Randomization Based Instrumental Variables Methods for Binary Outcomes with an Application to the IMPROVE Trial*

Luke Keele† Dylan Small‡ Richard Grieve§

August 28, 2015

Abstract

In randomized controlled trials with nonadherence, instrumental variable methods are frequently used to report the complier average causal effect. With binary outcomes, many of the available IV estimation methods impose distributional assumptions. We develop a randomization-inference based method of IV estimation for binary outcomes. The method is nonparametric, based on Fisher's exact test, and estimates can be easily calculated from a set of 2×2 or $2 \times 2 \times 2$ tables. While we retain the standard IV identification assumptions for confidence regions and point estimates, the IV estimand under randomization inference is sample-specific, and does not assume the RCT participants are a random sample from the target population. We illustrate the method with the IMPROVE trial that compares emergency endovascular versus open surgical repair for patients with ruptured aortic aneurysms.

*We thank Paul Rosenbaum and Paul Clarke for comments and discussion.

†Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16802 Email: ljk20@psu.edu, corresponding author.

‡Professor, Department of Statistics, 400 Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104. E-mail: dsmall@wharton.upenn.edu

§Professor of Health Economics Methodology, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, Email: richard.grieve@lshtm.ac.uk

1 Introduction

Well-conducted Randomized Controlled Trials (RCTs), can provide estimates of causal effects for treatments, but a common challenge is that participants depart from their assigned treatment. For example, under "two-way" noncompliance, some RCT participants assigned to the treatment receive the control, and vice-versa. A general concern, is that departures from the randomized treatment are non-ignorable; they reflect prognostic factors, only some of which are observed. Instrumental variable (IV) approaches can provide estimates of the causal effect of the treatment received, the complier average causal effect (CACE) (Angrist et al. 1996). See Baiocchi et al. (2014) for a recent review of IV and its assumptions.

In this paper, we develop a randomization inference approach to IV methods with unpaired binary outcomes. Fisher (1935) advocated randomization as the "reasoned basis for inference" in experiments because it only assumes that the randomization was conducted properly and does not require any additional assumptions. The randomization inference test of no effect can be inverted to provide distribution-free confidence intervals, and the Hodges-Lehmann method produces point estimates, see Rosenbaum (2002, ch .2) for details. Recent work demonstrates that randomization inference methods for binary outcomes, which we consider here, attain the narrowest width of confidence interval while maintaining nominal coverage (Rigdon and Hudgens 2015). Rosenbaum (1996) first derived a randomization inference approach to IV for continuous outcomes.

IV methods in the randomization inference framework have several attractive properties. First, the hypothesis test that treatment received has a zero effect on the outcome is the same for the intention-to-treat (ITT) estimand and the IV estimand, as both test the same null hypothesis. As such, the test that the IV estimand is zero can made with the same weaker assumptions as for the ITT estimand. Second, Imbens and Rosenbaum (2005) prove that randomization inference provides correct coverage for confidence intervals when instruments are weak, unlike the more standard two-stage least squares (2SLS) estimator. Moreover, they also show that when the

outcome distribution is heavy-tailed, randomization inference IV methods have higher power than 2SLS estimators. Third, randomization inference provides point estimates and confidence regions that only pertain to the units included in the RCT. This approach recognizes that the RCT design only justifies causal inferences for the RCT participants unless some additional sampling mechanism is present. Hence, unlike most parametric and semiparametric approaches which assume that the RCT participants are a random sample from the target population, this approach makes explicit that further assumptions will be required to generalize the IV estimand to other populations.

While randomization inference based IV methods are well developed for continuous outcomes, there has been comparatively little work for binary outcomes. Yang et al. (2014) outline an IV method for paired binary outcomes, but a randomization inference based IV method for unpaired binary outcomes has not yet been developed. Moreover, the method developed by Yang et al. (2014) does not fully account for uncertainty and thus is not exact. We use Rosenbaum's (2001) concept of attributable effects to develop IV tests and estimates for unpaired binary outcomes that are nonparametric and exact. The method, which is based on Fisher's exact test, is easy to interpret and to implement; the required statistics can all be calculated from a set of contingency tables. While we develop the approach in a setting where both outcomes and non-compliance are binary, the approach has general application; it is invariant to both the distribution of the outcome and the measure of compliance. As the method is exact, it provides valid inferences even in small samples, and will provide valid confidence statements even when the instrument is weak.

The nonparametric method we develop in this paper contrasts with many available IV methods for binary outcomes which require parametric assumptions. One approach uses maximum likelihood methods, which tend to be highly sensitive to violations of the distributional assumptions and difficult to estimate (Copas 1988; Freedman and Sekhon 2010; Little 1985). Another popular method is based on "plug-in" estimators that plug-in quantities estimated from first stage models into a second stage model to estimate IV models with binary outcomes (Terza et al. 2008; Cai

et al. 2011). Plug-in approaches rely on distributional assumptions (Clarke and Windmeijer 2012), and analytic and simulation evidence has shown that plug-in methods tend to have higher levels of bias compared to semiparametric methods (Cai et al. 2011, 2012; Vansteelandt et al. 2011). Several semiparametric methods have been developed as alternatives to the plug-in approach (Tan 2006, 2010; Vansteelandt and Goetghebeur 2003; Abadie 2003). However, identification for one semiparametric method, the double logistic SMM, depends on a no effect modification assumption (Clarke and Windmeijer 2010). This assumption is implausible in RCTs with two-sided noncompliance, exemplified by the example we consider, the IMPROVE trial (investigators 2014). Recent work has shown that semiparametric methods are not immune to the problems of weak instruments (Burgess et al. 2014).

The article is organized as follows. Section 2 reviews the IMPROVE trial which forms the empirical application that motivates our method. In Section 3 we review randomization inference methods for unpaired binary outcomes. Here, we also extend the IV method for paired binary outcomes proposed by Yang et al. (2014) to unpaired data. Next, we outline the proposed method in Section 4. We demonstrate the use of our method with the IMPROVE trial data in Section 5. Section 6 concludes.

2 Motivating Example: IMPROVE Trial

The IMPROVE trial assessed the effectiveness of emergency endovascular (EVAR) versus open surgical repair strategies for patients with a clinical diagnosis of ruptured aortic aneurism (investigators 2014). The trial was a multi-center RCT with a parallel design, whereby individual patients were randomized to either the EVAR strategy or Open repair. An independent contractor provided telephone randomization, with computer generated assignment of patients in a 1:1 ratio, using variable block size and stratified by centre. Date and time of randomization together with type of initial consent (written/verbal/other) were recorded automatically. As this was a surgical trial, neither investigators nor patients could be masked to the treatment allocation. Adherence to the allocated treatment group was reinforced wherever possible by onsite training and newsletters

(investigators 2013).

The primary outcome was binary—whether or not a patient was alive at 30 days post-randomization. The primary estimand was intention–to–treat (ITT); and all 613 randomized patients recruited to 30 clinical centers (29 UK, 1 Canada) were included in this analysis. The study also pre-specified subgroup analyses according to gender, age and a baseline measure of prognosis (Hardman Index). The main findings from the ITT analyses were that, overall the 30 day mortality was similar between the randomized arms (EVAR strategy 35.4%; Open strategy 37.4%), but women were more likely than men to benefit from the EVAR strategy; with corresponding odds ratios of 1.18 (95% CI 0.80 to 1.75) (men) and 0.44 (95% CI 0.22 to 0.91) (women). As there was substantial non-compliance in both randomized arms, there is also interest in the IV estimand. As in a previous causal analysis (investigators 2014), we limited the RCT sample to those patients who had a clinical diagnosis of a ruptured aortic aneurism, who actually received either EVAR or Open Repair (n=501, 259 randomized to EVAR and 242 to open repair).

The randomization inference approach to IV estimation requires the standard identification assumptions outlined by Angrist et al. (1996) to hold. Though as we note below, the test of the null hypothesis under randomization inference does relax some of these assumptions. The IV identification assumptions are: (1) ignorable (as-if random) instrument status; (2) the stable unit treatment value assumption (SUTVA); (3) no direct effect of the instrument on the outcome, also known as the exclusion restriction; (4) monotonicity; and (5) the instrument must have a nonzero effect on the treatment.

How realistic are these assumptions in the IMPROVE trial? Assumption (1) is satisfied by design, as accepted randomization procedures were followed in assigning patients treatment arms (investigators 2014). The SUTVA assumption appears plausible since for both randomized arms, all 30 clinical centers followed similar clinical protocols, and there was little evidence of between-center heterogeneity in the care process or health outcomes; i.e. it is unlikely that there were multiple versions of either form of surgery (investigators 2014). Second, randomization was at the level of the individual not the cluster, and relatively few patients were randomized within each

clinical center; it seems plausible that the management of patient A did not affect that of patient B.

In the IMPROVE trial the exclusion restriction also appears plausible. Patients receiving either form of surgery (EVAR or Open repair) had similar management (e.g. delay to surgery) whichever arm they were randomly assigned to. In other RCTs of EVAR versus open repair for elective surgery, the exclusion restriction may well be violated, since amongst those patients who receive EVAR, the delay to operation may be much longer in those who were originally randomized to receive Open repair versus EVAR, and so randomization may have a direct effect on outcome. The monotonicity assumption, which rules out the presence of defiers, also appears reasonable. The main reason that patients switched was due to clinical judgement as to the suitability of patients if receipt of either procedure which would have been invariant to the randomized arms. Generally in such encouragement designs the presence of defiers is regarded as unlikely (Bellamy et al. 2007). Finally, overall, 76% of the IMPROVE trial participants fully complied with their treatment assignment, so the effect of the instrument on assignment is nonzero. Next we provide a basic outline of randomization inference, before providing details for binary outcomes.

3 Randomization Inference for Binary Outcomes: A Review and Extension

Here we review randomization inference for a binary outcome. The appendix contains a more detailed summary of these methods. These methods can be used to construct an estimate of the ITT estimand in a RCT with non-compliance. This is useful since the ITT estimate is necessary to construct the IV estimate.

3.1 Notation

There are n subjects, $i = 1, \ldots, n$, and subject i has two potential outcomes: the outcome Y_{1i} that would be observed if i were assigned to treatment and the outcome, Y_{0i} that would be observed if i were assigned to control (Neyman 1923; Rubin 1974). In the IMPROVE trial,

$Y_{zi} = 1$ if subject i was alive at 30 days post-randomization, and $Y_{zi} = 0$ otherwise. Of the n subjects, m are randomly assigned to receive treatment, and the remaining $n - m$, control. Each of the possible $\binom{n}{m}$ treatment assignments has the same probability $\binom{n}{m}^{-1}$. For each RCT participant i, the indicator $Z_i = z \in \{0, 1\}$, records the randomly assigned treatment status. Combining potential outcomes with the treatment indicator, we can define the observed outcome for each unit i: $Y_i = Z_i Y_{1i} + (1 - Z_i)Y_{0i}$.

The ITT estimand is defined as the counterfactual contrast: $Y_{1i} - Y_{0i}$, however, this contrast cannot be observed, since, the treatment assignment only reveals one of the potential outcomes. Identification of the ITT estimand requires only the first two of the IV identification assumptions (Angrist et al. 1996). First, we assume that potential outcomes are independent of treatment assignment: $Y_{1i}, Y_{0i} \perp\!\!\!\perp Z_i$, which should by design in the IMPROVE trial. Second, we assume that the SUTVA holds (Rubin 1986) which has the two following components: 1) there are no hidden forms of treatment, which implies that for unit i under $Z_i = z$, we assume that $Y_{zi} = Y_i$ and 2) a subject's potential outcome is not affected by other subjects' exposures. The first component of SUTVA is often referred to as the consistency assumption in the epidemiological literature. Under randomization inference, the only stochastic quantity is the treatment assignment (Fisher 1935). The potential outcomes Y_{1i}, Y_{0i} are fixed features of the finite population of n subjects. Randomization creates the distribution used for inference. The observed response, Y_i varies with treatment assignment and thus is not fixed.

3.2 Review: Randomization Inference for the ITT Estimand with a Binary Outcome

Under the sharp null hypothesis, we test whether the treatment effect is zero for all units. In potential outcomes notation, if the sharp null hypothesis holds then $Y_{1i} = Y_{0i}$ for every i. Note that the test of the sharp null only depends on random assignment of the treatment, SUTVA need not hold. To test the sharp null, we define a test statistic as a function of the data. First we denote $\mathbf{Z} = (Z_1 \dots Z_n)^T$ and $\mathbf{Y} = (Y_1 \dots Y_n)^T$. The test statistic is then defined as a function of

the observed outcomes and the treatment assignment, $t(\mathbf{Z}, \mathbf{Y})$. We collect all possible realizations of $\mathbf{Z}$ in the set Ω. Under H_0 the only randomness is from $\mathbf{Z}$, whose distribution is known. The sharp null hypothesis is tested by calculating the observed value of a test-statistic and comparing it to the randomization distribution.

For a binary treatment and outcome, Fisher's exact test is a randomization test (Rosenbaum 2002, ch. 5). The test statistic, $t(\mathbf{Z}, \mathbf{Y})$, is the total number of units in the treatment group with a response equal to 1. Under the null hypothesis of no treatment effect, the randomization distribution of this test statistic follows a hypergeometric distribution. When n is large, the χ^2 test is an approximation to the exact p-value.

At this point, we note that the test of the sharp null for the IV estimand is equivalent to the test of the sharp null for the ITT estimand, because under the IV assumptions, the compliers are the only subjects for whom there could possibly be an effect of being assigned to treatment (Rosenbaum 1996). Therefore, Fisher's exact test is also the test of the sharp null for the IV estimand. Clearly, the test can only find an effect among compliers if there is an ITT effect in the study population.

We now review methods for producing a point estimate and confidence statements for the ITT estimand. Unlike the test of the sharp null, to obtain point estimates and confidence regions for the ITT estimand, we must now assume that SUTVA holds, but the remaining IV identification assumptions are not required. In addition, confidence intervals and point estimates also require an assumption about the nature of response to treatment or a model of effects (Rosenbaum 2002). The most common model of effects is the constant-additive effects. This model of effects has the following form: $Y_{1i} = Y_{0i} + \tau$ for every $i = 1, 2, \cdots, n$. However, when (Y_{1i}, Y_{0i}) are each binary, 1 or 0, say dead or alive, if we adopt a constant-additive model of effects, $Y_{1i} = Y_{0i} + \tau$, then τ must be 1, 0, or -1. If $\tau = -1$ then every subject must die under treatment and survive under control, which is unlikely. For binary outcomes, we adopt a model of effects that assumes the treatment assignment has a nonnegative effect, $Y_{1i} \geq Y_{0i}$ for all i (Rosenbaum 2002, ch .5). Later, we relax this assumption.

Under this model of effects, the effect of treatment assignment on subject i is $\tau_i = Y_{1i} - Y_{0i}$. We collect all treatment effects in the vector $\boldsymbol{\tau} = (Y_{11} - Y_{00}, Y_{12} - Y_{02}, \cdots, Y_{1n} - Y_{0n})$, which is an n-dimensional vector with 1 or 0 values that describes the pattern of responses. The parameter $\boldsymbol{\tau}$ is n-dimensional and thus an unwieldy summary of the treatment effect, therefore, we focus on A, the number of successes, $(Y_i = 1)$, attributable to treatment assignment (Rosenbaum 2001). That is for subjects assigned to treatment, A is the number of events which would not have occurred had those subjects been assigned to control. The quantity A is defined as $A = \sum_i^n Z_i(Y_{1i} - Y_{i0}) = \sum_i^n Z_i\tau_i$, as such it is a random variable, but Rosenbaum (2001) notes that A can serve as a summary for the n-dimensional treatment effect. In the IMPROVE RCT, A is the number of patients alive at 30 days, whose vital status is attributable to assignment to the EVAR arm of the trial. To summarize response to treatment, we wish to produce both a point estimate and confidence region for A.

While Rosenbaum (2002, ch. 5) develops confidence regions for A, he does not consider how to produce a point estimate for A. However, one can easily invoke the method of Hodges-Lehmann (Hodges and Lehmann 1963) to produce a point estimate for A. Yang et al. (2014) use the method of Hodges-Lehmann to produce a point estimate for A with paired binary outcomes. Here, we outline how to apply the Hodges-Lehmann principle to unpaired binary outcomes.

Under the Hodges-Lehmann method, the point estimate for A is the single most plausible value for A which is the value such that in the adjusted table of outcomes, the treatment has no effect on the binary outcome. Hence, we require a metric to summarize the null expectation for the 2×2 table. A simple summary metric for whether the treatment has a zero effect on a binary outcome, is the odds-ratio. Here, the point estimate for A is the value that makes the estimated odds-ratio equal one in the adjusted table (equivalently, one could adjust the table until the calculated risk ratio is one). Since values of A must be integers, we may generally find that a value of A may not make the odds-ratio exactly one. In this situation we can average the values of A around the odds-ratio value of one consistent with the general Hodges-Lehmann principle (Rosenbaum 2002, ch. 2). Alternatively one could adjust the table until the p-value from Fisher's

exact test is maximized. This value, $\hat{A}$, serves as the estimator for A, the number of patients who survived to 30 days attributable to being assigned to EVAR. When noncompliance is present $\hat{A}$ is the estiamtor for the ITT estimand.

Rosenbaum (2002, ch. 5) demonstrated that the corresponding confidence region for A is the set of values of $\boldsymbol{\tau}$ that are not rejected by an α level test, but reporting this set is impractical because it is n-dimensional. However, Rosenbaum (2001) proved that all compatible hypotheses $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ that yield the same attributable effect $A_0 = \sum Z_i \tau_{i0}$, are simultaneously all either included or excluded from a confidence set $\boldsymbol{\tau}$. Therefore we focus on the set for $\boldsymbol{\tau}$ that contains all $\boldsymbol{\tau}_0$ compatible with $A_0 = \sum Z_i \tau_{0i} \geq a$. This is the set of those treatment effects $\boldsymbol{\tau}$ for those assigned to treatment, i.e. those that are *attributable* to treatment assignment. In this way, an n-dimensional confidence set for $\boldsymbol{\tau}$ becomes a 1-dimensional set of plausible values for A. In a RCT with noncompliance, these methods provide estimates and tests for the ITT estimand.

4 Randomization Inference for the IV Estimand with Binary Outcomes

4.1 Notation

First, we define some additional notation. Let D_{zi} be the potential treatment status for subject i if assigned to treatment z, $D_{zi} = 1$ (0) means subject i would (would not) receive treatment if assigned level z of the treatment. We also define the potential outcome $Y_{z,di}$ as the outcome subject i would have if she were assigned level z of the treatment and received level d of the treatment. Z_i, D_i and Y_i are the observed treatment assigned, treatment received and outcome respectively so that $D_i = D_{Z_i}$ and $Y_i = Y_{Z_i,D_i i}$. The finite population IV estimand is $\frac{1}{n_d} \sum_i \{(Y_{1i} - Y_{i0}) \mathbb{1}(D_{1i} = 1, D_{0i} = 0)\}$ where $n_d = \sum_{i=1}^{n} \mathbb{1}(D_{1i} = 1, D_{0i} = 0)$ and $\mathbb{1}[\cdot]$ is the indicator function. Identification of the IV estimand requires the three additional assumptions outlined in Section 2. Formally, we must assume monotonicity: $D_{1i} \geq D_{0i}$ for all $i = 1, \cdots, N$. Second we assume the exclusion restriction that Y_{1,d_i} is exchangeable with Y_{0,d_i}. Finally, we assume that

Z_i has a nonzero causal effect on D_i. Below we re-define the IV estimand in terms of a set of attributable effects that is subject to the same identification assumptions. Table 1 records the observed $2 \times 2 \times 2$ contingency table for an RCT with noncompliance and a binary responses.

Table 1: The observed $2 \times 2 \times 2$ contingency table for binary responses as a function of treatment assignment and receipt.

	Treatment Assignment ($Z_i = 1$)	
Outcome	Treatment Receipt ($D_i = 1$)	Treatment Receipt ($D_i = 0$)
$Y_i = 0$	$\sum Z_i D_i (1 - Y_{1i})$	$\sum Z_i (1 - D_i)(1 - Y_{0i})$
$Y_i = 1$	$\sum Z_i D_i Y_{1i}$	$\sum Z_i (1 - D_i) Y_{0i}$
	Treatment Assignment ($Z_i = 0$)	
Outcome	Treatment Receipt ($D_i = 1$)	Treatment Receipt ($D_i = 0$)
$Y_i = 0$	$\sum (1 - Z_i) D_i (1 - Y_{1i})$	$\sum (1 - Z_i)(1 - D_i)(1 - Y_{0i})$
$Y_i = 1$	$\sum (1 - Z_i) D_i Y_{1i}$	$\sum (1 - Z_i)(1 - D_i) Y_{0i}$

4.2 Extending IV Methods for Paired Binary Outcomes to the Unpaired Case

Yang et al. (2014) develop IV methods within the randomization framework for paired binary data. The method in Yang et al. (2014) can be easily adapted for a unpaired binary outcomes, which we do here. However, this method is not exact. Later, we develop an exact estimator. First, we note that Fisher's exact test remains a valid test of the sharp null even when the focus is on the IV rather than the ITT estimand. We focus on point and interval estimates. For now, we leave the model of effects for a binary outcome, $Y_{1i} \geq Y_{0i}$ for all i, unaltered.

IV estimates are usually constructed from a ratio estimator often referred to as the Wald estimator. See Angrist et al. (1996) for a formal motivation of the Wald estimator. Following Yang et al. (2014), we define a randomization inference IV estimator based on the principle of the Wald estimator. We use $\hat{A}$ as the numerator in the IV estimator. Using D_i, we define the observable random variable $U = \sum_{i=1}^{n} Z_i D_i - (1 - Z_i) D_i$, which is the number of units assigned to treatment that receive the treatment relative to the number of units not assigned to treatment.

Therefore U is an estimate of the number of compliers assigned to treatment when subjects are assigned to treatment or control with equal probabilities.

We define an IV estimator as $\hat{A}/U$, which is the rate of successes for each additional unit that complied with the assigned treatment. In the IMPROVE trial, this quantity is an estimate of the number of patients alive at 30 days for each additional person who actually underwent the EVAR procedure. Confidence regions can be formed for this point estimate using bounds on A and U. This IV estimator is closely related to one developed in Hansen and Bowers (2009) based on randomization inference but using sample theoretic arguments designed to allow for regression adjustment and cluster level treatment assignments.

Strictly speaking, the quantity $\hat{A}/U$ is somewhat different than more standard IV estimators due to the fact that A is an unobservable random variable. However, Yang et al. (2014) show that this quantity converges in probability to the effect among compliers if treatment is randomized. Although the quantity $\hat{A}/U$ is a valid IV estimate, it has one disadvantage. The associated confidence set does not account for the uncertainty in the estimate of U. Failure to account for this estimation uncertainty will result in confidence sets that are too short in finite samples. While the confidence set for A/U will have the correct coverage asymptotically, one reason for using randomization inference is the exact nature of the inferences. Moreover, the methods in Hansen and Bowers (2009) also rely on asymptotic approximations for inference. Next, we develop an exact IV method for binary outcomes.

4.3 An Exact IV Estimator for Binary Outcomes

Next, we generalize attributable effects for binary outcomes to settings with noncompliance. This more general set of attributable effects will produce confidence statements with exact coverage and also allow us to produce an IV estimate on a risk ratio scale. Under our generalization, we define the attributable effect of treatment assignment on both receipt of treatment and the outcome. That is, we define attributable effects for the number of subjects that received the treatment attributable to being assigned to the treatment. In the IMPROVE trial, this

is the number of subjects that received EVAR attributable to being assigned to EVAR. This generalization of the attributable effects will allow us to treat the number of compliers as an unknown quantity and fully account for this uncertainty in the confidence sets.

We define three effects attributable to Z_i on both treatment receipt and survival. The first of these attributable effects is $A_1 = \sum_i^n Z_i(D_{1i} - D_{0i})\mathbb{1}[Y_{1i} = 1, Y_{0i} = 0]$. This is the number of subjects assigned to treatment for whom both the treatment receipt and survival were attributable to being assigned to the treatment. In the IMPROVE trial A_1 is the unobserved additional number of patients that receive the treatment and survive that is attributable to being randomly assigned to the EVAR arm, rather than being assigned to the control arm and not receiving EVAR.

The second attributable effect is $A_2 = \sum_i^n Z_i(D_{1i} - D_{0i})\mathbb{1}[Y_{1i} = Y_{0i} = 1]$. A_2 is the number who receive treatment (in this case EVAR) that is attributable to treatment assignment (in this case the intention to receive EVAR) amongst those who would have survived irrespective of treatment assignment status (whether they were randomized to EVAR versus control). The final attributable effect is $A_3 = \sum_i^n Z_i(D_{i1} - D_{i0})\mathbb{1}[Y_{1i} = Y_{0i} = 0]$, which is the number who receive treatment attributable to treatment assignment among those that would not have survived irrespective of treatment assignment status (those in IMPROVE who would have survived whether they were assigned to EVAR or open repair). While we could make inferences about each attributable effect individually, we are primarily interested in the occurrence of A_1 relative to A_2 and A_3 which are instances where assignment induces patients to be exposed to the treatment.

Next, we define two estimable causal quantities based on these three attributable effects. The first is

$$ACCE = \frac{A_1}{(A_1 + A_2 + A_3)}.$$

This quantity, which we denote as the attributable complier causal effect or $ACCE$, is the proportion of subjects exposed to the treatment among those assigned to the treatment whose outcome was changed to 1. In the IMPROVE trial, the ACCE is the fraction of survivals among those exposed to the treatment attributable to being assigned to the treatment. The next quantity

we define is the attributable complier risk ratio:

$$ACRR = \frac{(A_1 + A_2)}{A_2}.$$

We denote the quantity above as the attributable complier risk ratio or $ACRR$. This quantity recasts the $ACCE$ as the number of survivals and exposures attributable to treatment relative to the number of subjects exposed to the treatment because they were assigned to the treatment.

The chief drawback to the quantities above is that we must maintain a model of effects that rules out negative outcome effects. This assumption may be overly restrictive in many settings including the IMPROVE trial, where it is possible that exposure to the treatment could cause harm. To relax this assumption, we define one additional attributable effect: $A_4 = \sum_i^n Z_i(D_{1i} - D_{0i})\mathbb{1}[Y_{1i} = 0, Y_{0i} = 1]$. This is the number of subjects assigned to treatment who both received treatment and did not survive, and for whom both the treatment receipt and death were attributable to treatment assignment. In the IMPROVE trial, A_4 is the unobserved number of patients that received the EVAR and whose death is attributable to assignment to EVAR rather than open repair. Using A_4 we can redefine the $ACCE$ and $ACRR$ to allow for the possibility of negative treatment effects:

$A_4 = \sum_i^n Z_i(D_{1i} - D_{0i})\mathbb{1}[(Y_{1i} - Y_{0i}) = 0]$

$$ACCE = \frac{A_1 - A_4}{(A_1 + A_2 + A_3 + A_4)}$$

$$ACRR = \frac{(A_1 + A_2)}{A_2 + A_4}.$$

Thus, using A_4 we can relax the assumption of nonnegative effects and both quantities allow the possibility that treatment exposure is harmful. Note that if we relax the assumption of nonnegative effects, the ACCE represents the IV estimand on a risk difference scale. As such, estimates of the ACCE will generally be similar to more standard estimates based on the Wald estimator. However, as we detail next, inferences will be exact.

With these estimable quantities defined, we now outline an estimation and testing strategy. First, we might test the sharp null hypothesis that the ACCE and ACRR are zero. Above we noted that the test of the sharp null for the IV estimate is equivalent to the test of the sharp null for the ITT estimate. Note, however, that for the ACCE, if $A_1 = A_4$ the ACCE will equal zero but Fisher's null may not hold. The formation of point estimates and confidence regions proceeds by testing hypotheses about A_1, A_2, A_3 and A_4 if we wish to relax the assumption of nonnegative effects. To test a specific hypothesis about these attributable effects, we adjust the observed data to make it consistent with that hypothesis. This is possible since we can test the adjusted data against the null distribution (Rosenbaum 2001).

Table 2: Table of potential outcomes and the pattern of adjustment for attributable effects.

	Treatment Assignment ($Z_i = 1$)	
Outcome	Treatment Receipt ($D_{0i} = 1$)	Treatment Receipt ($D_{0i} = 0$)
$Y_{0i} = 0$	$\sum Z_i D_i (1 - Y_i) - A_3(-A_4)$	$\sum Z_i (1 - D_i)(1 - Y_i) + A_1 + A_3$
$Y_{0i} = 1$	$\sum Z_i D_i Y_i - A_1 - A_2$	$\sum Z_i (1 - D_i) Y_i + A_2(+A_4)$
	Treatment Assignment ($Z_i = 0$)	
Outcome	Treatment Receipt ($D_{0i} = 1$)	Treatment Receipt ($D_{0i} = 0$)
$Y_{0i} = 0$	$\sum (1 - Z_i) D_i (1 - Y_i)$	$\sum (1 - Z_i)(1 - D_i)(1 - Y_i)$
$Y_{0i} = 1$	$\sum (1 - Z_i) D_i Y_i$	$\sum (1 - Z_i)(1 - D_i) Y_i$

Note: For each attributable effect, we must subtract it from one cell and add it to another to maintain the table margins. For a set of values for A_1, A_2, A_3, and A_4, we adjust the observed data using following the pattern above and apply Fisher's exact test. A 95% confidence region is obtained for the set of values of A_1, A_2, A_3, and A_4 for which Fisher's exact test yields a p-value equal to or greater than 0.05.

Point estimates and confidence regions are formed by testing a series of hypotheses for a range of values for A_1, A_2, A_3, and A_4. For example, we might first test the hypothesis that $A_1 = 1$, $A_2 = 0$, $A_3 = 0$, and $A_4 = 0$. To test this specific hypotheses, we adjust the observed data in Table 1 using the hypothesized values for the four attributable effects to compute a table the contains the potential outcomes consist with that specific hypothesis. Table 2 contains the pattern of adjustment based on the four attributable effects. For example, to test the hypothesis

above, we subtract A_1 from the lower left cell within the $Z_i = 1$ strata. We must then add it to the upper right cell to maintain the table margins. No other adjustments are needed since the other attributable effects are zero under this hypothesis. If the other values of the attributable effects are nonzero, we must adjust the table accordingly. After, we have adjusted the table, we then apply Fisher's exact test and retain the p-value from this test.

We then iterate over the full range of values for A_1, A_2, A_3, and A_4 testing the entire range of hypotheses for each attributable effect. The range of values is known since these attributable effects cannot be less than zero and cannot exceed their observed values. Each adjustment corresponds to a possible configuration of the potential outcomes under a specific hypothesis based on the possible values of A_1, A_2, A_3, and A_4. Alternatively, if we maintained the assumption of nonnegative effects, we would omit A_4 from this process.

We obtain a confidence region by inverting this series of tests. The 95% confidence region is the set of values of A_1, A_2, A_3, A_4 for which Fisher's exact test yields a p-value greater than or equal to α. For a 95% confidence region this would be the set of values of A_1, A_2, A_3, and A_4 for which Fisher's exact test yields a p-value equal to or greater than 0.05. More precisely, for a fixed α, conventionally $\alpha = 0.05$, it is possible to find an observed set of random variables $\tilde{A}_1$, $\tilde{A}_2$, $\tilde{A}_3$, and $\tilde{A}_4$ so that $A_1 \geq \tilde{A}_1$, $A_2 \geq \tilde{A}_2$, $A_3 \geq \tilde{A}_3$, $A_4 \geq \tilde{A}_4$ holds with 95% confidence such that the unobserved attributable effects A_1, A_2, A_3, A_4 are at least equal to $\tilde{A}_1$, $\tilde{A}_2$, $\tilde{A}_3$, and $\tilde{A}_4$ except in at most 100α% of experiments. See Rosenbaum (2002) and Weiss (1955) for a general discussion of forming confidence sets for unobserved random variables in terms of observed random variables.

Under the principle of Hodges-Lehmann, the point estimates for the ACCE and ACRR will be the values of A_1, A_2, A_3, and A_4 that leaves the adjusted table exactly without treatment effect. In practice, the point estimate corresponds to the set of values for A_1, A_2, A_3, and A_4 that maximize the p-value from Fisher's exact test in the series of tests.

5 Example with the IMPROVE Trial

Next, we illustrate the methods described above using data from the IMPROVE trial. Parallel to the original study, we present results overall and then by gender subgroup (investigators 2014). Table 3 contains three 2×2 contingency tables of treatment assignment and whether or not the patient survived to 30 days post randomization for the entire study population and for the two gender subgroups. Table 4 contains three 2×2 contingency tables of compliance status by treatment assignment status for the same groups. We first report results for the entire study population.

Table 3: Observed contingency table for survival in the IMPROVE Trial

		Treatment Assignment $(Z_i = 0)$	Treatment Assignment $(Z_i = 1)$
Outcome	Dead	87	84
	Alive	155	175
		242	259
Males			
		Treatment Assignment $(Z_i = 0)$	Treatment Assignment $(Z_i = 1)$
Outcome	Dead	59	69
	Alive	135	140
		194	209
Females			
		Treatment Assignment $(Z_i = 0)$	Treatment Assignment $(Z_i = 1)$
Outcome	Dead	28	15
	Alive	20	35
		48	50

We begin by applying Fisher's exact test of the sharp null. The test of the sharp null is a test of no effect for both the ITT and the IV estimates. If we cannot reject the null of no effect for the

ITT estimate, there cannot be any effect for the IV estimate. For the overall study population, the p-value from Fisher's exact test is 0.231, and so we cannot rule out that randomization to EVAR was without effect. We also cannot rule out the receipt of EVAR was without effect. Next, we calculate an estimate for A which is the number of events attributable to EVAR assignment, that is the number of patients alive at 30 days who would not have survived if they had been assigned to the open surgery arm. To calculate this quantity, we adjust Table 3 until it is exactly without treatment effect. We adjust the data in Table 3, by adding units to the upper right cell of Table 3 and subtracting them from the lower right cell of Table 3. To estimate A, we apply this adjustment until the estimated odds-ratio is 1. In Table 3 the estimated odds-ratio is 1.17, which reduces to 1.0019 when $A = 9$. Therefore, the ITT estimate of the attributable effect, A indicates that assignment to the EVAR arm of the study caused an additional 9 patients to be alive at 30 days post randomization, that is 2.7% of survivals are attributable to treatment assignment $(9/(155 + 175))$.

We place an upper-bound on the attributable effect using a 95% confidence set. To find the 95% confidence set, we again adjust the table and find the value of A such that the p-value from Fisher's exact test equals 0.05 for one-sided test, and .025 for a two-sided test. For example if $A = 33$ the p-value from Fisher's exact test is 0.022, and if $A = 32$ the p-value from Fisher's exact test is 0.027. Therefore, we can be approximately 95% confident that as many as 32 survivals were attributable to assignment to EVAR. Of course, we can calculate a lower bound as well, which for these data is zero.

Next we turn to estimation and interpretation of the three different estimates of the IV estimand. First, from the data in Table 4, we calculate the compliance rate; the rate at which assignment to EVAR caused patients to receive EVAR, which is $149 - 32 = 117$. Therefore the estimate of $\hat{A}/U$ is $9/117 \approx 0.08$, which implies that for each additional person who complied, an additional .08 patients survived to 30 days. The 95% confidence set for this rate is 0 and 0.27. This confidence set, however, does not account for the fact that U is an estimate.

We now present IV estimates based on the four attributable effects: A_1, A_2, A_3, and A_4.

Table 4: Observed contingency table for compliance in the IMPROVE Trial

		Treatment Assignment $(Z_i = 0)$	$(Z_i = 1)$
Treatment Receipt	Open	210	110
	EVAR	32	149
		242	259
Males			
		Treatment Assignment $(Z_i = 0)$	$(Z_i = 1)$
Treatment Receipt	Open	166	84
	EVAR	28	125
		194	209
Females			
		Treatment Assignment $(Z_i = 0)$	$(Z_i = 1)$
Treatment Receipt	Open	44	26
	EVAR	4	24
		48	50

These estimates fully account for this uncertainty. Table 5 contains $2 \times 2 \times 2$ contingency table of survival by treatment assignment and receipt for the overall IMPROVE study population. The first estimate is for the $ACCE$, which is, among the compliers assigned to the EVAR treatment, the proportion who survived because of assignment to EVAR. This estimate is 0.07 with an associated 95% confidence set of 0 and 0.36 if we assume only positive effects and confidence set of -0.18 and 0.36 if we relax this assumption using A_4. This implies that we estimate that approximately 7% of the survivals among compliers is due to assignment to the EVAR treatment; however based on the confidence set, we cannot rule out that no compliers survived due assignment to EVAR. Next, we report the results for $ACRR$, which is on a risk ratio scale. The estimate for the $ACRR$ is 1.11 with a confidence interval of 1 and 1.93 assuming positive effects and 0.78 and 1.93 if we do not.

As a comparison, we also used two more standard IV estimators. We applied 2SLS and a plug-in estimator suggested by Palmer et al. (2008) that produces estimates on a risk ratio scale. The estimate based on 2SLS is 0.08 with a 95% confidence interval of $[-0.11, 0.26]$. The estimate using the plug-in estimator 1.13 with a 95% confidence interval of 0.85 and 1.50. The point estimates for both methods are quite similar, but the $ACRR$ estimate is based on weaker distributional assumptions. Moreover, the width of the CI based on randomization inference is longer, as it is exact and fully accounts for uncertainty due to noncompliance and the sample size.

We now conduct subgroup analyses. For the male subgroup, the p-value from Fisher's exact test is 0.748, and so we are unable to reject the sharp null hypothesis. The number of male patients that survived to 30 days attributable to treatment assigned is 0 with a 95% confidence set of 0 and 15, and hence the estimate for A/U is also 0. Table 6 contains $2 \times 2 \times 2$ contingency table of survival by treatment assignment and receipt for males in the IMPROVE study. The estimate for the $ACCE$ based on the males only subpopulation is 0 with a 95% confidence set of 0 to 0.24 under nonnegative effects, and -0.03 with a 95% confidence set of -0.33 and 0.24 if we allow for negative effects. The estimate for the $ACRR$ is 1 with a 95% confidence set of 1 to 1.48. If we allow for nonnegative effects, the point estimate is 0.92 with a 95% confidence interval of 0.66 and 1.48. The estimate based on 2SLS is -0.06 with a 95% confidence interval of $[-0.26, 0.14]$, and the estimate based on the plug-in principle is 0.92 $[0.69, 1.23]$. As before, the exact confidence intervals are wider.

The p-value from Fisher's exact test amongst females is 0.002, so we can reject the sharp null hypothesis. After adjusting the cells in Table 3 until they are consistent with no treatment effect, we find that the estimate for A is 14.5, which implies that 14.5 (25%) of the survivals among females was attributable to assignment to the EVAR arm of the study. The two-sided 95% confidence set for A is 3 to 24. Amongst females, the number of patients who complied with EVAR assignment is $24-4 = 20$, so the estimate for $\hat{A}/U$ is $14.5/20 = .725$. For each additional woman who received EVAR because of assignment to EVAR, an additional 0.725 women survived

Table 5: Observed contingency table for survival by both treatment assignment and receipt in the IMPROVE trial for the entire study population.

Treatment Assignment: EVAR			
		Treatment Receipt	
		EVAR	Open
Outcome	Dead	42	42
	Alive	107	68
Treatment Assignment: Open			
		Treatment Receipt	
		EVAR	Open
Outcome	Dead	8	79
	Alive	24	131

at 30 days post randomization. The 95% confidence set for the estimate of A/U is .15 to 1.2. This confidence set, however, fails to fully account for uncertainty in the estimate of U. We next report the estimates based on A_1, A_2, A_3, and A_4. Table 5 contains the $2 \times 2 \times 2$ contingency table of survival by treatment assignment and receipt for females in the IMPROVE study.

In the female subpopulation, the $ACCE$ is estimated to be 0.72, with a 95% confidence set of 0.10 to 1 regardless of what we assume about the presence of nonnegative effects. In the female subgroup, there were several values of the estimated ACRR that maximized the p-value. Following the principle in Rosenbaum (2002, ch. 2) we take the average of the minimum and maximum values that achieve the maximum p-value. However, this produced a point estimate of infinity since the largest value of the $ACRR$ was infinity. In a slight abuse of the Hodges-Lehman principle, instead we averaged the largest point estimate less than infinity with the smallest value with the same maximum p-value. Under this method, the point estimate for the $ACRR$ is 15 with a 95% confidence set of 1.15 to ∞. Again we used 2SLS and a plug-in method as a comparison. The estimate based on 2SLS was 0.71 with 95% confidence interval $[0.19, 1.24]$, and the plug-in estimate was 3.70 with a 95% confidence interval of 1.41 and 9.66. In both cases, the asymptotic approximations required for inference produce confidence intervals that are too short.

When the sample size is small, it is notable that we witness substantial differences across

Table 6: Observed contingency table for survival by both treatment assignment and receipt in the IMPROVE Trial for males

Treatment Assignment: EVAR			
		Treatment Receipt	
		EVAR	Open
Outcome	Dead	36	33
	Alive	89	51
Treatment Assignment: Open			
		Treatment Receipt	
		EVAR	Open
Outcome	Dead	7	52
	Alive	21	114

methods.The confidence sets for the $ACCE$ and the $ACRR$ are exact and here become longer and must include positive infinity in one case to maintain exact coverage. The confidence sets for both the plug-in method and $\hat{A}/U$, however, overstate the information present in the data. This is not unusual for IV estimators (Imbens and Rosenbaum 2005).

Table 7: Observed contingency table for survival by both treatment assignment and receipt in the IMPROVE Trial for females

Treatment Assignment: EVAR			
		Treatment Receipt	
		EVAR	Open
Outcome	Dead	6	9
	Alive	18	17
Treatment Assignment: Open			
		Treatment Receipt	
		EVAR	Open
Outcome	Dead	1	27
	Alive	3	17

6 Discussion

This paper develops new IV methods for binary outcomes. Our randomization inference approach provides an exact inference, and avoids the distributional assumptions required by many estimation methods, such as plug-in methods. Moreover, our method also relaxes the assumption that rules out the possibility of negative treatment effect that were invoked in past work.

A further strength of randomization inference is in making explicit that the IV estimate only pertains to the RCT participants. Any attempt to obtain IV estimates for a larger population, will require further assumptions to be explicitly defined. By contrast, under common parametric (Baiocchi et al. 2014; Cai et al. 2011, 2012; Terza et al. 2008), and semi-parametric approaches (Clarke and Windmeijer 2012; Vansteelandt et al. 2011), the target parameter is the population CACE. Such approaches typically invoke asymptotic assumptions, and imply that the RCT participants are a random sample from the target population. Compared to other alternatives for binary outcomes, such as semi-parametric methods, (Clarke and Windmeijer 2012; Vansteelandt et al. 2011), randomization inference is an accessible method; each of the required quantities can be calculated from contingency tables. The supplemental appendix contains R code to implement our method.

One drawback of randomization inference is that it does not easily allow for covariate adjustment. Rosenbaum (2002) demonstrates how covariate adjustment can be achieved with continuous endpoints, by applying randomization inference to the residuals from a model with the outcome regressed on a set of covariates. However, with binary outcomes this method does not provide residuals on the requisite 0 to 1 scale. Hence further research is warranted to develop randomization inference methods for binary outcomes that would include covariates to gain precision.

7 Supplementary Materials

The reader is referred to the on-line Supplementary Materials for annotated R programs and an expanded discussion of randomization inference methods for binary outcomes.

Acknowledgments

The authors thank Paul Rosenbaum and Paul Clarke for comments. This report is independent research supported by the National Institute for Health Research (Senior Research Fellowship, Dr Richard Grieve, SRF-2013-06-016). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Conflict of Interest: None declared.

References

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

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

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

Bellamy, S., Lin, J., and Have, T. T. (2007), "An introduction to causal modeling in clinical trials," *Clinical Trials*, 48, 58–73.

Burgess, S., Granell, R., Palmer, T. M., Sterne, J. A., and Didelez, V. (2014), "Lack of identification in semiparametric instrumental variable models with binary outcomes," *American journal of epidemiology*, 180, 111–119.

Cai, B., Hennessy, S., Flory, J. H., Sha, D., Have, T. R. T., and Small, D. S. (2012), "Simulation Study of Instrumental Variable Approaches With An Application to a Study of the Antidiabetic Effect of Bezafibrate," *Pharmacoepidemiology and Drug Safety*, 21, 114–120.

Cai, B., Small, D. S., and Have, T. R. T. (2011), "Two-stage instrumental variable methods for estimating the causal odds ratio: Analysis of bias," *Statistics in Medicine*, 30, 1809–1824.

Clarke, P. S. and Windmeijer, F. (2010), "Identification of Causal Effects on Binary Outcomes Using Structural Mean Models," *Biostatistics*, 11, 756–770.

— (2012), "Instrumental Variable Estimators for Binary Outcomes," *Journal of the American Statistical Association*, 107, 1638–1652.

Copas, J. (1988), "Binary Regression Models for Contaminated Data," *Journal of The Royal Statistical Society B*, 50, 225–265.

Fisher, R. A. (1935), *The Design of Experiments*, London: Oliver and Boyd.

Freedman, D. A. and Sekhon, J. S. (2010), "Endogeneity in Probit Response Models," *Political Analysis*, 18, 138–150.

Hansen, B. B. and Bowers, J. (2009), "Attributing Effects to A Clustered Randomized Get-Out-The-Vote Campaign," *Journal of the American Statistical Association*, 104, 873–885.

Hodges, J. L. and Lehmann, E. (1963), "Estimates of Location Based on Ranks," *The Annals of Mathematical Statistics*, 34, 598–611.

Imbens, G. W. and Rosenbaum, P. (2005), "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education," *Journal of The Royal Statistical Society Series A*, 168, 109–126.

investigators, I. t. (2013), *Improve Trial Website*, `http://www.improvetrial.org` (accessed August 26, 2015).

— (2014), "Endovascular or open repair strategy for ruptured abdominal aortic aneurysm: 30 day outcomes from IMPROVE randomised trial," *BMJ: British Medical Journal*, 348.

Little, R. J. (1985), "A Note About Models for Selectivity Bias," *Econometrica*, 53, 1469–1474.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, 5, 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

Palmer, T. M., Thompson, J. R., Tobin, M. D., Sheehan, N. A., and Burton, P. R. (2008), "Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses," *International Journal of Epidemiology*, 37, 1161–1168.

Rigdon, J. and Hudgens, M. G. (2015), "Randomization inference for treatment effects on a binary outcome," *Statistics in medicine*, 34, 924–935.

Rosenbaum, P. R. (1996), "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association*, 91, 465–468.

— (2001), "Effects Attributable To Treatment: Inference In Experiments And Observational Studies With A Discrete Pivot," *Biometrika*, 88, 219–231.

— (2002), *Observational Studies*, New York, NY: Springer, 2nd ed.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 6, 688–701.

— (1986), "Which Ifs Have Causal Answers," *Journal of the American Statistical Association*, 81, 961–962.

Tan, Z. (2006), "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 101, 1607–1618.

— (2010), "Marginal and Nested Structural Models Using Instrumental Variables," *Journal of the American Statistical Association*, 105, 157–169.

Terza, J., Basu, A., and Rathouz, P. (2008), "Two-stage residual inclusion estimation: addressing endogeneity in health econometric Two-stage residual inclusion estimation: addressing endogeneity in health econometric Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling," *Journal of Health Economics*, 27, 531–543.

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. (2011), "On Instrumental Variables Estimaton of Causal Odds Ratios," *Statistical Science*, 26, 403–422.

Vansteelandt, S. and Goetghebeur, E. (2003), "Causal Inference with Generalized Structural Mean Models," *Journal of the Royal Statistical Society, Series B*, 65, 817–835.

Weiss, L. (1955), "A note on confidence sets for random variables," *The Annals of Mathematical Statistics*, 26, 142–144.

Yang, F., Zubizaretta, J., Small, D. S., Lorch, S., and Rosenbaum, P. (2014), "Dissonant Conclusions When Testing the Validity of an Instrumental Variable," *The American Statistician*, 68, 253–263.