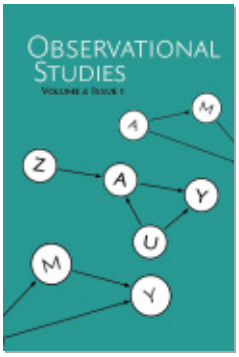
The Validity and Efficiency of Hypothesis Testing in Observational Studies with Time-Varying Exposures



Harlan Campbell, Paul Gustafson

Observational Studies, Volume 4, Issue 1, 2018, pp. 260-291 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2018.0010*

Observational Studies 4 (2018) 260-291 Submitted 6/17; Published 8/18

The Validity and Efficiency of Hypothesis Testing in Observational Studies with Time-Varying Exposures

Harlan Campbell **harlan.campbell@stat.ubc.ca**
Department of Statistics
University of British Columbia
Vancouver, Canada

Paul Gustafson **gustaf@stat.ubc.ca**
Department of Statistics
University of British Columbia
Vancouver, Canada

Abstract

The fundamental obstacle of observational studies is that of unmeasured confounding. If all potential confounders are measured within the data, and treatment occurs at but a single time-point, conventional regression adjustment methods provide consistent estimates and allow for valid hypothesis testing in a relatively straightforward manner. In situations for which treatment occurs at several successive timepoints, as in many longitudinal studies, another type of confounding is also problematic: even if all confounders are known and measured in the data, time-dependent confounding may bias estimates and invalidate testing due to collider-stratification. While "causal inference methods" can adequately adjust for time-dependent confounding, these methods require strong and unverifiable assumptions. Alternatively, instrumental variable analysis can be used. By means of a simple illustrative scenario and simulation studies, this paper sheds light on the issues involved when considering the relative merits of these two approaches for the purpose of hypothesis testing in the presence of time-dependent confounding.

Keywords: Observational studies; g-computation; Instrumental variable analysis; Statistical power; Time-varying confounding; Causal inference.

1. Introduction

The fundamental obstacle of observational studies is that of unmeasured confounding. If all potential confounders are measured within the data, and treatment occurs at but a single time-point, standard regression adjustment methods provide consistent estimates and valid hypothesis testing in a relatively straightforward manner. Alternatively, techniques such as propensity-score adjustment and inverse probability weighted estimation (IPW) can be employed. For lack of a better umbrella term, these techniques are subsequently referred to as *causal inference methods*.

The relative merits of causal inference methods and conventional regression techniques in the single time-point scenario have been previously investigated and widely discussed.

According to some, causal inference methods do not provide better control for unmeasured confounding relative to conventional regression, see for example the conclusions of Senn et al. (2007), Stürmer et al. (2006), and Biondi-Zoccai et al. (2011). In certain situations, causal inference methods may in fact be problematic for hypothesis testing. Austin et al. (2015) note that, among many exploratory simulation studies conducted, "the empirical type I error rate of the IPW design was often significantly different from the advertised rate of 0.05." Schuster et al. (2016) express similar concerns.

An alternative approach often considered is that of instrumental variable (IV) analysis. If one can identify a valid *instrument*, then valid testing and consistent estimates may be possible while circumventing the requirement that all confounders are known and measured; see Greenland (2000) and Baiocchi et al. (2014). A number of observational studies compare results obtained by IV analysis, causal inference methods, and conventional regression adjustment (e.g. Brookhart, Wang, et al. (2006) and Hadley et al. (2010)).

For situations in which treatment occurs at several successive timepoints, as in many longitudinal studies, another type of confounding is also problematic. Even if all confounding variables are known and measured, *time-dependent confounding* can occur when there exists a factor, causally influenced by past treatment, that impacts future treatment while also impacting the outcome of interest; see Hernán et al. (2000) and Daniel et al. (2011). Time-dependent confounding will invalidate standard methods (i.e. regression adjustment) and can lead to what is known as "collider-stratification bias" (Daniel et al., 2013).

Unlike regression adjustment, causal inference methods for time-varying exposures (Robins & Hernán, 2009), such as g-computation and IPW, can adequately adjust for time-dependent confounding. However, these methods require rather strong and unverifiable assumptions, namely that all confounders are measured in the data and that the model is correctly specified. IV analysis avoids the problems of time-dependent confounding altogether, but is "poorly equipped" (Hernán & Robins, 2006) to address time-varying exposures. Despite limitations inherent to both approaches, each may be used for valid and efficient testing, depending on the particulars of the data. By means of a simple illustrative scenario and simulation studies, this paper aims to shed light on the issues involved when considering the relative merits of these two approaches for the purposes of testing in the presence of time-dependent confounding.

One specific motivation for this work arises in the evaluation of the long-term effectiveness of treatments for chronic diseases. For instance, consider β-interferon therapy for the treatment of relapsing-remitting multiple sclerosis (MS). This drug therapy gained regulatory approval on the basis of randomized controlled trials (RCTs) using outcome measures such as short-term risk of relapse. However, it is not practical to run a RCT for longer-term outcomes such as reaching irreversible disability; see Cohen & Rudick (2011). Hence, one must study the observational data of patients, who, in consultation with their physicians, start and stop β-interferon therapy as they see fit.

One strategy is to apply causal inference methods for time-varying exposures. For instance, Karim et al. (2014) use IPW estimation of marginal-structural models to study the long-term efficacy of β-interferon therapy in the MS context. A completely different

strategy is to view regulatory approval of the therapy as creating a quasi-natural experiment. Or, phrased slightly differently, a putative instrumental variable such as calendar period of disease onset can be employed. We are not aware of this approach being applied in the MS context, though it has been considered for other exposure-disease relationships, e.g. Cain et al. (2009) and Johnston et al. (2008).

So we have two rather different approaches to consider, each of which requires strong but different assumptions. Yet there seems to be little or no consideration of how the two approaches might fare relative to one another in a given setting. Even basic questions such as how much more power one approach might have, should both sets of assumptions be valid, has not received attention. Hence, one motivation for this work is to fill this void. This paper is structured as follows. In Section 2, we introduce the simple scenario referenced throughout the paper. Given that each approach requires a different set of assumptions, understanding the trade-offs in these is essential and in Section 3, we provide an overview. We also take the opportunity to discuss more nuanced considerations such as the "g-null paradox" of Robins (1986), and Martens et al. (2006)'s "limit on the strength of instruments."

An additional consideration is that of statistical power. Not only does a study with low statistical power have a reduced chance of detecting a true effect, low power reduces the likelihood that a statistically significant result reflects a true effect (Button et al., 2013). In Section 4, we investigate the power of causal inference methods relative to IV analysis and in Section 5, we investigate the type I error rate given modest violations to the assumptions required of each approach. We also consider a joint assessment of the type I error rate and power by comparing the two approaches in terms of the Bayes risk for selecting the correct hypothesis. In Section 6, motivated by the fact that more and more observational studies are using IV analysis and causal inference methods concurrently -often leading to contradictory results (Laborde-Castérot et al., 2015)– we briefly consider the implications of this type of multiple testing strategy.

2. The simple scenario

The scenario considered, based on examples introduced by Robins & Wasserman (1997), is the simplest possible in which the problem of time-dependent confounding can be examined: there are two timepoints and all variables are binary; see the directed acyclic graph (DAG) displayed in Figure 1. Dashed lines represent paths that would violate necessary assumptions of at least one method we consider. Consider the variables as unfolding in the temporal order (V, H, U, A_0, L_1, A_1, Y). For ease of exposition, regard $A_j = 1$ ($A_j = 0$) as being "on" ("off") treatment at the j-th time-point, while $Y = 1$ is a given outcome status. The variable V is an unmeasured factor influencing both the measured confounder L_1 and the outcome variable Y. The variable U is a possible unmeasured confounder and H is a potential instrument. Note that, while we only consider a binary instrument, the methods and trade-offs we discuss should be applicable to non-binary instruments as well.

The covariate L_1 is a "time-dependent confounder" since it lies on the causal pathway between A_0 and A_1 and is also a predictor of Y. Thus, despite being known and measured, L_1 has the potential to confound the causal relationship between the time-varying treatment (A_0, A_1) and the outcome of interest (Y). In addition, because V and A_0 are both causes of L_1 (i.e. L_1 is a "collider" on the path $A_0 \rightarrow L_1 \leftarrow V \rightarrow Y$), conditioning on L_1 in any standard model will induce a non-causal association between A_0 and V. As a result, "collider-stratification bias" will occur since the causal impact of the unmeasured factor V on Y can be mistaken for a non-existent causal impact of A_0 on Y; see Daniel et al. (2013).

In the drug therapy for MS context mentioned earlier, H could be later versus earlier calendar period at subject entry into the study cohort (with entry defined by disease onset), while A_0 and A_1 could indicate the use of drug therapy in the first and second years post-entry, with L_1 indicating relapse during the first year post-entry. The outcome Y could indicate whether or not the subject has undergone irreversible disease progression by the end of their second year in the cohort. The variable V could reflect subject frailty. If in play, the variable U could be some other confounding variable such as absence/presence of a concomitant symptom.

We denote $Pr(Y|do(A_0 = a_0, A_1 = a_1))$ as the distribution of Y given an intervention by which A_0 and A_1 are fixed at the values a_0 and a_1. In the presence of confounding, this differs from $Pr(Y|A_0 = a_0, A_1 = a_1)$, the distribution of Y, given A_0 and A_1 are *observed* to be a_0 and a_1. We define the counterfactual mean outcome as $\theta_{a_0a_1} := \mathrm{E}(Y|do(A_0 = a_0, A_1 = a_1))$, and the total effect of treatment as $\mathrm{TE} := \theta_{11} - \theta_{00}$. While the formalism of causal inference supports the estimation of many different inferential targets, the total effect contrasting the "always-treat" and "never-treat" strategies is an intuitive and general descriptor of the population.

As stated in the introduction, throughout this paper we focus our attention on testing (rather than estimation), and this is framed in terms of the total effect. We consider the specific average causal null hypothesis test: H_0: TE = 0 vs. H_1: TE $\neq$ 0. Such a pair of hypotheses involves five potential causal pathways between exposure and outcome (1) $A_0 \rightarrow Y$, (2) $A_1 \rightarrow Y$, (3) $A_0 \rightarrow A_1 \rightarrow Y$, (4) $A_0 \rightarrow L_1 \rightarrow Y$, and (5) $A_0 \rightarrow L_1 \rightarrow A_1 \rightarrow Y$ and can be tested with either a causal inference approach or an IV analysis. Fully non-parametric g-computation is employed as our causal inference method of choice. However, in fully non-parametric applications such as ours, g-computation and IPW approaches produce identical results (Daniel et al., 2013). As such, any conclusions based on g-computation are entirely applicable to the IPW situation.

3. Methods

3.1 g-computation

The theoretical properties of g-computation have been previously examined, see Taubman et al. (2009); and applications exist across a wide range of disciplines: determining the effect of pillbox organizer use on adherence and viral suppression, Petersen et al. (2007); assessing

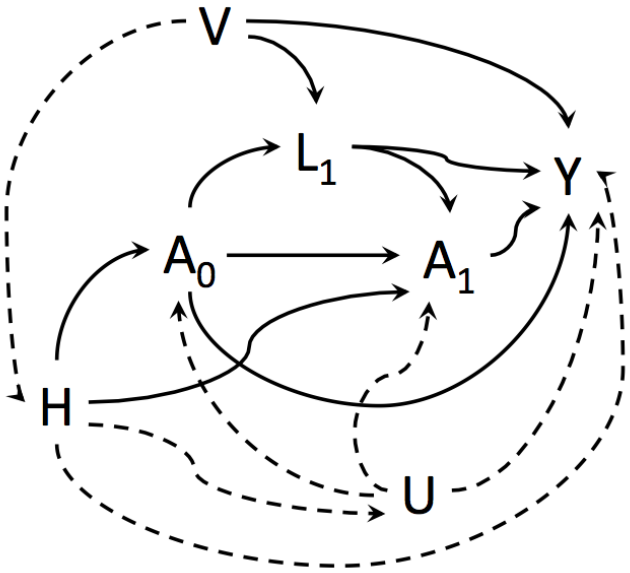


Figure 1: The simple scenario as represented by a DAG. Consider the variables- V: an unmeasured factor influencing both the measured confounder, L_1, and the outcome, Y; U: a potential unmeasured confounder; H: a potential instrument; A_0: exposure to treatment at time 0; and A_1: exposure to treatment at time 1.

the effect of antiretroviral therapy on incident AIDS, Cain et al. (2009); and evaluating time trends in associations between ozone concentrations and hospitalization for asthma, Moore et al. (2008). The *Bayesian* g-computation, while not as prevalent in the literature, shows much promise; see Gustafson (2015) and Keil et al. (2017).

Whether one uses the frequentist or Bayesian g-computation framework (or IPW for that matter), the validity of the causal approach relies on the assumptions that (1) the model is correctly specified; (2) there is a nonzero probability of each treatment regimen within all subgroups for which the effect of treatment is being assessed (the positivity assumption); and (3) all confounders are known and measured in the data. In our simple scenario, a violation of the assumption of no unmeasured confounders could be represented in the causal diagram with arrows: $U \to Y$, and either $U \to A_0$ and/or $U \to A_1$.

Consider a bootstrap-based implementation of g-computation in the context of our simple scenario. For testing the null hypothesis of no TE, a two-sided p-value is obtained using non-parametric bootstrap resampling (with large B) as follows:

1. For b in 1 to B:
   (a) Sample with replacement n observations from the data: $D^{[b]} = [Y, A_0, A_1, L_1, H]_b$.
   (b) For $l_1 = 0, 1$, obtain empirical estimates:
   - $\hat{\Pr}(L_1 = l_1|a_0)^{[b]} = \frac{\sum_{i \in D^{[b]}} \mathbb{1}(L_{1i}=l_1)\mathbb{1}(A_{0i}=a_0)}{\sum_{i \in D^{[b]}} \mathbb{1}(A_{0i}=a_0)}$,
   - $\hat{\mathrm{E}}(Y|a_0, a_1, l_1)^{[b]} = \frac{\sum_{i \in D^{[b]}} \mathbb{1}(Y_i=1)\mathbb{1}(A_{0i}=a_0, A_{1i}=a_1, L_{1i}=l_1)}{\sum_{i \in D^{[b]}} \mathbb{1}(A_{0i}=a_0, A_{1i}=a_1, L_{1i}=l_1)}$.
   (c) Apply the the *g-formula*: $\hat{\theta}^{[b]}_{a_0 a_1} = \sum_{l_1=0}^{1} \hat{\mathrm{E}}(Y|a_0, a_1, l_1)^{[b]} \cdot \hat{\Pr}(L_1 = l_1|a_0)^{[b]}$.
   (d) Calculate: $\hat{\mathrm{TE}}^{[b]} = \hat{\theta}^{[b]}_{11} - \hat{\theta}^{[b]}_{00}$.
2. Report the two-sided p-value $= 2 \cdot min\left(\frac{1+\sum_{b=1}^{B} \mathbb{1}(\hat{\mathrm{TE}}^{[b]}>0)}{1+B}, \frac{1+\sum_{b=1}^{B} \mathbb{1}(\hat{\mathrm{TE}}^{[b]}\leq 0)}{1+B}\right)$.

In the foundational work of Robins (1986) (see also Robins & Wasserman (1997)), an alternative testing scheme to g-computation was suggested. The *g-null test* relies on the same three assumptions listed above, but is based on a score statistic and as such, is computationally simpler than g-computation. However, as others have discussed (e.g. Buckley et al. (2015) and Hertz-Picciotto et al. (2000)), relative to g-computation, the g-null test suffers from low power. With current computational capacities, there are no longer any reasons to prefer the g-null test over g-computation for hypothesis testing.

3.1.1 A note on the g-null paradox

The g-computation method above estimates the required probabilities and expectations by means of empirical estimates. This *non-parametric* g-computation is only possible due to the fact that the model is saturated. In a more complex scenario (e.g. in the presence of continuous variables), one may require a *parametric* g-computation for which the "g-null

paradox" may be problematic, see Robins & Wasserman (1997). We describe the problem as it applies to the simple scenario. Consider a parametric estimate of $\theta_{a_0a_1}$, with the required probabilities and expectations estimated by logistic regression (with main effect parameters β_0, β_1, β_2, β_3, γ_0 and γ_1):

$$\begin{aligned}\theta_{a_0a_1} &= \sum_{l_1=0}^{1} \mathbf{E}(Y|a_0,a_1,l_1)\cdot \Pr(L_1=l_1|a_0) \qquad (1)\\ &= expit(\beta_0+\beta_1a_0+\beta_2a_1)+\\ &\quad expit(\gamma_0+\gamma_1a_0)\Big(expit(\beta_0+\beta_1a_0+\beta_2a_1+\beta_3)-expit(\beta_0+\beta_1a_0+\beta_2a_1)\Big).\end{aligned}$$

As such, the null hypothesis will hold ($\theta_{11}-\theta_{00}=0$), if and only if:

1. $\beta_1=\beta_2=\beta_3=0$; or:
2. $\beta_1=\beta_2=0$ and $expit(\gamma_0+\gamma_1a_0)=0$, for $a_0=0,1$; or:
3. $$\begin{aligned}expit(\beta_0) - expit(\beta_0+\beta_1+\beta_2) =&\\ expit(\gamma_0+\gamma_1)\Big(expit(\beta_0+\beta_1+\beta_2+\beta_3)-expit(\beta_0+\beta_1+\beta_2)\Big)&\\ -expit(\gamma_0)\Big(expit(\beta_0+\beta_3)-expit(\beta_0)\Big).& \qquad (2)\end{aligned}$$

If L_1 impacts Y, and/or if both variables are correlated by means of an unmeasured common cause (e.g. the unmeasured variable V), the limiting value of β_3 will not equal zero. Similarly for γ_1, if A_0 impacts L_1. Also, if L_1 is a collider on the causal path $A_0 \rightarrow L_1 \leftarrow V \rightarrow Y$, then conditioning on L_1 will result in β_1 not being equal zero. As such, it may only be possible to "parametrize the null" with the delicate equality of equation (2). For time-to-event data, Young et al. (2014) arrive at a similar conclusion upon finding but a single parametrization in which it is mathematically possible to generate null data from standard parametric models. (Note that, in our example, a parametric g-computation with fully saturated logistic regression (i.e. all two- and three-way interaction terms included) would produce estimates identical to those from the non-parametric g-computation, and therefore would not suffer from the g-null paradox).

Due to the g-null paradox, Robins & Hernán (2009) conclude that, since parametric g-computation will, given sufficient data, reject the null hypothesis even when true, it should be avoided for testing. While this is a most sensible recommendation, the extent to which the g-null paradox will inflate type I error has (to the best of our knowledge) never been investigated by means of simulation. In Section 4, we pursue this by incorporating the parametric g-computation (as outlined in equation (1)) into our simulation study.

3.2 Instrumental variable analysis

The essential idea of IV analysis is to exploit a natural experiment. Whereas in a RCT, exposure is determined by random allocation, in an IV analysis exposure is influenced by a natural event (e.g. a policy change, the presence of a genetic variant). As such, one might

consider IV analysis as circumventing the requirement that all confounders are known and measured, at the cost of using an imperfect proxy for exposure (Cain et al., 2009).

Consider the three assumptions (i.e. "core conditions" (Didelez & Sheehan, 2007)) required for IV analysis. Note that, while the first assumption can be empirically validated (to a certain degree), the second and third assumptions cannot.

1. **The instrument is valid:** the instrument is associated with exposure (we require $H \to A_0$ or $H \to A_1$). If the instrument's association with exposure is only small, the instrument, while valid, is considered "weak."

2. **The exclusion restriction** (or the "main assumption"): the instrument affects the outcome only indirectly through the exposure. We require $H \not\to Y$, i.e. $Pr(Y = 1|do(H = 0), A_0 = a_0, A_1 = a_1) = Pr(Y = 1|do(H = 1), A_0 = a_0, A_1 = a_1)$. Note that, while indirect, the causal pathway $H \to U \to Y$ would also violate this "main assumption."

3. **There are no common causes:** the instrument does not share common causes with the outcome Y, i.e. no confounding for the effect of H on Y, (we require $V \not\to H$).

In our simple scenario, the first assumption has that $H \to A$. However, note that a valid instrument H need not necessarily have a direct causal effect on exposure (Didelez & Sheehan, 2007). An indirect association will suffice (e.g. $H \leftarrow H_* \to A$, where H_* is some unknown intermediary).

In addition to the three conditions above, in order to test our specific average causal null hypothesis (H_0: TE = 0), we also require:

4. The **"monotonic treatment condition"** (Swanson et al., 2018) that specifies that either:

$$E[Y|do(A_0 = 0, A_1 = 0)] \leq E[Y|do(A_0 = i, A_1 = j)] \leq E[Y|do(A_0 = 1, A_1 = 1)], \forall i \neq j.$$
$$or:$$
$$E[Y|do(A_0 = 0, A_1 = 0)] \geq E[Y|do(A_0 = i, A_1 = j)] \geq E[Y|do(A_0 = 1, A_1 = 1)], \forall i \neq j.$$

Without this additional condition, having that $E[Y|Z = 1] \neq E[Y|Z = 0]$ would not necessarily be evidence against our average causal null hypothesis. For example, consider a situation in which the causal impact of A_0 would be to increase the probability of Y=1 by 10% and the causal impact of A_1 would be to decrease the probability of $Y = 1$ by 10%. Then, we may have that $\mathrm{E}(Y|do(A_0 = 0, A_1 = 0)) = \mathrm{E}(Y|do(A_0 = 1, A_1 = 1))$, while also having that $E[Y|Z = 0] \neq E[Y|Z = 1]$.

For testing the null hypothesis of no TE, a p-value is obtained by simply testing the association between the two binary variables H and Y (for which a multitude of tests are available, (Lydersen et al., 2009)). While our main focus is testing (see the discussion of Swanson et al. (2018)), we wish to mention four delicate issues that will arise for estimation:

(1) Only a small violation of the second assumption may result in highly biased estimates in the case of a weak instrument, see Bound et al. (1995).

(2) In general, the target of inference is different when using IV and non-IV methods. With IV analysis, the estimated parameter is typically the "complier average causal effect" and additional assumptions must be made in order to identify the "average causal effect" (Baiocchi et al., 2014).

(3) IV analysis for binary outcomes requires additional care to obtain unbiased estimates due to the non-collapsibility of the logistic regression estimates, see Clarke & Windmeijer (2012); Burgess (2013); Vansteelandt et al. (2011).

(4) IV analysis for time-varying exposure cannot identify the effect of different treatment regimens but rather is testing for the "observational analog of the intention-to-treat effect commonly estimated from randomized experiments" (Hernán & Robins, 2006). As such, interpretation of IV estimates in time-varying settings is not straightforward. Indeed, estimating the effects of time-varying exposures with instrumental variables remains "relatively understudied" (Brookhart et al., 2010).

3.2.1 A note on the limits of the strength of instruments

Martens et al. (2006) arrive at the intriguing conclusion that: "the presence of considerable confounding and a strong instrument will probably indicate a violation of the main assumption." Let ψ_{XY} be the correlation between two variables, X and Y. Figure 2 illustrates a DAG for a basic single time-point scenario. As before, H represents the potential instrument, A represents exposure to treatment, U is an unmeasured confounder, and Y is the outcome of interest. The argument of Martens et al. (2006) is based on the observation that, for continuous variables, ψ_{HA} is bounded by the inequality $|\psi_{HA}| \leq |\psi_{HU}||\psi_{AU}| + \sqrt{1-\psi^2_{HU}}\sqrt{1-\psi^2_{AU}}$. Should the "main assumption" hold, then $H \not\to U$ and therefore $\psi_{HU} = 0$, and we have that $|\psi_{HA}| \leq \sqrt{1-\psi^2_{AU}}$. Thus, the "strength" of a valid instrument is bounded by a function of the degree of confounding (ψ_{AU}), with high levels of confounding resulting in at most weak to modest instruments.

In our simple time-varying scenario (Figure 1), things are made more complicated due to all variables being binary, as well as the sequential exposures, and the potential for time-varying confounding. As such, we will only consider how the argument of Martens et al. (2006) applies empirically. Among scenarios within our simulation study that have considerable confounding, we will see if those for which there is a violation of the main assumption show higher levels of observed correlation between exposure and instrument. Note that a violation of the main assumption may occur either with $H \to Y$, or with $H \to U \to Y$. Each possibility may have a different impact on limiting the observed strength of the instrument.

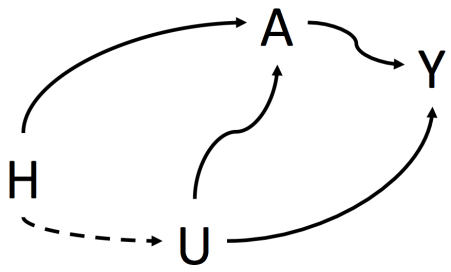


Figure 2: A simple single time-point scenario as represented by a DAG. Consider the variables: U, an unmeasured confounder; A, the exposure to treatment; Y, the outcome of interest; and H, the instrumental variable. The dashed line indicates a causal relationship between H and U that would violate the "main assumption."

3.3 Competing assumptions

As an alternative to conventional methods, IV analyses are often considered due to the fact that they rely on an "entirely different" (Brookhart et al., 2010) set of assumptions. In our simple time-varying scenario, the assumptions required of g-computation and IV analysis are not entirely exclusive, see the first two columns of Table 1. While there are two complimentary assumptions ((1) the assumption of no unmeasured confounders, required for g-computation, is not necessary for IV analysis; and (2) the assumption that the instrument is correlated with exposure, required for IV analysis, is not required for g-computation), there are also two assumptions necessary for both methods. Should H impact exposure ($H \to A_0$, $H \to A_1$), than either (1) a common cause for instrument and outcome ($H \leftarrow V \to Y$) or (2) a causal path between instrument and outcome (direct: $H \to Y$, or indirect: $H \to U \to Y$), will invalidate both methods. The g-computation approach can be made robust to both these possibilities by simply conditioning on H (i.e. including H in the adjustment set), see the third column of Table 1, labelled 'g-comp$|H$'. In the simple scenario, this requires a second summation in the g-formula:

$$\hat{\theta}_{a_0 a_1} = \sum_{h=0}^{1} \sum_{l_1=0}^{1} \hat{\mathbf{E}}(Y|a_0, a_1, l_1, h) \cdot \hat{\Pr}(L_1 = l_1|a_0, h) \cdot \hat{\Pr}(H = h) \quad ,$$

$$\text{with: } \hat{\Pr}(L_1 = l_1|a_0, h) = \frac{\sum_i \mathbb{1}(L_{1i} = l_1)\mathbb{1}(A_{0i} = a_0, H_i = h)}{\sum_i \mathbb{1}(A_{0i} = a_0, H_i = h)} \quad ,$$

$$\text{and: } \hat{\Pr}(H = h) = \sum_i \frac{\mathbb{1}(H_i = h)}{n}.$$

While conditioning on H is advantageous in terms of robustness, it may be detrimental with regards to efficiency; see Rotnitzky et al. (2010). Our simulation study provides an opportunity to examine this trade-off.

Table 1: Checkmarks indicate the assumptions required for each of the approaches considered. Violations of assumptions refer to the simple scenario, see Figure 1.

Assumption	**Violations of assumption**	**IV**	***g*-comp**	***g*-comp**$\vert H$
Instrument does not share common causes with outcome	$H \leftarrow V \rightarrow Y$	✓	✓	×
Instrument independent of outcome given exposure	$H \rightarrow U \rightarrow Y$ $H \rightarrow Y$	✓	✓	×
Instrument correlated with exposure	$H \not\rightarrow A_0$ $H \not\rightarrow A_1$	✓	×	×
No unmeasured confounders	$Y \leftarrow U \rightarrow A_0$ $Y \leftarrow U \rightarrow A_1$	×	✓	✓

4. Simulation Study I - Power with time-varying exposures

Methods for determining power and sample size for IV analysis in the single time-point setting can be implemented relatively easily; see for example Burgess (2014) and Walker et al. (2017). However, in the case of a time-varying treatment (and a single instrument), standard methods for determining power are unavailable. While it is well known that IV analysis tends to have reduced efficiency, the degree to which power is diminished relative to causal inference methods in the presence of time-varying confounding is not well understood. Power calculations for causal inference methods are also difficult (regardless of the number of timepoints) and typically require simulation studies (Austin et al., 2015). We designed our Monte Carlo simulations to be as simple as possible so that as many of the elements potentially effecting power can be examined while restricting the overall number of scenarios to a manageable number. Data were simulated from the following logistic models in turn:

$$Pr(V = 1) = 0.5, \quad Pr(H = 1) = expit(\alpha_H + \beta_{V\rightarrow H}V), \quad Pr(U = 1) = expit(\alpha_U + \beta_{H\rightarrow U}H),$$
$$Pr(A_0 = 1) = expit(\alpha_{A_0} + \beta_{H\rightarrow A_0}H + \beta_{U\rightarrow A_0}U),$$
$$Pr(L_1 = 1) = expit(\alpha_{L_1} + \beta_{A_0\rightarrow L_1}A_0 + \beta_{V\rightarrow L_1}V),$$
$$Pr(A_1 = 1) = expit(\alpha_{A_1} + \beta_{H\rightarrow A_1}H + \beta_{A_0\rightarrow A_1}A_0 + \beta_{L_1\rightarrow A_1}L_1 + \beta_{U\rightarrow A_1}U),$$
$$Pr(Y = 1) = expit(\alpha_Y + \beta_{A_0\rightarrow Y}A_0 + \beta_{A_1\rightarrow Y}A_1 + \beta_{U\rightarrow Y}U + \beta_{L_1\rightarrow Y}L_1 + \beta_{V\rightarrow Y}V + \beta_{H\rightarrow Y}H),$$

where $\alpha_H = -0.5$, $\alpha_U = -0.5$ and all assumptions outlined in Table 1 are met by setting: $\beta_{H\rightarrow Y} = 0$, $\beta_{H\rightarrow U} = 0$, $\beta_{V\rightarrow H} = 0$, $\beta_{U\rightarrow A_1} = 0$, and $\beta_{U\rightarrow A_0} = 0$.

We sought to examine the impact of the following eight factors on the statistical power:

- **f1 - The number of observations** was allowed to take three values, $n = 750, 1000, 1400$.
- **f2 - The proportion of subjects with $Y = 1$.** The intercept parameter, α_Y, took on values corresponding to three possibilities: (1) a *rare* outcome ($Pr(Y = 1) = 0.05$), (2) a *low* outcome ($Pr(Y = 1) = 0.10$), and (3) a *balanced* outcome ($Pr(Y = 1) = 0.50$). For comparison in the MS context, Karim et al. (2014) report that approximately 8% of subjects within the study (n=1,697) reached the outcome of irreversible progression of disability.
- **f3 - The strength of the instrumental variable.** The parameters $\beta_{H\rightarrow A_0}$ and $\beta_{H\rightarrow A_1}$ took together two possible values, 0.75 and 1.5.
- **f4 - The proportion of subjects exposed to treatment.** In order to modify the level of exposure while keeping all other parameters constant, the intercept parameters, α_{A_0} and α_{A_1}, took on values corresponding to two possibilities: (1) exposure levels of $Pr(A_0 = 1) = 0.31$ and $Pr(A_1 = 1) = 0.63$ ("more"); and (2) $Pr(A_0 = 1) = 0.19$, $Pr(A_1 = 1) = 0.38$ ("less"). In a perfectly balanced design, the proportion of subjects exposed at each time-point would equal 0.5. In this sense, the *more* exposed scenario is slightly *more* balanced than the *less* exposed scenario. In the *more* exposed scenario, approximately 33% of subjects remain untreated at both time-points. In the *less* exposed scenario approximately 57% of subjects remain untreated at both time-points. For comparison in the MS context, Karim et al. (2014) report that approximately 48% of subjects remained untreated throughout the course of the study. While efforts were made to achieve the stated marginal distributions, readers should appreciate the degree to which one is limited when simulating correlated binary variables, see Table 2.
- **f5 - The level of potential collider-stratification bias.** The degree to which the unmeasured variable, V, impacts both a measured confounder, L_1, and the outcome, Y, was allowed to change by setting parameters $\beta_{V\rightarrow L_1} = \beta_{V\rightarrow Y} = 0$ and 2.
- **f6 - The magnitude of the direct effect.** The parameter $\beta_{A_1\rightarrow Y}$ took on three values: 0 (no direct effect), 0.35 (mild direct effect), 0.75 (strong direct effect) and $\beta_{A_0\rightarrow Y}$ was set to equal half the value of $\beta_{A_1\rightarrow Y}$.
- **f7 - The magnitude of the indirect effect.** The degree to which a time-dependent factor (L_1) that impacts subsequent treatment (A_1) also impacts the outcome (Y) was allowed to change by setting $\beta_{A_0\rightarrow L_1} = 2$, while the parameter $\beta_{L_1\rightarrow Y}$ took on three values: 0, 0.75, and 2.
- **f8 - The correlation between successive exposures.** The degree to which initial exposure (A_0) impacts subsequent exposure (A_1) was allowed to change by setting $\beta_{A_0\rightarrow A_1} = 0.75$, or $\beta_{A_0\rightarrow A_1} = 1.25$. On average this change results in an increase in the correlation between A_0 and A_1 from $cor(A_0, A_1) = 0.26$ to $cor(A_0, A_1) = 0.34$.

f8	f4	$Pr(A_0 = 1)$	$Pr(A_1 = 1)$	$cor(A_0, A_1)$	$Pr(A_0 = 0 \text{ and } A_1 = 0)$
0.75	more	0.312	0.625	0.272	0.319
0.75	less	0.187	0.375	0.257	0.556
1.25	more	0.312	0.625	0.339	0.334
1.25	less	0.187	0.375	0.336	0.572

Table 2: Marginal distributions of A_0 and A_1 across all scenarios simulated given different levels of f8 and f4.

We used a full factorial design and therefore considered $1,296 = 3 \cdot 3 \cdot 2 \cdot 2 \cdot 3 \cdot 3 \cdot 2 \cdot 2$ different scenarios. The intercept parameter, α_{L_1}, was set such that the $Pr(L_1 = 1) = 0.5$. Other parameters were fixed: $\beta_{L_1 \to A_1} = 1$, and $\beta_{A_0 \to L_1} = 2$. The true TE was determined for each combination of parameters and ranged between 0 and 0.41. For each given scenario, we simulated 1,500 independent datasets. Once a simulated dataset was created, two-sided p-values were computed using the following methods:

1. The standard logistic regression model $Pr(Y = 1|A_0, A_1, L_1) = expit(\beta_0 + \beta_1 A_0 + \beta_2 A_1 + \beta_3 L_1)$ is fit and a likelihood ratio test determines the significance of exposure (i.e. if the increase in likelihood obtained by including A_0 and A_1 in the model is significant), ('mv-reg');

2. g-computation as described in Section 3.1 with B=150, ('g-comp');

3. g-computation conditioning on H as described in Section 3.3 with B=150, ('g-comp|H');

4. IV analysis whereby a logistic regression model is fit: $Pr(Y = 1|H) = expit(\alpha_0 + \beta H)$, ('IVA').

We also implemented parametric g-computation ('g-comp-p' and 'g-comp|H-p') as outlined in Section 3.1.1 in order to observe the effects of the g-null paradox.

Note that we include standard regression ('mv-reg') simply to confirm the damage inflicted by adjusting for a variable on the causal pathway from exposure to outcome. As mentioned earlier, including L_1 in the model will have the detrimental effect of blocking any indirect treatment effect in which L_1 is an intermediate (e.g. $A_0 \to L_1 \to Y$, or $A_0 \to L_1 \to A_1 \to Y$). Note also that the option of regressing Y on only A_0 and A_1 is not pursued, since excluding L_1 (a measured confounder) would make the model susceptible to suggesting the existence of a treatment effect in cases when no such effect exists (e.g. when $A_1 \leftarrow L_1 \to Y$, along with $A_0 \not\to L_1$, $A_0 \not\to Y$, and $A_1 \not\to Y$). With regards to adjusting for L_1 in a standard regression model, neither available option is satisfactory.

For small samples, a bootstrap resampling of the data may have zero-cell counts (i.e. within a bootstrap sample of the data, there are no observations required to estimate the necessary conditional probabilities and expectations.) Such bootstrap resamples were discarded (as in the related work of Schnitzer et al. (2016)), potentially biasing results. The

empirical estimate of statistical power is the proportion of simulated datasets in which the null hypothesis was rejected at α-level 0.05.

Finally, note that the factors varied in the simulation can be thought of in the drug therapy for MS context discussed previously. For instance, f3 describes the magnitude of increased drug uptake in the later calendar period relative to the former period, f6 reflects the drug efficacy in directly reducing the risk of disease progression given relapse status, while f7 reflects the drug efficacy in reducing the risk of relapse which itself predisposes subjects to disease progression.

4.1 Results of simulation study I - Type I error

See Table 3. Before reviewing the simulation results with regards to statistical power, consider the 144 scenarios for which the true TE = 0. Several findings merit comment.

1. Type I error rates for standard multivariable regression are well-above the desired 0.05 level in the presence of collider-stratification bias ($\beta_{V\rightarrow L_1,Y} = 2$) with specific scenarios yielding type I error rates ranging from 0.06 to 0.66 (summarized in rows 7-12, 'mv-reg'). In all other scenarios, standard regression analysis exhibits desired type I error.

2. When sample size is low, a substantial proportion of simulations do not allow for g-computation when one conditions on H (see '% NA' columns to the right of 'g-comp' and 'g-comp$|H$' respectively). This is due to the fact that for small samples, sparsity is such that the observations required to estimate the necessary conditional probabilities and expectations are not present within the sample. IPW methods can encounter a similar problem: with a small number of observations, variance estimation can be problematic due to the positivity assumption being "practically violated" (Xiao et al., 2013). Different estimators are affected differently and while potential solutions exist (Porter, 2011), sparsity will no doubt restrict the number of variables upon which one can condition.

3. The g-computation with H ('g-comp$|H$') shows higher than desired type I error rates in those situations where there were a substantial number of simulated datasets for which the method could not be applied due to sparsity.

4. IV analysis ('IVA') appears to have desired type I error rate across all scenarios.

4.1.1 Effects of the g-null paradox

Overall, we do not observe higher empirical type I error rate for parametric g-computation than for non-parametric g-computation. This suggests that, while the g-null paradox is of theoretical concern, it may not be overly consequential in practice. However, it is impossible to generalize this to all cases. To the extent that there was a difference in empirical type I error rate, this appears to be the result of the sparsity issue which is problematic for non-parametric g-computation but is not for parametric g-computation.

$\beta_{V\to L_1,Y}$	$f4$	n	mv-reg	g-comp	(% NA)	g-comp$\|H$	(% NA)	IVA	g-comp-p	g-comp$\|H$-p
0	more	750	0.05	0.04	0.00	0.04	0.12	0.05	0.04	0.04
0	more	1000	0.05	0.04	0.00	0.04	0.03	0.05	0.04	0.04
0	more	1400	0.05	0.04	0.00	0.04	0.00	0.05	0.04	0.04
0	less	750	0.05	0.05	0.01	0.10	17.76	0.05	0.05	0.05
0	less	1000	0.05	0.05	0.01	0.09	10.75	0.05	0.04	0.04
0	less	1400	0.05	0.04	0.00	0.08	5.52	0.05	0.04	0.04
2	more	750	0.20	0.04	0.00	0.05	0.03	0.05	0.04	0.04
2	more	1000	0.24	0.04	0.00	0.04	0.00	0.05	0.04	0.04
2	more	1400	0.32	0.04	0.00	0.04	0.00	0.05	0.04	0.04
2	less	750	0.15	0.05	0.00	0.09	11.47	0.05	0.05	0.04
2	less	1000	0.18	0.05	0.00	0.09	6.25	0.05	0.05	0.04
2	less	1400	0.24	0.04	0.00	0.07	2.49	0.05	0.04	0.04

Table 3: Results of Simulation Study I - Average type I error rate obtained by each of the six methods, by varying levels for $\beta_{V\to L_1,Y}$, $f4$, and n, across scenarios for which the true TE = 0. Columns to the right of g-comp and g-comp$|H$ (% NA) indicate the percentage of simulation runs for which there was insufficient data for g-computation.

4.2 Results of simulation study I - Power

Each panel plotted within Figure 3 contrasts different levels of factors f3 and f4 (i.e. one panel for each combination of the settings). Each panel plotted within Figure 4 contrasts different levels of factors f5 and f6 (i.e. one panel for each combination of the settings). In both figures, within each panel, power results are plotted against true TE for each of three sample sizes (n = 750, 1000, 1400). Smoothed curves are the result of fitting a strictly monotone and smooth regression; see Dette & Scheder (2006). Several findings merit comment.

1. As expected, power of the IV analysis increases with increasing instrument strength (IS); compare the lower row to the upper row in Figure 3. The results support the caution of Brookhart et al. (2010): should the instrument be weak, an IV analysis will be underpowered for "anything less than a very strong effect, even with large samples."

2. In every scenario examined, the power of the standard g-computation analysis surpasses that of the IV analysis and the power of the g-computation analysis surpasses that of the g-computation with H. This confirms well established findings cautioning that including instruments within the causal model will be detrimental to power; see Brookhart, Schneeweiss, et al. (2006), Myers et al. (2011). In some scenarios, this difference is negligible: consider the case in which the instrument is weak and exposure to treatment is more balanced, see Figure 3, Panel B. In other scenarios the difference can be substantial: consider the case in which exposure to treatment is less balanced and instrument strength is high, see Figure 3, Panel C. In general, the cost

(in terms of efficiency) of conditioning on H increases with instrument strength and with imbalance of exposure.

3. Balance in exposure increases the power of all methods of analysis, see the left column versus the right column of Figure 3.

4. When there is no collider-stratification bias, ($\beta_{V\to L_1} = \beta_{V\to Y} = 0$), standard multivariable regression has about the same power as g-computation when the effect is direct ($\beta_{A_1\to Y}$=1) and when exposure is more balanced. The power of standard multivariable regression is reduced when exposure is less balanced. When the effect is indirect, due entirely to the $A_0 \to L_1 \to Y$ pathway (i.e. $\beta_{A_1\to Y}$=0, and $\beta_{A_0\to Y} = 0$), multivariable regression has negligible power, whereas the other methods still show moderate power, see Figure 4 panel A. While, in Figure 4 panel D, the standard multivariable regression appears to show some power despite the absence of any direct effect, this is entirely the result of collider-stratification bias ($\beta_{V\to L_1} = \beta_{V\to Y} = 2$). In order to obtain power equivalent to that of a standard regression analysis, the simulation results suggest that a causal inference method does not require a larger sample size as might be thought (McCulloch, 2015).

5. There does not appear to be a substantial difference in the power of IV analysis with varying levels of outcome rarity (f2) for a given treatment effect size as measured on the additive scale. For a given effect size on the multiplicative scale (in terms of the odds ratio), power is indeed substantially lower for rarer outcomes, as discussed in Brookhart et al. (2010).

6. There does not appear to be any substantial effect of $\beta_{A_0\to A_1}$ (f8) with regards to the power of g-computation or IV analysis.

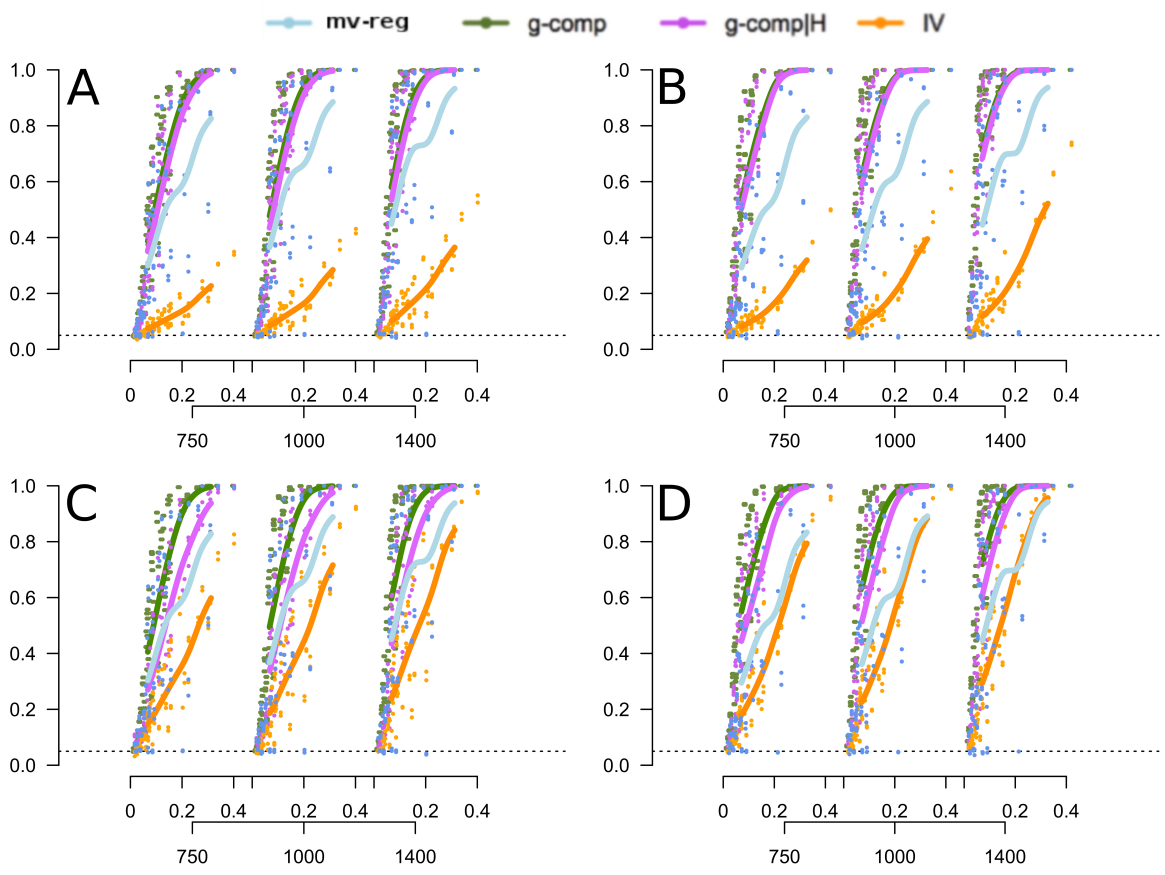


Figure 3: Each panel shows empirical power plotted against effect size (true TE) for the four methods given sample sizes of 750, 1000 and 1400. Top row (lower row) panels show scenarios for which $\beta_{H\to A_0} = \beta_{H\to A_1} = 0.75$ ($\beta_{H\to A_0} = \beta_{H\to A_1} = 1.5$). Left column (right column) panels show scenarios representing 'less exposure' ('more exposure'), (the f4 factor).

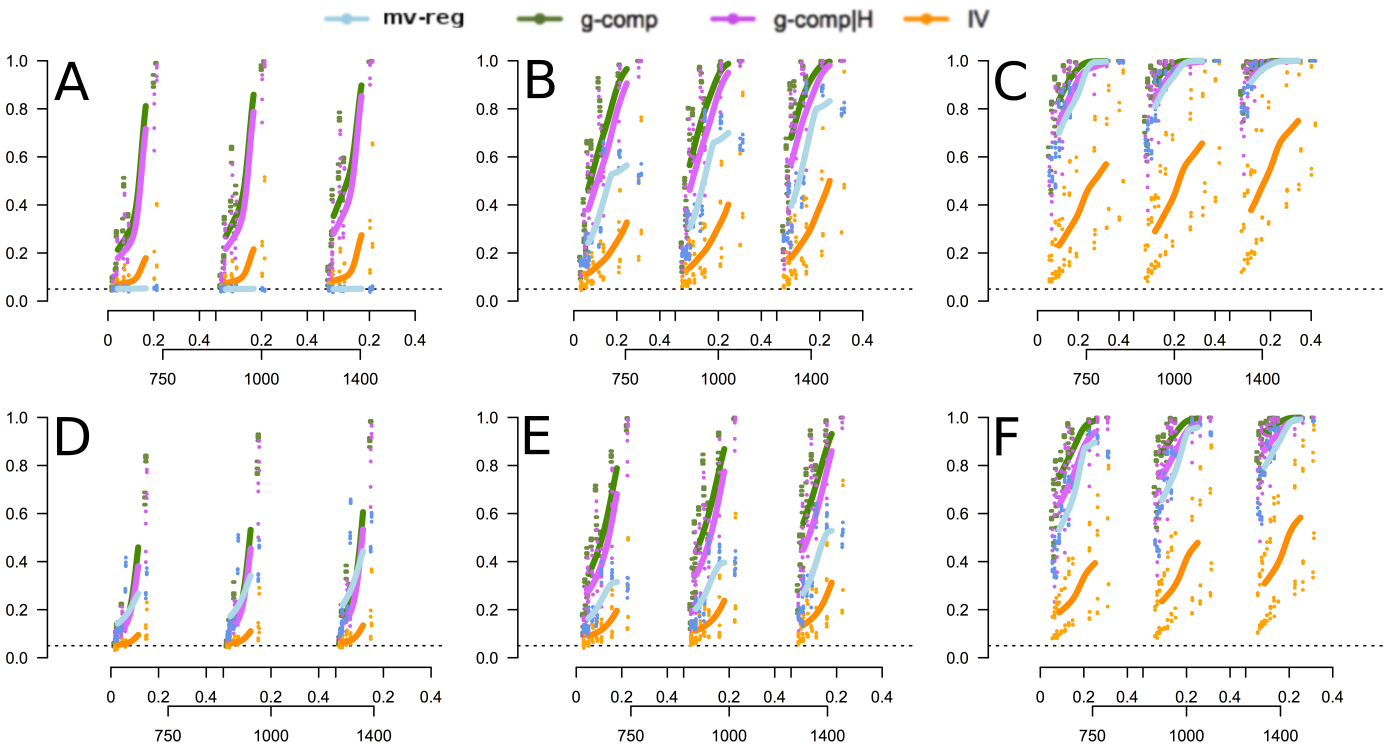


Figure 4: Each panel shows empirical power plotted against effect size (true TE) for the four methods given sample sizes of 750, 1000 and 1400. Top row (lower row) panels show scenarios across the f5 factor, for which $\beta_{V\to L_1} = \beta_{V\to Y} = 0$ ($\beta_{V\to L_1} = \beta_{V\to Y} = 2$). Left column, middle column, and right column panels show scenarios across the f6 factor, representing no direct effect, mild direct effect, and strong direct effect respectively. These correspond to $\beta_{A_1\to Y} = 0, \beta_{A_1\to Y} = 0.25$ and, $\beta_{A_1\to Y} = 0.75$, with $\beta_{A_0\to Y}$ always equal half the value of $\beta_{A_1\to Y}$.

5. Simulation study II - The impact of modest violations to model assumptions

With extensions to the simulation studies described in Section 4, we investigated how the empirical type I error changes with modest violations to the various assumptions required of each method (as summarized in Table 1). We also considered a joint assessment of the type I error and power in terms of the Bayes risk for selecting the correct hypothesis. For this, data were simulated from the logistic models detailed in Section 4. In addition to the eight factors (f1-f8) of simulation study I, we examined the impact of the following four factors on type I error:

- **f9 - How the instrument shares a common cause with the outcome.** The parameter $\beta_{V\to H}$ took on values 0, 0.75 and 1.5. Non-zero values of $\beta_{V\to H}$ represent violations to the required assumptions of IV analysis and g-computation methods in scenarios for which $\beta_{V\to L_1} = \beta_{V\to Y} = 2$.
- **f10 - How the instrument, given exposure, impacts the outcome.** The parameter $\beta_{H\to Y}$ took on values 0, 0.2 and 0.5. Non-zero values of $\beta_{H\to Y}$ represent violations to the required assumptions of IV analysis and g-computation methods.
- **f11 - The degree of unmeasured confounding.** Three parameters took on four equal values: $\beta_{U\to A_0} = \beta_{U\to A_1} = \beta_{U\to Y} = 0$, 0.25, 0.75 and 1.5. Non-zero values represent violations to the required assumptions of the g-computation and g-computation with H methods.
- **f12 - How the instrument impacts a potential unmeasured confounder.** The parameter $\beta_{H\to U}$ took on three values: $\beta_{H\to U} = 0, 0.25, 0.75$. Should $\beta_{U\to Y} \neq 0$, non-zero values of $\beta_{H\to U}$ represent violations to the required assumptions of IV analysis and g-computation methods.

Other parameters remained fixed as in Section 4. Not all violations were considered concurrently. As such, we considered a total of only 6,430 different scenarios. As before, for each given scenario, we simulated 1,500 independent datasets and obtained two-sided p-values.

5.1 Results of simulation study II - Type I error

See Table 4. We considered all scenarios in which both the direct and indirect effects were null (i.e. scenarios for which $\beta_{L_1\to Y} = 0$, $\beta_{A_0\to Y} = 0$ and $\beta_{A_1\to Y} = 0$) and in which $\beta_{V\to L_1} = \beta_{V\to Y} = 2$. Overall, results are as expected given the trade-offs in the assumptions outlined in Table 1. Two findings merit comment.

1. Bound et al. (1995) observed that the bias in estimation resulting from a violation of the "main assumption" is amplified in the presence of a weak instrument. With respect to testing, we do not observe a similar phenomenon. Recall that the IV p-value is based solely on the association between H and Y. Since any association between

H and A does not contribute to the p-value calculation, we should not expect any bias amplification with weaker instruments. The inflated type I error resulting from a violation in "main assumption" (i.e. when $\beta_{H\to Y} \neq 0$, or when both $\beta_{H\to U} \neq 0$ and $\beta_{U\to Y} \neq 0$) is not further increased when the instrument is weak relative to when the instrument is strong ; see Table 4, column IVA, lines 11-12, 19-20, 29-34, and 37-42.

2. Brookhart et al. (2010) recommended that "if only a small amount of unmeasured confounding is expected," it may not be necessary to turn to IV analysis (given its lower efficiency). Our results are mostly in agreement with this sentiment. When unmeasured confounding is low ($\beta_{U\to A_0,A_1,Y}$ =0.25) and no other assumptions were violated, the average (across scenario) type I error rate for g-computation was 0.045 (for g-computation with H, 0.068); see Table 4, column g-comp (g-comp|H), lines 5-6. No such single scenario ever exceeded an empirical type I error rate of 0.067 (for g-computation with H, 0.165). It appears that, while conditioning on H will protect against the possibility that H is a confounder (i.e. against $V \to H$, or $H \to Y$, or $H \to U \to Y$), it will also increase the risk of obtaining a type 1 error, should there be any unknown variables that are unmeasured confounders (i.e. should $\beta_{U\to A_0,A_1,Y} \neq 0$).

5.1.1 Limits on the strength of instruments

Martens et al. (2006) make the case that in the presence of considerable confounding, the strength of a valid instrument is "bound to a relatively small value," and as such, "the presence of considerable confounding and a strong instrument will probably indicate a violation of the main assumption and thus a biased estimate." Is the observed strength of a valid instrument bounded by the degree of confounding? Among the various scenarios from which we simulated data, we do not see any evidence to suggest that the degree of unmeasured confounding ($\beta_{U\to A_0,A_1,Y}$) is a predictor of the observed strength of a valid instrument ($\hat{cor}(H, A_0)$ and $\hat{cor}(H, A_1)$); see rows 1-8 in Table 5. In the presence of a high degree of confounding ($\beta_{U\to A_0,A_1,Y} = 1.50$), the observed instrument strength is not a predictor of whether or not, and to what extent, the main assumption has been violated by a *direct* effect of the instrument on the outcome (i.e. is not a predictor of $\beta_{H\to Y}$). However, the observed instrument strength is indeed a predictor of whether or not, and to what extent, the main assumption has been violated by the instrument's impact on a potential unmeasured confounder (i.e. is indeed a predictor of $\beta_{H\to U} \neq 0$); see rows 9-18 in Table 5. This is due to the fact that when the variable U acts as both a confounder for the $(A_0, A_1) \to Y$ relationship, and as an intermediate variable between H and Y, (i.e. when both $\beta_{H\to U} \neq 0$ and $\beta_{U\to A_0,A_1,Y} \neq 0$), the observed instrument strength will increase with increasing $\beta_{H\to U}$ and/or increasing $\beta_{U\to A_0,A_1,Y}$.

5.2 Results of simulation study II - Bayes Risk

We considered Bayes risk (BR) with respect to a zero-one loss function for selecting the correct hypothesis of no treatment effect vs. treatment effect. We set independent discrete prior distributions over the parameters corresponding to f2, f3, f4, f5, f6 and f7 in such a way

	$\beta_{H\to U}$	$\beta_{V\to L_1,Y}$	$\beta_{V\to H}$	$\beta_{H\to Y}$	$\beta_{U\to A_0,A_1,Y}$	IS	mv-reg	g-comp	g-comp$\vert H$	IVA
1	0	2	0.75	0	0	0.75	0.17	0.05	0.05	0.41
2	0	2	0.75	0	0	1.5	0.15	0.09	0.08	0.41
3	0	2	1.5	0	0	0.75	0.15	0.09	0.05	0.76
4	0	2	1.5	0	0	1.5	0.15	0.21	0.07	0.76
5	0	2	0	0	0.25	0.75	0.21	0.04	0.05	0.05
6	0	2	0	0	0.25	1.5	0.21	0.04	0.08	0.05
7	0	2	0	0	0.75	0.75	0.14	0.09	0.09	0.05
8	0	2	0	0	0.75	1.5	0.14	0.08	0.10	0.05
9	0	2	0	0	1.5	0.75	0.44	0.58	0.57	0.05
10	0	2	0	0	1.5	1.5	0.36	0.53	0.49	0.05
11	0	2	0	0.2	0	0.75	0.19	0.05	0.05	0.17
12	0	2	0	0.2	0	1.5	0.17	0.05	0.08	0.16
13	0	2	0	0.2	0.25	0.75	0.18	0.05	0.05	0.17
14	0	2	0	0.2	0.25	1.5	0.16	0.06	0.07	0.17
15	0	2	0	0.2	0.75	0.75	0.14	0.13	0.09	0.16
16	0	2	0	0.2	0.75	1.5	0.14	0.15	0.10	0.16
17	0	2	0	0.2	1.5	0.75	0.49	0.63	0.57	0.15
18	0	2	0	0.2	1.5	1.5	0.46	0.62	0.49	0.15
19	0	2	0	0.5	0	0.75	0.15	0.07	0.05	0.62
20	0	2	0	0.5	0	1.5	0.15	0.13	0.07	0.61
21	0	2	0	0.5	0.25	0.75	0.15	0.07	0.05	0.62
22	0	2	0	0.5	0.25	1.5	0.14	0.14	0.07	0.61
23	0	2	0	0.5	0.75	0.75	0.15	0.19	0.09	0.61
24	0	2	0	0.5	0.75	1.5	0.18	0.29	0.09	0.61
25	0	2	0	0.5	1.5	0.75	0.56	0.68	0.58	0.58
26	0	2	0	0.5	1.5	1.5	0.59	0.73	0.50	0.58
27	0.25	2	0	0	0	0.75	0.22	0.04	0.05	0.05
28	0.25	2	0	0	0	1.5	0.22	0.04	0.08	0.05
29	0.25	2	0	0	0.25	0.75	0.22	0.04	0.05	0.05
30	0.25	2	0	0	0.25	1.5	0.21	0.05	0.08	0.05
31	0.25	2	0	0	0.75	0.75	0.14	0.10	0.09	0.05
32	0.25	2	0	0	0.75	1.5	0.13	0.10	0.10	0.05
33	0.25	2	0	0	1.5	0.75	0.46	0.60	0.57	0.07
34	0.25	2	0	0	1.5	1.5	0.40	0.57	0.48	0.07
35	0.75	2	0	0	0	0.75	0.22	0.04	0.05	0.05
36	0.75	2	0	0	0	1.5	0.22	0.04	0.08	0.05
37	0.75	2	0	0	0.25	0.75	0.20	0.04	0.05	0.06
38	0.75	2	0	0	0.25	1.5	0.20	0.05	0.08	0.06
39	0.75	2	0	0	0.75	0.75	0.14	0.11	0.09	0.11
40	0.75	2	0	0	0.75	1.5	0.13	0.12	0.10	0.10
41	0.75	2	0	0	1.5	0.75	0.50	0.62	0.55	0.25
42	0.75	2	0	0	1.5	1.5	0.47	0.62	0.46	0.25

Table 4: Results of simulation study II - Type I error: Average across scenarios of empirical type I error for different levels of instrument strength (IS).

that 50% prior probability is assigned to scenarios for which the null is true. The remaining 50% prior probability on scenarios for which the alternative is true is evenly attributed to

	$\beta_{H\to U}$	$\beta_{H\to Y}$	$f4$	$\hat{P}(A_0)$	$\hat{P}(A_1)$	$\beta_{U\to A_0,A_1,Y}$	$\hat{cor}(H,A_0)$	$\hat{cor}(H,A_1)$
1	0.00	0.00	more	0.31	0.62	0.00	0.25	0.29
2	0.00	0.00	more	0.31	0.62	0.25	0.25	0.29
3	0.00	0.00	more	0.31	0.62	0.75	0.24	0.28
4	0.00	0.00	more	0.31	0.62	1.50	0.22	0.25
5	0.00	0.00	less	0.19	0.38	0.00	0.20	0.28
6	0.00	0.00	less	0.19	0.38	0.25	0.20	0.28
7	0.00	0.00	less	0.19	0.38	0.75	0.20	0.27
8	0.00	0.00	less	0.19	0.38	1.50	0.19	0.25
9	0.00	0.50	more	0.31	0.62	1.50	0.22	0.25
10	0.00	0.20	more	0.31	0.62	1.50	0.22	0.25
11	0.00	0.00	more	0.31	0.62	1.50	0.22	0.25
12	0.25	0.00	more	0.31	0.62	1.50	0.24	0.27
13	0.75	0.00	more	0.31	0.62	1.50	0.28	0.32
14	0.00	0.50	less	0.19	0.38	1.50	0.19	0.25
15	0.00	0.20	less	0.19	0.38	1.50	0.19	0.25
16	0.00	0.00	less	0.19	0.38	1.50	0.19	0.25
17	0.25	0.00	less	0.19	0.38	1.50	0.21	0.27
18	0.75	0.00	less	0.19	0.38	1.50	0.23	0.31

Table 5: Limits on the strength of instrument: Average across scenarios of observed instrument strength ($\hat{cor}(H,A_0)$ and $\hat{cor}(H,A_1)$).

the levels considered in our simulation study. In other words, we look at the combined type I and type II error rates averaged over scenarios considered in our simulation study. Because we are interested in determining how BR varies with instrument strength, sample size and degree of unmeasured confounding, we compute the BR for different fixed levels of these factors, see Table 6. We only considered scenarios for which all the assumptions required for IV analysis are satisfied (i.e. scenarios where $\beta_{V\to H}=0$, $\beta_{H\to U}=0$, and $\beta_{H\to Y}=0$). Consider the following:

1. The greatest BR for g-computation is seen for the greatest level of confounding. The relationship between unmeasured confounding and the BR for g-computation however, remains uncertain, as the BR tends to slightly decrease as the unmeasured confounding increases up to 0.75. In the absence of any unmeasured confounding, the BR for g-computation decreases slightly with increasing sample size, see Table 6 lines 1-6, column "g-comp."

2. Given the specific settings of our simulation study, the BR for IV analysis is higher than the BR for g-computation for all scenarios considered. The stronger the instrument, the lower the BR is for IV analysis. In addition, the BR for IV analysis decreases with increasing sample size. As such, when unmeasured confounding is high, and the instrument is strong, the BR for IV analysis will decrease with increasing sample size while the BR for g-computation will increase with increasing sample size. This trade-off suggests that, at a certain point, the testing accuracy will be comparable for both

	n	$\beta_{U\to A_0,A_1,Y}$	IS	g-comp	g-comp|H	IVA
1	750	0.00	0.75	0.26	0.27	0.48
2	750	0.00	1.50	0.25	0.32	0.41
3	1000	0.00	0.75	0.22	0.24	0.47
4	1000	0.00	1.50	0.21	0.29	0.39
5	1400	0.00	0.75	0.18	0.20	0.45
6	1400	0.00	1.50	0.18	0.24	0.35
7	750	0.25	0.75	0.25	0.27	0.47
8	750	0.25	1.50	0.24	0.32	0.41
9	1000	0.25	0.75	0.21	0.23	0.47
10	1000	0.25	1.50	0.21	0.28	0.39
11	1400	0.25	0.75	0.17	0.19	0.45
12	1400	0.25	1.50	0.17	0.24	0.35
13	750	0.75	0.75	0.21	0.23	0.48
14	750	0.75	1.50	0.21	0.27	0.42
15	1000	0.75	0.75	0.18	0.19	0.47
16	1000	0.75	1.50	0.18	0.23	0.39
17	1400	0.75	0.75	0.16	0.16	0.46
18	1400	0.75	1.50	0.15	0.19	0.36
19	750	1.50	0.75	0.30	0.30	0.48
20	750	1.50	1.50	0.28	0.29	0.43
21	1000	1.50	0.75	0.33	0.32	0.48
22	1000	1.50	1.50	0.31	0.30	0.41
23	1400	1.50	0.75	0.36	0.36	0.46
24	1400	1.50	1.50	0.34	0.33	0.38

Table 6: Results of simulation study II - Average Bayes Risk all across scenarios for which $\beta_{V\to H}=0$, $\beta_{H\to U}=0$ and $\beta_{H\to Y}=0$.

approaches, see Table 6 line 24. These observations are in line with those of Boef et al. (2014).

3. Adjusting for an instrument H is detrimental for g-computation in terms of BR when unobserved confounding is absent or is low. However, with a higher degree of unobserved confounding, the BR for unadjusted g-computation and H-adjusted g-computation is comparable. The association between the strength of the instrument and the BR for H-adjusted g-computation is also dependent on the amount of unobserved confounding.

6. The 'test-both' strategy

Given the increasing number of observational studies that simultaneously use both IV analysis and causal inference methods (Laborde-Castérot et al., 2015), our investigation would not be complete without some consideration of how these methods work in conjunction. When the results of both analyses are in agreement, there should be reason for additional confidence in one's conclusions. However, when results disagree, correct interpretation is

n	IS	at least one significant p-value (IV and g-comp)	methods disagree (IV and g-comp)	at least one significant p-value (IV and g-comp| H)	methods disagree (IV g-comp| H)
750	0.75	0.09	0.09	0.10	0.10
750	1.5	0.09	0.08	0.13	0.13
1000	0.75	0.09	0.09	0.10	0.09
1000	1.5	0.09	0.08	0.12	0.12
1400	0.75	0.09	0.09	0.10	0.09
1400	1.5	0.09	0.08	0.12	0.11

Table 7: Proportion of simulated 'null' datasets in which at least one method (IV and/or g-computation) established significance (p-value < 0.05) and in which results disagreed.

more difficult. Brookhart et al. (2010) caution that divergent results may be due to treatment effect heterogeneity and Laborde-Castérot et al. (2015) suggest that such an outcome may be the result of confounding. Disagreement could also indicate a simple lack of power for one method relative to the other.

Another concern is that of potential type I error inflation due to multiple comparisons. Suppose the null hypothesis is rejected should either the causal inference method or the IV analysis produce a p-value ≤ 0.05. Among the various null scenarios (true TE=0) from which we simulated data, the empirical type I error for this 'test-both strategy' is, on average, nearly twice the desired level, see Table 7. This is most problematic for small sample sizes when the instruments is weak, unless g-computation conditions on the instrument. If g-computation does condition on the instrument, type I error inflation increases with instrument strength.

7. Conclusion

The validity of observational studies rests entirely on the proper use of appropriate statistical methods (Sauerbrei et al., 2014), and understanding these methods is now more important than ever. Failure to acknowledge and adequately adjust for "systematic sources of variation" inherent to observational data (Gustafson & McCandless, 2010) has resulted in a situation where "the majority of observational studies would declare statistical significance when no effect is present" (Schuemie et al., 2014). For testing the effect of a time-varying exposure, two rather different approaches are available, each of which requires strong but different assumptions. A wide range of issues must be considered in order to determine which tack is most suitable, and how best it can be followed.

While we discussed the important assumptions required of each approach, we must also address more practical considerations. The availability of data is primary. While IV analysis can require data on as few as two variables, finding a plausibly valid and sufficiently strong instrument is oftentimes a barrier. As is well known, with a weak instrument, an IV analysis will likely be underpowered. To mitigate the potential bias in estimation due to weak instruments, robust methods are recommended (Kleibergen, 2007; Hansen et al.,

2008), though these will not improve power for testing. In order to increase power, one could consider redefining the study population to "build a stronger instrument" (Baiocchi et al., 2010). Despite the fact that this strategy is likely to reduce one's sample size, this maybe a worthwhile trade-off. Ertefaie et al. (2018) conclude that "gains in instrument strength can more than compensate for the loss of sample size."

On the other hand, a causal inference method can require (complete) data on a large number of variables in order to protect against confounding. Appropriate means of tackling sparsity and missing data in the presence of time varying confounding is a current topic of research, e.g. Doretti et al. (2016). Another practical consideration is whether or not, in addition to testing, estimation of effect size is important. Causal inference methods provide consistent estimates in the presence of time-varying exposures. Not so for IV analysis. Most recently, Swanson et al. (2018) discuss the difficulties encountered when instrumental variable methods are applied to time-varying exposures.

In the simple scenario we considered, the assumptions required of the causal approach were relatively straightforward. With increased complexity, additional assumptions are required, depending on the model and method of estimation (e.g. double robust estimators of a marginal structural model or g-estimation of a structural nested model). We considered the benefits and drawbacks of "unnecessary adjustment" (Schisterman et al., 2009) for a potential instrument in the time-varying setting. While there are drawbacks with regards to power, there can be advantages in terms of robustness. With regards to the assumptions required of IV analysis, Wang et al. (2017) have proposed a clever test to determine whether a binary IV model is compatible with the observed data (in the single timepoint setting).

As in all simulation studies, results apply only to the specific scenarios studied. We have four comments on this point. First, our scenarios were characterized by data with only two exposure timepoints. Biases identified may be amplified or attenuated under conditions with more than two timepoints. Second, our simulations were limited to binary outcomes for simplicity. There are recent related simulation studies concerning time-dependent confounding with time-to-event outcomes, see Naimi et al. (2011) and Karim et al. (2016). Third, while we only considered a binary instrument, the methods and trade-offs discussed should be applicable to non-binary instruments as well. Finally, our scenarios included but a single unmeasured confounder. While a single unmeasured confounder may result in considerable bias should its effect be large, a multitude of unmeasured confounders with small effect may prove equally problematic (Fewell et al., 2007).

This work was motivated in part by an application where both analysis strategies could have been employed, but adjustment for time-varying confounding was pursued (Karim et al., 2014). This application, investigating the impact of β-interferon therapy on irreversible disease progression in MS patients, has a more complex structure than considered in this paper. For instance, it involves treatment at three timepoints, and additional adjustment for baseline confounders. So we cannot directly export present findings to this subject-area problem. Nonetheless, the present work suggests a role for more customized simulations in the planning phase for an observational study, when a putative instrumental variable is indeed available.

Using educated guesses about relevant parameters, one can assess how much power an IV analysis would have relative to an analysis which successfully adjusts for time-dependent confounding. Of course an "apples and oranges" comparison cannot be avoided. If indeed the IV assumptions seem more defensible than the unmeasured confounding assumptions, then a judgement is required on what amount of lost power is acceptable for the sake of this improved defensibility.

Acknowledgments

We gratefully acknowledge support from Natural Sciences and Engineering Research Council of Canada.

References

Austin, P. C., Schuster, T., & Platt, R. W. (2015). Statistical power in parallel group point exposure studies with time-to-event outcomes: an empirical comparison of the performance of randomized controlled trials and the inverse probability of treatment weighting (iptw) approach. *BMC Medical Research Methodology*, *15*(1), 87.

Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, *33*(13), 2297–2340.

Baiocchi, M., Small, D. S., Lorch, S., & Rosenbaum, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, *105*(492), 1285–1296.

Biondi-Zoccai, G., Romagnoli, E., Agostoni, P., Capodanno, D., Castagno, D., D'Ascenzo, F., ... Modena, M. G. (2011). Are propensity scores really superior to standard multivariable analysis? *Contemporary Clinical Trials*, *32*(5), 731–740.

Boef, A. G., Dekkers, O. M., Vandenbroucke, J. P., & le Cessie, S. (2014). Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and level of confounding. *Journal of Clinical Epidemiology*, *67*(11), 1258–1264.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450.

Brookhart, M. A., Rassen, J. A., & Schneeweiss, S. (2010). Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety*, *19*(6), 537–554.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149–1156.

Brookhart, M. A., Wang, P., Solomon, D. H., & Schneeweiss, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology (Cambridge, Mass.)*, *17*(3), 268.

Buckley, J. P., Keil, A. P., McGrath, L. J., & Edwards, J. K. (2015). Evolving methods for inference in the presence of healthy worker survivor bias. *Epidemiology*, *26*(2), 204–212.

Burgess, S. (2013). Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine*, *32*(27), 4726–4747.

Burgess, S. (2014). Sample size and power calculations in mendelian randomization with a single instrumental variable and a binary outcome. *International Journal of Epidemiology*, *43*, 922–929.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Cain, L. E., Cole, S. R., Greenland, S., Brown, T. T., Chmiel, J. S., Kingsley, L., & Detels, R. (2009). Effect of highly active antiretroviral therapy on incident aids using calendar period as an instrumental variable. *American Journal of Epidemiology*, *169*(9), 1124–1132.

Clarke, P. S., & Windmeijer, F. (2012). Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, *107*(500), 1638–1652.

Cohen, J. A., & Rudick, R. A. (2011). *Multiple sclerosis therapeutics*. Cambridge University Press.

Daniel, Cousens, S., De Stavola, B., Kenward, M., & Sterne, J. (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine*, *32*(9), 1584–1618.

Daniel, R., De Stavola, B. L., Cousens, S. N., et al. (2011). gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata Journal*, *11*(4), 479.

Dette, H., & Scheder, R. (2006). Strictly monotone and smooth nonparametric regression for two or more variables. *Canadian Journal of Statistics*, *34*(4), 535–561.

Didelez, V., & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, *16*(4), 309–330.

Doretti, M., Geneletti, S., & Stanghellini, E. (2016). Tackling non-ignorable dropout in the presence of time varying confounding. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *65*(5), 775–795.

Ertefaie, A., Small, D. S., & Rosenbaum, P. R. (2018). Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. *Journal of the American Statistical Association*, 1–13.

Fewell, Z., Smith, G. D., & Sterne, J. A. (2007). The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *American Journal of Epidemiology*, *166*(6), 646–655.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, *29*(4), 722–729.

Gustafson, P. (2015). Discussion of 'on bayesian estimation of marginal structural models'. *Biometrics*, *71*(2), 291–293.

Gustafson, P., & McCandless, L. C. (2010). Probabilistic approaches to better quantifying the results of epidemiologic studies. *International Journal of Environmental Research and Public Health*, *7*(4), 1520–1539.

Hadley, J., Yabroff, K. R., Barrett, M. J., Penson, D. F., Saigal, C. S., & Potosky, A. L. (2010). Comparative effectiveness of prostate cancer treatments: evaluating statistical adjustments for confounding in observational data. *Journal of the National Cancer Institute*, *102*(23), 1780–93.

Hansen, C., Hausman, J., & Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, *26*(4), 398–422.

Hernán, Brumback, & Robins, J. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 561–570.

Hernán, & Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, *17*(4), 360–372.

Hertz-Picciotto, I., Arrighi, H. M., & Hu, S.-W. (2000). Does arsenic exposure increase the risk for circulatory disease? *American Journal of Epidemiology*, *151*(2), 174–181.

Johnston, K., Gustafson, P., Levy, A., & Grootendorst, P. (2008). Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, *27*(9), 1539–1556.

Karim, M. E., Gustafson, P., Petkau, J., Zhao, Y., Shirani, A., Kingwell, E., ... Tremlett, H. (2014). Marginal structural cox models for estimating the association between β-interferon exposure and disease progression in a multiple sclerosis cohort. *American Journal of Epidemiology*, *180*(2), 160–171.

Karim, M. E., Petkau, J., Gustafson, P., Platt, R. W., Tremlett, H., Group, B. S., et al. (2016). Comparison of statistical approaches dealing with time-dependent confounding in drug effectiveness studies. *Statistical Methods in Medical Research*, 0962280216668554.

Keil, A. P., Daza, E. J., Engel, S. M., Buckley, J. P., & Edwards, J. K. (2017). A bayesian approach to the g-formula. *Statistical methods in medical research*, 0962280217694665.

Kleibergen, F. (2007). Generalizing weak instrument robust iv statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *Journal of Econometrics*, *139*(1), 181–216.

Laborde-Castérot, H., Agrinier, N., & Thilly, N. (2015). Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review. *Journal of Clinical Epidemiology*, *68*(10), 1232–1240.

Lydersen, S., Fagerland, M. W., & Laake, P. (2009). Recommended tests for association in 2× 2 tables. *Statistics in Medicine*, *28*(7), 1159–1175.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V., & Klungel, O. H. (2006). Instrumental variables: application and limitations. *Epidemiology*, *17*(3), 260–267.

McCulloch, C. E. (2015). Editorial: Observational studies, time-dependent confounding, and marginal structural models. *Issue: Arthritis and Rheumatology*, *67*(3), 609–611.

Moore, K., Neugebauer, R., Lurmann, F., Hall, J., Brajer, V., Alcorn, S., & Tager, I. (2008). Ambient ozone concentrations cause increased hospitalizations for asthma in children: an 18-year study in Southern California. *Environmental Health Perspectives*, *116*(8), 1063–70.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., ... Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, *174*(11), 1213–1222.

Naimi, A. I., Cole, S. R., Westreich, D. J., & Richardson, D. B. (2011). A comparison of methods to estimate the hazard ratio under conditions of time-varying confounding and nonpositivity. *Epidemiology*, *22*(5), 718-723.

Petersen, M. L., Wang, Y., Van Der Laan, M. J., Guzman, D., Riley, E., & Bangsberg, D. R. (2007). Pillbox organizers are associated with improved adherence to hiv antiretroviral therapy and viral suppression: a marginal structural model analysis. *Clinical Infectious Diseases*, *45*(7), 908–915.

Porter, K. E. (2011). *The relative performance of targeted maximum likelihood estimators under violations of the positivity assumption* (Unpublished doctoral dissertation). Univ. of California, Berkeley.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9-12), 1393–1512.

Robins, J. M., & Hernán, M. A. (2009). Estimation of the causal effects of time-varying exposures. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Advances in longitudinal data analysis* (pp. 553–599). New York: Chapman and Hall/CRC Press.

Robins, J. M., & Wasserman, L. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the thirteenth conference on uncertainty in artificial intelligence* (pp. 409–420).

Rotnitzky, A., Li, L., & Li, X. (2010). A note on overadjustment in inverse probability weighted estimation. *Biometrika*, *97*(4), 997–1001.

Sauerbrei, W., Abrahamowicz, M., Altman, D. G., Cessie, S., & Carpenter, J. (2014). Strengthening analytical thinking for observational studies: the stratos initiative. *Statistics in Medicine*, *33*(30), 5413–5432.

Schisterman, E. F., Cole, S. R., & Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, *20*(4), 488–495.

Schnitzer, M. E., Steele, R. J., Bally, M., & Shrier, I. (2016). A causal inference approach to network meta-analysis. *Journal of Causal Inference*, *4*(2), 2193-3677.

Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A., & Madigan, D. (2014). Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine*, *33*(2), 209–218.

Schuster, T., Lowe, W. K., & Platt, R. W. (2016). Propensity score model overfitting led to inflated variance of estimated odds ratios. *Journal of Clinical Epidemiology*, *80*, 97–106.

Senn, S., Graf, E., & Caputo, A. (2007). Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine*, *26*(30), 5529–5544.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, *59*(5), 437.e1–437.e24.

Swanson, S. A., Labrecque, J., & Hernán, M. A. (2018). Causal null hypotheses of sustained treatment strategies: What can be tested with an instrumental variable? *European Journal of Epidemiology*, 1–6.

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

Vansteelandt, S., Bowden, J., Babanezhad, M., Goetghebeur, E., et al. (2011). On instrumental variables estimation of causal odds ratios. *Statistical Science*, *26*(3), 403–422.

Walker, V. M., Davies, N. M., Windmeijer, F., Burgess, S., & Martin, R. M. (2017). Power calculator for instrumental variable analysis in pharmacoepidemiology. *International Journal of Epidemiology*, *46*(5), 1627–1632.

Wang, L., Robins, J. M., & Richardson, T. S. (2017). On falsification of the binary instrumental variable model. *Biometrika*, *104*(1), 229–236.

Xiao, Y., Moodie, E. E., & Abrahamowicz, M. (2013). Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods*, *2*(1), 1–20.

Young, J. G., Hernán, M. A., & Robins, J. M. (2014). Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic Methods*, *3*(1), 1–19.