficients and outputs the posterior predictive distribution of covariates) for a fixed (β, η, g) is given as

$$p\left(\tilde{y}^{g}(t) \mid o, \beta, \eta\right) =$$

$$\int_{\breve{y}(t-1)} \int_{\breve{\ell}(t)} \prod_{j=0}^{t} \left[p\left(\tilde{y}(j) \mid g, \breve{\ell}(j), \breve{y}(j-1), o, \beta\right) \times p\left(\tilde{\ell}(j) \mid g, \breve{\ell}(j-1), \breve{y}(j-1), o, \eta\right)\right] d\breve{y}\, d\breve{\ell}$$

Outputs from this function can be recursively drawn in order of time. For example, at a given value of (β, η, g) we would first draw a value from $p(\tilde{\ell}(0), /o, \eta)$, next, we would draw from $p(\tilde{y}(0)/g, \tilde{\ell}(0), o, \beta)$, followed by a draw from $p(\tilde{\ell}(1)/g, \tilde{y}(0), \tilde{\ell}(0), o, \beta)$, and so on through time t. The conditioning on o is implicit because we draw values of (β, η) from the posterior distribution $p(\beta, \eta/o)$.

It follows by definition A1 that the posterior predictive distribution of the potential outcome $Y^g(t)$ under intervention $\bar{g}$ is given by the Bayesian g-formula as proportional to A12.

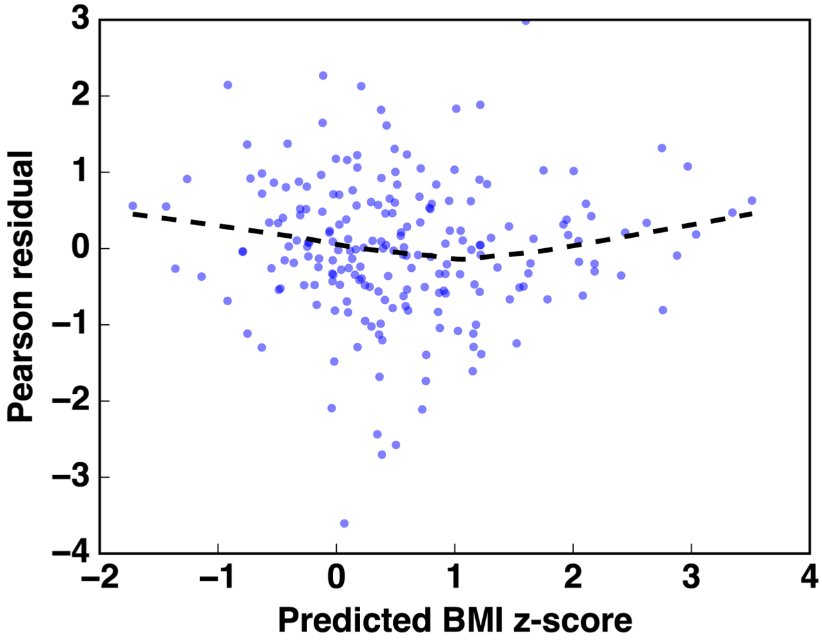

Figure A.1.

Pearson residuals versus predicted BMI z-score from the linear model to predict BMI z-score in the example given in §5. Dashed line represents a LOESS line-of-fit to the data shown in the figure.

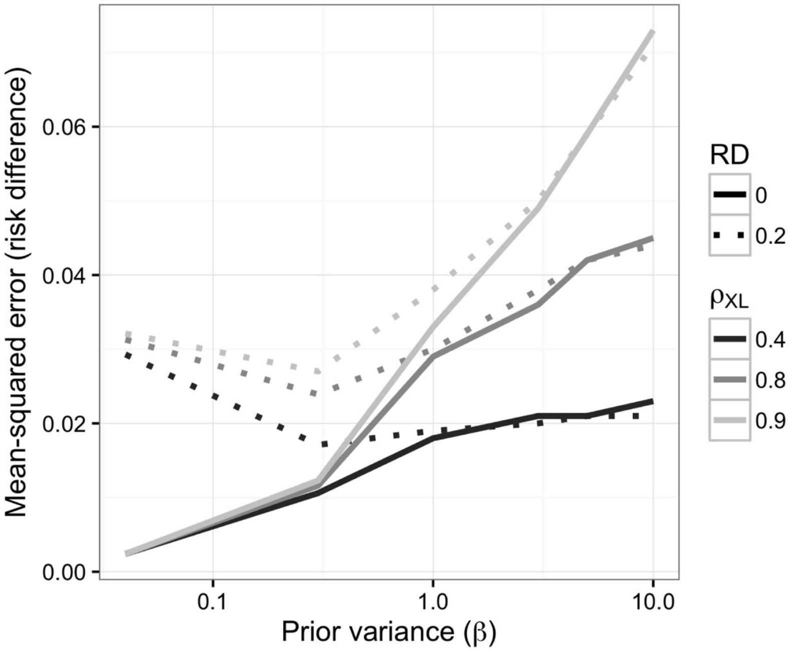

Figure 1.

Mean-squared error for the risk difference as a function of correlation between X and L and of the prior variance on regression coefficients (log-odds ratios) for simulation given in §4.1

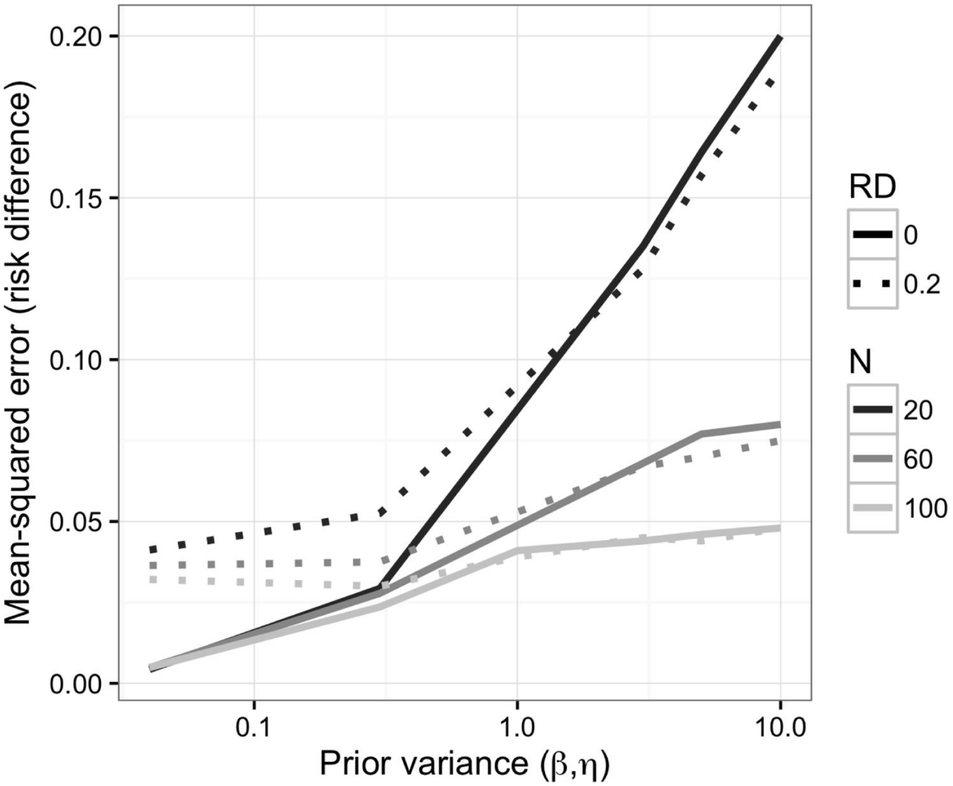

Figure 2.

Mean-squared error as a function of sample size and the prior variance on regression coefficients for simulation given in §4.2

Table 1

Simulation scenario 1: Two correlated exposures, one time-fixed confounder model intercept priors were vague ($\mathcal{N}(ln(0.5), 1000)$) and other coefficients were null centered and moderately informative ($\mathcal{N}(0, 3)$)

Method	Correlation(X,L)	True RD	Mean bias	SD bias	MSE	Coverage	MSE ratio
Standard	0.9	0	−0.01	0.24	0.12	0.92	1
Bayes	0.9	0	−0.01	0.14	0.05	0.98	0.42
Standard	0.9	0.2	0.00	0.23	0.11	0.89	1
Bayes	0.9	0.2	−0.06	0.14	0.05	0.98	0.47
Standard	0.8	0	0.00	0.17	0.06	0.94	1
Bayes	0.8	0	−0.01	0.13	0.04	0.97	0.59
Standard	0.8	0.2	−0.01	0.17	0.06	0.93	1
Bayes	0.8	0.2	−0.04	0.13	0.04	0.96	0.63
Standard	0.4	0	0.00	0.11	0.02	0.94	1
Bayes	0.4	0	0.00	0.10	0.02	0.95	0.88
Standard	0.4	0.2	0.01	0.11	0.02	0.94	1
Bayes	0.4	0.2	−0.01	0.10	0.02	0.95	0.87

Table 2

Simulation scenario 2: One exposure, one time-varying confounder that depends on exposure. Model intercept priors were vague ($\mathcal{N}(ln(0.5), 1000)$) and other coefficients were null centered and moderately informative ($\mathcal{N}(0, 3)$)

Method	N	True RD	Mean bias	SD bias	MSE	Coverage	MSE ratio
Standard	20	0	−0.01	0.41	0.36	0.92	1
Bayes	20	0	−0.01	0.25	0.14	0.97	0.37
Standard	20	0.2	0.02	0.37	0.33	0.94	1
Bayes	20	0.2	−0.06	0.23	0.13	0.98	0.39
Standard	60	0	0.00	0.21	0.09	0.95	1
Bayes	60	0	0.00	0.18	0.07	0.96	0.76
Standard	60	0.2	0.00	0.21	0.09	0.94	1
Bayes	60	0.2	−0.02	0.18	0.07	0.96	0.77
Standard	100	0	−0.01	0.16	0.05	0.96	1
Bayes	100	0	−0.01	0.15	0.04	0.95	0.85
Standard	100	0.2	−0.01	0.17	0.05	0.94	1
Bayes	100	0.2	−0.02	0.15	0.05	0.94	0.87

Table 3

Comparing childhood BMI z-scores after potential interventions on environmental tobacco smoke exposure in 69 children, New York, USA 2004 to 2008.

	Observed	Intervention		
Visit (age[a])	Mean z-score (95% CI)	Never exposed Mean z-score (95% CrI)	Always exposed Mean z-score (95% CrI)	Difference Mean (95% CrI)
1 (5)	0.52 (0.22, 0.81)	0.35 (0.13, 0.57)	0.64 (0.27, 1.02)	0.29 (−0.14, 0.73)
2 (6)	0.55 (0.29, 0.80)	0.46 (0.19, 0.73)	0.88 (0.32, 1.43)	0.42 (−0.22, 1.05)
3 (7)	0.59 (0.32, 0.85)	0.44 (0.10, 0.79)	0.80 (0.07, 1.54)	0.36 (−0.49, 1.21)

[a] Mean age at visit