Introduction to the Parametric G Formula

Ashley I. Naimi, PhD

June 20, 2017

Abstract

Applied health scientists are increasingly dealing with complex data structures to answer questions about exposure effects and mediation. In such settings, feedback between confounders, exposures, and mediators render standard adjustment methods (regression, restriction, stratification, matching) inappropriate. The parametric g formula—one of three "g" methods—is a versatile tool that can be used to quantify a variety of exposure effects with complex data structures. This workshop will provide a comprehensive overview of the g formula for identifying and estimating causal effects. After a brief introduction to the potential outcomes framework, we will review obstacles to effect estimation and mediation analysis with complex longitudinal data. The g formula will then be introduced with three examples using actual data and software code: (i) a simple simulated analysis that minimizes technical details and emphasizes core concepts; (ii) a mediation analysis setting where interest lies in direct/indirect effects; and (iii) a complex longitudinal data setting where interest lies in estimating the total effect of an exposure measured repeatedly over many months of follow-up. The goal of this workshop will be to enable participants to implement the parametric g formula in a range of settings, to articulate and evaluate key assumptions/limitations, and to implement critical model validation techniques. No prior knowledge of causal modeling, counterfactuals, or g methods is required.

Outline

Causal Inference

- Introduction
- Complex Longitudinal Data
- Notation
- Estimand, Estimator, Estimate
- Identifiability
    a. Counterfactual Consistency
    b. No Interference
    c. Excheangability
    d. Correct Model Specification
    e. Positivity

The Parametric G-Formula

- Example 0: Model-Based Standardization with Time Fixed Data
- Example 1: Model-Based Standardization with Time-Varying Data (g formula)
- Example 2: National Survey of Family Growth Mediation analysis
- Example 3: Effects of Aspirin on Gestation and Reproduction Per Protocol Effect Estimation

Appendix

- Abstraction of Steps Required to Implement Parametric G formula
- Bootstrap for CI estimation

Causal Inference

Introduction

"Causal inference" deals primarily with the formal mechanisms by which we can combine data, assumptions, and models to interpret a correlation (or association) as a causal relation.[1] The framework by which we define what we mean by "causal relation" or "causal effect" is the **potential outcomes framework**.

[1] There are a number of excellent introductory books and articles on causal inference in the empirical sciences. Here are some excellent options: M. A. Hernán and Robins (Forthcoming), Pearl, Glymour, and Jewell (2016), Imbens and Rubin (2015)

A central notion in the potential outcomes framework is the counterfactual. This notion stems from the intuitive and informal practice of interpreting cause-effect relations as **circumstances (e.g., health outcomes) that would have arisen had things (e.g., exposures) been different**.

While this intuition serves an important purpose, it is not sufficient for doing rigorous science. Suppose we ask: "what is the effect of smoking on CVD risk, irrespective of smoking's effect on body weight?" This question seems clear and intiutive. To answer this question, we would do a study in which we collect data, enter these into a computer, perform some calculations, and obtain a number (the "effect").

But there is a problem.[2] The calculations performed by the computer are **rigorously defined mathematical objects**. On the other hand, **english language sentences about cause effect relations are ambiguous**. For example, the "effect of smoking" can mean many different things:

[2] This problem was articulated by Robins (1987), and I am using the example from his paper.

- All people smoke any tobacco ever versus no people smoke tobacco ever.
- All people smoke 3 cigarettes per day versus all people smoke 2 cigarettes per day.
- All people who have smoked any tobacco in the last 15 years cease to smoke any tobacco whatsoever.

Similarly, "irrespective of" can mean a number of things:

- The effect of smoking on CVD risk that would be observed in a hypothetical world where smoking did not affect body mass?

- The effect of smoking on CVD risk if everyone were set to "normal" body mass?
- The effect of smoking on CVD risk if everyone were held at the body mass they had in the month prior to study entry?

But the numerical strings of data and the computer algorithms applied to these data are well defined mathematical objects, which do not admit such ambiguity. Depending on several choices, including the data, how variables are coded, and the modeling strategy, the computer is being told which question to answer. There is a lot of potential uncertainty in the space between the English language sentences we use to ask causal questions, and the computer algorithms we use to answer those questions. Causal inference is about clarifying this uncertainty.

Complex Longitudinal Data

This short course is about complex longitudinal data, so let's define that here. We will be dealing with data from a cohort study, individuals sampled from a well-defined target population, and clear study start and stop times (i.e., closed cohort). Data from such a cohort are **longitudinal** when they are measured repeatedly over time.[3]

[3] Another such form is when data are measured repeatedly across space. We will not be dealing with these data here.

Different scenarios can lead to longitudinal data:

1. exposure and covariates do not vary over time, but the study outcome can occur more than once
2. exposure and covariates vary over time, but the study outcome can only occur once
3. exposure and covariates vary over time, and the study outcome can occur more than once

We will deal with data that from scenario 2 (however, it is not difficult to generalize the logic to scenario 3). Repeated exposure, covariate, and/or outcome measurement is what leads to "longitudinal" data. But why complex?

Repeated measurement over time opens up the possibility of complex causal relations between past and future covariates. Suppose we measure an expsoure twice over follow-up, a covariate once,

and the outcome at the end of follow-up (Figure 1). If we can assume that past exposure/covariate values do not affect future exposure/covariate values (usually a very risky assumption), we might not consider these data "complex," becuase we can use many standard methods we already know to analyze these data.

Figure 1: Longitudinal data that might not be considered 'complex' because there is no feedback between exposure and covariates.

On the other hand, if past exposure/covariates affect future exposure/covariates in such a way that prior exposures or covariates confound future exposures (Figure 2), more advanced analytic techniques are needed.

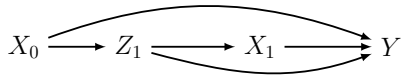


Figure 2: The simplest kind of complex longitudinal data. Note that the exposure at time zero affects the covariate at time 1 which affects the exposure at time 1. This feedback leads to confounding of the time 1 expsoure by a covariate that is affected by the prior exposure. Analysis of these data require more general methods to account for this complex form of confounding.

In this short course, we will learn how to use the parametric g formula to account for this type of complex time-varying confounding.

Notation

The building blocks for causal inference are **potential outcomes** (Rubin 2005). These are conceptually distinct from **observed outcomes**. Potential outcomes are functions of exposures. For a given exposure x, we will write the potential outcome as Y^x.[4] **This is interpreted as "the outcome (Y) that would be observed if X were set to some value x"**. For example, if X is binary [denoted $X \in (0,1)$], then Y^x is the outcome that would be observed if $X = 0$ or $X = 1$. If we wanted to be specific about the value of x, we could write $Y^{x=0}$ or $Y^{x=1}$ (or, more succinctly, Y^0 or Y^1).

[4] Alternate notation includes: Y_x, $Y(x)$, $Y \mid Set(X = x)$, and $Y|do(X = x)$.

Study Question 1: Suppose you collect data from a single person

and find that they are exposed. Can you interpret their outcome to be the potential outcome that would have been observed had they been exposed? Why or why not?

When the exposoure and/or outcome are measured repeatedly over follow-up, notation must account for that. We thus use subscripts to denote when the variable was measured. For example, if the exposure is measured twice, we can denote the first measurement X_0 and the second X_1. Additionally, we use overbars to denote the history of a variable over follow-up time. For example, $\overline{X}_1$ denotes the set $\{X_0, X_1\}$. More generally, for some arbitrary point over follow-up m, $\overline{X}_m$ denotes $\{X_0, X_1, X_2, \ldots X_m\}$. We can then define potential outcomes as a function of these exposure histories: For two exposure measurements, $\overline{X}_j = \{1, 1\}$, $Y^{\overline{x}_j = \overline{1}}$ is the outcome that would be observed if X_0 were set to 1 and X_1 were set to 1.

Estimand, Estimator, Estimate

Causal inference starts with a clear idea of the effect of interest (the target causal parameter). To do this, it helps to distinguish between estimands, estimators, and estimates.

STUDY QUESTION 2A: You are familar with the well known odds ratio equation for a 2×2 table: (ab/cd). Is this an estimand, estimator, or estimate?

The **estimand** is the (mathematical) object we want to quantify. It is, for example, the causal risk difference, risk ratio, or odds ratio for our exposure and outcome of interest. In our smoking CVD example, we might be interested in:

$$P(Y^1 = 1) - P(Y^0 = 1), \quad \frac{P(Y^1 = 1)}{P(Y^0 = 1)}, \quad \frac{Odds(Y^1 = 1)}{Odds(Y^0 = 1)},$$

where $Odds(Y^x = 1) = P(Y^x = 1)/(Y^x = 0)$. There are many others besides these.

STUDY QUESTION 2B: List some estimators that can be used to quantify the odds ratio.

The estimand is the object we want to estimate. The **estimator** is an equation that allows us to use our data to quantify the estimand. Suppose, for example, we were explicitly intersted in quantifying the causal risk difference for the relation between smoking and CVD risk. To do this, we have to start by quantifying the associational risk difference, but there are many ways to do this, including ordinary least squares, maximum likelihood, or the method of moments.

To be specific, let's simulate some hypothetical data on the relation between smoking and CVD. Let's look at ordinary least squares and maximum likelihood as estimators:

```
### CODE SET 1
# define the expit function
expit<-function(z){1/(1+exp(-(z)))}
set.seed(123)
n<-1e6
confounder<-rbinom(n,1,.5)
smoking<-rbinom(n,1,expit(-2+log(2)*confounder))
CVD<-rbinom(n,1,.1+.05*smoking+.05*confounder)


round(mean(confounder),3)

## [1] 0.499

round(mean(smoking),3)

## [1] 0.166

round(mean(CVD),3)

## [1] 0.133

#OLS
round(coef(lm(CVD~smoking+confounder)),4)
```

```
## (Intercept)     smoking  confounder
##      0.1000      0.0485      0.0501
```

```
#ML1
round(coef(glm(CVD~smoking+confounder,family=poisson("identity"))),4)
```

```
## (Intercept)     smoking  confounder
##      0.0999      0.0487      0.0502
```

```
#ML2
round(coef(glm(CVD~smoking+confounder,family=binomial("identity"))),4)
```

```
## (Intercept)     smoking  confounder
##      0.1000      0.0487      0.0501
```

```
### END CODE SET 1
```

In our simple setting with 1 million observations, ordinary least squares and maximum likelihood yielded the same associational risk difference (as expected) even though they are different **estimators**. Finally, the values obtained from each regression approach are our **estimates**.

Identifiability

In our simulation example, we estimated the associational risk difference using three different estimators. Estimating associations is all we can do with empirical data. But we want to use the associational risk difference to quantify the causal risk difference. We can only do so if the causal risk difference is **identified**. A parameter (e.g., causal risk difference) is identified if we can write it as a function of the observed data.

The causal risk difference is defined as a contrast of potential outcomes. Referring back to our simulated example, we want to estimate the causal risk difference conditional on C:

$$P(Y^1 = 1 \mid C) - P(Y^0 = 1 \mid C),$$

where Y^1, Y^0 are the potential CVD outcomes that would be observed if smoking were set to 1 and 0, respectively. On the other

hand, the associational risk difference is defined as a contrast of observed outcomes:

$$P(Y = 1 \mid X = 1, C) - P(Y = 1 \mid X = 0, C),$$

where each term in this equation is interpreted as the risk of CVD **among those who had** $X = x$. The causal risk difference is identified if the following equation holds:[5]

$$P(Y^x = 1 \mid C) = P(Y = 1 \mid X = x, C)$$

which says that the risk of CVD that would be observed if everyone were set to $X = x$ is equal to the risk of CVD that we observe among those with $X = x$. In this equation, the right hand side equation is written entirely in terms of observed data ($Y = 1$). The left hand side is a function of unobserved potential outcomes ($Y^x = 1$). This equivalence will only hold if we can make some assumptions.

The first is **counterfactual consistency**, which states that the potential outcome that would be observed if we set the exposure to the observed value is the observed outcome (Miguel A. Hernán 2005,Hernan and Taubman (2008),Miguel A Hernán and VanderWeele (2011),VanderWeele and Hernán (2013)).[6] Formally, counterfactually consistency states that:

$$\text{if } X = x \text{ then } Y^x = Y$$

The status of this assumption remains unaffected by the choice of analytic method (e.g., standard regression versus g methods). Rather, this assumption's validity depends on the nature of the exposure assignment mechanism.

We must also assume **no interference**, which states that the potential outcome for any given individual does not depend on the exposure status of another individual (M. G. Hudgens and Halloran 2008,Ashley I. Naimi and Kaufman (2015)). If this assumption were not true, we would have to write the potential outcomes as a function of the expsoure status of multiple individuals. For example, for two different people indexed by i and j, we might write: $Y_i^{x_i, x_j}$.[7] Notation

[5] Throughout this course, we will assume that the target parameter of interest is a causal contrast of potential outcomes. Sometimes, the target parameter of interest is an associational contrast, and the assumptions needed are less demanding. See, e.g., Ashley I. Naimi et al. (2016).

[6] While somewhat convoluted, this assumption is about legitimizing the connection between our observational study, and future interventions in actual populations. In our observational study, we **see** people with with a certain value of the exposure. In a future intervention, we **set** people to a certain value of the exposure.

[7] Together, counterfactual consistency and no interference make up the stable-unit treatment value assumption (SUTVA), first articulated by D. B. Rubin (1980).

and methods that account for interference can be somewhat complex (E. J. Tchetgen Tchetgen and VanderWeele 2012,M. E. Halloran and Hudgens (2016)), and we will not consider the impact of interference here.

Together, counterfactual consistency and no interference allow us to make some progress in writing the potential risk $P(Y^x = 1 \mid C)$ as a function of the observed risk $P(Y = 1 \mid X = x, C)$. Specifically, by counterfactual consistency and no interference, we can do the following:

$$P(Y = 1 \mid X = x, C) = P(Y^x = 1 \mid X = x, C)$$

A third assumption is **exchangeability**, which implies that the potential outcomes under a specific exposure (Y^x) are independent of the observed exposures X (Greenland and Robins 1986,Greenland, Robins, and Pearl (1999),Greenland and Robins (2009)). If this holds, then we have:

$$P(Y^x = 1 \mid X = x, C) = P(Y^x = 1 \mid C)$$

If there is any confounding, selection, or information bias, the potential outcome will be associated with the observed exposure, and we cannot remove $X = x$ from the conditioning statement.[8] What this means is that the exposure is predictive of prognosis, independent of it's actual effect on the outcome.

[8] For an excellent discussion of why the potential outcomes are independent of the observed exposure under exchangeability, see Chapter 2 of M. A. Hernán and Robins (Forthcoming)

Study Question 3: Why is the word "exchangeable" used to describe this concept? What, precisely, is being "exchanged"?

Although it seems that we have successfully written the potential risk as a function of the observed data, we are in need of two more assumptions. The first is **correct model specification**. This assumption is required when we rely on models to estimate effects, but can be minimized or avoided by using semi- or non-parametric approaches. There are several ways in which this assumption can be

violated, and these include the omission of relevant interaction terms, or adjusting for continuous covariates using linear terms only.

STUDY QUESTION 4: Can you think of a relation between correct model misspecification and exchangeability?

The second is **positivity**,[9] and requires exposed and unexposed individuals within all levels of the confounder (Mortimer et al. 2005,Westreich and Cole (2010)). There are two kinds of positivity violations (non-positivity): structural (or deterministic) and stochastic[10] (or random). Structural non-positivity occurs when individuals with certain covariate values cannot be exposed. For example, in occupational epidemiology work-status (employed/unemployed in workplace under study) is a confounder, but individuals who leave the workplace can no longer be exposed to a work-based exposure. Alternatively, stochastic non-positivity arises when the sample size is not large enough to populate all confounder strata with observations. When faced with positivity violations, methods must be used that are not affected.[11] These include g estimation of a structural nested model, collaborative targeted minimum loss-based estimation, and the parametric g formula.

[9] Also known as the experimental treatment assignment assumption.

[10] The word **stochastic** is derived from the greek word "to aim," as in "to aim for a target."

[11] Warning: one cannot simply "avoid" positivity. In an extreme setting, non-positivity means that those who were exposed in the sample are very unlikely to be exposed (and vice versa). In such a situation, it may not make sense to estimate the average treatment effect, because there is a subset of the population who may never actually be exposed. In this case, g estimation, cTMLE, and the parametric g formula actually estimate parameters that differ slightly from the ATE.

The Parametric G Formula

In this section, we will illustrate implementation of the parametric g formula using four examples with simulated and empirical data. The first will be a very simple setting with one exposure, one confounder, and one outcome. This example will demonstrate model-based standardization, which is essentially what the parametric g formula does with complex longitudinal data. However, the data from the first example are neither complex nor longitudinal.

The second example will be identical to the first, except the exposure will be measured twice (time-varying). It will also include a time-varying confounder measured once, but that creates a feedback loop between the first and second exposure measurement. This is the simplest complex longitudinal data scenario in which one can implement the g formula, and we will use it to emphasize core concepts.

In the first two examples, we will establish a series of procedures to implement the g formula in a wide range of settings. Specifically, we will discuss problem setup, implementation, validation, and interpretation. The setup stage is about what you need to write down and organize to implement the parametric g formula. In the implementation stage, I will show you what models you need to fit based on the setup. After fitting these models, we need to evaluate quality (validation stage). Finally we must interpret in light of the assumptions we covered in the previous section.

The third example will be with actual data from the National Survey of Family Growth. We will answer questions about total, direct, and indirect effects (causal mediation).

Our final example will be the most complex. We will use data based on a randomized trial of daily low dose aspirin on pregnancy outcomes. We will deal with multiple time-points, and multiple competing outcomes. I will show you how to tailor implementation of the the g formula to a given study design.

The parametric g formula is the first of three "g" methods developed by James Robins beginning in the mid-1980s. The other g methods are: g estimation of a structural nested model, and inverse probability weighted marginal structural models.

Inverse probability weighted marginal structural models consist of two important parts: the marginal structural model, which is a model for potential outcomes (structural) averaged over the entire population (marginal). Inverse proabaility weights are a tool that enable estimation of the MSM parameters (e.g., weighted least squares or weighted maximum likelihood).

G estimation of a structural nested model also consist of two parts: the structural nested model, which is a model for a contrast of potential outcomes (structural) within levels of past time-varying and baseline covariates (nested). G estimation is an **estimator** that takes advantage of the independence between the potential outcomes and the observed expsoure (i.e., exchangeability) to solve for the parameters of a SNM.

Marginal structural and structural nested models target very different estimands. As we will see, the g formula is simply an equation that links potential outcomes to observed data (i.e., outcomes, exposures, confounders). It can be used to target the quantities defined in either marginal structural or structural nested models. As it turns out, if we are willing to model each of the terms in the (potentially lengthy) equation, can also use it to estimate the effects quantified by these models.

Example 0: Model-Based Standardization

Let's start with a simple simulated example, and presume it represents data to answer questions about the effect of treatment for HIV on CD4 count. The causal diagram representing this scenario is depicted in Figure 3.

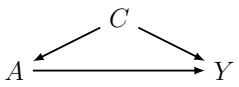


Figure 3: Causal diagram representing the relation between anti-retroviral treatment (A), HIV viral load just prior to treatment (C), and CD4 count measured at the end of follow-up (Y).

Table 1 presents data from this simulated observational cohort study ($A = 1$ for treated, $A = 0$ otherwise).

The CD4 outcome in Table 1 is summarized (averaged) over the

C	A	Y	N
0	0	94.3	344052
0	1	119.2	154568
1	0	130.6	154560
1	1	155.7	346820

Table 1: Example data illustrating the number of subjects (N) within each possible combination of treatment (A) and HIV viral load (C). The outcome column (Y) corresponds to the mean of Y within levels of A and C.

participants at each level of the treatments and covariate. Becuase the continuous outcome is summarized over each treatment $\times$ covariate level, we cannot estimate standard errors but will rather focus on estimating the parameter of interest.[12] We will analyze these data using model-based standardization, which is equivalent to the parametric g formula in a time-fixed exposure setting.

[12] The bootstrap will be demonstrated in the Appendices.

Setup

We first start with the **setup**, where we define our estimand, order our variables causally, write down our models, and "tie" them together into the g formula. In this simple setting, our estimand of interest is the marginal average causal effect on the difference scale:

$$E(Y^{a=1} - Y^{a=0})$$

This estimand tells us that we need to quantify two outcome averages: one that would be observed if everyone were exposed, and one if everyone were unexposed.

Next, we examine our causal diagram to order our variables causally. The causal sequence of variables is: C (first), A (second), and Y (third). To see why, note that in Figure 3 there are no variables that cause C, A is caused by C, and Y is caused by both A and C. Becuase of this, A cannot come before C (an effect cannot precede its cause), nor can Y come before A or C. The causal ordering of our variables is therefore C, A, and Y.

We then write down models for each variable.[13] How do we know which models to specify? We regress each variable against everything that comes before it.

[13] Recall: The "expit" function is the inverse of the logit: $\text{expit}(a) = 1/[(1 + \exp(-a)]$.

However, we must ensure that we do not break the **cardinal rule: do not adjust for the future.**

Variable	Model
Y	$E(Y \mid A, C) = \alpha_0 + \alpha_1 A + \alpha_2 C$
A	$P(A \mid C) = \text{expit}(\beta_0 + \beta_1 C)$
C	$P(C) = \text{expit}(\gamma_0)$

Finally, we tie each of these models together to give us a precursor to the g formula. To do this, we invoke the law of total probability, which states that the $P(A) = \sum_B P(A \mid B)P(B)$. This allows us to "average over" a conditional to obtain a marginal. In our case, the relevant conditional is the regression model for the outcome, and we have to average over the distributions of A and C:

$$E(Y) = \sum_A \sum_C E(Y \mid A, C)P(A \mid C)P(C)$$

To obtain the g formula from this expression, we replace all instances of A with $A = a$ and remove $P(A \mid C)$

$$E(Y^a) = \sum_C E(Y \mid A = a, C)P(C)$$

which holds under our identifiability assumptions.

Implementation and Validation

We're now ready for **implementation**. Suppose we wanted to estimate the unconditional (i.e., marginal) mean outcome in the sample. There are two ways we can do this. The easy way would be to simply take the average in the sample:

```
## CODE SET 2
# arrange into long data
C<-c(0,0,1,1);A<-c(0,1,0,1);Y<-c(94.3,119.2,130.6,155.7)
N<-c(344052,154568,154560,346820)
D<-NULL
for(i in 1:4){
  d<-data.frame(cbind(rep(C[i],N[i]),rep(A[i],N[i]),rep(Y[i],N[i])))
  D<-rbind(D,d)
}
names(D)<-c("C","A","Y")
# take the mean of Y
mean(D$Y)
```

```
## [1] 125.054

## END CODE SET 2
```

But we could also compute the marginal mean using the law of total probability. To do this, we can estimate our models using the data, and then predict from each in sequence:

```
## CODE SET 3
# fit models
mC<-glm(C~1,data=D,family=binomial("logit"))
mA<-glm(A~C,data=D,family=binomial("logit"))
mY<-glm(Y~A+C,data=D,family=gaussian("identity"))

## obtain predictions
# obtain C predictions
pC<-predict(mC,type="response")
# use predicted C to obtain predicted A
pA<-predict(mA,newdata=data.frame(C=pC),type="response")
# use predicted A and C to obtain predicted Y
pY<-predict(mY,newdata=data.frame(A=pA,C=pC),type="response")

# compute marginal mean of predicted Y
mean(pY)

## [1] 125.0584

## END CODE SET 3
```

The key is that C is predicted, then A is predicted using the C predictions, and then Y is predicted using the A and C predictions.

SIDE NOTE: To see why this works, suppose we're interested in the marginal (i.e., averaged over C) mean of Y if $A = 0$, and let's assume for illustrative purposes that $P(C = 1) = 0.2$ (it's not in our example): Note that, in the second line of the above, $E(Y \mid A = 0, C = 0)$ and $E(Y \mid A = 0, C = 1)$ are just the averages of Y among those with $A = 0, C = 0$ and $A = 0, C = 1$, respectively. We can therefore replace

$$\begin{aligned} E(Y \mid A=0) &= \textstyle\sum_C E(Y \mid A=0, C) P(C) \\ &= E(Y \mid A=0, C=0) P(C=0) + E(Y \mid A=0, C=1) P(C=1) \\ &= \alpha_0 \times 0.8 + (\alpha_0 + \alpha_2) \times 0.2 \end{aligned}$$

these with the parameters from our model. In a dataset of 100 people with $A = 0$, $\sim$ 80 would have $C = 0$ and $\sim$ 20 would have $C = 1$. Among those 100, the true average outcome for those with $C = 0$ would be α_0, and the true average outcome for those with $C = 1$ would be $\alpha_0 + \alpha_2$. Therefore, the average of Y among these 100 people with $A = 0$ would be precisely the weighted combination of averages that we need: $\alpha_0 \times 0.8 + (\alpha_0 + \alpha_2) \times 0.2$. This is why we can use our data and/or predictions to implement the law of total probability.

Back to our original example, we have two versions of our outcome: the actual data (Y) and the predictions based on our models (pY). The mean of both these versions is the same: 125.0. This **validation** step tells us that our models are doing a decent job at recreating the averages that result from our actual data generating mechanisms.

Continuing with our **implementation**, we can also use this code to predict Y if $A = 1$ for everyone or if $A = 0$ for everyone. We must just replace "A=pA" with "A=1" and "A=0" in the last line of code that yields the predictions we want. Replacing "A=pA" with "A=a" is tantamount to replacing all instances of A in the above equations with $A = a$, and removing the $P(A \mid C)$ term:

```
## CODE SET 4
# for A=1
pY_1<-predict(mY,newdata=data.frame(A=1,C=pC),type="response")
mY_1<-mean(pY_1)

#for A=0
pY_0<-predict(mY,newdata=data.frame(A=0,C=pC),type="response")
mY_0<-mean(pY_0)
## END CODE SET 4
```

The difference between these two means of interest is 25, which we must **interpret**.

Interpretation

The basic question is whether we can interpret this difference as the causal effect of ART on CD4 count. To do this, we must refer back to the set of assumptions discussed in the section on identifiability. For counterfactual consistency, we must ask two key questions: 1) how many different ways are there to assign someone to ART?; and 2) will these different assignment mechanisms lead to different outcomes? Suppose, for instance, that 1/2 of the sample took ART with ibuprofen. Suppose further that ibuprofen reduces the efficacy of ART. We then have a situation where counterfactual consistency may be violated, becuase assigning someone to ART (without ibuprofen) will not lead to the same effect that was quantified in our study. If we assume that all of the different ways in which one can take ART will not really lead to different outcomes, we can assume counterfactual consistency.

For interference, we must ask whether giving someone ART will affect the CD4 count of another person. In this case, it seems reasonable to assume no such interference occurs. Exchangeability is something we often consider in epidemiology, and requires no uncontrolled confounding, information, or selection bias.

Becuase of the small number of variables in this example, correct model specification is not likely to pose any problems. If, for example, an interaction between A and C in the model for Y is required, our model would be mis-specified. With a small number of categorical variables, we can saturate all the models to estimate things nonparametrically. However, this is often not possible when there are many categorical confounders, or any continuous confounders.

Finally, for positivity, we must ask whether there are exposed and unexposed individuals in each confounder level. In our simple setting, it is easy to verify this with a 2×2 table:

```
table(D$A, D$C)
```

```
##
##          0      1
##   0 344052 154560
##   1 154568 346820
```

Becuase there are no empty cells in this table, we can assume positivity is met. Additionally, becuase we are willing to make all these identifiability assumptions, we infer that the causal effect of A on Y is 25.

Example 1: ART effect on CD4 Count (Simulated)

In the previous example, we dealt with data that was neither longitudinal nor complex. We did not need to analyze these data using the g formula. In fact, a simple standard regression would have given us the same result. Here, we extend our previous example by adding an additional exposure, and converting our time-fixed confounder C to a time-dependent confounder Z. Our research question again deals with the effect of treatment for HIV on CD4 count.[14] The causal diagram representing this scenario is depicted in Figure 4.

[14] This example was taken from Ashley I. Naimi et al. (2016)

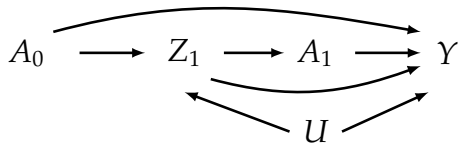


Figure 4: Causal diagram representing the relation between anti-retroviral treatment at time 0 (A_0), HIV viral load just prior to the second round of treatment (Z_1), anti-retroviral treatment status at time 1 (A_1), the CD4 count measured at the end of follow-up (Y), and an unmeasured common cause (U) of HIV viral load and CD4.

Study Question 5: Does the fact that U is unmeasured in Figure 4 create problems for our analysis? Why or why not?

Table 1 presents data from a hypothetical observational cohort study ($A = 1$ for treated, $A = 0$ otherwise). Treatment is measured at baseline (A_0) and once during follow up (A_1). The sole covariate is elevated HIV viral load ($Z = 1$ for those with > 200 copies/ml, $Z = 0$ otherwise), which is constant by design at baseline ($Z_0 = 1$) and measured once during follow up just prior to the second treatment (Z_1). The outcome is CD4 count measured at the end of follow up in units of cells/mm^3. Again, the CD4 outcome in Table 1 is summarized (averaged) over the participants at each level of the treatments and covariate.

A_0	Z_1	A_1	Y	N
0	0	0	87.29	209,271
0	0	1	112.11	93,779
0	1	0	119.65	60,654
0	1	1	144.84	136,293
1	0	0	105.28	134,781
1	0	1	130.18	60,789
1	1	0	137.72	93,903
1	1	1	162.83	210,527

Table 2: Prospective study data illustrating the number of subjects (N) within each possible combination of treatment at time 0 (A_0), HIV viral load just prior to the second round of treatment (Z_1), and treatment status for the 2nd round of treatment (A_1). The outcome column (Y) corresponds to the mean of Y within levels of A_0, Z_1, A_1. Note that HIV viral load at baseline is high ($Z_0 = 1$) for everyone by design.

Setup

The number of participants is provided in the rightmost column of Table 1. In this hypothetical study of one million participants we ignore random error (i.e., we will not focus on confidence interval estimation). Let's again start with the problem **setup**, where we define our estimand, order our variables causally, write down our models, and "tie" them together into the g formula. Here, we focus on the average causal effect of always taking treatment, $(a_0 = 1, a_1 = 1) \equiv \bar{a}_1 = 1$, compared to never taking treatment, $(a_0 = 0, a_1 = 0) \equiv \bar{a}_1 = 0$:

$$\psi = E(Y^{\bar{a}_1=1}) - E(Y^{\bar{a}_1=0}).$$

This average causal effect consists of the joint effect of A_0 and A_1 on Y (Daniel et al. 2013). Here, $Y^{\bar{a}_1}$ represents a potential outcome value that would have been observed had the exposures been set to specific levels a_0 and a_1.

The causal order of our observed variables is: A_0, Z_1, A_1, and Y.[15] For each of these variables, we can write down the following models:

[15] Note that we ignore U in this step becuase it is not measured.

Variable	Model
Y	$E(Y \mid A_1, Z_1, A_0) = \alpha_0 + \alpha_1 A_1 + \alpha_2 Z_1 + \alpha_3 A_0$
A_1	$P(A_1 \mid Z_1) = \text{expit}(\beta_0 + \beta_1 Z_1)$
Z_1	$P(Z_1 \mid A_0) = \text{expit}(\gamma_0 + \gamma_1 A_0)$
A_0	$P(A_0) = \text{expit}(\theta_0)$

Again, these models are obtained by regressing each variable against everything that comes before. Next, we tie each of these equations together to give us a precursor to the g formula. As in the previous example, we use the law of total probability to do this, which

yields:

$$E(Y) = \sum_{A_1}\sum_{Z_1}\sum_{A_0} E(Y \mid A_1, Z_1, A_0)P(A_1 \mid Z_1)P(Z_1 \mid A_0)P(A_0).$$

We get the g formula when we replace all instances of A_0 and A_1 with a_0 and a_1, respectively, and remove the models for A_0 and A_1:

$$E(Y^{a_0,a_1}) = \sum_{Z_1} E(Y \mid A_1 = a_1, Z_1, A_0 = a_0)P(Z_1 \mid A_0 = a_0).$$

which holds under our identifiability assumptions.

Implementation

Let's now **implement** the g formula in our software programs. We will again start by estimating the unconditional (i.e., marginal) mean outcome in the sample, by first taking the sample average:

```
## CODE SET 5
# arrange into wide data
a0<-c(0,0,0,0,1,1,1,1);z1<-c(0,0,1,1,0,0,1,1);a1<-c(0,1,0,1,0,1,0,1)
y<-c(87.29,112.11,119.65,144.84,105.28,130.18,137.72,162.83)
N<-c(209271,93779,60654,136293,134781,60789,93903,210527)
D<-NULL
for(i in 1:8){
  d<-data.frame(cbind(rep(a0[i],N[i]),rep(z1[i],N[i]),rep(a1[i],N[i]),rep(y[i],N[i])))
  D<-rbind(D,d)
}
names(D)<-c("a0","z1","a1","y")
# take the mean of Y
mean(D$y)

## [1] 125.0948

## END CODE SET 5
```

Next, we compute the marginal mean using the law of total probability by estimating our models using the data, and then predicting from each in sequence:

```
## CODE SET 6
# fit models
mA0<-glm(a0~1,data=D,family=binomial("logit"))
mZ1<-glm(z1~a0,data=D,family=binomial("logit"))
mA1<-glm(a1~z1,data=D,family=binomial("logit"))
mY<-glm(y~a1+z1+a0,data=D,family=gaussian("identity"))

## obtain predictions
# obtain A0 predictions
pA0<-predict(mA0,type="response")
# use predicted A0 to obtain predicted Z1
pZ1<-predict(mZ1,newdata=data.frame(a0=pA0),type="response")
# use predicted Z1 to obtain predicted A1
pA1<-predict(mA1,newdata=data.frame(z1=pZ1),type="response")
# use predicted A0, Z1 and A1 to obtain predicted Y
pY<-predict(mY,newdata=data.frame(a0=pA0,z1=pZ1,a1=pA1),type="response")

# compute marginal mean of predicted Y
mean(pY)

## [1] 125.102

## END CODE SET 6
```

Validation

Once again, we have two versions of our outcome: the actual data (Y) and the predictions based on our models (pY). These latter predictions are obtained under a very specific scenario: by consistency and no interference, it is the outcome distribution that would be observed if the exposure distribution was what actually occurred in our data. This scenario, called the **natural course**, is in contrast to what might have been observed if everyone were exposed/unexposed at both time-points. Estimating the natural course is an important **validation step** when using the parametric g formula. If the empirical results align closely with the natural course, this offers some assurance that our models are not grossly mis-specified. On the other hand, if our empirical and natural course results differ substantially,

[16] Note the evasive language ("some assurance", "suggests", etc). This is because unbiased causal effect estimation is still possible if the natural course and empirical results are very different. It is also possible that a parameter estimate is biased if the natural course and empirical results are identical. Thus, this validation step provides evidence that is neither necessary nor sufficient for valid estimation. However, becuase these scenarios are unlikely to occurr in practice, the evidence provided by this validation step is informative.

this suggests that something may be wrong.[16]

In our example, the empirical and natural course means are again the same: 125.1.

Continuing with our **implementation**, we can also use this code to predict Y if $A = 1$ for everyone or if $A = 0$ for everyone:

```
## CODE SET 7
# for A=1
pZ_1<-predict(mZ1,newdata=data.frame(a0=1),type="response")
pY_1<-predict(mY,newdata=data.frame(a0=1,z1=pZ_1,a1=1),type="response")
mY_1<-mean(pY_1)

#for A=0
pZ_0<-predict(mZ1,newdata=data.frame(a0=0),type="response")
pY_0<-predict(mY,newdata=data.frame(a0=0,z1=pZ_0,a1=0),type="response")
mY_0<-mean(pY_0)
## END CODE SET 7
```

Interpretation

The difference between these two means is 50 cells/mL (a 25 cell/mL difference for each time-point, which corresponds to the true effect in our simulated scenario). If we make the same assumptions as in the previous example (counterfactual consistency, no interference, exchangeability, no model mis-specification, positivity), we can interpret this as our causal effect of interest.

SIDE NOTE: The parametric g formula is subject to what is known as the "g null paradox," which arises when the true exposure effect is null. In this setting, it is possible that the parametric g formula will estimate a non-null effect. Not much is known about the g null paradox, but it is currently the topic of active research by several groups.

Before moving on to our next examples, let's take another look at our second simulated example. According to the causal diagram in Figure 4, we should be able to obtain an unbiased estimate of the A_0

and A_1 effects using simple regression models. For example, if we adjust for Z_1, there is no open back-door path from A_1 to Y. If we run the code to do this, we find this is actually the case:

```
# CODE SET 8
round(coef(glm(y~a1+z1,data=D,family=gaussian("identity"))),1)
```

```
## (Intercept)          a1          z1
##        94.3        25.0        36.4
```

```
# END CODE SET 8
```

Similarly, because there are no confounders of the relation between A_0 and Y, the causal diagram seems to suggest that simply regressing Y against A_0 will give us an unbiased effect estimate (the true effect is 25.0 cells/mL):

```
# CODE SET 9
round(coef(glm(y~a0,data=D,family=gaussian("identity"))),1)
```

```
## (Intercept)          a0
##       111.6        27.1
```

```
# END CODE SET 9
```

However, doing this overestimates the true effect by 2.1 cells/mL. Why? This is a consequence of feedback between A_0 and A_1. Becuase A_0 affects A_1 indirectly through Z_1, this regression model is estimating the overall effect of A_0 on Y. Thus, the estimate of 27.1 is not wrong *per se*. It is simply quantifying the direct effect of A_0 on Y, **plus** the indirect effect of A_0 on Y via A_1.

Note that while this estimate is not incorrect by itself, if we were intersted in estimating $E(Y^{\bar{a}_1=1} - Y^{\bar{a}_1=0})$, and we added the two estimates from these simple regression models to do this, we would be wrong because we'd be counting a portion of the A_1 effect twice.

Example 2: Mediation Analysis Using the Natural Survey of Family Growth

In this second example, we will look at how to conduct mediation analyses using the parametric g formula with data from the publicly available Natural Survey of Family Growth ("Centers for Disease Control and Prevention, National Survey of Family Growth" Updated Jan 22 2014. Accessed May 27 2014.). For our purposes, we will ignore the complex survey sampling design. In this example, we will have the opportunity to get a close look at several controversial issues in causal inference. Notably, we will talk about the challenges in identifying, estimating, and interpreting natural direct and indirect effects (Ashley I. Naimi, Kaufman, and Maclehose 2014).[16] We will also look at the issues involved in estimating the effects of so-called "nonmanipulable" expsoures (Hernan and Taubman 2008).

[16] There has recently been explosion in the number of estimands that can be defined to conduct mediation analyses. Among these include standardized direct effects (Didelez, Dawid, and Geneletti 2006), randomized interventional analogues to natural direct and indirect effects (T. J. VanderWeele, Vansteelandt, and Robins 2014), "organic" direct and indirect effects (Lok 2016), and stochastic mediation contrasts (Ashley I. Naimi et al. 2014).

Two types of effects are commonly used in mediation analysis. The first are controlled direct effects, which quantify the exposure effect under an intervention that sets the mediator to a specific value for all individuals in the population. Controlled indirect effects are notably difficult to conceptualize, and instead are defined as a contrast between the total and controlled direct effect in the absence of exposure-mediator interactions (Kaufman, Maclehose, and Kaufman 2004). The second type of direct effects are "natural" (which include "total" or "pure") effects. The natural direct effect quantifies exposure effect that would be observed holding the mediator fixed at the value it would have taken under some referent exposure value. Similarly, natural indirect effect quantifies the effect the exposure has through the mediator by changing the mediator from the value it would have taken under some referent exposure to the value it would have taken under some specific alternate exposure (all while holding the exposure fixed).

Here, we will estimate natural and controlled effects to evaluate the extent to which the racial disparity in preterm birth is "explained" by the manner in which a woman pays for the delivery of a pregnancy. As we will see, this question is itself problematic, but was chosen specifically to illustrate some rather subtle issues. On top

of the problematic nature of the question itself,[17] we will have the opportunity to go over some of the challenges in identifying, interpreting, and estimating natural direct and indirect effects. Estimating controlled effects are also problematic in this setting, and we will see why.

[17] While the example here involves questions that might typically be asked in social epidemiology, these issues are also routinely encountered in other areas, including lifecourse epidemiology, research on the effects of growth, weight, obesity, or age.

In spite of the challenges (which are not insignificant) inherent in using the NSFG data to answer this question, its didactic value is clear. It will give us plenty of opportunity to discuss the reasons for which natural effects are not identifiable,[18] and illustrate the challenges in estimating any "effects" of exposures such as race or payment method for delivery.

[18] Note that Lin et al have developed a "parametric mediational g formula" algorithm for estimating what are known as randomzied interventional analogues to natural direct and indirect effects (Lin et al. 2017). This is a useful contribution that resolves the problems related to identifying and estimating natural effects.

Setup: Natural and Controlled Effects

Let's begin with **setting up** the problem. In this setup section, we will again define our estimands of interest, causally order our variables, specify parametric models for each, and then link them together using the law of total probability. In our NSFG data, we will let X denote race, M denote the method of payment for delivery, and Y denote preterm birth (1 if less than 37 weeks, 0 otherwise). We will also assume that the only confounders of the relation between race and the method of payment, or between race and preterm birth is maternal age.[19] We will also let C_{MY} denote confounders of the relation bewteen preterm birth and the method of delivery payment. In our analysis, these will include maternal age, prenatal care, the wantedness of a pregnancy, marital status, educational level, and the gestational age of a prior pregnancy.

[19] This is actually a realistic and defensible assumption, provided we are interested in quantifying disparities, and not estimating the "effects" of race (Ashley I. Naimi et al. 2016)

The causal diagram that will be motivating our analysis will be as follows:

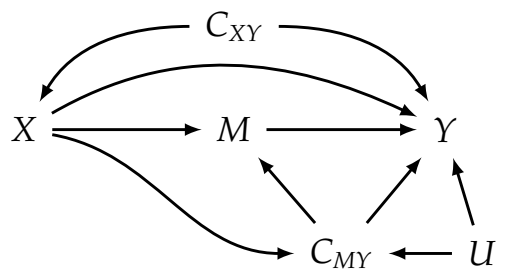


Figure 5: Causal diagram representing the relation between race (X), method of payment for delivery (M), and preterm birth (Y). In this diagram, C_{MY} represent M-Y confounders that are also associated with race.

STUDY QUESTION 6: Can you identify the structural similarities between Figures 4 and 5?

Conceptually, the natural direct effect is the expected change in the outcome if we were to "freeze" the mediator value for each person at the level it would have taken had the person's exposure been some referent level (but when, in actuality, the person's exposure status changes). Similarly, the natural indirect effect is the expected change in the outcome when the mediator changes as though the exposure had (but when, in actuality, the exposure doesn't change). These natural effects are divided into two types: pure and total.

The pure direct effect is the exposure effect if the mediator were held at its potential value under $X = 0$:

$$PDE = E\left(Y^{1,M^0}\right) - E\left(Y^{0,M^0}\right),$$

The total direct effect is identical to the pure direct effect, except that the mediator is held at its potential value under $X = 1$:

$$TDE = E\left(Y^{1,M^1}\right) - E\left(Y^{0,M^1}\right),$$

The controlled direct effect is also very similar, except that the mediator is held fixed at a given level uniformly in the population:

$$CDE(m) = E\left(Y^{x,m}\right) - E\left(Y^{x^*,m}\right)$$

STUDY QUESTION 7: Under what conditions will $TDE = PDE = CDE(m = 0) = CDE(m = 1)$?

For all three estimands, the outcome $Y^{x,M^{x^*}}$ represents the potential outcome that would have been observed if X were set to x *and* under the potential mediator value that would have been observed if

X were set to some other value x^*. Notice how the pure and total direct effects require a union of two logically incompatible states?: the outcome under exposure x with the mediator set to what it would have been under x^*. Because no single individual can ever exist with exposure values x and x^*, this composite counterfactual requires information that can only exist in two separate "worlds," and has thus been referred to as a "cross-world" counterfactual (Richardson and Robins 2013).

In a similar vein, the pure and total indirect effects are:

$$PIE = E\left(Y^{0,M^1}\right) - E\left(Y^{0,M^0}\right),$$

and,

$$TIE = E\left(Y^{1,M^1}\right) - E\left(Y^{1,M^0}\right).$$

These are the estimands that we will seek to estimate with the parametric g formula. Continuing our **setup**, we must next causally order our variables, including all of the C_{MY}.

SIDE NOTE: Thus far, we've been causally ordering **all** of the variables in our causal diagrams. In actuality, we only have to order the variables that come after the exposure. We do not have to order the baseline variables, or the variables at the first time-point that are not affected by the exposure. Becuase there are no baseline variables in our mediation example (no confounders of the race-preterm birth relation), we will also order all our variables here too. But in the next example, we will see how to deal with pre-exposure variables.

In Table 3, the race variables as coded as 1="non-Hispanic Black" and 0="non-Hispanic white". The mediator was payment for delivery, and was coded as 1="Medicaid/government assistance", and 0="insurance only, own income and insurance, or other combinations of payment method." Thus, for example, if we set $M = 0$, then the CDE is interpreted as the effect of race that would be observed if everyone paid for delivery using insurance only, own income and insurance, or other combinations of payment method.[20]

[20] Warning: as we will see in the interpretation section, we will not be able to interpret any of these results as "effects".

Order	Notation	Variable
9	Y	Preterm birth
8	M	Payment for delivery
7	$C_{(7)MY}$	Prenatal care
6	$C_{(6)MY}$	Wantedness of the pregnancy
5	$C_{(5)MY}$	Gestational age at birth of prior pregnancy
4	$C_{(4)MY}$	Marital status at conception
3	$C_{(3)MY}$	Maternal education at conception
2	X	Race
1	C_{XY}	Maternal age

Table 3: Assumed causal ordering of variables in the National Survey of Family Growth, 2006-2010 and 2011-2013 waves.

Now that we have chosen a causal order, we to specify models for each variable in Table 3. Instead of writing each model out as before, we will simply note that each variable in Table 3 is binary, prompting use of logistic regression. Furthermore, each variable will be regressed against everything that comes before it in Table 3.[21] Finally, we have to tie all these models together using the law of total probability to obtain the g formula for our mediation questions. Doing so, we get:

[21] If this is not clear, the code implementing this below should help.

$$E(Y^{x,M^{x^*}}) = \sum_{C_{MY}} E(Y \mid X = x, M^{x^*}, C_{MY}, C_{XY}) P(M \mid X = x^*, C_{MY}) \prod_{j=7:3} P(C_{(j)MY} \mid X = x, C_{(j-1)MY}, \ldots, C_{(3)MY}) P(C_{XY}),$$

for the natural effects, and

$$E(Y^{x,m}) = \sum_{C_{MY}} E(Y \mid X = x, M = m, C_{MY}, C_{XY}) \prod_{j=7:3} P(C_{(j)MY} \mid X = x, C_{(j-1)MY}, \ldots, C_{(3)MY}) P(C_{XY}),$$

for the controlled effect.

Side Note: There has been some controversy over natural effects. There are several reasons for this controversy. One of them is the fact that defining natural effects requires outcomes under two logically incompatible exposure states. The other is more complicated, but arises when mediator outcome confounders are affected by the exposure. In

this situation, mediator-outcome confounders affected by the exposure are known as "recanting witnesses." On the one hand, they signal to the mediator that they are in one state (unexposed) while simultaneously signalling to the outcome that they are in another state (exposed). With the exception of Naimi et al (2014), this issue has not been discussed in the health sciences literature.

Implementation

Now that we have setup the problem, lets begin with implementation. We first import and explore the NSFG data:

```
setwd("~/Dropbox/Documents/Research/Presentations/SER 2017/SER Workshop/pgf_course-master/")
nsfg<-read_sas("n2.sas7bdat")
head(nsfg,3);tail(nsfg,3)
```

```
## # A tibble: 3 x 12
##   `_id` HISPRACE pnc_d20 educ_d marit_d
##   <dbl>    <dbl>   <dbl>  <dbl>   <dbl>
## 1     5        0       0      0       0
## 2     6        0       0      0       0
## 3     7        0       0      0       0
## # ... with 7 more variables: want_d <dbl>,
## #   pay_d <dbl>, mage_c1 <dbl>,
## #   wksg_d <dbl>, mage_c2 <dbl>,
## #   mage_c3 <dbl>, ptb <dbl>
```

```
## # A tibble: 3 x 12
##   `_id` HISPRACE pnc_d20 educ_d marit_d
##   <dbl>    <dbl>   <dbl>  <dbl>   <dbl>
## 1 30025        0       0      0       0
## 2 30031        0       0      0       1
## 3 30035        1       0      0       1
## # ... with 7 more variables: want_d <dbl>,
## #   pay_d <dbl>, mage_c1 <dbl>,
## #   wksg_d <dbl>, mage_c2 <dbl>,
## #   mage_c3 <dbl>, ptb <dbl>
```

```r
# sample size
nrow(nsfg)
```

```
## [1] 13611
```

```r
# verify no missing
missFunc<-function(x){sum(is.na(x))}
sum(apply(nsfg,2,missFunc))
```

```
## [1] 0
```

```r
# tabular analyses
table(nsfg$pay_d)
```

```
##
##    0    1
## 7299 6312
```

```r
table(nsfg$HISPRACE)
```

```
##
##    0    1
## 8662 4949
```

```r
table(nsfg$ptb)
```

```
##
##     0     1
## 11733  1878
```

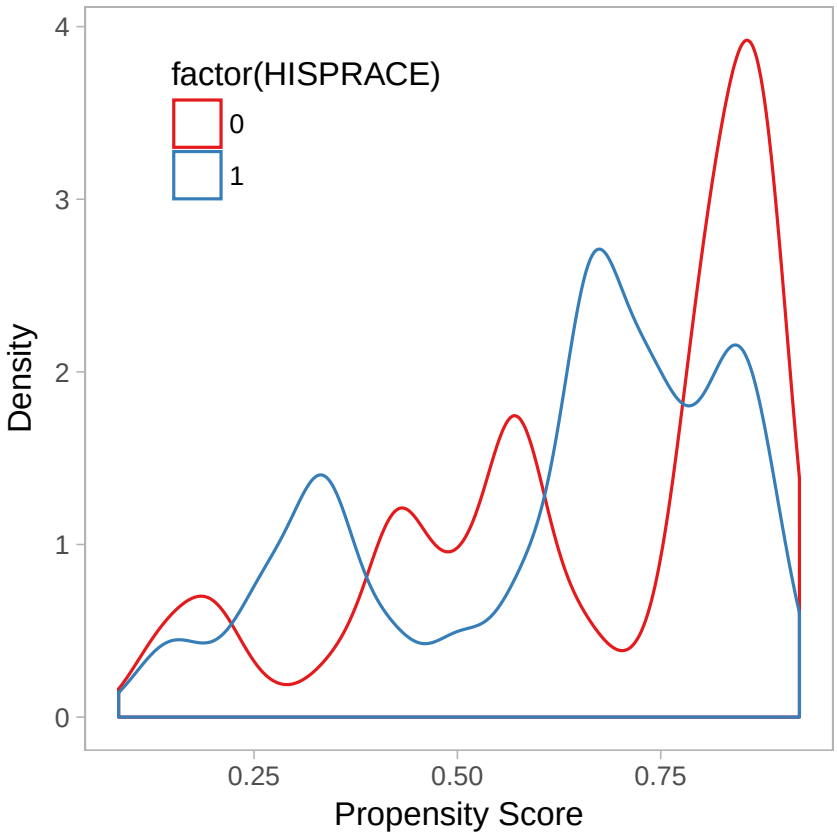
We should also look at the crude associations between the exposure, mediator, and outcome:

```r
# risk differences between exposure, mediator, and outcome
## preterm birth and race
coef(lm(ptb~HISPRACE,data=nsfg))[2]*100
```

```
## HISPRACE
## 5.561264
```

```r
## preterm birth and delivery payment
coef(lm(ptb~pay_d,data=nsfg))[2]*100
```

```
##    pay_d
## 3.961511

## delivery payment and race
coef(lm(pay_d~HISPRACE,data=nsfg))[2]*100

## HISPRACE
## 26.89094

# look at propensity score overlap for mediator (pay_d)
mod<-glm(pay_d~HISPRACE+pnc_d20+educ_d+marit_d
         +want_d+mage_c1+mage_c2+mage_c3+wksg_d,data=nsfg,family=binomial("logit"))
nsfg$pM<-predict(mod,type="response")*nsfg$pay_d+
  (1-predict(mod,type="response"))*(1-nsfg$pay_d)

ggplot(nsfg,aes(pM,color=factor(HISPRACE))) +
  theme_light() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(x = "Propensity Score",y = "Density") +
  scale_colour_brewer(palette="Set1") +
  geom_density() + theme(
  legend.position = c(.5, .95),
  legend.justification = c("right", "top"),
  legend.box.just = "right",
  legend.margin = margin(6, 6, 6, 6)
)
```



Now that we've explored the data, let's fit models for each equation in the g formula.

```
# logistic models for joint distribution of observed data
## preterm birth
m9<-glm(ptb~pay_d+pnc_d20+want_d+wksg_d+
          marit_d+educ_d+HISPRACE+pay_d*HISPRACE+
          mage_c1+mage_c2+mage_c3,data=nsfg,family=binomial("logit"))
## mediator
m8<-glm(pay_d~pnc_d20+want_d+wksg_d+
          marit_d+educ_d+HISPRACE+
          mage_c1+mage_c2+mage_c3,data=nsfg,family=binomial("logit"))
## pnc
m7<-glm(pnc_d20~want_d+wksg_d+
          marit_d+educ_d+HISPRACE+
          mage_c1+mage_c2+mage_c3,data=nsfg,family=binomial("logit"))
## wantedness
```

```
m6<-glm(want_d~wksg_d+
          marit_d+educ_d+HISPRACE+
          mage_c1+mage_c2+mage_c3,data=nsfg,family=binomial("logit"))
## prior gestational age at birth
m5<-glm(wksg_d~marit_d+educ_d+HISPRACE+
          mage_c1+mage_c2+mage_c3,data=nsfg,family=binomial("logit"))
## marital status
m4<-glm(marit_d~educ_d+HISPRACE+
          mage_c1+mage_c2+mage_c3,data=nsfg,family=binomial("logit"))
## educational status
m3<-glm(educ_d~HISPRACE+mage_c1+mage_c2+mage_c3,data=nsfg,family=binomial("logit"))
```

Next we select a Monte Carlo sample of 50,000, and keep only the baseline data (maternal age and race). We then predict the follow up under the exposure and mediator scenarios we need to quantify our estimands:

```
index<-sample(1:nrow(nsfg),5e4,replace=T)
mc<-nsfg[index,c("HISPRACE","mage_c1","mage_c2","mage_c3")]
head(mc);nrow(mc)
```

```
## # A tibble: 6 x 4
##   HISPRACE mage_c1 mage_c2 mage_c3
##      <dbl>   <dbl>   <dbl>   <dbl>
## 1        0       1       0       0
## 2        1       0       1       0
## 3        1       0       0       1
## 4        0       0       1       0
## 5        1       0       1       0
## 6        0       0       1       0
```

```
## [1] 50000
```

```
pgfm<-function(exposure,mediator){
  if(!is.null(exposure)){
    mc$HISPRACE<-exposure
  }
  e<-as.numeric(predict(m3,newdata=mc,type="response")>
```

```
                runif(nrow(mc)));ND<-data.frame(mc,educ_d=e)
  m<-as.numeric(predict(m4,newdata=ND,type="response")>
                runif(nrow(mc)));ND<-data.frame(ND,marit_d=m)
  w<-as.numeric(predict(m5,newdata=ND,type="response")>
                runif(nrow(mc)));ND<-data.frame(ND,wksg_d=w)
  wnt<-as.numeric(predict(m6,newdata=ND,type="response")>
                runif(nrow(mc)));ND<-data.frame(ND,want_d=wnt)
  pnc<-as.numeric(predict(m7,newdata=ND,type="response")>
                runif(nrow(mc)));ND<-data.frame(ND,pnc_d20=pnc)
  if(is.null(mediator)){
      pay<-as.numeric(predict(m8,newdata=ND,type="response")>
                runif(nrow(mc)));ND<-data.frame(ND,pay_d=pay)
  } else if(mediator=="n1"){
      ND<-data.frame(ND,HISPRACE=1)
      pay<-as.numeric(predict(m8,newdata=ND,type="response")>
                runif(nrow(mc)));ND<-data.frame(ND,pay_d=pay)
  } else if(mediator=="n0"){
      ND<-data.frame(ND,HISPRACE=0)
      pay<-as.numeric(predict(m8,newdata=ND,type="response")>
                runif(nrow(mc)));ND<-data.frame(ND,pay_d=pay)
  } else{
      pay<-mediator;ND<-data.frame(ND,pay_d=pay)
  }
  ptb<-as.numeric(predict(m9,newdata=ND,type="response")>
                runif(nrow(mc)))
  return(ptb)
}

Yn<-pgfm(NULL,NULL) # natural course
Y1<-pgfm(1,NULL) # NH Black
Y0<-pgfm(0,NULL) # NH White
Y1mn0<-pgfm(1,"n0") # NH Black, NH White method of payment
Y0mn0<-pgfm(0,"n0") # NH White, NH White method of payment
Y1mn1<-pgfm(1,"n1") # NH Black, NH Black method of payment
Y0mn1<-pgfm(0,"n1") # NH White, NH Black method of payment
```

```
Y1m0<-pgfm(1,0) # NH Black, Insurance, Own Income, or Other method of payment
Y0m0<-pgfm(0,0) # NH White, Insurance, Own Income, or Other method of payment

# disparity
(mean(Y1)-mean(Y0))*100
```

```
## [1] 5.154
```

```
# PDE
(mean(Y1mn0)-mean(Y0mn0))*100
```

```
## [1] 4.99
```

```
# PIE
(mean(Y0mn1)-mean(Y0mn0))*100
```

```
## [1] -0.028
```

```
# TDE
(mean(Y1mn1)-mean(Y0mn1))*100
```

```
## [1] 5.144
```

```
# TIE
(mean(Y1mn1)-mean(Y1mn0))*100
```

```
## [1] 0.126
```

```
# CDE
(mean(Y1m0)-mean(Y0m0))*100
```

```
## [1] 5.53
```

Validation and Interpretation

After quantifying these associations, we must now validate, and interpret them in light of the assumptions listed above. First we compare the natural course preterm birth generated from our g formula algorithm to the observed preterm birth outcomes in the NSFG. We can do this using several measures, but to keep things simple we will look at means. The natural course mean for preterm birth was 0.137 while the preterm birth mean in the empirical data was 0.138.

Let's start with interpreting the controlled direct effect estimate of 5.53. From the point of view of the estimand alone, the CDE is interpreted as (for example) the exposure effect that would be observed if everyone's exposure was set to 1 and the mediator was set to 0, versus if everyone's exposure was set to 1 and the mediator was set to 0:

$$CDE(0) = E\left(Y^{1,0}\right) - E\left(Y^{0,0}\right)$$

Our exposure was "race," which was measured as a combination of a response to a questionnaire that asked "what is your race?" and "are you Hispanic or Latina, or of Spanish origin?" To interpret the association we quantified as a controlled direct effect, we must assume counterfactual consistency, which allows us to state that the observed outcome among those who classified themselves as non-Hispanic black is, in fact, what we would have observed had we (somehow) set them to be non-Hispanic black. Unfortunately, there is no conceivable intervention that we could conjure up to set individuals to be, e.g., "non-Hispanic black." For this reason, self or other reported measures of race cannot be construed as counterfactually causal.

SIDE NOTE: This has been a somewhat controversial issue. But the controversy stems (in large part) from a misunderstanding. "Race," measured as the response to questions about whether one considers oneself non-Hispanic Black, White, Hispanic, etc, cannot be causal. But these measures do not quantify the fundamental issues that underlie health disparities. Race relations, on the other hand, can be affected by social policy and a host of other interventions. Such interventions can easily be incorporated into the causal modelling framework without issue. Numerous examples exits, and include the Civil Rights Act of the mid 1960s, the desegregation of hospitals, and the repeal of unjust laws, which are all (policy) interventions, far removed from the simplistic classification schemes commonly used to represent "race."

Counterfactual consistency is a fundamental assumption in that most other assumptions depend on it. Consider, for example, the arrow from C_{XY} to X in Figure 5. This arrow implies that there is

some mechanism by which we can affect race. Translated into the counterfactual world, this means that we would somehow be able to intervene so as to set X ("race") to some specified value. But we have just argued that there is no such intervention. This raises the question about what it means for self reported race to be confounded. A confounder is a variable that affects both the exposure and the outcome. But if the expsoure can't be affected, how can it be confounded?

Points for further discussion:

- Natural Effects: Cross World Counterfactuals and Recanting Witness
- Causal inference in social epidemiology
- Other

Example 3: Per protocol effect of aspirin on fetal loss

In this final example, we will look at how to estimate the per protocol effect of daily low dose aspirin on pregnancy outcomes using data from a randomized trial. When seeking to estimate per protocol effects in RCTs, adjusting for non-compliance is imperative. Non-compliance occurs when study participants fail to take the assigned treatment. When the assigned treatment is a point intervention (e.g., vaccine, surgery), adjusting for noncompliance is straightforward (Dunn and Lovrić 2011).[22] However, when treatment consists of a series of actions over a long period (Ford and Norrie 2016), accounting for the consequences of noncompliance becomes more complex.

[22] There are several "compliance adjusted" effects one can estimate. These include as treated, per protocol, and complier average causal effects. The latter are an example of principal strata effects, and share similarities with natural direct and indirect effects discussed in the mediation section. As with natural effects, complier average causal effects are not identifiable using the g formula.

Common practice in randomized trials is to estimate the intent-to-treat (ITT) parameter (Piantadosi 2005). This parameter is identified (i.e., computable) in a trial because under randomization the randomized treatment indicator is independent of all the potential outcomes. However, the ITT parameter is often not the parameter of primary interest (Prentice, Pettinger, and Anderson 2005). Deviation from study protocol can lead to biases in treatment effect estimates that are not resolved by estimating the intent-to-treat (ITT) parameter. The ITT effect can still be used to quantify the effect of **treatment assignment**, quantifying the effect of compliance with the treatment plan, or of exposure to the treatment itself is more complex. Though unadjusted "as treated" or "per protocol" estimators are often used (Miguel A Hernán and Hernández-Díaz 2012,Shrier2014), they can be biased by post-randomization confounding and selection bias (Hernán, Hernández-Díaz, and Robins 2013).

Simple covariate adjustment methods (e.g., regression, stratification, matching) will often fail to properly account for post-randomization confounding that results from noncompliance in randomized trials (Robins and Hernán 2009). At any point over follow-up, the decision to comply with protocol may depend on the history of compliance prior to that point. For example, a participant may decide not to take study medications after three months of perfect compliance because of adverse side effects, yet resume complying with protocol after a month of noncompliance. When

treatment compliance affects post-randomization variables that are common causes of subsequent compliance and the outcome (e.g., the post-randomization variable is a mediator, or there is a common cause of the variable and the outcome), standard adjustment methods are inconsistent (Robins and Hernán 2009,Daniel2013,Naimi2016b). This is exactly the scenario depicted in the causal diagram in Figure 4.

While such patterns of partial noncompliance are common, they are often either ignored, or simply classified into dichotomous compliant or non-compliant categories and analyzed via naive (unadjusted) as treated or per protocol techniques. This latter tactic discards important information on the nature of compliance in a given trial, and fails to account for post-randomization confounding (Miguel A Hernán and Hernández-Díaz 2012,Shrier2014). Here, we will demonstrate how to use the parametric g formula to adjust for noncompliance in the Effects of Aspirin on Gestation and Reproduction (EAGeR) trial.

The EAGeR trial sought to evaluate the role of aspirin on pregnancy outcomes among 1,228 women. The ITT parameter of aspirin assignment on the risk of live birth was estimated as 5.1% [95% CI -0.8 to 11.0]). However, there was a non-trivial degree of noncompliance with study protocol. In addition, predictors of the trial outcome (e.g., bleeding, gastrointenstinal discomfort) both affected and were affected by aspirin use, and were thus deemed time-varying confounders affected by prior exposure. Referring to Figure 4, A becomes our measure of compliance, and Z becomes these time-varying confounders.

We'll again start with the problem **setup**, in which we define our estimands of interest. In a typical RCT with randomization indicator R and outcome Y, the ITT effect is defined on the difference scale as:

$$\begin{aligned}\Delta_{ITT} &= P(Y \mid R=1) - P(Y \mid R=0) \\ &= P(Y^{r=1}) - P(Y^{r=0}),\end{aligned}$$

where $P(Y \mid R = r)$ is the probability of the outcome among those assigned to $R = r$, while $P(Y^r)$ is the counterfactual probability that would be observed if R were set to r. The marginal counterfac-

tual probability $P(Y^r)$ equals the observed conditional probability $P(Y \mid R = r)$ if the counterfactual outcome is well defined and exchangeability holds (Ashley I Naimi, Cole, and Kennedy 2016). Under randomization, exchangeability holds in expectation. This ITT parameter quantifies the effect of assignment to a treatment intervention.

Several causal estimands can be defined in the presence of a time-dependent noncompliance measure $\overline{X}_j$, where j indexes time since randomization and overbars denote history (e.g., $\overline{X}_j = \{X_0, X_2, \cdots, X_j\}$). The per protocol effect is often of interest along with the ITT, and is defined as:

$$\Delta_{PP} = P(Y^{r=\overline{x}=1} = 1) - P(Y^{r=\overline{x}=0} = 1)$$

where $Y^{r,\overline{x}}$ denotes the counterfactual outcome that would be observed if compliance $\overline{x}$ was set to the assigned value r over the entire course of follow-up. This equation defines the per protocol effect via potential outcomes, as it represents the effect of taking aspirin if assigned (Miguel A Hernán and Hernández-Díaz 2012,Shrier et al. (2014)). In contrast to Δ_{ITT},

$$\Delta_{PP} \neq P(Y = 1 \mid R = \overline{X} = 1) - P(Y = 1 \mid R = \overline{X} = 0)$$

unless compliance is unassociated with study withdrawal and with any measured or unmeasured predictors of compliance and the outcome.

Empirical Data We use data from 1,228 women enrolled in the EAGeR study to illustrate the use of the parametric g formula to adjust for noncompliance. The design and protocols used for the EAGeR study have been documented (Schisterman et al. 2013). Briefly, EAGeR data were obtained from a multicenter, block randomized, double-blind, placebo-controlled trial. Eligible women were between 18-40 years of age, actively trying to conceive, and experienced one or two documented pregnancy losses. Women were randomized to receive 81 mg aspirin per day ($N = 615$) or placebo ($N = 613$); all received 400 mcg folic acid in addition.

Follow-up occurred for ≤ 6 menstrual cycles while attempting

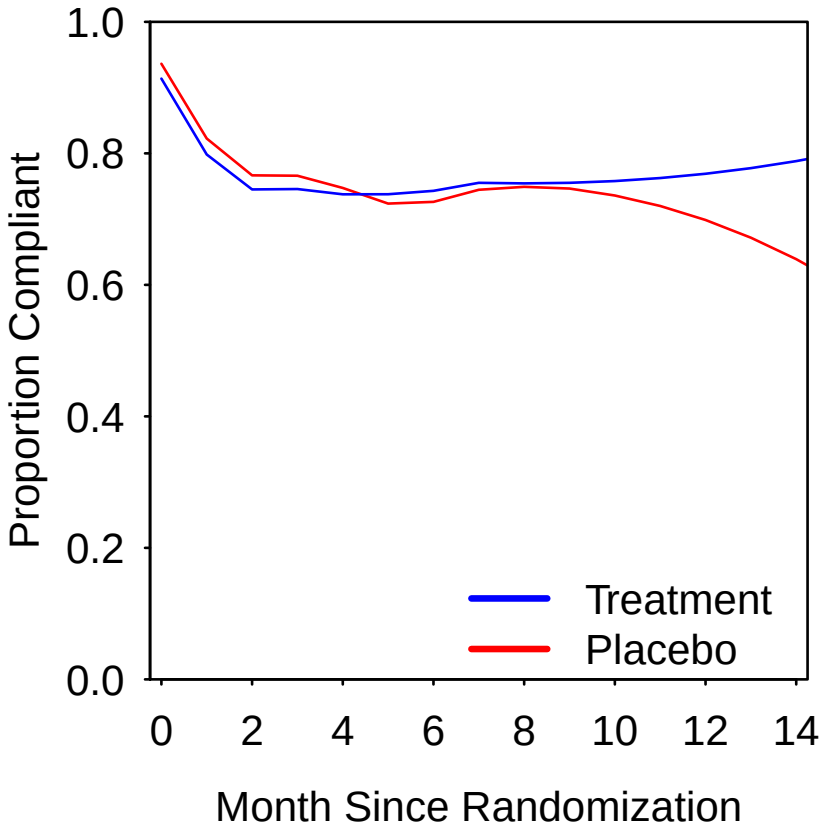
to conceive, and throughout pregnancy for those who conceived. Planned study visits occurred at least monthly. Conception was determined via hCG and confirmed by 6.5 week ultrasound. For women who conceived, treatment was continued until 36 weeks gestation. A total of 286,844 person-days of data from daily dairies, study questionnaires, and clinical and telephone evaluations were collected from the EAGeR participants. From these sources, we compiled a dataset with information on baseline, time-dependent, and outcome covariates. Baseline covariates included randomization indicator, randomization date, age and body mass index at randomization, income, race/ethnicity, education, marital status, employment status, study site. Time-dependent covariates included compliance, vaginal bleeding, gastro-intestinal symptoms, and time to conception. The primary study outcome for the trial was the proportion of live births. Women were followed from randomization through pregnancy for live birth or pregnancy loss. To facilitate open access to the data and software programs, here we present results based on a synthetic copy of the EAGeR data. This synthetic copy was naturally created as a by-product of the g computation algorithm fit to the original EAGeR trial data. We can download this synthetic copy of the EAGeR data from GitHub (aspirin.txt). The data are arranged in long form (one row for each person month).

```
## CODE SET NN
# import synthetic EAGeR data
aspirin<-read.table("aspirin.txt",header=T,sep="\t")
head(aspirin[names(aspirin) %in% c("id","study_month","treatment","final_outcome")],3)

##   id final_outcome treatment study_month
## 1  1    live birth         0           1
## 2  1    live birth         0           2
## 3  1    live birth         0           3

tail(aspirin[names(aspirin) %in% c("id","study_month","treatment","final_outcome")],3)

##         id  final_outcome treatment
## 10170 1228 pregnancy loss         1
```

```
## 10171 1228 pregnancy loss         1
## 10172 1228 pregnancy loss         1
##       study_month
## 10170           5
## 10171           6
## 10172           7

## END CODE SET NN
```

There were a total of 10172 person-months, and 1228 women in the sample we will use to implement the g formula. Bottle-weight measurements taken at regular intervals over follow-up were used to quantify compliance. Let's look at how compliance patterns differed between the treatment and control group:

```
aspirin$mm1<-aspirin$study_month-1;span<-.6
par(mar = c(5,5,2,2),mgp = c(3, .5, 0))
lo<-loess(compliance~mm1,data=aspirin,span=span)
j <- order(aspirin$mm1)
plot(aspirin$mm1[j],lo$fitted[j],type="l",las=1,lwd=2,col="white",axes=F,
     ylim=c(0,1),xlim=c(-0.25,14.25),xlab=NA,ylab=NA,xaxs="i",yaxs="i")#
for(k in 1:2){
  lo<-loess(compliance~mm1,data=aspirin[aspirin$treatment==k-1,],span=span)
  j <- order(aspirin[aspirin$treatment==k-1,]$mm1)
  lines(aspirin[aspirin$treatment==k-1,]$mm1[j],lo$fitted[j],ylim=c(0,1),
        lwd=1,col=c("red","blue")[k])#
}
mtext(side=1,line=2,"Month Since Randomization")
mtext(side=2,line=2,"Proportion Compliant")
axis(side=2,las=1,col="black",lwd=1,tcl=-.1)
axis(side=1,las=1,lwd=1,tcl=-.1)
legend(6,.2,legend=c("Treatment","Placebo"),lty=c(1,1),lwd=c(2.5,2.5),
       col=c("blue","red"),bty="n")
box()
```



This Figure shows a moderate degree of noncompliance that was slightly higher in the placebo group over early follow-up. Let's also look at the outcome distributions including live birth, pregnancy loss, and end of follow-up without pregnancy (no pregnancy) among all women in the EAGeR trial:

```
t1a<-addmargins(table(aspirin[aspirin$last==1,]$final_outcome,
                      aspirin[aspirin$last==1,]$treatment))
attributes(t1a)$dimnames[[1]]<-c("Live Birth (Y)","Pregnancy Loss (D)",
                                 "Withdrawal (C)","No Pregnancy (S)","Total")
attributes(t1a)$dimnames[[2]]<-c("Placebo","Aspirin","Total")
t1a
```

```
##
##                       Placebo Aspirin Total
##   Live Birth (Y)          173     134   307
##   Pregnancy Loss (D)      286     309   595
##   Withdrawal (C)           92      97   189
```

```
##   No Pregnancy (S)        62      75   137
##   Total                  613     615  1228
```

This Table shows the total number of each outcome, overall and within treatment and placebo groups, and can be used to compute the ITT risk differences/ratios. In our synthetic data, the overall risk differences of aspirin assignment on live birth was 3.59 (95% CI: -1.58,9.36). Because of the design of the EAGeR study, there were no live births prior to 5 months on study. Follow up was stopped if conception did not occur within 6 months of study start. These design features will be used when we implement the parametric g formula, as explained below.

The Parametric G Formula for EAGeR

We will again rely on the Monte Carlo method to estimate the expectations needed to quantify the per protocol effect. To do this, we need to **setup** the problem, which starts with assuming a causal ordering of each relevant variable for any given month on study:

Order	Notation	Variable
8	Y	Live birth
7	D	pregnancy loss
6	S	No pregnancy
5	C	Withdrawal
4	Z	Conception
3	X	Compliance
2	N	GI Symptoms
1	B	Bleeding

Table 4: Assumed causal ordering of variables in any given month among 1,228 women in the EAGeR study, 2006-2012.

This ordering implies that, at any given month, a woman's decision to comply with her assigned treatment will depend on that month's bleeding and GI symtoms. This decision to comply will then affect conception. Bleeding, GI symptoms, compliance, and conception will then affect the decision to remain in the study or withdraw. If the woman remains on study, and did not yet conceive, her follow-up may end without pregnancy. If she does conceive, the conception may end in pregnancy loss, or live birth. If, at any given time point, the true ordering of compliance with respect to other variables is unknown, sensitivity analyses in which different orderings are evaluated is required (see validation section).

Even though we assume bleeding and GI symptoms are temporally (causally) prior to compliance in month j, the g formula allows compliance in month j to affect bleeding and GI symptoms in all months $> j$. That is, this presumed ordering does not preclude the possibility that variables in prior months may affect variables anywhere in the ordering in subsequent months, with the exception of terminal events (e.g., live birth, pregnancy loss). Furthermore, though the variables in this Table cannot affect baseline variables (denoted V), they may be affected by them.

Now that we've specified a causal ordering, let's continue our **setup** and write down the g formula for the per protocol averages. We are trying to quantify:

$$P(Y^{r=\overline{x}} = 1),$$

which can be interpreted as the risk of the outcome (in this case, live birth) that would be observed if someone were assigned to treatment level r, and their complete course of compliance over follow-up matched what they were randomized to. Under our selected causal ordering, we can define the marginal probability of Y as a function of the conditional probabilities of all relevant variables over all months of follow-up via the law of total probability. Applying this to all variables over all months yields the following compact equation:

$$P(Y_k^{\overline{x}_k, r, \overline{c}=0} = 1) = \sum_{m=0}^{k} \sum_{\overline{w}_m} P(Y_m = 1 \mid \overline{X}_m := \overline{x}_m, R := r, \overline{C}_m := 0, \overline{W}_m = \overline{w}_m) \prod_{j=0}^{m} f(w_j \mid \overline{X}_m := \overline{x}_m, R := r, \overline{C}_m := 0, \overline{W}_{j-1} = \overline{w}_{j-1}), \tag{1}$$

where $:=$ denotes "setting" the random variable to a fixed value, and $f(w_j \mid \bullet)$ denotes the joint distribution of all time-dependent and baseline covariates w. We quantify the probability in equation 1 twice: first setting assignment to $R = 1$, and all compliance values to $\overline{x} = 1$ and then setting assignment to $R = 0$, and all compliance values to $\overline{x} = 0$. Taking the difference or ratio of these two probabilities

provides the causal effect of interest.

This equation is essentially an application of the law of total probability to the variables in our EAGeR data. If we expanded this equation and used the same notation as in the causal ordering table, we would obtain:

$$
\begin{aligned}
P(Y_k = 1) = \sum_{m=1}^{k} \sum_{x,z,b,v} \Bigg\{ & P\big(Y_m = 1 \mid \overline{X}_m = \overline{x}_m, \overline{Z}_m = \overline{z}_m, \overline{B}_m = \overline{b}_m, \overline{V} = \overline{v}, \overline{Y}_{m-1} = \overline{D}_m = \overline{S}_m = \overline{C}_m = 0\big) \\
& \prod_{j=0}^{m} \Bigg[\begin{array}{l}
P\big(D_j = 0 \mid \overline{X}_j = \overline{x}_j, \overline{B}_j = \overline{b}_j, \overline{V} = \overline{v}, \overline{Y}_{j-1} = \overline{D}_{j-1} = \overline{S}_j = \overline{C}_j = 0, Z_j = 1\big) \times \\
P\big(S_j = 0 \mid \overline{X}_j = \overline{x}_j, \overline{B}_j = \overline{b}_j, \overline{V} = \overline{v}, \overline{Y}_{j-1} = \overline{D}_{j-1} = \overline{S}_{j-1} = \overline{C}_j = \overline{Z}_j = 0\big) \times \\
P\big(C_j = 0 \mid \overline{X}_j = \overline{x}_j, \overline{Z}_j = \overline{z}_j, \overline{B}_j = \overline{b}_j, \overline{V} = \overline{v}, \overline{Y}_{j-1} = \overline{D}_{j-1} = \overline{S}_{j-1} = \overline{C}_{j-1} = 0\big) \times \\
\Big[P\big(Z_j = 1 \mid \overline{X}_j = \overline{x}_j, \overline{B}_j = \overline{b}_j, \overline{V} = \overline{v}, \overline{Y}_{j-1} = \overline{D}_{j-1} = \overline{S}_{j-1} = \overline{C}_{j-1} = 0\big)(1 - Z_{j-1}) + Z_{j-1} \Big] \times \\
P\big(B_j = 1 \mid \overline{X}_j = \overline{x}_j, \overline{B}_{j-1} = \overline{b}_{j-1}, \overline{Z}_{j-1} = \overline{z}_{j-1}, \overline{V} = \overline{v}, \overline{Y}_{j-1} = \overline{D}_{j-1} = \overline{S}_{j-1} = \overline{C}_{j-1} = 0\big) \times \\
P\big(X_j = 1 \mid \overline{X}_{j-1} = \overline{x}_{j-1}, \overline{B}_{j-1} = \overline{b}_{j-1}, \overline{Z}_{j-1} = \overline{z}_{j-1}, \overline{V} = \overline{v}, \overline{Y}_{j-1} = \overline{D}_{j-1} = \overline{S}_{j-1} = \overline{C}_{j-1} = 0\big) \times \\
P\big(Y_{j-1} = 1 \mid \overline{X}_k = \overline{x}_k, \overline{Z}_k = \overline{z}_k, \overline{B}_k = \overline{b}_k, \overline{V} = \overline{v}, \overline{Y}_{j-2} = \overline{D}_{j-1} = \overline{S}_{j-1} = \overline{C}_{j-1} = 0\big) \times f(v)
\end{array} \Bigg] \Bigg\}
\end{aligned}
$$

If this lengthy equation still doesn't help clarify how to implement the g formula to estimate the per protocol effect, we can interpret it as an algorithm depicted in the following Figure:

This Figure shows the process by which the g formula takes as

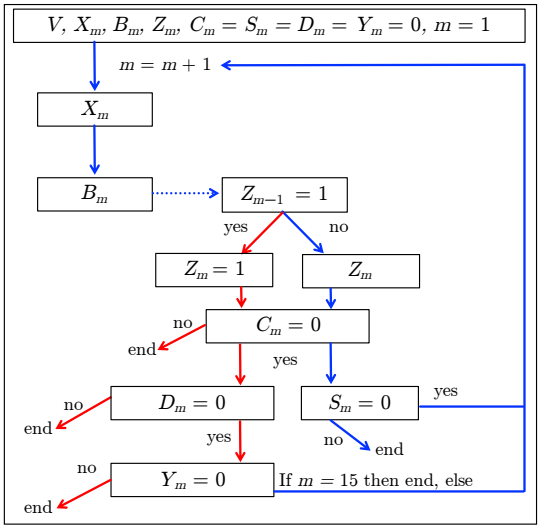


Figure 6: Algorithmic description of the parametric g formula displayed in equation efgform. The algorithm begins with a random resample of all baseline covariates (V) and all time-varying covariates (X, B, Z) at the first month on study ($m = 1$) from the observed data. These covariates are used to predict compliance and then bleeding at the second month on study (X_2, B_2). Each observation's conception value at the previous month ($m = 1$) on study is then examined. If conception occurred in $m = 1$, then in $m = 2$ the observation is at risk pregnancy loss (D_m) or live birth (Y_m). If conception did not occur, then the observation is at risk of ending follow-up without pregnancy (S_m). If any terminal events occur, follow-up is ended. If not, the the month is advanced by one and the algorithm is re-iterated until a terminal event occurs, or m reaches 15. In this manner, a simulated follow-up for each observation is generated. All predictions are simulated from regression models fit in the observed data.

input all baseline variables (V) and each time-dependent variable at the first month on study ($m = 1$, rectangular box at top of diagram), and generates a predicted compliance value for the second month on study ($m = 2$). The predicted bleeding value is then generated using predicted compliance at $m = 2$, observed time-dependent covariates at $m = 1$, and all baseline covariates. If, in the previous time-point ($m = 1$), conception occurred, then the conception value at the subsequent time-point ($m = 2$) will be set to one. Otherwise it will be predicted using the bleeding and compliance values generated for $m = 2$, the time-dependent covariate values at $m = 1$, and all baseline covariates. If Z_m is predicted to be one, then the algorithm proceeds down the left path in which withdrawal, pregnancy loss, and live birth are predicted to occur. Otherwise, the algorithm proceeds down the right path in which end of follow-up without pregnancy is predicted. If no terminal events are predicted to occur (pregnancy loss, live birth, end of follow-up without pregnancy), and $m \leq 15$, then m is advanced by one and the algorithm iterated.

As mentioned above, certain EAGeR design features can (should) be used when writing the code to implement the parametric g formula. We can improve the performance of the algorithm by refining models to account for the natural history of pregnancy. For example, there were no live births prior to 8 months on study, as all women

were randomized pre-conception. Therefore, we set Y to zero for $m < 8$. Second, no pregnancy loss occurred in the first two months on study, so we set D to zero for $m <= 2$. Finally, follow-up ended if a woman did not become pregnant in the first six months on study. Therefore, if no conception was predicted to occur within $m = 6$, the algorithm was stopped by setting S to one.

This algorithmic approach to solving for the g formula in equation 1 relies on Monte Carlo integration (Metropolis and Ulam 1949), which is subject to variation from both empirical and simulation sources. To reduce simulation variation, we input into the algorithm a random resample (with replacement) of 10,000 women from the original sample of 1,228 participants. We chose 10,000 as our target resample size because no notable improvements occurred with a target sample of 15,000. In the context of a randomized trial, empirical variation can be reduced by accounting for study design, which can be accomplished by resampling (with replacement) within levels of the randomization indicators (5,000 in each group).

To estimate the per-protocol effect in the absence of withdrawal, we run the algorithm twice. First by setting both randomization and compliance to one, and setting withdrawal to zero for all observations over all follow-up, and again by setting randomization and compliance and withdrawal to zero. These yield outcomes distributed as they would if all women (first) and no women (second) took aspirin over follow-up, and no one withdrew from the study. We then take the difference and ratio of averages of these outcomes to estimate the causal risk differences and ratios of interest.

Each of the blue arrows connecting one set of covariates to the next in the above Figure represents a parametric model in which the variable to be predicted is modeled conditional on all variables that precede it. These models must be fit in the original sample of 1,228 women, and not a larger resample. Furthermore, assuming that conditional on all covariates in each model, withdrawal from the study occurs at random, these models are fit among the subset of women who remained on study.

For example, the arrow from all baseline covariates and time-

dependent covariates at time $m = 1$ to X_m in the Figure represents a pooled logistic regression model for the probability of $X_j = 1$ in long equation above. In our implementation, this model was specified as:

$$\begin{aligned}\text{logit}\, P\big(X_j = 1 \mid \overline{X}_{j-1}, \overline{N}_{j-1}, \overline{B}_{j-1}, \overline{N}_{j-1}, \overline{Z}_{j-1}, \overline{V}, \overline{C}_J = 0\big) \\ = \beta_j + \beta_1 X_{j-1} + \beta_2 N_{j-1} + \beta_3 N_{j-2} \\ + \beta_4 B_{j-1} + \beta_5 B_{j-2} + \beta_6 Z_{j-1} + \beta_7 Z_{j-2} + \boldsymbol{\beta}_V^{\intercal} V,\end{aligned} \tag{2}$$

where $\text{logit}\bullet \equiv \frac{\log(\bullet)}{\log(1-\bullet)}$, V represents a vector of baseline covariates that includes study site, randomization date, eligibility stratum, age at study entry, income, race/ethnicity, education, marital status, employment status, and where the study time-scale $j = 1 \ldots J$ reflects each person's month since randomization. The intercept β_j is month specific, which can be accomplished by including the month since randomization variable, and some terms (e.g., splines or fractional polynomials) that account for potential nonlinearity.

In R, this model would look something like:

```
library(splines)
fitX<-glm(X~V1+V2+V3+V4+V5+V6+V7+ns(V8,df=3)+V9+V10+V11+V12+V13+ns(V14,df=3)+
                Xl+Xl1+B+Bl+Bl1+N+Nl+Nl1+Z+Zl+Zl1+ns(jj,df=3),
                 family=binomial,data=boot,subset=R==k)
```

Note this model assumes that the histories $\overline{X}_{j-1}$, $\overline{B}_{j-1}$, $\overline{N}_{j-1}$, and $\overline{Z}_{j-1}$ in the conditioning statement of the left hand side of equation 2 are adequately represented by the values of these variables in the previous two time-points (i.e., at months $j-1$ and $j-2$). These correspond to "Xl" and "Xl1" in the R code above. This is an assumption whose potential effects could be evaluated by adding lagged terms for each variables to the right hand side of equation 2, and by including relevant interactions.

The following code implements the parametric g formula to estimate the per protocol effect:

```
# load required libraries
library(data.table)
```

```
library(splines)
# #################

setwd("~/Dropbox/Documents/Research/Presentations/SER 2017/SER Workshop/pgf_course-master/")

## CODE SET NN
# import data
aspirin<-read.table("aspirin2.txt",header=T,sep="\t")

## look at data, re-assign to a2
a2<-aspirin
#head(a2)

## create variables: baseline confounders
SS<-data.frame(model.matrix(~factor(a2$site))[,-1]);names(SS)<-c(paste("V",1:6,sep=""))
a2<-cbind(a2,SS)
names(a2)[names(a2)=="eligibility"]<-"V7"
names(a2)[names(a2)=="age"]<-"V8"
names(a2)[names(a2)=="income"]<-"V9"
names(a2)[names(a2)=="education"]<-"V10"
names(a2)[names(a2)=="white"]<-"V11"
names(a2)[names(a2)=="marital"]<-"V12"
names(a2)[names(a2)=="employed"]<-"V13"
names(a2)[names(a2)=="BMI"]<-"V14"
names(a2)[names(a2)=="treatment"]<-"R"

## time-dependent variables
names(a2)[names(a2)=="study_month"]<-"j"
  a2$jj<-scale(a2$j);mean_j<-attributes(a2$jj)$‘scaled:center‘;sd_j<-attributes(a2$jj)$‘scaled:scale‘
names(a2)[names(a2)=="compliance"]<-"X"
names(a2)[names(a2)=="bleeding"]<-"B"
names(a2)[names(a2)=="gastro"]<-"N"
names(a2)[names(a2)=="conceived"]<-"Z"

# outcomes
```

```
names(a2)[names(a2)=="efuwp"]<-"S"
names(a2)[names(a2)=="pregnancy_loss"]<-"D"
names(a2)[names(a2)=="live_birth"]<-"Y"

# lag variables
a2<-data.table(a2)
a2[, Xl:=c(0, X[-.N]), by=id]
a2[, Bl:=c(0, B[-.N]), by=id]
a2[, Nl:=c(0, N[-.N]), by=id]
a2[, Zl:=c(0, Z[-.N]), by=id]
a2[, Xl1:=c(0, Xl[-.N]), by=id]
a2[, Bl1:=c(0, Bl[-.N]), by=id]
a2[, Nl1:=c(0, Nl[-.N]), by=id]
a2[, Zl1:=c(0, Zl[-.N]), by=id]
head(a2)

##    id j V7         V8 V9 V10 V11 V12 V13
## 1:  1 1  1 -0.7402309  0   1   1   0   1
## 2:  1 2  1 -0.7402309  0   1   1   0   1
## 3:  1 3  1 -0.7402309  0   1   1   0   1
## 4:  1 4  1 -0.7402309  0   1   1   0   1
## 5:  1 5  1 -0.7402309  0   1   1   0   1
## 6:  1 6  1 -0.7402309  0   1   1   0   1
##          V14 X R B N Z S D Y last site V1 V2
## 1: -1.040349 1 0 0 0 0 0 0 0    0    3  0  0
## 2: -1.040349 1 0 0 0 0 0 0 0    0    3  0  0
## 3: -1.040349 1 0 0 0 0 0 0 0    0    3  0  0
## 4: -1.040349 0 0 0 0 1 0 0 0    0    3  0  0
## 5: -1.040349 1 0 1 0 1 0 0 0    0    3  0  0
## 6: -1.040349 0 0 0 0 1 0 0 0    0    3  0  0
##    V3 V4 V5 V6         jj Xl Bl Nl Zl Xl1
## 1:  1  0  0  0 -1.3469120  0  0  0  0   0
## 2:  1  0  0  0 -1.0506832  1  0  0  0   0
## 3:  1  0  0  0 -0.7544544  1  0  0  0   1
## 4:  1  0  0  0 -0.4582256  1  0  0  0   1
## 5:  1  0  0  0 -0.1619968  0  0  0  1   1
```

```
## 6:  1  0  0  0  0.1342321  1  1  0  1   0
##    Bl1 Nl1 Zl1
## 1:   0   0   0
## 2:   0   0   0
## 3:   0   0   0
## 4:   0   0   0
## 5:   0   0   0
## 6:   0   0   1
```

```
# indicator of last record
# a2$last<-as.numeric(!duplicated(a2$id,fromLast=T))

# look at summaries of data
# summary(a2)

# function to bootstrap the g formula
seed=0
set.seed(seed)
clusters <- names(table(a2$id))
index <- sample(1:length(clusters), length(clusters), replace=TRUE)
bb <- table(clusters[index])
boot <- NULL
if(seed==0){
  boot<-a2
} else{
for(jj in 0:max(bb)){
  cc <- a2[a2$id %in% names(bb[bb %in% c(jj:max(bb))]),]
  cc$bid<-paste0(cc$id,jj)
  boot <- rbind(boot, cc)
}}
identical(boot,a2)
```

```
## [1] TRUE
```

```
# fit models
Rfit<-function(k){
  ## time-varying confoudner 1 (bleeding)
```

```
fitB<-glm(B~V1+V2+V3+V4+V5+V6+V7+ns(V8,df=3)+V9+V10+V11+V12+V13+ns(V14,df=3)+
            Xl+Xl1+Bl+Bl1+Nl+Nl1+Zl+Zl1+ns(jj,df=3),
             family=binomial,data=boot,subset=R==k)
  ## time-varying confoudner 2 (gastro)
  fitN<-glm(N~V1+V2+V3+V4+V5+V6+V7+ns(V8,df=3)+V9+V10+V11+V12+V13+ns(V14,df=3)+
            Xl+Xl1+B+Bl+Bl1+Nl+Nl1+Zl+Zl1+ns(jj,df=3),
             family=binomial,data=boot,subset=R==k)
  #time-varying confounder 3 (conception)
  fitZ<-glm(Z~V1+V2+V3+V4+V5+V6+V7+ns(V8,df=3)+V9+V10+V11+V12+V13+ns(V14,df=3)+
              Xl+B+Bl+N+Nl+ns(jj,df=3),
                family=binomial,data=boot,subset=Zl==0&R==k)
  ## compliance
  fitX<-glm(X~V1+V2+V3+V4+V5+V6+V7+ns(V8,df=3)+V9+V10+V11+V12+V13+ns(V14,df=3)+
              Xl+Xl1+B+Bl+Bl1+N+Nl+Nl1+Z+Zl+Zl1+ns(jj,df=3),
               family=binomial,data=boot,subset=R==k)
  #outcome 1 (end of follow-up without pregnancy)
  fitS<-glm(S~V1+V2+V3+V4+V5+V6+V7+ns(V8,df=3)+V9+V10+V11+V12+V13+ns(V14,df=3)+
              X+Xl+B+Bl+N+Nl+jj,
               family=binomial,data=boot,subset=Z==0&R==k)
  #outcome 2 (pregnancy loss)
  fitD<-glm(D~V1+V2+V3+V4+V5+V6+V7+ns(V8,df=3)+V9+V10+V11+V12+V13+ns(V14,df=3)+
              X+Xl+B+Bl+N+Nl+ns(jj,df=3),
               family=binomial,data=boot,subset=Z==1&R==k)
  #outcome 3 (live birth)
  fitY<-glm(Y~V1+V2+V3+V4+V5+V6+V7+ns(V8,df=3)+V9+V10+V11+V12+V13+ns(V14,df=3)+
              X+Xl+X*Xl+B+Bl+N+Nl+ns(jj,df=1),
               family=binomial,data=boot,subset=Z==1&R==k)
  return(list(fitX,fitB,fitN,fitZ,fitS,fitD,fitY))
}
fitR<-lapply(0:1,function(x) Rfit(x))

#create object to hold results from pgf function
cols<-c("boot","id","j","Y")
res.g <- data.frame(matrix(nrow=1,ncol=length(cols))); colnames(res.g) <- cols
```

```
# PREDICT FOLLOW-UP BASED ON G FORMULA USING PGF FUNCTION
pgf<-function(ii, mc_data, length, randomization = NULL, exposure = NULL){

    # define generic prediction function
    pFunc<-function(mod,ndat){as.numeric(
      predict(mod,newdata=ndat,type="response")>runif(1))}
    d <- mc_data
    d<-d[d$id==ii,]
    lngth <- length
    Vp <- Rp <- Bp <- Np <- Zp <- Xp <- Sp <- Dp <- Yp <- mm <- numeric()
    mm[1] <- j <- 1
    id <- d$id
    Vp <- d[, c("V1", "V2", "V3", "V4", "V5", "V6", "V7","V8","V9","V10","V11","V12","V13","V14")]
    if (is.null(randomization)) {
      Rp <- d$R
    } else {
      Rp <- randomization
    }
    if (is.null(exposure)) {
      Xp[1] <- d$X
    } else {
      Xp[1] <- exposure
    }
    Bp[1] <- d$B
    Np[1] <- d$N
    Zp[1] <- d$Z
    Sp[1] <- Yp[1] <- 0
    dDp <- data.table(Vp, X = Xp[j], Xl = 0, Xl1 = 0,
                      B = Bp[j], Bl = 0, Bl1 = 0, N = Np[j],
                      Nl = 0, Nl1 = 0,jj=((j-mean_j)/sd_j))
    Dp[1] <- ifelse(Zp[1] == 1,
                    pFunc(fitR[[Rp+1]][[6]], dDp),
                    0)

    for (j in 2:lngth) {
```

```
if (Sp[j - 1] == 0 & Dp[j - 1] == 0 & Yp[j - 1] == 0) {
  if (j==2){
    Xl1<-Bl1<-Nl1<-Zl1<-0
  } else{
    Xl1 <- Xp[j - 2]
    Bl1 <- Bp[j - 2]
    Nl1 <- Np[j - 2]
    Zl1 <- Zp[j - 2]
  }

  dBp <- data.table(Vp, X = Xp[j], Xl = Xp[j - 1], Xl1,
                    Zl=Zp[j-1], Zl1,Bl = Bp[j - 1], Bl1,
                    Nl = Np[j - 1], Nl1, jj=((j-mean_j)/sd_j))
  Bp[j] <- pFunc(fitR[[Rp+1]][[2]], dBp)

  dNp <- data.table(Vp, X = Xp[j], Xl = Xp[j - 1], Xl1,
                    B = Bp[j], Bl = Bp[j - 1], Bl1,
                    Nl = Np[j - 1], Nl1, Zl=Zp[j-1],
                    Zl1, jj=((j-mean_j)/sd_j))
  Np[j] <- pFunc(fitR[[Rp+1]][[3]], dNp)

  dZp <- data.table(Vp, X = Xp[j], Xl = Xp[j - 1],
                    B = Bp[j], Bl = Bp[j - 1],
                    N = Np[j], Nl = Np[j - 1],jj=((j-mean_j)/sd_j))
  if (Zp[j - 1] == 0){
    Zp[j] <- pFunc(fitR[[Rp+1]][[4]], dZp)
  } else {
    Zp[j] <- 1
  }

  if (is.null(exposure)) {
    dXp <- data.table(Vp, Xl = Xp[j - 1], Xl1, Zl=Zp[j-1],
                      Zl1, B=Bp[j], Bl = Bp[j - 1], Bl1,
                      N=Np[j],Nl = Np[j - 1], Nl1,Z=Zp[j],jj=((j-mean_j)/sd_j))
    Xp[j] <- pFunc(fitR[[Rp+1]][[1]], dXp)
```

```
  }
  else {
    Xp[j] <- exposure
  }

  dSp <- data.table(Vp, X = Xp[j], Xl = Xp[j - 1],
                    B = Bp[j], Bl = Bp[j - 1],
                     N = Np[j], Nl = Np[j -1],jj=((j-mean_j)/sd_j))
   if (j > 4 & Zp[j] == 0) {
     Sp[j] <- pFunc(fitR[[Rp+1]][[5]], dSp)
   } else {
     Sp[j] <- 0
   }

   dDp <- data.table(Vp, X = Xp[j], Xl = Xp[j - 1],
                     Xl1, B = Bp[j], Bl = Bp[j - 1], Bl1,
                     N = Np[j], Nl = Np[j - 1], Nl1,jj=((j-mean_j)/sd_j))
   if (j < 11 & Sp[j] == 0 & Zp[j] == 1){
     Dp[j] <- pFunc(fitR[[Rp+1]][[6]], dDp)
   } else {
     Dp[j] <- 0
   }
   dYp <- data.table(Vp, B = Bp[j], Bl = Bp[j - 1],
                     N = Np[j], Nl = Np[j - 1],
                     X=Xp[j],Xl=Xp[j-1],jj=((j-mean_j)/sd_j))
   if (j > 8 & Dp[j] == 0 & Sp[j] == 0 & Zp[j] == 1){
     Yp[j] <- pFunc(fitR[[Rp+1]][[7]], dYp)
   } else {
     Yp[j] <- 0
   }
}
else {
  break
}
mm[j] <- j
```

```
      if(Sp[j]==1|Dp[j]==1|Yp[j]==1|j==15){
        res.g$boot<-seed;res.g$id<-ii
        res.g$j<-j;res.g$Y<-Yp[j]
      }
    }
    #print(ii)
    return(res.g)
}

## create monte carlo dataset
# select first obs for each person to obtain joint empirical distribution of baseline covariates
MC0<-boot[boot$j==1,]
montecarlo<-500

# sample with replacement from each treatment arm
spl <- split(MC0, list(MC0$R))
samples <- lapply(spl, function(x) x[sample(1:nrow(x), montecarlo/length(spl), replace=T),])
MC <- rbindlist(samples)
MC$id<-1:montecarlo
bn<-lapply(1:montecarlo,function(i) pgf(ii=i,mc_data=MC,length=15,randomization=NULL,exposure=NULL))
bn<-rbindlist(bn)
head(bn,3);tail(bn,3)

##    boot id j Y
## 1:    0  1 4 0
## 2:    0  2 6 0
## 3:    0  3 7 0

##    boot  id  j Y
## 1:    0 498 13 1
## 2:    0 499 10 1
## 3:    0 500 11 1

b1<-lapply(1:montecarlo,function(i) pgf(ii=i,mc_data=MC,length=15,randomization=1,exposure=1))
b1<-rbindlist(b1)
head(b1,3);tail(b1,3)

##    boot id j Y
```

```
## 1:    0  1 2 0
## 2:    0  2 7 0
## 3:    0  3 6 0
```

```
##    boot  id  j Y
## 1:    0 498  7 0
## 2:    0 499 10 1
## 3:    0 500  9 1
```

```
b0<-lapply(1:montecarlo,function(i) pgf(ii=i,mc_data=MC,length=15,randomization=0,exposure=0))
b0<-rbindlist(b0)
head(b0,3);tail(b0,3)
```

```
##    boot id  j Y
## 1:    0  1  6 0
## 2:    0  2  7 0
## 3:    0  3 12 1
```

```
##    boot  id  j Y
## 1:    0 498  7 0
## 2:    0 499 13 1
## 3:    0 500  6 0
```

```
mean(bn$Y,na.rm=T);mean(a2[a2$last==1,]$Y)
```

```
## [1] 0.5381526
```

```
## [1] 0.517101
```

```
mean(b1$Y,na.rm=T)
```

```
## [1] 0.511022
```

```
mean(b0$Y,na.rm=T)
```

```
## [1] 0.448
```

```
(mean(b1$Y,na.rm=T)-mean(b0$Y,na.rm=T))*100
```

```
## [1] 6.302204
```

```
coef(lm(Y~R,data=a2[a2$last==1,]))[2]*100
```

```
##        R
## 6.351792
```

References

"Centers for Disease Control and Prevention, National Survey of Family Growth." Updated Jan 22 2014. Accessed May 27 2014. `http://www.cdc.gov/nchs/nsfg.htm`.

Daniel, R.M., S.N. Cousens, B.L. De Stavola, M. G. Kenward, and J. A. C. Sterne. 2013. "Methods for Dealing with Time-Dependent Confounding." *Stat Med* 32 (9): 1584–1618.

Didelez, Vanessa, AP Dawid, and S Geneletti. 2006. "Direct and Indirect Effects of Sequential Treatments." In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, 138–46. Arlington, VA: AUAI Press.

Dunn, G., and M. Lovrić. 2011. "'Complier-Average Causal Effect (CACE) Estimation'." In *International Encyclopedia of Statistical Science*. Germany: Springer Verlag.

Ford, Ian, and John Norrie. 2016. "Pragmatic Trials." *!! No Full Title or Abbreviation for Line:* 375 (5): 454–63.

Greenland, Sander, and James Robins. 2009. "Identifiability, Exchangeability and Confounding Revisited." *Epidemiol Perspect Innov* 6 (1): 4.

Greenland, Sander, and JM Robins. 1986. "Identifiability, Exchangeability, and Epidemiological Confounding." *Int J Epidemiol* 15 (3): 413–19.

Greenland, Sander, James M. Robins, and Judea Pearl. 1999. "Confounding and Collapsibility in Causal Inference." *Stat Sci* 14 (1): 29–46.

Halloran, M Elizabeth, and Michael G Hudgens. 2016. "Dependent Happenings: A Recent Methodological Review." *Curr Epidemiol Rep* 3 (4): 297–305.

Hernan, M A, and S L Taubman. 2008. "Does Obesity Shorten Life? The Importance of Well-Defined Interventions to Answer Causal Questions." *Int J Obes* 32 (S3): S8–S14.

Hernán, M. A., and JM Robins. Forthcoming. *Causal Inference*. Boca Raton, FL: Chapman/Hall.

Hernán, Miguel A, and Sonia Hernández-Díaz. 2012. "Beyond the

Intention-to-Treat in Comparative Effectiveness Research." *Clin Trials* 9 (1): 48–55.

Hernán, Miguel A, and Tyler J VanderWeele. 2011. "Compound Treatments and Transportability of Causal Inference." *Epidemiol* 22 (3): 368–77. doi:10.1097/EDE.0b013e3182109296.

Hernán, Miguel A. 2005. "Invited Commentary: Hypothetical Interventions to Define Causal Effects—Afterthought or Prerequisite?" *Am J Epidemiol* 162 (7): 618–20.

Hernán, Miguel A., Sonia Hernández-Díaz, and James M. Robins. 2013. "Randomized Trials Analyzed as Observational Studies." *Ann Intern Med* 159 (8): 560–62.

Hudgens, M. G., and M. E. Halloran. 2008. "Toward Causal Inference with Interference." *J Am Stat Assoc* 103 (482): 832–42.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Kaufman, Jay S, Richard F Maclehose, and Sol Kaufman. 2004. "A Further Critique of the Analytic Strategy of Adjusting for Covariates to Identify Biologic Mediation." *Epidemiol Perspect Innov* 1 (1): 4.

Lin, Sheng-Hsuan, Jessica Young, Roger Logan, Eric J. Tchetgen Tchetgen, and Tyler J. VanderWeele. 2017. "Parametric Mediational G-Formula Approach to Mediation Analysis with Time-Varying Exposures, Mediators, and Confounders." *Epidemiology* 28 (2). `http://journals.lww.com/epidem/Fulltext/2017/03000/Parametric_Mediational_g_Formula_Approach_to.16.aspx`.

Lok, Judith J. 2016. "Defining and Estimating Causal Direct and Indirect Effects When Setting the Mediator to Specific Values Is Not Feasible." *Statistics in Medicine* 35 (22): 4008–20. doi:10.1002/sim.6990.

Metropolis, N, and S Ulam. 1949. "The Monte Carlo method." *J Am Stat Assoc* 44 (247): 335–41.

Mortimer, Kathleen M, Romain Neugebauer, Mark van der Laan, and Ira B Tager. 2005. "An Application of Model-Fitting Procedures for Marginal Structural Models." *Am J Epidemiol* 162 (4): 382–88. doi:10.1093/aje/kwi208.

Naimi, Ashley I, Stephen R Cole, and Edward H Kennedy. 2016.

"An Introduction to G Methods." *Int J Epidemiol* In Press.

Naimi, Ashley I., and Jay S. Kaufman. 2015. "Counterfactual Theory in Social Epidemiology: Reconciling Analysis and Action for the Social Determinants of Health." *Curr Epidemiol Reports* 2 (1): 52–60.

Naimi, Ashley I., Jay S. Kaufman, and Richard F. Maclehose. 2014. "Mediation Misgivings: Ambiguous Clinical and Public Health Interpretations of Natural Direct and Indirect Effects." *Int J Epidemiol* 43 (5): 1656–61.

Naimi, Ashley I., Erica EM. Moodie, Nathalie Auger, and Jay S Kaufman. 2014. "Stochastic Mediation Contrasts in Epidemiologic Research: Interpregnancy Interval and the Educational Disparity in Preterm Birth." *Am J Epidemiol* 180 (4): 436–45.

Naimi, Ashley I., Mireille E. Schnitzer, Erica E. M. Moodie, and Lisa M. Bodnar. 2016. "Mediation Analysis for Health Disparities Research." *American Journal of Epidemiology* 184 (4): 315–24. doi:10.1093/aje/kwv329.

Pearl, Judea, Madelyn R Glymour, and Nicholas Jewell. 2016. *Causal Inference in Statistics: A Primer*. United Kingdom: Wiley.

Piantadosi, Steven. 2005. *Clinical Trials: A Methodologic Perspective*. Hoboken, NJ: Wiley-Interscience.

Prentice, Ross L., Mary Pettinger, and Garnet L. Anderson. 2005. "Statistical Issues Arising in the Women's Health Initiative." *Biometrics* 61 (4): 899–911.

Richardson, Thomas S, and James M Robins. 2013. "Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality." Number 128. http://www.csss.washington.edu/Papers/wp128.pdf. Accessed Aug 26th, Center for Statistics; the Social Sciences, University of Washington.

Robins, James M, and Miguel Á Hernán. 2009. "Estimation of the Causal Effects of Time-Varying Exposures." In *Advances in Longitudinal Data Analysis*, edited by G Fitzmaurice, M. Davidian, G Verbeke, and G Molenberghs, 553–99. Boca Raton, FL: Chapman & Hall.

Robins, JM. 1987. "Addendum to 'a New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period– Application to Control of the Healthy Worker Survivor Effect'." *Comp*

Math Appl 14 (9-12): 923–45.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes." *J Am Stat Assoc* 100 (469): 322–31.

Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *J Am Stat Assoc* 75 (371): 591–93.

Schisterman, Enrique F, Robert M Silver, Neil J Perkins, Sunni L Mumford, Brian W Whitcomb, Joseph B Stanford, Laurie L Lesher, et al. 2013. "A Randomised Trial to Evaluate the Effects of Low-Dose Aspirin in Gestation and Reproduction: Design and Baseline Characteristics." *Paediatr Perinat Epidemiol* 27 (6): 598–609. doi:10.1111/ppe.12088.

Shrier, Ian, Russell J Steele, Evert Verhagen, Rob Herbert, Corinne A Riddell, and Jay S Kaufman. 2014. "Beyond Intention to Treat: What Is the Right Question?" *Clin Trials* 11 (1): 28–37. doi:10.1177/1740774513504151.

Tchetgen Tchetgen, Eric J, and Tyler J VanderWeele. 2012. "On Causal Inference in the Presence of Interference." *Stat Methods in Med Res* 21 (1): 55–75.

VanderWeele, Tyler J, and Miguel Ángel Hernán. 2013. "Causal Inference Under Multiple Versions of Treatment." *Journal of Causal Inference* 1 (1): 1–20.

VanderWeele, Tyler J., Stijn Vansteelandt, and James M. Robins. 2014. "Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder." *Epidemiol* 25 (2): 300–306.

Westreich, Daniel, and Stephen R. Cole. 2010. "Invited Commentary: Positivity in Practice." *Am J Epidemiol* 171 (6): 674–77.