University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 157

Estimation of Treatment Effects in Randomized Trials with Noncompliance and a Dichotomous Outcome

Mark J. van der Laan* Alan E. Hubbard†

Nicholas P. Jewell‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, hubbard@stat.berkeley.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, jewell@uclink.berkeley.edu

Estimation of Treatment Effects in Randomized Trials with Noncompliance and a Dichotomous Outcome

Mark J. van der Laan, Alan E. Hubbard, and Nicholas P. Jewell

Abstract

We propose a class of estimators of the treatment effect on a dichotomous outcome among the treated subjects within covariate and treatment arm strata in randomized trials with non-compliance. Recent articles by Vansteelandt and Goethebeur (2003) and Robins and Rotnitzky (2004) have presented consistent and asymptotically linear estimators of a causal odds ratio, which rely, beyond correct specification of a model for the causal odds ratio, on a correctly specified model for a potentially high dimensional nuisance parameter. In this article we propose consistent, asymptotically linear and locally efficient estimators of a causal relative risk and a new parameter – called a switch causal relative risk – which only rely on the correct specification of a model for the parameter of interest. As in Vansteelandt and Goethebeur (2003) and Robins and Rotnitzky (2004) our estimators are always consistent, asymptotically linear at the null hypothesis of no-treatment effect, thereby providing valid testing procedures. We examine the finite sample properties of these instrumental variable-based estimators and the associated testing procedures in simulations and a data analysis of decaffeinated coffee consumption and miscarriage.

1 Introduction and the statistical model.

To motivate the estimators using instrumental variables, consider a randomized trial with non-compliance in which the observed data on a randomly sampled subject consists in chronological order of a vector of baseline covariates V, a randomly assigned treatment arm R, an actual treatment received A, and a binary outcome Y. Given a sample of n i.i.d. observations $O_i = (V_i, R_i, A_i, Y_i)$, $i = 1, \ldots, n$, corresponding with n randomly sampled subjects, this article concerns methods for estimation of a causal effect of the actual received treatment A on the outcome Y within a subpopulation defined by $V = v, R = r, A = a$. A particular example of this type is presented in Hirano et al. (2000)).

Many other important examples are covered by allowing R to represent any random variable that is conditionally independent of the characteristics of the subject, given V, but is predictive of the actual treatment A. Such a variable R is often referred to as an instrumental variable. Although their application to estimation of causal effects in epidemiologic studies has been limited, clever choices of instrumental variables can rescue estimation of potentially causal associations in the presence of significant unmeasured confounding. For instance, a particular environmental application is a recreational swimming study for establishing effects of pathogens in the water on the occurrence of illness among the swimmers. In this case, V are measured baseline characteristics of a sampled subject, R is the concentration of pathogens in the ocean on the day the subject swims, T is the amount of time the subject has spent in the water, $A = R * T$ is a measure of exposure to the pathogens, and Y is an outcome such as the occurance of diarrhea (Kay et al. (1994)). In this case, the amount of swimming a subject does is plausibly related to their overall health, which could also be related to their underlying rates of illness. Thus, the simple empirical association of pathogen exposure and illness can be confounded by these unmeasured measures of health. Another application are studies with potentially strong unmeasured confounding for treatments of interest, but where there is a strong predictor of treatment that is not related to prognosis. For instance, Johnston et al. (2002) report a study of different treatments (standard surgical clipping versus endovascular techniques) for ruptured cerebral aneurysms (where particular hospitals are prone to one treatment versus another). In this case, V are baseline covariates, R is the patient's hospital, A is the treatment the patient received, and Y is an outcome (mortality). Finally, we consider a

study of miscarriage and decaffeinated coffee consumption during pregnancy, originally reported in Fenster et al. (1997). In this case, the reported consumption of coffee before pregnancy is a potential instrumental variable as it is related to consumption of decaf during pregnancy, but (in theory) it should have no independent effect on pregnancy outcomes (this example is explored in more detail in the Data Analysis Section below). In the first example R and A are both continuous variables; in the second and third examples R is a categorical variable, having more outcomes than treatment A. For all, R is (plausibly at least) a valid instrumental variable because it is 1) assigned (or chosen) independently of the prognosis of the individual, and 2) strongly predictive of actual treatment or exposure.

In order to formally define the targeted causal parameter, we assume the counterfactual framework for causal inference (Neyman (1990), Rubin (1978), Robins (1986)). That is, the full data structure on a randomly sampled subject (i.e., the experimental unit) is defined as $X \equiv ((Y_{ra} : r, a), V)$, where Y_{ra} is the counterfactual outcome one would have observed, possibly contrary to the fact, if the subject would have been assigned $(R = r, A = a)$, and (r, a) vary over the support of (R, A). The observed data structure is now defined as a missing data structure on X:

$$O \equiv (V, R, A, Y = Y_{RA}). \tag{1}$$

This assumption is often referred to as the consistency assumption implying a subject's observed outcome is equal to the potential outcome associated with the assigned and received treatments, R and A. We also assume that R is randomized:

$$P(R = r \mid X) = P(R = r \mid (Y_{ra} : r, a), V)) = P(R = r \mid V), \tag{2}$$

that is, R is conditionally independent of the subject-specific counterfactual outcomes $(Y_{ra} : r, a)$, given $V = v$. Since our methods rely on having a consistent estimator of $P(R = r \mid V)$, we assume that we have a correctly specified model for these randomization probabilities:

$$P(R = r \mid V) = g_{\eta_0}(r \mid V) \tag{3}$$

for some parametrization $\{g_\eta : \eta\}$. In randomized trials this distribution of the assigned treatment arm is controlled by the experimenter and therefore known. Finally, a typical assumption for instrumental variable estimators,

which also applies to our estimators is the so-called exclusion restriction, which states that $Y_{ra} = Y_a$ with probability 1 for all (r, a) (Angrist et al. (1996), Abadie (2003), and Hirano et al. (2000)). In fact, our estimator requires a somewhat weaker restriction, that $E(Y_{R0} \mid V, R) = \tau(V)$ for any arbitrary function τ.

The causal parameter ψ_0 we wish to estimate is now defined as a particular difference between the conditional probability $m(V, R, A) = E(Y \mid V, R, A)$ of an event in the observed data world and the conditional probability $m_0(V, R, A) = E(Y_0 \mid V, R, A)$ of an event in the counterfactual world in which treatment is set to zero (baseline), within strata defined by the randomized treatment arm $R = r$, treatment $A = a$, and baseline covariates $V = v$. We note that, by the consistency assumption, $m_0(V, R, 0) = m(V, R, 0)$ with probability 1. Specifically, this paper concerns estimation of (i) the (adjusted) causal relative risk ψ_{0RR} of having an event (Robins (1989) and Robins (1994)), (ii) the causal additive risk ψ_{0AR}, and (iii) a newly defined (adjusted) switch causal relative risk ψ_{0SRR}, defined, respectively, by

$$\begin{aligned}
\psi_{0RR}(v, r, a) &= \frac{m_0(v, r, a)}{m(v, r, a)} \\
\psi_{0AR}(v, r, a) &= m_0(v, r, a) - m(v, r, a) \\
\psi_{0SRR}(v, r, a) &= \left(I_{\mathcal{A}_0}(v, r, a) \frac{m_0(v, r, a)}{m(v, r, a)} + I_{\mathcal{A}_0^c}(v, r, a) \frac{1 - m_0(v, r, a)}{1 - m(v, r, a)}, I_{\mathcal{A}_0}(v, r, a) \right),
\end{aligned}$$

where $\mathcal{A}_0 \equiv \{(v, r, a) : m_0(v, r, a) \leq m(v, r, a)\}$ identifies the sub-populations for which treatment is not harmful relative to control , $I_{\mathcal{A}_0}$ denotes the indicator function for the set $\mathcal{A}_0$, and $\mathcal{A}_0^c$ denotes the complement of the set $\mathcal{A}_0$. Since the causal relative risk $\psi_{0RR}^- \equiv \frac{1-m_0(v,r,a)}{1-m(v,r,a)}$ of having no event is nothing else than the causal relative risk ψ_{0RR} of $Y' \equiv 1 - Y$, our results for ψ_{0RR} directly imply the results for ψ_{0RR}^- with appropriate modification. Similarly, redefining the null/baseline value 0 for A used to define m_0, provides other causal relative risk and switch causal relative risks of interest, and are therefore captured by the methodology presented in this article.

The switch causal relative risk, ψ_{0SRR}, yields the causal relative risk, ψ_{0RR}, only for those values of R and A for which $\psi_{0RR} \leq 1$; where $\psi_{0RR} > 1$, it yields ψ_{0RR}^- instead. Note that which region is which is identified as part of the parameter ψ_{0SRR}. In a randomized trial with non-compliance with two treatment arms, there can only be two possible values of R and A where ψ_{0RR} can differ from 1: either $R = A = 1$ or $R = 0, A = 1$, otherwise

$m_0 = m$. We also remark that this parameter reduces to a marginal causal effect within strata of V in the special case that (A, R) is jointly randomized (i.e., $(A, R) \perp X \mid V$), since in that case $m_0(v, r, a) = P(Y_{0r} = 1 \mid V)$ and $m(v, r, a) = P(Y_{ar} = 1 \mid V)$.

The first important issue to discuss is the identification of ψ_0 in the above model for the observed data distribution. As noted in the literature, without making further assumptions, ψ_0 cannot be identified: see, for example, Balke and Pearl (1994) and Balke and Pearl (1997), who establish bounds for the additive risk. We also refer to (Angrist et al. (1996), and Abadie (2003) for discussions on the identification of causal effects based on instrumental variables.

From an estimating function point of view (see e.g., Robins (1989), Robins (1994)), each function $h(R, V)$ with conditional mean zero, given V, maps into an unbiased estimating function for ψ_0, which we use to propose estimating functions in Section 3. For example, if all variables are discrete, then one can identify for each value v of V, $\mid \mathcal{R} \mid -1$ number of parameters, where $\mid \mathcal{R} \mid$ denotes the number of categories of R. However, since ψ_0 can be any function of (v, r, a) which equals 1 for $a = 0$, it typically follows that the data generating distribution does not completely identify ψ_0.

On the other hand, if R has 2 or more categories, A is binary, and, for at least one value of R, A is determined (e.g., $P(A = 0 \mid R = 0) = 1$), then we have the wished non-parametric identifiability. An example is a randomized trial for comparing two treatment arms in which everybody in the control group complies.

In order to deal with the curse of dimensionality and/or the fact that ψ_0 is typically not fully identifiable from the observed data, we assume a correctly specified model for the parameter of interest ψ_0 of the distribution of (X, A, R):

$$\psi_0(v, r, a) = \gamma(v, r, a \mid \beta_0) \tag{4}$$

for some parametrization $\beta \to \gamma(\cdot \mid \beta)$ respecting the constraint $\gamma(v, r, 0 \mid \beta) = 1$ for all β.

In the next section, we will describe such models for the causal relative risk, causal additive risk, and switch causal relative risk parameter. In this article we are concerned with estimation of β_0 in the above model defined by (1), (2), (3), and (4).

To end this section, we will review the immediately relevant literature on this model and estimation problem. Robins (1989) and Robins (1994) pro-

vide the class of robust unbiased estimating functions (and thereby of corresponding estimators) for the causal additive and relative risk ψ_0 in the above model for *continuous* and *count* outcomes. Robins refers to these models as additive and multiplicative structural nested mean models, where each particular structural nested mean model corresponds with a link function Φ: i.e., $\Phi(x) = x$ and $\Phi(x) = exp(x)$, respectively. Robins and Rotnitzky (2004) remark that, for dichotomous outcomes, additive and multiplicative structural nested mean models cannot generally be used, because these models may fail to guarantee response probabilities in the interval $(0, 1)$. As a consequence, Vansteelandt and Goethebeur (2003) and Robins and Rotnitzky (2004) focus on a logistic structural nested mean model so that ψ_0 denotes the causal odds ratio, given V, R, A:

$$\psi_0(v, r, a) = \log \left\{ \frac{P(Y_{ra} = 1 \mid V = v, R = r, A = a)}{P(Y_{ra} = 0 \mid V = v, R = r, A = a)} \Big/ \frac{P(Y_{r0} = 1 \mid V = v, R = r, A = a)}{P(Y_{r0} = 0 \mid V = v, R = r, A = a)} \right\}$$

In contrast to the multiplicative and additive link functions, Robins and Rotnitzky (2004) show that, in this logistic structural nested mean model, consistent (and asymptotically linear) estimation of ψ_0 requires, beyond correct specification of models for $P(R \mid V)$ and ψ_0, also correct specification of models for nuisance parameters (e.g., $E(Y \mid V, R, A)$). Vansteelandt and Goethebeur (2003) and Robins and Rotnitzky (2004) show, however, that when the null hypothesis of no treatment effect is true, estimators based on their class of estimating functions remain consistent at misspecified nuisance parameters. Thus, consistent tests of treatment effects can be derived that do not rely on correct specification of the nuisance parameters.

We will show that the switch causal relative risk can be directly modelled in terms of (e.g.) a logistic model, so that the concern expressed above for modelling the causal relative risk does not apply to this newly defined parameter. We also show that, by using a particular modelling strategy, it is possible to obtain appropriate models for the causal relative and additive risk as well. In this manner, in this article we are able to provide locally efficient estimators of the causal relative risk, causal additive risk, and causal switch relative risk in the above model for the observed data without the need to rely on correct specification of a nuisance parameter.

1.1 Organization.

In Section 2 we present our models for the causal relative risk which is known to be bounded by $1/\delta$ for some $\delta \in (0,1)$, a general causal relative risk, and the switch causal relative risk. In Section 3 we present for each of these causal parameters and corresponding models the class of estimating functions, and the corresponding asymptotically linear (locally efficient) estimators. In particular, we discuss statistical inference for the switch causal relative risk parameter that addresses the irregular behavior of estimators at data generating distributions in which for certain strata (different from $a = 0$) the true causal relative risk equals 1: that is, $\{(a,r,v) : a \neq 0, m_0(a,r,v)/m(a,r,V) = 1\}$ is a set with positive probability. Finally, in Section 4 we conclude with a simulation study and in Section 5 a data analysis to illustrate the practical performance of our estimators; the article finishes with a discussion.

2 Modelling the causal risks.

2.1 Multiplicative structural nested mean model for causal relative risk.

The primary objective of this subsection is to describe a class of (multiplicative structural nested mean) models for ψ_{0RR} that possesses two fundamental properties: Property I and Property II. Property I states that one can always choose a sufficiently flexile model in this class so that it contains the true causal relative risk function ψ_{0RR}. Property II states that, even when the model is misspecified, the model respects the fact that m times the misspecified fit of ψ_0 is contained in $[0,1]$. The verification of these two properties Property I and II is deferred to the Appendix.

In general, we have the following generic modelling strategy for modelling ψ_{0RR}. We specify a possibly misspecified working model (in fact, a singleton will represent an important special case)

$$P_\alpha(Y = 1 \mid V, R, A) = m(V, R, A \mid \alpha) = \frac{1}{1 + \exp(-f(V, R, A \mid \alpha))},$$

indexed by a parameter α. In case this working model is not a trivial singleton, we let α_n be the iteratively re-weighted least squares estimator (i.e., the

maximum likelihood estimator) defined by:

$$\alpha_n = \arg\min_{\alpha} \sum_{i=1}^{n} (Y_i - m(V_i, R_i, A_i \mid \alpha))^2 \frac{1}{m(V_i, R_i, A_i \mid \alpha)(1 - m(V_i, R_i, A_i \mid \alpha))}.$$

Let α_1 denote the limit of α_n. Secondly, we specify another (possibly misspecified) working model for the counterfactual conditional expectation

$$P_{\alpha,\beta}(Y_{R0} = 1 \mid V, R, A) = m_0(V, R, A \mid \alpha, \beta),$$

where we enforce the constraint that, for all (α, β), $m_0(V, R, 0 \mid \alpha, \beta) = m(V, R, 0 \mid \alpha)$. For example, a possible parametrization is

$$\begin{aligned} m(V, R, A \mid \alpha) &\equiv \frac{1}{1 + \exp(-f(V, R, A \mid \alpha) - C(V, R, A \mid \alpha))} \\ m_0(V, R, A \mid \alpha, \beta) &\equiv \frac{1}{1 + \exp(-f_0(V, R, A \mid \beta) - C(V, R, A \mid \alpha))}, \end{aligned} \quad (5)$$

where $f(V, R, 0 \mid \alpha) = f_0(V, R, 0 \mid \beta) = 0$ for all R, V, α, β. Thus, one could model $f(V, R, A \mid \alpha) = A * h(R, V \mid \alpha)$ and $f_0(V, R, A \mid \beta) = A * h_0(R, V \mid \beta)$ for certain parameterizations h and h_0.

We now assume the following multiplicative structural nested mean model, in terms of the working model for m_0 and limit $m(\cdot \mid \alpha_1)$, for the causal relative risk ψ_{0RR}:

$$\psi_{0RR} \in \left\{ \gamma_{\alpha_1}(V, R, A \mid \beta) \equiv \frac{m_0(V, R, A \mid \alpha_1, \beta)}{m(V, R, A \mid \alpha_1)} : \beta \right\}.$$

Let β_0 be the true parameter value: that is, $\psi_{0RR} = \gamma_{\alpha_1}(\cdot \mid \beta_0)$.

One should not view the parametrization $\{m_0(\cdot \mid \alpha_1, \beta) : \beta\}$ as a model for the true counterfactual response probability m_0, but one should view $\{m_0(\cdot \mid \alpha_1, \beta)/m(\cdot \mid \alpha_1) : \beta\}$ as a model for the causal relative risk $\psi_{0RR} = m_0/m$: this is particularly obvious in the important case that we choose a singleton as working model for m. The only reason for selecting $m(\cdot \mid \alpha_1)$ data adaptively (by fitting a model) is to guarantee that, if the collection $\{m_0(\cdot \mid \alpha_1, \beta) : \beta\}$ of functions mapping into $[0, 1]$ is chosen large enough, then our model for ψ_{0RR} always contains the truth. In the Appendix, we show this is true if $m_0/m \leq 1/m(\cdot \mid \alpha_1)$.

In particular, if it is known that $\psi_{0RR} \leq \frac{1}{\delta}$ for some $\delta \in (0,1)$, one can set $m(\cdot \mid \alpha_1) = \delta$. This corresponds with the following multiplicative structural nested mean model

$$\gamma_\delta(V,R,A \mid \beta) = \frac{1/\delta}{1+\exp(-f_0(V,R,A \mid \beta) - C(\delta))}, \tag{6}$$

where $C(\delta) = \log(\delta/1-\delta)$ and $f_0(V,R,A \mid \beta)$ is a parametrization satisfying $f_0(V,R,0 \mid \beta) = 0$. Note that the null hypothesis $H_0 : \psi_{0RR} = 1$ corresponds with the test $H_0 : f_0(V,R,A \mid \beta_0) = 0$ a.e., which, for most parameterizations, is equivalent to $H_0 : \beta_0 = 0$.

One could also decide to let δ be a parameter of this multiplicative structural nested mean model, in which case $\gamma(V,R,A \mid \beta,\delta) \equiv \gamma_\delta(V,R,A \mid \beta)$ is our model with (β_0,δ_0) being the unknown parameter.

Model for additive causal risk. The same modeling strategy can be applied for additive structural nested mean models for $\psi_{0AR} = m_0 - m$. In order to have the wished model Properties I and (analogue of) II, the allowed level of misspecification of $m(\cdot \mid \alpha_1)$ is now on the more restrictive additive scale: for details, we refer to the Appendix.

2.2 Model for switch causal relative risk.

We will first adopt the same modeling strategy as for the causal relative risk, and subsequently point out that in this case we can always choose a singleton $m(\cdot \mid \alpha_1) = 0.5$ as working model for m. Thus, we assume the following parametrization for the switch causal relative risk ψ_{0SRR}

$$\gamma_{\alpha_1}(V,R,A \mid \beta) = (\gamma^1_{\alpha_1}(V,R,A \mid \beta), I_{\mathcal{A}(\alpha_1,\beta)}(V,R,A)), where \tag{7}$$

$$\mathcal{A}(\alpha_1,\beta) \equiv \left\{(V,R,A) : \frac{m_0(V,R,A \mid \alpha_1,\beta)}{m(V,R,A \mid \alpha_1)} \leq 1\right\},$$

and

$$\begin{aligned}\gamma^1_{\alpha_1}(V,R,A \mid \beta) \;\equiv\; & I_{\mathcal{A}(\alpha_1,\beta)}(V,R,A)\frac{m_0(V,R,A \mid \alpha_1,\beta)}{m(V,R,A \mid \alpha_1)} \\ & + I_{\mathcal{A}(\alpha_1,\beta)^c}(V,R,A)\frac{1-m_0(V,R,A \mid \alpha_1,\beta)}{1-m(V,R,A \mid \alpha_1)}.\end{aligned}$$

In the Appendix we verify that the two wished model properties I, II hold at any $m(\cdot \mid \alpha_1)$. Since our model for the switch causal relative risk is valid at any $m(\cdot \mid \alpha_1)$, there is truly no need to fit m at all. Instead, we can simply use the model implied by (e.g.) $m(\cdot \mid \alpha_1) = 0.5$. This yields the following parametrization $\gamma(V, R, A \mid \beta)$ for the switch causal relative risk:

$$\gamma(V, R, A \mid \beta) = \left(I_{\mathcal{A}(\beta)}(V, R, A)\frac{m_0(V,R,A|\beta)}{0.5} + I_{\mathcal{A}(\beta)^c}(V, R, A)\frac{1-m_0(V,R,A|\beta)}{0.5}, I_{\mathcal{A}(\beta)}(V, R, A)\right).$$

Here $\mathcal{A}(\beta) \equiv \{(V, R, A) : m_0(V, R, A \mid \beta)/0.5 \leq 1\}$, and $m_0(\cdot \mid \beta)$ is a $[0, 1]$-valued parametrization satisfying $m_0(V, R, 0 \mid \beta) = 0.5$. A possible parametrization is

$$m_0(\cdot \mid \beta) = \frac{1}{1 + \exp(-A * f_0(R, V \mid \beta))},$$

which indeed satisfies $m_0(V, R, 0 \mid \beta) = 0.5$ everywhere. Note that the null hypothesis $H_0 : \psi_{0SRR} = 1$ now corresponds with testing $H_0 : f_0(R, V \mid \beta_0) = 0$ a.e. For example, if $f_0(R, V) = \beta_{0o}R + \beta_{10}V + \beta_{20}RV$, then this is equivalent with testing $H_0 : \beta_0 = (\beta_{00}, \beta_{10}, \beta_{20}) = 0$.

Note, the key idea behind the switch causal relative risk and its estimators is the generalized (to discrete outcomes) quantile-quantile function, as proposed in Yu and van der Laan (2002) and we provide a detailed explanation of this relationship in the Appendix.

3 Estimation and Inference.

3.1 The class of estimating functions and corresponding estimators.

Let $H_0(Y, V, R, A \mid \alpha_1, \beta)$ be a function of the observed data structure $O = (Y, V, R, A)$ and the unknown parameters of our model $\{\gamma_{\alpha_1}(\cdot \mid \beta) : \beta\}$ of our parameter of interest ψ_0 which satisfies

$$E(H_0(Y, V, R, A \mid \alpha_1, \beta_0) \mid V, R, A) = m_0(V, R, A).$$

Specifically, depending on the parameter of interest ψ_0, this function $(O, \beta) \rightarrow H_0(O \mid \alpha_1, \beta)$ is defined as follows (note, in the models for the causal relative risk assuming that $\psi_{0RR} \leq 1/\delta$ for some known $\delta \in (0, 1)$, and the switch

causal relative risk, if one sets $m(\cdot \mid \alpha_1) = 0.5$), then $\gamma_{\alpha_1} = \gamma$ is known and in the this case we use the notation $H_0(O \mid \alpha_1, \beta) = H_0(0 \mid \beta)$)

$$\begin{aligned}
H_{0RR}(O \mid \alpha_1, \beta) &= I(Y=1)\gamma_{\alpha_1}(V, R, A \mid \beta) \\
H^{-}_{0RR}(O \mid \alpha_1, \beta) &= 1 - I(Y=0)\gamma_{\alpha_1}(V, R, A \mid \beta) \\
H_{0AR}(O \mid \alpha_1, \beta) &= I(Y=1) - \gamma_{\alpha_1}(V, R, A) \\
H_{0SRR}(O \mid \beta) &= I_{\{m_0(V,R,A|\beta)/0.5 \leq 1\}} I(Y=1) \frac{m_0(V, R, A \mid \beta)}{0.5} \\
&\quad + I_{\{m_0(V,R,A|\beta)/0.5 > 1\}} \left(1 - I(Y=0) \frac{1 - m_0(V, R, A \mid \beta)}{0.5}\right)
\end{aligned}$$

or, using an estimator of m

$$\begin{aligned}
H_{0SRR}(O \mid \alpha_1, \beta) &= I_{\{m_0(V,R,A|\alpha_1,\beta)/m(V,R,A|\alpha_1) \leq 1\}} I(Y=1) \frac{m_0(V, R, A \mid \alpha_1, \beta)}{m(V, R, A \mid \alpha_1)} \\
&\quad + I_{\{m_0(V,R,A|\alpha_1,\beta)/m(V,R,A|\alpha_1) > 1\}} \left(1 - I(Y=0) \frac{1 - m_0(V, R, A \mid \alpha_1, \beta)}{1 - m(V, R, A \mid \alpha_1)}\right).
\end{aligned}$$

For any user supplied function h of (R, V) and q of V, we have the following unbiased estimating function for β:

$$D_{h,q,\alpha_1}(O, \beta \mid \eta) = (h(R, V) - E_\eta(h(R, V) \mid V))\,(H_0(O \mid \alpha_1, \beta) - q(V))\,.$$

The estimating function D_{h,q,α_1} for β is indexed by the nuisance parameter η of our model g_η for $P_{R|V}$. For any h and q, this estimating function has expectation zero at the true β_0 and true η_0. This is shown by 1) first conditioning on R, V, 2) using that $E(H_0(O \mid \alpha_1, \beta_0) \mid R, V) = E(m_0(V, R, A) \mid R, V) = E(Y_0 \mid R, V)$ and by the exclusion restriction, this equals $E(Y_0 \mid V)$. Finally, note that for any function $f(V)$, $E(\{h(R, V) - E(h \mid V)\}f(V)) = 0$. In fact, the estimating functions are double robust in the following sense.

Result 1 *Let $O \sim P_0$. Consider the class of estimating functions $D_{h,q,\alpha_1}(O, \beta \mid \eta)$ indexed by nuisance parameters $\eta = P_{R|V}$ defined by:*

$$D_{h,q,\alpha_1}(O, \beta \mid \eta) \equiv (h(R, V) - E_\eta(h(R, V) \mid V))(H_0(O \mid \alpha_1, \beta) - q(V)).$$

If either $\eta(V) = \eta_0(V)$ (thus $P_{R|V}$ is correctly specified) or

$$\begin{aligned}
q &= q_{opt}(V) \\
&\equiv E_{P_0}(H_0(O \mid \alpha_1, \beta_0) \mid V) = E_{P_0}(Y_0 \mid V),
\end{aligned}$$

then $E_{P_0} D_{h,q,\alpha_1}(O, \beta_0 \mid \eta) = 0$.

This result can be directly verified.

Given an estimator η_n of η_0, a k-variate choice (possibly data dependent) $h_n = (h_{n1}, \ldots, h_{nk})$ and univariate q_n (estimating q_{opt}), we propose to estimate β_0 with the solution $\beta_n = \beta_n(h_n, q_n, \eta_n, \alpha_1)$ of the k-variate equation

$$0 = \sum_{i=1}^{n} D_{h_n,q_n,\alpha_1}(O_i, \beta \mid \eta_n).$$

If α_1 is unknown, then α_1 is replaced by the weighted least squares estimator α_n. If a solution β_n does not exist, then one can simply set β_n equal to the minimizer of the Euclidean norm of this estimating equation. Because the estimating equation is differentiable at all β, except at $\beta = 0$ for the switch causal relative risk model, one can use the Newton-Raphson algorithm to determine the solution (or minimum) with the usual line search to guarantee convergence. The non-differentiability at $\beta = 0$ discussed below does not cause the derivatives used in the Newton-Raphson algorithm to converge to infinity for $\beta \approx 0$, since the differential quotients at $\beta = 0$ are bounded, but does not converge to a unique limit.

Clearly, the efficiency of the estimator $\beta_n(h, q, \eta_0, \alpha_1)$ can be strongly affected by the choice (h, q). Therefore it is natural to use a data dependent (h_n, q_n) which is designed to locally estimate an optimal choice (h_{opt}, q_{opt}) for which we provide and derive the closed form formula in the Appendix.

3.2 Asymptotic linearity of the estimators.

In the next subsections, if the parameter of interest is the switch causal relative risk, then we make the assumption that

$$P_0\left((V, R, A) \in \{(v, r, a) : m_0(v, r, a)/m(v, r, a) = 1, a \neq 0\}\right) = 0, \qquad (8)$$

where we remind the reader that P_0 is the distribution of O. This assumption is not needed for asymptotic linearity and inference for the causal relative and additive risk. In the last subsection, we discuss statistical inference for the switch causal relative risk not relying on this assumption.

In our model with the true $g_{\eta_0} = P_{R|V}$ being known, under appropriate regularity conditions, we have that $\beta_n(h, q, \eta_0, \alpha_1)$ is an asymptotically linear estimator of β_0 with influence curve

$$IC_{h,q,\alpha_1}(O) \equiv -\frac{d}{d\beta} E_{P_0} D_{h,q,\alpha_1}(O, \beta \mid \eta_0)|_{\beta=\beta_0}^{-1} D_{h,q,\alpha_1}(O, \beta_0 \mid \eta_0). \qquad (9)$$

That is,

$$\sqrt{n}(\beta_n(h,q,\eta_0,\alpha_1)-\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC_{h,q,\alpha_1}(O_i)+o_P(1/\sqrt{n}),$$

and $\sqrt{n}(\beta_n(h,q,\eta_0,\alpha_1)-\beta_0)$ converges in distribution to the multivariate normal distribution $N(0,\mathrm{COV}(IC_{h,q,\alpha_1}(O)))$.

If η_0 is replaced by an efficient estimator η_n according to the model $\{g_\eta:\eta\}$ (e.g., $\eta_n=\arg\max_\eta \prod_i g_\eta(R_i \mid V_i)$ is the maximum likelihood estimator), then $\beta_n(h,q,\eta_n,\alpha_1)$ is asymptotically linear at P_0 with influence curve $IC_{h,q,\alpha_1}-\Pi(IC_{h,q,\alpha_1} \mid T_{\eta_0})$, where T_{η_0} denotes the linear subspace of $L_0^2(P_0)$ spanned by the scores of η, and $\Pi(\cdot \mid T_{\eta_0})$ denotes the projection operator onto this subspace in this Hilbert space $L_0^2(P_0)$ (see Theorem 2.3, page 135 van der Laan and Robins (2002)). That is, the efficiency of β_n is non-decreasing in the dimension of the model $\{g_\eta:\eta\}$ for η_0. In the special case that $q=q_{opt}$ the projection $\Pi(IC_{h,q_{opt},\alpha_1} \mid T_{\eta_0})=0$ for all nuisance tangent spaces T_{η_0}. Thus, the explicit influence curve IC_{h,q,α_1} can always be used as a conservative influence curve providing conservative confidence intervals.

Given asymptotic linearity uniformly in h and q, under the same regularity conditions,

$$\sqrt{n}(\beta_n(h_n,q_n,\eta_0,\alpha_1)-\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC_{h^*,q^*,\alpha_1}(O_i)+o_P(1/\sqrt{n}),$$

and $\sqrt{n}(\beta_n(h_n,q_n,\eta_0,\alpha_1)-\beta_0)$ converges in distribution to the multivariate normal distribution $N(0,\Sigma(h^*,q^*)=\mathrm{COV}(IC_{h^*,q^*,\alpha_1}(O)))$, where h^* and q^* denote the limits of h_n and q_n, respectively. Similarly as in the previous paragraph, efficient estimation of η_0 subtracts out the projection of IC_{h^*,q^*} onto the tangent space of $\{g_\eta:\eta\}$.

3.3 Local efficiency.

In the Appendix we show that the asymptotic covariance matrix $\Sigma(h,q)$ is optimal at an explicitly specified (h_{opt},q_{opt}). This proves that, if $(h^*,q^*)=(h_{opt},q_{opt})$, then $\beta_n(h_n,q_n,\eta_n,\alpha_1)$ is asymptotically optimal among our class of candidate estimators indexed by all (h,q) and is asymptotically efficient.

3.4 Confidence regions.

This asymptotic linearity result, under condition (8) for the switch causal relative risk, allows us now to construct Wald-type (conservative) asymptotic $1-\alpha$ confidence intervals based on the asymptotically valid working model $\beta_n \sim N(\beta_0, \Sigma_n/n)$, where

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} \hat{IC}_{h_n,q_n,\alpha_1}(O_i)^2$$

and $\hat{IC}_{h_n,q_n}$ is the substitution estimator of IC_{h_n,q_n}. If α_1 is estimated (not known), then the influence curve has an additional component, which can be explicitly derived.

Because of the smoothness of the estimating function in β, β_n will be a compactly differentiable functional of the empirical distribution so that the bootstrap is asymptotically consistent as well (van der Vaart and Wellner (1996)). Thus, one could also use the bootstrap to construct an asymptotic $1-\alpha$ confidence region for β_0, which is particularly attractive in the case that α_1 is estimated.

3.5 Asymptotic behavior and inference when α_1 is estimated.

Since there is no need to estimate α in the switch causal relative risk model, we will here only discuss the implications of estimating α_1 in the causal relative risk model. Above, we provided a locally efficient estimator $\beta_n(h_n, q_n, \eta_n, \alpha_1)$ of β_0 of the unknown parameters in our assumed model $\{\gamma_{\alpha_1}(\cdot \mid \beta) : \beta\}$ for the causal relative risk ψ_{0RR}. That is, given α_1, $\beta_n(h_n, q_n, \eta_n, \alpha_1)$ is a locally efficient estimator of the true β_0 satisfying $\gamma_{\alpha_1}(\cdot \mid \beta_0) = \psi_{0RR}$. If α_1 denotes the limit of the maximum likelihood estimator α_n (i.e., the iteratively re-weighted least squares estimator) according to a working model $m(\cdot \mid \alpha)$ for $E(Y \mid A, R, V)$, then α_1 is an unknown nuisance parameter. Since α_n is, by definition of α_1 as the limit of α_n, a consistent and asymptotically linear estimator of α_1, under regularity conditions, we also have that $\beta_n(h_n, q_n, \eta_n, \alpha_n)$ is an asymptotically linear estimator of β_0. In addition, its influence curve can be explicitly derived. If it can be argued that the iterative re-weighted least squares estimator α_n is an efficient estimator of α_1 in our causal model for the observed data, then it also follows that $\beta_n(h_n, q_n, \alpha_n)$ is locally efficient. This statement follows from the general result that a differentiable

function of an efficient estimator is efficient (van der Vaart (1991)). At minimal this argument suggests hat $\beta_n(h_n, q_n, \alpha_n)$ is, if not locally efficient, it will approach locally efficiency. For the purpose of inference, in order to avoid calculation of the influence curve, we recommend the bootstrap.

3.6 Irregularity of the switch causal relative risk at the null.

James Robins and a referee made us aware of the fact that the switch causal relative risk is not a path-wise differentiable parameter at a data generating distribution which violates assumption (8). This follows from the fact that

$$\beta \to E_0 D_{h,q,\alpha_1}(O \mid \beta, \eta_0),$$

is not differentiable at the true β_0 if $m_0(v, r, a \mid \beta_0, \alpha_1) = m(v, r, a \mid \alpha_1)$ on a set which has positive probability under P_0. At such β_0, the indicator function $\beta \to I(m_0(\cdot \mid \beta) \leq m(\cdot \mid \alpha_1))$ in the estimating function D_{h,q,α_1} can jump from 1 to 0 in any neighborhood of β_0, and thereby causes a discontinuity in the derivative at β_0: that is, one can calculate "left" and "right" derivatives of the expectation of the estimating function as a function of β at β_0, but they are not equal to each other. Interestingly enough, in the special case that $m(\cdot \mid \alpha_1) = m$, the derivative does exist, but this result is not useful since we wish to avoid correct specification of a model for m.

At a data generating distribution violating (8), by carrying out a generalized type of Taylor-expansion at β_0, noting that a derivative along a sequence β_n converging to β_0 is still bounded (but does not converge to a unique limit), and using empirical process theory, one can still show that β_n is a root-n consistent estimator of β_0. Unfortunately, we have not been able to prove weak convergence of β_n to a particular limiting distribution.

3.7 Inference for the switch causal relative risk at the null and testing.

We refer to Robins (2004) who discusses in detail inference in the case that 1) estimators solve estimating equations which are non-differentiable at null-values of the parameter of interest and 2) the parameter of interest is not path-wise differentiable at these null values. Robins points out that the irregularity of the estimators at these null values causes the Wald-type confidence

regions to not have the wished coverage *uniformly* over the whole model (assuming the model does not exclude a neighborhood of these null values), and testing at these null-values with the test statistic being a standardized version of the estimator β_n itself would require deriving the limit distribution of the standardized estimator at these null values, and the latter does not necessarily exist.

Therefore, Robins proposed inference and testing based on the multivariate normal limit distribution of the standardized estimating equation. Specifically, let

$$U_n(\beta) = P_n D_{h_n,q_n,\alpha_n}(\cdot \mid \beta, \eta_n),$$

and let $U_0(\beta) = P_0 D_{h^*,q^*,\alpha_1}(\cdot \mid \beta, \eta_0)$ denote its target, where we note that $U_0(\beta_0) = 0$. Under regularity conditions, we have that $\sqrt{n}(U_n - U_0)$ converges in distribution as a random function in β to Gaussian process, and, in particular,

$$\{U_n(\beta) - U_0(\beta)\}^\top \Sigma_n(\beta)^{-1} \{U_n(\beta) - U_0(\beta)\} \stackrel{D}{\Rightarrow} \mathcal{X}^2_k,$$

where $\Sigma_n(\beta)$ denotes a consistent estimator of the asymptotic covariance matrix of $U_n(\beta)$, and $\mathcal{X}^2_k$ denotes the Chi-square distribution with k degrees of freedom. As a consequence, an asymptotically valid confidence region for the true parameter value β_0 is given by:

$$\{\beta : U_n(\beta)^\top \Sigma_n(\beta_n)^{-1} U_n(\beta) \leq \mathcal{X}^2_{k,0.95}\},$$

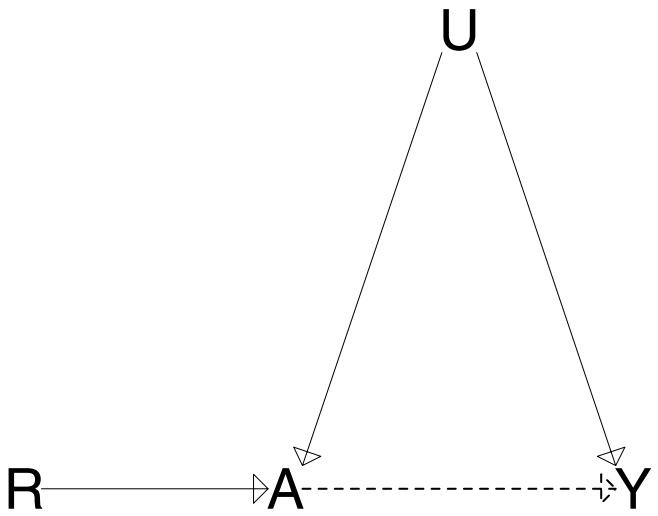
where $\mathcal{X}^2_{k,0.95}$ denotes the 0.95-quantile of the Chi-square distribution with k degrees of freedom. In general, we propose to use the bootstrap to estimate the covariance matrix $\Sigma_n(\beta)$ at a given β, but if η_0 and α_1 are known, then $\Sigma_n(\beta)$ can be trivially estimated with the empirical covariance matrix of $D_{h_n,q_n,\alpha_1}(O_i \mid \beta, \eta_0)$, $i = 1, \ldots, n$.

Similarly, we obtain a valid test for testing $H_0 : \beta_0 = 0$ by using as test-statistic $U_n(0)^\top \Sigma_n(\beta_n)^{-1} U_n(0)$, and rejecting the test if this test statistic exceeds $\mathcal{X}^2_{k,0.95}$.

4 Simulation Study

To investigate the consistency of the proposed estimators, we conducted several simulations based on the causal diagram displayed in Figure 1. The influence of an unmeasured prognostic variable, U, and the treatment as-

Figure 1: Causal diagram with instrumental variable(R), treatment(A), outcome (U) and unmeasured confounder (U)



signment on the choice of the treatment actual taken is implemented through the interaction of U and R on A. Our simulation is from a data-generating distribution where we can non-parametrically identify the parameters of interest, specifically the causal relative risk ψ_{ORR}. In this case, we generate both U and R uniformly over the integers $(0, 1, 2)$, $A \mid U, R$ for $R > 0$ is from a binary with probability model $\sim logit\left[P(A \mid U, R)\right] = b_0 + b_1 U + b_2 R$, with $(b_0, b_1, b_2) = (-5, 2, 2)$; A is deterministically 0 if $R = 0$. Finally Y is simulated from $logit\left[P(Y \mid U, A, R)\right] = a_0 + a_1 U + a_2 A$ with $(a_0, a_1, a_2) = (-5, 2, 2)$. In this case, there are 2 causal relative risks of interest (one for $R = 1$ and $R = 2$) since the relative risk, $\psi_{0RR}(R = 0, A = 1)$ is undefined. Given the data-generating model, one can easily calculate the true risk ratios from this data as:

$$\psi_{0RR}(R = 1, A = 1) = \frac{m_0(R = 1, A = 1)}{m(R = 1, A = 1)} = 0.34 \tag{10}$$

$$\psi_{0RR}(R = 2, A = 1) = \frac{m_0(R = 2, A = 1)}{m(R = 2, A = 1)} = 0.32. \tag{11}$$

One can fit a saturated model to this data of the form:

$$\psi_{0RR}(R = r, A = 1) = A * (\beta_0 + \beta_1 * I(r = 2)) + (1 - A), \tag{12}$$

which guarantees that $\psi_{0RR}(R = r, A = 0) = 1$. As estimating function, we use:

$$D_{h,q}(O, \beta) = \left(h(R) - E[h(R)]\right)\left(H_0(O \mid \beta)\right) - q\right), \tag{13}$$

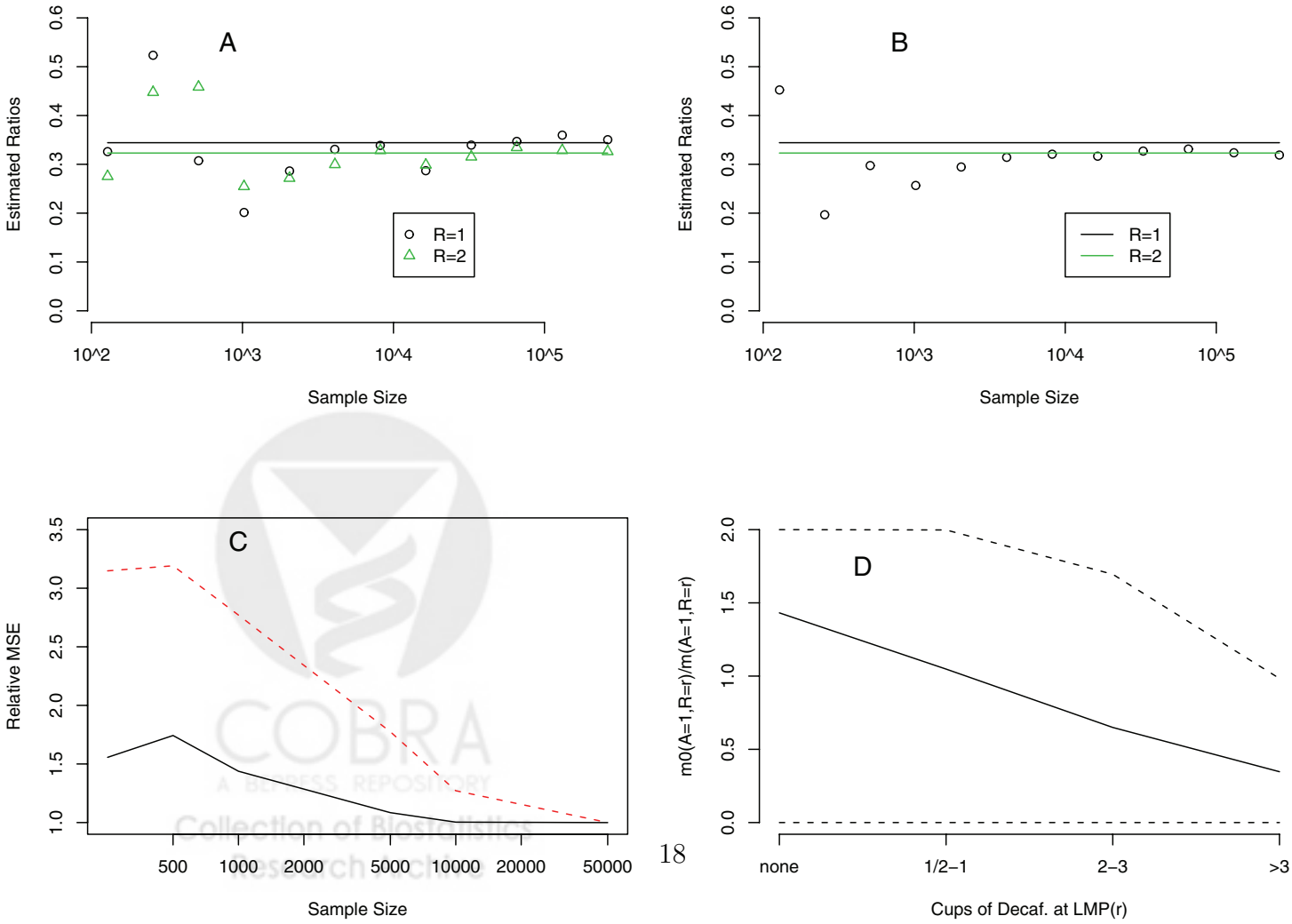
where

$$\begin{aligned} h(R) &= E\left(\frac{d}{d\beta}\epsilon(\beta) \mid_{\beta=\beta_0} \mid R\right) \\ \epsilon(\beta) &= H_0(O \mid \beta) - E(H_0(O \mid \beta)) \\ H_0(O \mid \beta) &= Y\psi_{0RR}(a) \\ q &= E_0(H_0(O \mid \beta)). \end{aligned}$$

This estimating equation is linear in the parameters and can be solved as a linear regression of the form $Z = \beta_0 X_0 + \beta_1 X_0$ where $Z = -\left[Y(1 - A) - \hat{E}[Y(1 - A)]\right]$, $X_0 = \hat{E}(Y * A \mid R) - \hat{E}(Y * A)$ and $X_1 = \hat{E}(Y * A * I(R = 2) \mid R) - \hat{E}(Y * A * I(R = 2))$.

The results are presented in figure 2A, where simulations at progressively larger sample sizes have been run and the true ratios (for $R = 1$ and $R = 2$)

Figure 2: Results of simulations and data analysis: A) estimates from single simulations versus sample size using a saturated model. Circles are the estimates of $\psi_{0RR}(R = r, A = 1)$ when $R = 1$ and triangles for $R = 2$. The lines represent the respective true ratios; B) as A) but fitting a misspecified model (estimates are open circles), which assumes $\psi_{0RR}(R = r, A = 1)$ is constant in R; C) the relative mean-squared error (RMSE) of $\psi_{0RR}(R = r, A = 1)$ for both $r = 1$ (solid line) and $r = 2$ (dashed line) versus sample size. RMSE is the MSE for estimating m over that setting m to a constant δ; D) Estimated causal relative risk ratios by R (and 95 percent pointwise confidence intervals) for the association of decaffeinated coffee consumption during pregnancy and miscarriage.



are shown as a horizontal line. As one can see, the estimator is converging to the truth, but has fairly substantial variability even at relatively large sample sizes.

We also want to examine the estimator when the estimating model is misspecified. We use the same simulation but now assume that the ratio, $\psi_{0RR}(R = r, A = 1)$ is constant in R. As opposed to estimating m we set m to a constant ($\delta = 0.5$) and model the ratio of interest as (6), where a model is used that respects that both m and m_0 are probabilities by 1) setting $m(A = 1, R = r) = \delta$ and 2) assuming a logistic model for m_0. In addition, the model also guarantees that $m/m_0 = 1$ when $A = 0$, precisely as described in section 2.1. Specifically, the model for the causal relative risk is:

$$\psi_{0RR}(R = r, A = 1) = \frac{1/\delta}{1 + exp(-(logit(\delta) + \beta * A))} \tag{14}$$

This model has one (non-nuisance) parameter of interest (β) and only defines one ratio for both $R = 1$ and $R = 2$ when $A = 1$. Using the same estimating equation approach outlined above, the estimates (of a single simulation) versus sample size are shown in figure 2B. In this case the estimated causal relative risk is converging to a value close to true causal relative risk when $R = 2$. Our experience from simulations is that the convergence of these misspecified models is not necessarily to an easily interpretable value (say, some weighted average of the two causal relative risks in this case) and so interpretation of the results in these misspecified, lower-dimensional models should be done with caution.

Finally, we compare the relative efficiency in finite samples to two consistent approaches to estimating the causal relative risk: 1) estimating m nonparametrically and using the parameterization (5) and 2) setting $m = \delta$ and using the model (6), similar to the simulation above, but now with a correctly specified model. Again, the same data-generating distribution is used as above and we perform repeated simulations at progressively larger sample sizes (1000 at each sample size). The results are estimated as the relative efficiency defined as the ratio of the estimated mean-squared error (RMSE) of the estimator using a non-parametric estimate of m divided by that using the estimator fixing m at δ. Figure 2C has the results plotted as relative efficiencies versus sample size (the solid line is the RMSE of the ratio when $R = 1$ and the dashed line for $R = 2$). In this simulation, one gains efficiency for both ratios in finite samples by using the approach that sets m at a fixed value δ except at very large sample sizes. This approach does not

work unless one can a priori set a reasonable bound on the causal relative risk (obviously $1/\delta$ is the upper bound on the ratio using this approach). However, in many practical situations, one might have good reason to either expect the ratio is, for instance, ≤ 1 for all R (e.g., A is a risk factor for some disease with no plausible benefit to the subject at any combination of A and R). When one can set a plausible upper bound on the causal relative risk, then at least this one simulation suggests the efficiency gains can be significant by using model (6).

5 Data analysis

To demonstrate the method on existing data, we used data from a published study examining the association of decaffeinated coffee and miscarriage (Fenster et al. (1997)). A significant association was found between women reporting drinking 2 or more cups of decaffeinated coffee when interviewed during the first trimester of pregnancy and the subsequent occurrence of miscarriage. The hypothesis was that this reflected not an actual risk from decaffeinated coffee, but was confounded by nausea. Specifically, women with nausea during early pregnancy have lower rates of miscarriage and these same women tend to drink less decaffeinated coffee for obvious reasons. The data was gathered from a questionnaire given shortly after a positive pregnancy test, which asked about behaviors both during this early pregnancy period and those same behaviors before their last menstrual period (LMP). A potential instrumental variable for decaffeinated coffee consumption during pregnancy is the amount of coffee (de- and caffeinated) the woman reported drinking before their LMP. Theoretically, this should be related to (and is) their future consumption during pregnancy, but should have no independent contribution to the outcome. Their are many potential reasons why this might not be an ideal instrumental variable, particularly given that it is based on recall and it is certainly possible a women will have trouble distinguishing her consumption before and after her LMP. Given those caveats and other potential weaknesses, we use the example as a demonstration of our method and believe at least it is a potential way to reduce the influence of unmeasured confounding.

The observed data is R (ordered categorical total coffee consumption (in cups) before LMP: $0 = 0$, $1 = 1/2 - 1$, $2 = 2 - 3$, and $3 => 3$; A is binary (0 if none, 1 otherwise) and Y is the binary miscarriage outcome (0 no, 1 yes).

We assume an unsaturated, two parameter causal relative risk model of the form:

$$\psi_{0RR}(R = r, A = 1) = \frac{1/\delta}{1 + exp(-(\frac{\delta}{1-\delta} + A(\beta_0 + \beta_1 * rr)))}. \qquad (15)$$

Using bootstrapping to derive the confidence intervals for the estimate coefficient, β and converting back to the estimated causal relative risk, we present the results and 95 percent non-parametric bootstrapped confidence intervals (CI) in figure 2D. The plot shows the suggestion that decaffeinated coffee consumption increases the risk of miscarriage, only among the highest those that consume the highest of coffee prior to their LMP, but the confidence intervals clearly overlap the null ($\psi_{0RR}(R = r, A = 1) = 1$). In fact, the variability is so high for the first two levels of R, there is essentially no information about the causal relative risk for those values, i.e., the CI hits both the minimum (0) and maximum ($1/\delta$) value possible for ψ_{0RR}. As a follow-up, we also test the association using the same model and the estimating equation-based chi-square test discussed in section 3.7, which results in $\chi^2 = 3.2$ (df=2), p-value=0.20. Looking at the naive approach assuming no confounding, results in a Pearson's $\chi^2 = 6.0$ (df=1), p-value=0.01. No obvious conclusions can be made from this contrast, beyond that properly accounting for the unmeasured confounding by this instrumental variable approach gives back inferences more commiserate with the actual knowledge of the data-generating distribution, rather than an approach that assumes no confounding.

6 Discussion

In this article we have provided various new results for estimation of the causal relative risk and a newly defined switch causal relative risk for binary outcomes, based on an instrumental variable assumption. In our general method for obtaining a model for the causal relative risk we pose working models for the two conditional response probabilities m_0 and m, which incorporate the constraint that the response probabilities are equal within strata of untreated sub-populations (i.e., $m_0/m = 1$ at $A = 0$ for all R, V). Our proposed model for the causal relative risk is now defined by the working model for the counterfactual conditional response probability m_0 divided by the asymptotic least-squares fit $m_1(A, R, V)$ of m according to the working

model $\{m(\cdot \mid \alpha) : \alpha\}$. By noting that, for given m_1, this model is a multiplicative structural nested mean model for the causal relative risk, we obtain immediately the class of unbiased estimating functions and corresponding asymptotically linear and locally efficient estimators (Robins (1989, 1994)). Substituting for m_1 the iteratively re-weighted least squares estimator of $E(Y \mid A, R, V)$ according to the possibly misspecified working model, results now in our proposed class of consistent and asymptotically linear estimators of the causal relative risk. An important special case is to set m_1 equal to a known constant (see next paragraph), so that it is not even necessary to fit $m(A, R, V) = E(Y \mid A, R, V)$.

We show that, if the model for the counterfactual response probability m_0 is left unspecified, then the true causal relative risk is always contained in this model, as long as the true causal relative risk is bounded by 1 divided by m_1 (i.e., $m_0/m \leq 1/m_1$). Based on this property of our class of models, given that it is known that the true causal relative risk m_0/m is bounded by $1/\delta$ for some $\delta \in (0, 1)$, we propose to set the working model for the observed conditional response probability m equal to a singleton δ: that is, $m_1 = \delta$. In this case, the model constraints the true causal relative risk to be between 0 and $1/\delta$. One can also decide to make δ an additional parameter in our model for the causal relative risk. Simulations presented in Section 4 suggest that this approach can significantly improve efficiency in finite samples relative to the approach where m is estimated.

Since 1 is always an element of our model for the causal relative risk, our estimator provides an asymptotically valid test of the null hypothesis of no treatment effect, even when our model for the causal relative risk is misspecified. Vansteelandt and Goethebeur (2003) and Robins and Rotnitzky (2004) highlight this as a fundamental and important property of their proposed estimator of the causal odds ratio.

Although our fit or choice m_1 for the observed conditional probability is allowed to be heavily misspecified (without affecting the sensibility of the implied multiplicative structural nested mean model), it would be preferable if such an assumption on a nuisance parameter can be avoided at all. (Just as it is not needed in the case that the causal relative risk is known to be smaller than $1/\delta$ for a known $\delta \in (0, 1)$). This motivates the introduction of the switch causal relative risk and the corresponding estimators. The above modeling strategy for the switch causal relative risk allows now arbitrary miss-specification of m_1, so that one can simply set (e.g.) $m_1 = 0.5$, and thereby avoid fitting m at all.

In this case, our estimating functions are based on a generalized causal quantile-quantile function proposed in Yu and van der Laan (2002). As already noted in Yu and van der Laan (2002), this generalized quantile-quantile function provides us also with a generalization of structural nested models of Robins (e.g., Robins (1997)) to general types of outcomes, including discrete valued outcomes. In particular, it shows that our methods for estimation of the causal quantile-quantile function for binary outcomes presented in this article can be straightforwardly generalized to categorical outcomes. An interesting issue is the irregularity of the regression parameters for such discrete structural nested models at null values, and the practical and theoretical implications are of interest and worth further study. This irregularity and its practical implications have been discussed in Robins (2004) in the context of a general class of structural nested models for modeling and estimation of optimal dynamic treatment regimens.

Acknowledgement.

We thank James Robins for pointing out that the switch causal risk parameter is not path-wise differentiable if there is no treatment effect, and we thank the referees for their help, which has improved the presentation.

APPENDIX

Verification of properties I and II for our models of causal relative risk.

In order to understand if the above strategy of formulating a model $\gamma_{\alpha_1}(\cdot \mid \beta)$ results in a sensible multiplicative structural nested mean model, two desirable properties are investigated.

Property I: Consider the maximal size model $\mathcal{M}_0(m_1) \equiv \{\tilde{m}_0/m_1 : \tilde{m}_0\}$ for the causal relative risk, where $m_1 = m(\cdot \mid \alpha_1)$, and $\tilde{m}_0$ ranges over all $[0,1]$-valued functions satisfying $\tilde{m}_0(V,R,0) = m_1(V,R,0)$ a.e. This model at $m_1 = m(\cdot \mid \alpha_1)$ corresponds with our model for the causal relative risk if we choose a saturated model for $m_0(\cdot \mid \beta, \alpha_1)$. This model for the causal relative risk should contain the true causal relative risk $\psi_{0RR} = m_0/m$.

Solving $\tilde{m}_0/m_1 = \psi_{0RR} = m_0/m$ w.r.t. $\tilde{m}_0$ shows that

$$\tilde{m}_0 = \frac{m_1}{m} m_0,$$

but we need to make sure it maps into $[0, 1]$. Thus, under the assumption that

$$\frac{m_1(V, R, A)}{m(V, R, A)} \leq \frac{1}{m_0(V, R, A)} \text{ with probability 1,} \tag{16}$$

or equivalently,

$$\psi_{0RR}(V, R, A) = \frac{m_0(V, R, A)}{m(V, R, A)} \leq \frac{1}{m_1(V, R, A)} \text{ with probability 1,} \tag{17}$$

the nonparametric model $\mathcal{M}_0(m_1)$ always yields a correctly specified model for ψ_{0RR}. Note that assumption (16) states that one can misspecify the true observed conditional response probability $m(V, R, A)$ by a factor $1/m_0(V, R, A)$, or equivalently, this assumption holds whenever it is known that the true causal risk m_0/m is bounded by $1/m_1$.

Property II: Let $\tilde{\psi}_{0RR} = m_0(\cdot \mid \alpha_1, \tilde{\beta}_0)/m(\cdot \mid \alpha_1)$ be an approximation of the true ψ_{0RR}, representing the limit of our estimator of ψ_{0RR} under our possibly misspecified model for ψ_{0RR}. One would like to have that, even when the model for ψ_{0RR} is misspecified, it still respects that

$$m * \tilde{\psi}_{0RR} \leq 1. \tag{18}$$

We have

$$m * \tilde{\psi}_{0RR} = m_0 \frac{\tilde{\psi}_{0RR}}{\psi_{0RR}}.$$

Thus, (18) holds if and only if

$$\frac{\tilde{\psi}_{0RR}}{\psi_{0RR}} \leq \frac{1}{m_0} \text{ with probability 1.} \tag{19}$$

Thus, even when our model for the causal relative risk missspecifies the true causal relative risk by a factor $1/m_0$, its asymptotic fit still respects that it represents a ratio of probabilities.

Note that (19) puts no constraints on the level of missspecification of $m(\cdot \mid \alpha_1)$ as an approximation of m. Therefore, if Property I holds under (17) and it is known that $\psi_{0RR} \leq \frac{1}{\delta}$ for some $\delta \in (0, 1)$, then one can simply set $m(\cdot \mid \alpha_1) = \delta$. This corresponds with the multiplicative structural nested mean model (6).

Verification of properties I and II for models of switch causal relative risk.

Verification of Property I for model (7) for Switch Causal Relative Risk: In order to understand if the model $\gamma_{\alpha_1}(\cdot \mid \beta)$ is a sensible model, one must first verify if a flexible model $\mathcal{M}_0(\alpha_1)$ for $\beta \to m_0(\cdot \mid \alpha_1, \beta)$ yields a correctly specified model for $\psi_{0SRR} = I_{\mathcal{A}_0} m_0/m + I_{\mathcal{A}_0^c}(1-m_0)/(1-m)$, where $\mathcal{A}_0 = \{(V,R,A) : m_0/m(V,R,A) \leq 1\}$ (ie., property I).

Solving $m_0(\cdot \mid \alpha_1, \beta_0)/m(\cdot \mid \alpha_1) = m_0/m$ on $\mathcal{A}_0$ w.r.t. β_0 demonstrates that

$$m_0(\cdot \mid \alpha_1, \beta_0) = \frac{m(\cdot \mid \alpha_1)}{m} m_0 \text{ on } \mathcal{A}_0.$$

Similarly, solving $(1 - m_0(\cdot \mid \alpha_1, \beta_0))/(1 - m(\cdot \mid \alpha_1)) = (1-m_0)/(1-m)$ on $\mathcal{A}_0^c$ w.r.t. β_0 shows that

$$m_0(\cdot \mid \alpha_1, \beta_0) = \frac{1 - m(\cdot \mid \alpha_1)}{1-m}(1-m_0) \text{ on } \mathcal{A}_0^c.$$

Thus,

$$m_0(\cdot \mid \alpha_1, \beta_0) = I_{\mathcal{A}_0} \frac{m(\cdot \mid \alpha_1)}{m} m_0 + I_{\mathcal{A}_0^c} \frac{1 - m(\cdot \mid \alpha_1)}{1-m}(1-m_0).$$

Now, note that for any α_1 (i.e., whatever the level of missspecification of m is), the right-hand side is bounded by 1; use $m_0/m \leq 1$ on $\mathcal{A}_0$ and $(1 - m_0)/(1-m) \leq 1$ on $\mathcal{A}_0^c$. This proves that by choosing a saturated model $\{m_0(\cdot \mid \alpha_1, \beta) : \beta\}$ our model for the switch causal relative risk will be correctly specified, at any $m(\cdot \mid \alpha_1)$.

Verification of Property II: It also follows that, at any $m(\cdot \mid \alpha_1)$, we have $m * \gamma^1_{\alpha_1} \leq 1$ on $\mathcal{A}(\alpha_1, \beta)$, and $(1-m) * \gamma^1_{\alpha_1} \leq 1$ on its complement $\mathcal{A}(\alpha_1, \beta)$.

To conclude, the model for the switch causal relative risk satisfies the wished two properties I and II at any $m(\cdot \mid \alpha_1)$.

In terms of our general parametrization (5) for $m_0(\cdot \mid \beta, \alpha) = 1/(1 + \exp(-f_0(\cdot \mid \beta) + C(\cdot \mid \alpha)))$, and $m(\cdot \mid \alpha) = 1/(1 + \exp(-f(\cdot \mid \alpha) + C(\cdot \mid \alpha)))$, with $f_0(V, R, 0 \mid \beta) = f(V, R, 0 \mid \alpha, \beta) = 0$ everywhere, we have that

$$\begin{aligned} f_0(\cdot \mid \beta_0) + C(\cdot \mid \alpha_1) &= I_{A_0} \log \left(\frac{m(\cdot \mid \alpha_1) m_0/m}{1 - m(\cdot \mid \alpha_1) m_0/m} \right) \\ &\quad + I^c_{A_0} \log \left(\frac{(1 - m(\cdot \mid \alpha_1))(1-m_0)/(1-m)}{1 - (1 - m(\cdot \mid \alpha_1))(1-m_0)/(1-m)} \right). \end{aligned}$$

Note that indeed, $f_0(V, R, 0 \mid \beta_0) = 0$ as required by noting $m_0(V, R, 0)/m(V, R, 0) = 1$ and $p \to \log(p/(1-p))$ is the inverse of $x \to 1/(1+\exp(-x))$.

Verification of Property I and II for proposed models for additive risk.

Using the same general modeling strategy as in Section 2, one could assume the following additive structural nested mean model for the additive risk ψ_{0AR}:

$$\psi_{ARR} \in \{\gamma_{\alpha_1}(V, R, A \mid \beta) \equiv m_0(V, R, A \mid \alpha_1, \beta) - m(V, R, A \mid \alpha_1) : \beta\}. \tag{20}$$

However, in this case verification of Properties I and II puts serious restrictions on the allowed level of misspecification of $m(\cdot \mid \alpha_1)$.

Verification of Property I: Let β_0 be the true parameter value. Solving $m_0(\cdot \mid \alpha_1, \beta_0)) - m(\cdot \mid \alpha_1) = m_0 - m$ w.r.t. the true parameter β_0 yields

$$m_0(\cdot \mid \alpha_1, \beta_0) = m(\cdot \mid \alpha_1) - m + m_0.$$

Thus, under the assumption that

$$-m_0(V, R, A) \leq m(V, R, A \mid \alpha_1) - m(V, R, A) \leq 1 - m_0 \text{ with probability 1}, \tag{21}$$

a nonparametric model $\mathcal{M}_0(\alpha_1)$ for the causal additive risk always yields a correctly specified model for ψ_{0AR}. In this case, both small values of m_0 as well as small values of $1 - m_0$ only allow minor levels of misspecification of the working model for m.

Therefore, we feel that this assumption (21) needs to be seriously considered before applying the estimators of the causal additive risk. Similarly, it follows that the condition $m + \gamma_{\alpha_1}(\cdot \mid, \beta) \in [0, 1]$ for a fit $\gamma_{\alpha_1}(\cdot \mid \beta)$ does not easily hold for misspecified $m(\cdot \mid \alpha_1)$: this is the analogue of Property II. Consequently, for estimating the causal additive risk we recommend a sincere attempt at estimating the true m in order to establish the wished sensibility of the corresponding model (20).

Identification and estimation of marginal additive causal risk.

It is of interest to note that for a subset $V_1 \subset V$ of the baseline covariates, we have

$$\theta_0(V_1) \equiv P(Y_0 = 1 \mid V_1) - P(Y = 1 \mid V_1) = E(\psi_{0AR}(V, R, A) \mid V_1).$$

Thus, identification of the additive causal risk ψ_{0AR} does imply identification of a causal effect of setting $A = 0$ (relative to the population mean) within strata $V_1 = v_1$. Given a model $\{\theta(\cdot \mid \beta) : \beta\}$ for this marginal additive risk $\theta_0(\cdot) = m(\cdot \mid \beta_0)$, one could estimate the unknown β_0 by regressing an (possibly highly nonparametric) estimator ψ_n of ψ_{0AR} on V_1:

$$\beta_n = \arg\min_{\beta} \sum_i (\psi_n(V_i, R_i, A_i) - m(V_{1i} \mid \beta))^2.$$

Generalized quantile-quantile function for general discrete distributions

For the interested reader we provide here also the formula for the generalized quantile-quantile function for general discrete distributions, which provides us with a structural nested model for categorical outcomes, using (e.g.) multinomial logistic models. In that case F_1 and F_2 play the role of $F_{Y_0|V,R,A}$ and $F_{Y|V,R,A}$, respectively, and they would be modelled with multinomial logistic regression models satisfying the constraint that they are equal at $A = 0$.

Result 2 *Let X_1, X_2 be discrete random variables on ordered outcomes $\{x_0, \ldots, x_K\}$ with corresponding probabilities $p_1(x_j)$, $p_2(x_j)$, $j = 0, \ldots, K$. Let $F_1(x) = \sum_{j=0}^{K} I(x_j \leq x) p_1(x_j)$, and $F_2(x) = \sum_{j=0}^{K} I(x_j \leq x) p_2(x_j)$ be the two cumulative distribution functions of X_1 and X_2, respectively. For notational convenience, we define $F_1(x_{-1}) = F_2(x_{-1}) = 0$. We have the following formula for the generalized quantile-quantile function*

$$F_1^{-1} F_2^{\Delta}(X_2) = \sum_{j=0}^{K} x_j I_{A_j}(X_2) I_{B_j}(\Delta, X_2),$$

where

$$\begin{aligned} A_j &\equiv \{x_2 : F_1(x_{j-1}) < F_2(x_2) \leq F_1(x_j) + p_2(x_2)\} \\ B_j &\equiv \left\{(\delta, x_2) : \frac{F_1(x_{j-1}) - F_2(x_2-)}{p_2(x_2)} \leq \delta \leq \frac{F_1(x_j) - F_2(x_2-)}{p_2(x_2)}\right\}. \end{aligned}$$

In particular, it follows that

$$E_\Delta(F_1^{-1}F_2^\Delta(X_2) \mid X_2) \;=\; \sum_{j=0}^{K} x_j I_{A_j}(X_2) d_j(X_2),$$

where

$$d_j(X_2) = \min\left(1, \frac{F_1(x_j) - F_2(X_2-)}{p_2(X_2)}\right) - \max\left(0, \frac{F_1(x_{j-1}) - F_2(X_2-)}{p_2(X_2)}\right).$$

Finally, in order to provide the reader with an understanding of the generalized quantile-quantile function, we provide here a direct simple proof for the pure discrete case.

Result 3 *(Special case of result in Yu and van der Laan (2002)) Let F be a discrete distribution function with support $\{x_0, \ldots, x_K\}$, and let $X \sim F$. We have*

$$F^\Delta(X) \sim U(0,1).$$

Consequently, for any cumulative distribution function F_1, we have

$$F_1^{-1}F^\Delta(X) \sim F_1.$$

Proof. We have $F^\Delta(X) = F(X-) + (1-\Delta)p(X)$, where we define $p(x) = F(x) - F(x-)$. Let $x_0 \in (0,1)$ and let $t(x_0) \in \{x_0, \ldots, x_K\}$ be the unique point for which $F(t(x_0)-) \leq x_0$ and $F(t(x_0)) > x_0$. Now,

$$\begin{aligned}
Pr(F^\Delta(X) \leq x_0) &= Pr\left(\Delta \leq \frac{x_0 - F(X-)}{p(X)}\right) \\
&= E\left\{I(0 \leq \frac{x_0 - F(X-)}{p(X)} < 1)\frac{x_0 - F(X-)}{p(X)} + I(F(X) \leq x_0)\right\} \\
&= E\left\{I(F(X) > x_0, F(X-) \leq x_0)\frac{x_0 - F(X-)}{p(X)} + I(F(X) \leq x_0)\right\} \\
&= \sum_{j=0}^{K} I(F(x_j) > x_0, F(x_j-) \leq x_0)(x_0 - F(x_j-)) \\
&\quad + \sum_{j=0}^{K} I(F(x_j) \leq x_0)p(x_j) \\
&= (x_0 - F(t(x_0)-)) + F(t(x_0)-) \\
&= x_0. \square
\end{aligned}$$

The relation between switch causal relative risk and the binary quantile-quantile function.

The key idea behind the switch causal relative risk and its estimators is the generalized (to discrete outcomes) quantile-quantile function, as proposed in Yu and van der Laan (2002). Their result states that, given two cumulative distribution functions F_1 and F_2, we have

$$X_1 \equiv F_1^{-1} F_2^{\Delta}(X_2) \sim F_1,$$

where $X_2 \sim F_2$,

$$F_2^{\Delta}(X_2) \equiv \Delta F_2(X_2) + (1-\Delta) F_2(X_2-), \tag{22}$$

$F_2(x-) \equiv P(X_2 < x)$, Δ is an external standard uniformly distributed random variable (i.e., $\Delta \sim U(0,1)$), and $F_1^{-1}(x) \equiv \inf\{y : F_1(y) \geq x\}$. Here F_1 and F_2 are allowed to be any cumulative distribution function, which thus includes the case that they are stepwise constant cumulative distributions (corresponding with discrete random variables), or, more general, that they have discontinuity points.

This result is proved by showing that for any cumulative distribution function F_2, we have $F_2^{\Delta}(X_2) \sim U(0,1)$. We also note that if F_2 is continuous, then $F_2^{\Delta}(X_2) = F_2(X_2)$ with probability 1, which shows that this quantile-quantile function indeed generalizes the quantile-quantile function for continuous random variables. The proof of this result is presented in Yu and van der Laan (2002). In the same spirit as in the structural nested models of Robins (see, e.g., Robins (1997)), this motivates us to define the quantile-quantile function

$$H_0(V,R,A,Y,\Delta) = F^{-1}_{Y_{0R}|V,R,A} F^{\Delta}_{Y|V,R,A}(Y), \tag{23}$$

where $F_{Y_{0R}|V,R,A}$ and $F_{Y|V,R,A}$ denote the cumulative distribution functions of the binary random variables $Y_{0R} \sim Bernoulli(m_0(V,R,A \mid \alpha,\beta))$ and $Y \sim Bernoulli(m(V,R,A \mid \alpha))$, conditional on V,R,A. We remind the reader that structural nested models as introduced by Robins model the quantile-quantile function of $F^{-1}_{Y_{0R}|V,R,A} F_{Y|V,R,A}$ for *continuous* outcomes Y (see also van der Laan and Robins (2002), chapter 6, for a detailed description and references).

In the case that X_1, X_2 are both binary random variables, the generalized quantile-quantile function has a simple explicit form provided in the next result.

Result 4 *Consider two binary random variables $X_j \in \{0,1\}$ with $P(X_j = 1) = p_j$, $j = 1,2$. Let F_j denote the cumulative distribution function of X_j: $F_j(x) = I(x \leq 0)(1-p_j) + I(x \leq 1)$, and $F_j^{-1}(u) = I((1-p_j) < u)$, $j = 1,2$. Then,*

$$F_1^{-1}F_2^{\Delta}(x) = I\left(1 - p_1 < I(x=0)\Delta(1-p_2) + I(x=1)(1-p_2+p_2\Delta)\right).$$

This Result 4 provides us with a closed form expression for $H_0(\Delta) = H_0(Y,R,V,A,\Delta)$. Application of the Result 4 with $F_1 = F_{Y_0R|V,R,A}$, $F_2 = F_{Y|V,R,A}$, $p_1 = m_0(V,R,A)$, $p_2 = m(V,R,A)$, and $x = Y$, results in:

$$\begin{aligned} H_0(\Delta) &= H_0(Y,R,V,A,\Delta) \\ &= I\left((1-m_0) < I(Y=0)\Delta(1-m) + I(Y=1)(1-m+\Delta m)\right) \\ &= I\left(\Delta > \frac{(1-m_0(V,R,A)) - I(Y=1)(1-m(V,R,A))}{I(Y=0)(1-m(V,R,A)) + I(Y=1)m(V,R,A)}\right) \end{aligned} \quad (24)$$

where we used short-hand notation at the second equality.

The unbiasedness of our estimating functions for the quantile-quantile function only relies on $E(H_0(\Delta) \mid V,R,A) = m_0(V,R,A)$. Since $E(H_0(\Delta) \mid V,R,A) = E(E_{\Delta}(H_0(\Delta) \mid Y,V,R,A) \mid V,R,A)$, it suffices to work with the expectation of $H_0(\Delta)$, given Y,V,R,A. We have

$$\begin{aligned} H_0(Y,V,R,A) &\equiv E_{\Delta}(H_0(\Delta) \mid Y,V,R,A) \\ &= I_{\{m_0(V,R,A)/m(V,R,A) \leq 1\}} I(Y=1)\frac{m_0(V,R,A)}{m(V,R,A)} \\ &\quad + I_{\{m_0/m(V,R,A) > 1\}}\left(1 - I(Y=0)\frac{1-m_0(V,R,A)}{1-m(V,R,A)}\right) \end{aligned} \quad (25)$$

Now, note that the switch causal relative risk ψ_{0SRR} and H_0 are equivalent parameters in the sense that H_0 is a function of ψ_{0SRR} and ψ_{0SRR} is a function of H_0. Thus, if we define a generalized structural nested model as a model for the generalized quantile-quantile function, $E_{\Delta}F^{-1}_{Y_0|V,R,A}F^{\Delta}_{Y|V,R,A}$, then our model on the switch causal relative risk is a generalized structural nested model.

The key to construction of unbiased estimating functions for switch causal relative risk.

It is easy to show that $E(H_0(Y,V,R,A) \mid V,R,A) = m_0(V,R,A)$. Since this is fundamental to the unbiasedness of our estimating functions $(h(R,V) -$

$E(h(R,V) \mid V))H_0(Y,V,R,A)$ presented in the next section, we will state this as a formal result.

Result 5 *We have* $E(H_0(Y,V,R,A) \mid V,R,A) = E(Y_0 \mid V,R,A)$.

Proof. Consider the expression (25) for $H_0(Y,V,R,A)$. First condition on V,R,A, and note that $E(I(Y=1) \mid V,R,A)) = m(V,R,A)$ and $E(I(Y=0) \mid V,R,A) = 1 - m(V,R,A)$ so that we obtain $I_{m_0/m\leq 1}m_0 + I_{m_0/m\geq 1}m_0 = m_0$. □

Local efficiency.

Our models for the causal relative risks and causal additive risk are just the structural nested mean models of Robins (1989,1994). He provides the efficient choice (h_{opt}, q_{opt}) in these models. In Section 3 we already specified the optimal choice q_{opt}, which follows by a simple projection argument. In our next result we provide the optimal choice in our class of estimating functions for general H_0, which is in agreement with Robins results, but also provides us with the optimal choice of estimating function for the switch causal relative risk.

Result 6 *Let* $\Sigma(h,q) \equiv COV(IC_{h,q,\alpha_1}(O))$ *be the covariance matrix for our estimator implied by the choice* h,q. *We define*

$$\begin{aligned}
\epsilon(\beta_0) &\equiv H_0(O \mid \alpha_1,\beta_0) - E_0(H_0(O \mid \alpha_1,\beta_0) \mid V) \\
\epsilon'(\beta_0 \mid R,V) &\equiv \frac{d}{d\beta}E_0(\epsilon(\beta) \mid R,V)\Big|_{\beta=\beta_0} \\
\sigma^2(R,V) &\equiv E_0(\epsilon^2(\beta_0) \mid R,V).
\end{aligned}$$

Let

$$\begin{aligned}
q_{opt}(V) &\equiv E_0(H_0(O \mid \alpha_1,\beta_0) \mid V) \\
h_{opt}(R,V) &= \frac{1}{\sigma^2(R,V)}\left\{\epsilon'(\beta_0 \mid R,V) - \frac{\int \frac{\epsilon'(\beta_0|r,V)}{\sigma^2(r,V)}dP_0(r \mid V)}{\int \frac{1}{\sigma^2(r,V)}dP_0(r \mid V)}\right\}.
\end{aligned}$$

(Note that $E_0(h_{opt}(R,V) \mid V) = 0$.*) For any vector* c *we have that*

$$c^\top \Sigma(h_{opt},q_{opt})c \leq c^\top \Sigma(h,q)c$$

for all possible choices $h(R,V)$ *and* $q(V)$.

Proof. Given q_{opt}, the optimal choice h_{opt} can be determined as a straightforward application of theorem 2.9, page 159, in van der Laan and Robins (2002)). Specifically, consider estimating functions of the form $h_0(R,V)\epsilon(\beta_0)$, where $h_0(R,V) = h(R,V) - E(h(R,V) \mid V)$. Note $d/d\beta E(h_0(R,V)\epsilon(\beta))|_{\beta=\beta_0} = \langle h, \epsilon'(\beta_0)\rangle_H$, where $\langle g_1, g_2\rangle_H = E_{F_{R,V}} g_1(R,V)g_2(R,V)$ is an inner product in the Hilbert space $H \equiv L^2(F_{R,V})$. Let $\tilde{A} : H \to L_0^2(P_0)$ be the Hilbert space operator defined by $\tilde{A}(h) = h_0(R,V)\epsilon(\beta_0)$. Its adjoint $\tilde{A}^\top : L_0^2(P_0) \to H$ is given by $\tilde{A}^\top g = E(\epsilon(\beta_0)g \mid R,V) - E(\epsilon(\beta_0)g \mid V)$. Thus, $\tilde{A}^\top\tilde{A}(h) = h_0(R,V)\sigma^2(R,V) - E(h_0(R,V)\sigma^2(R,V) \mid V)$. By Theorem 2.9 in van der Laan and Robins (2002), the optimal solution h_{opt} is characterized as the solution of $\tilde{A}^\top\tilde{A}(h) = \epsilon'(\beta_0)$. This equation has the explicit solution provided in the result. To see this, one first rewrites the equation as

$$h_0(R,V) = \frac{1}{\sigma^2(R,V)}\left(\epsilon'(\beta_0 \mid R,V) + E(h_0(R,V)\sigma^2(R,V) \mid V)\right).$$

Subsequently, take (on both sides of the equation) the conditional expectation w.r.t. R, given V. Since the conditional expectation on the left equals zero, this immediately yields the closed form solution for $E(h_0(R,V)\sigma^2(R,V) \mid V)$, and thereby of the complete solution h_{opt}. □

Our model for the observed data can be reformulated as $\{P : E_P(H_0(O \mid \beta(P)) \mid R,V) = E_P(H_0(O \mid \beta(P)) \mid V)\}$. That is, our model can be viewed as a semi-parametric regression model $H_0(O \mid \beta) = g(V) + \epsilon$, where g is arbitrary and $E(\epsilon \mid R,V) = 0$. Completely analogue as in Robins (1989,1994), it now follows that the class $\{h(R,V) - E_0(h(R,V) \mid V)\}(H_0(O \mid \beta_0) - E_{P_0}(H_0(O \mid \beta_0) \mid V)\}$ contains the efficient influence function at P_0. This proves that $D_{h_{opt},q_{opt}}(O \mid \beta_0, \eta_0)$ actually equals the efficient influence function, and that $\Sigma(h_{opt}, q_{opt})$ equals the covariance matrix of the efficient influence function.

Estimation of optimal index of estimating function.

In order to estimate h_{opt} and q_{opt}, one will first need an initial estimator β_{n0} of β, which can be based on a simple choice (possibly data dependent choice) (h,q). Given this estimator β_{n0}, one can estimate q_{opt} by regressing $H_0(O \mid \alpha_n, \beta_{n0})$ on V according to a working model. This results now in an estimate $\epsilon(\beta_{n0})$. Regressing $\epsilon(\beta_{n0})^2$ on R,V according to a working model results in an estimator of $\sigma^2(R,V)$. Finally, by regressing $d/d\beta H_0(O \mid \alpha_n, \beta)|_{\beta=\beta_{n0}}$ on R,V and V, one obtains an estimator of $\epsilon'(\beta_0 \mid R,V)$. These estimators provide us now with an estimator h_n of h_{opt} and q_n of q_{opt}. The estimator

$\beta_n(h_n, q_n, \eta_n, \alpha_1)$ is locally efficient in the sense that it is always consistent and asymptotically linear, and, if the guessed working models used to estimate h_{opt} and q_{opt} happen to be correct, then $\beta_n(h_n, q_n, \eta_n, \alpha_1)$ is asymptotically efficient.

References

A. Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113:231–263, 2003.

J. Angrist, G. Imbens, and D.B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–555, 1996.

A.A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In R. L. de Mantaras and D. Poole, editors, *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 46–54, 1994.

A.A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *JASA*, 92:1172–6, 1997.

L. Fenster, A. Hubbard, S.H. Swan, G.C. Windham, K. Waller, R.A. Hiatt, and N. Benowitz. Caffeinated beverages, decaffeinated coffee and spontaneous abortion. *Epidemiology*, 8:515–523, 1997.

K. Hirano, G. Imbens, D.B. Rubin, and X-H. Zhou. Assessing the effects of influenza vaccine in an encouragement design. *Biostatistics*, 1:69–88, 2000.

S.C. Johnston, T. Henneman, C.E. McCulloch, and M.J. van der Laan. Modeling treatment effects on binary outcomes with grouped-treatment variables and inividual covariates. *American Journal of Epidemiology*, 156: 753–760, 2002.

D. Kay, J.M. Fleisher, R.L. Salmon, F. Jones, M.D. Wyer, A.F. Godfree, Z. Zelenauch-Jacquotte, and R.Shore. Predicting the likelihood of gastroenteritis from sea bathing: results from a randomised exposure. *Lancet*, 344:905–909, 1994.

J. Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5:465–480, 1990.

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

J.M. Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach in causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley, editors, *Health Service Methodology: A Focus on AIDS*, pages 113–159. U.S. Public Health Service, National Center for Health SErvices Research, Washington D.C., 1989.

J.M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics*, 23:2379–2412, 1994.

J.M. Robins. Causal inference from complex longitudinal data. In M. Berkane, editor, *Latent Variable Modeling and Applications to Causality*, pages 69–117. Springer Verlag, New York, 1997.

J.M. Robins. Optimal structural nested models for optimal sequential decisions. In D.Y. Lin and P.J. Haegerty, editors, *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer Verlag, New York, 2004.

J.M. Robins and A. Rotnitzky. Estimation of treatment effects in randomized trials with noncompliance and a dichotomous outcome using structural nested mean models. *Biometrika*, 91(4):763–83, 2004.

D.B. Rubin. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6:34–58, 1978.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2002.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

A.W. van der Vaart. On differentiable functionals. *Annals of Statistics*, 19: 178–204, 1991.

S. Vansteelandt and E. Goethebeur. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society, Series B*, 65: 817–35, 2003.

Z. Yu and M.J. van der Laan. Construction of counterfactuals and the g-computation formula. Technical report, Division of Biostatistics, UC Berkeley, 2002.