PHARMACOEPIDEMIOLOGY AND DRUG SAFETY (2017)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.4158

REVIEW

A tutorial on the use of instrumental variables in pharmacoepidemiology

Ashkan Ertefaie[1,2,3]*, Dylan S. Small[2], James H. Flory[4] and Sean Hennessy[3]

[1] *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA*
[2] *Department of statistics, University of Pennsylvania, Philadelphia, PA, USA*
[3] *Center for Pharmacoepidemiology Research and Training, Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA*
[4] *Weill-Cornell School of Medicine, New York, USA*

ABSTRACT

Purpose Instrumental variable (IV) methods are used increasingly in pharmacoepidemiology to address unmeasured confounding. In this tutorial, we review the steps used in IV analyses and the underlying assumptions. We also present methods to assess the validity of those assumptions and describe sensitivity analysis to examine the effects of possible violations of those assumptions.
Methods Observational studies based on regression or propensity score analyses rely on the untestable assumption that there are no unmeasured confounders. IV analysis is a tool that removes the bias caused by unmeasured confounding provided that key assumptions (some of which are also untestable) are met.
Results When instruments are valid, IV methods provided unbiased treatment effect estimation in the presence of unmeasured confounders. However, the standard error of the IV estimate is higher than the standard error of non-IV estimates, e.g., regression and propensity score methods. Sensitivity analyses provided insight about the robustness of the IV results to the plausible degrees of violation of assumptions.
Conclusions IV analysis should be used cautiously because the validity of IV estimates relies on assumptions that are, in general, untestable and difficult to be certain about. Thus, assessing the sensitivity of the estimate to violations of these assumptions is important and can better inform the causal inferences that can be drawn from the study.

KEY WORDS—assumptions; instrumental variables; observational studies; sensitivity analysis; unmeasured confounders; pharmacoepidemiology

Received 24 November 2015; Revised 18 October 2016; Accepted 29 November 2016

INTRODUCTION

Estimating the beneficial or adverse effects of medications is the primary goal of many pharmacoepidemiologic studies. While randomized trials often provide the least biased evidence for causation, their high costs, need for clinical equipoise and the need for evidence that reflects the effect of the intervention as applied in real-world settings often lead to the conduct of observational studies, especially for rare adverse effects.[1] Moreover, the short duration of randomized trials often prevents detection of delayed treatment effects. In observational studies, the treatment is not assigned at random. As a result, the straightforward comparison of outcomes between different treatment groups can be biased because of confounding, which is the presence of systematic differences between patients in different treatment groups on factors that affect outcome. Confounding by measured factors can be adjusted for using approaches such as multivariable regression, propensity scores and inverse probability of treatment weighting.[2–5] However, the validity of the estimates obtained from such methods relies on the assumption that all confounders have been measured and adjusted for. This untestable and frequently implausible assumption is known as “exchangeability” or “no unmeasured confounders”.

Instrumental variable (IV) analysis is an approach to obtain unbiased treatment effect estimates even in the presence of unmeasured confounders, provided that certain assumptions are met.[6–8] The key assumptions for a pretreatment variable to be a valid instrument are:

A1) the IV is associated with the treatment
A2) the IV is not associated with unmeasured confounders after conditioning on measured confounders

*Correspondence to: A. Ertefaie, 265 Crittenden Bulv, Rochester, NY 14642, USA. Email: ashkan_ertefaie@urmc.rochester.edu

(i.e., after controlling for measured confounders by regression or matching)
A3) the IV affects the outcome only through the treatment (i.e., there is no direct effect of the IV on outcome; this assumption is known as "exclusion restriction").

The key advantage of IV method is that it allows relaxation of the "no unmeasured confounders" assumption. However, this advantage comes at the cost of reliance on alternative assumptions and increased variance of the estimate of treatment effect. Particularly, when IVs are weakly associated with the treatment, IV analyses may lead to highly variable estimates, i.e., wide confidence intervals. In this tutorial, we are discussing the setting where the IV is reasonably strong. Depending on the context, the IV assumptions may be more plausible than the "no unmeasured confounders" assumption. Because of the reduced precision, IV analysis is recommended as the primary analysis only when unmeasured confounding is a major concern.[9] In addition to their use in primary analyses, IV methods can be considered for secondary or sensitivity analyses of conventionally analyzed studies.[10]

In this tutorial, we illustrate the use of IV methods in the context of an applied example examining the effect on body mass index (BMI) of metformin versus sulfonylureas as initial therapy for diabetes mellitus. In this example, the outcome (BMI) is continuous rather than dichotomous (e.g., occurrence of a major adverse cardiovascular event). Nevertheless, the approach outlined in this tutorial is readily applied to dichotomous outcomes as well. See Discussion section for more detailed information of IV analyses with binary outcome. We restrict the study to new antidiabetic drug users,[11] who we define as persons who are present in the study database for at least 180 days before receiving any antidiabetic drugs, and then were started on an initial therapy with either metformin or a sulfonylurea, with a baseline glycosylated hemoglobin (HbA1c) of ≥7%. The outcome is the first measurement of BMI after two years of subsequent follow-up. Note that because we are adjusting for the baseline BMI, our analysis is conceptually equivalent to an analysis of the outcome of change in BMI. The applied study was conducted using The Health Improvement Network (THIN), a UK-based medical record database that is collected from over 500 practices.[12] In general, medical records are not collected for scientific purposes, and it is likely that some important confounders are unmeasured. Specifically, other studies in the UK primary care setting have found metformin versus sulfonylurea to be highly confounded by indication, with an overall tendency for recipients of sulfonylurea to have a higher burden of comorbidities. For example, in one cohort study, 10% of metformin users had a history of cancer versus 14% of sulfonylurea users; similarly, 10% of metformin users had a history of major adverse cardiovascular events versus 16% of sulfonylurea users, and the Charlson comorbidity index of metformin users was lower than that of sulfonylurea users (1.3 vs. 1.7) ($p < 0.001$ for all comparisons). Such a higher baseline burden of disease, particularly cancer, could plausibly also be associated with decline in BMI over time. Therefore, baseline disease burden represents a potential confounder for this comparison that would be difficult to fully measure and adjust for using conventional means.[13] This helps to motivate the use of IV methods in our example. The measured baseline characteristics include BMI, glycosylated hemoglobin (HbA1c), gender and marital status. After excluding patients with missing baseline values for one of these variables, we identified 44 517 eligible patients. Eighty-eight percent (39 102) initiated metformin, while 12% (5415) initiated a sulfonylurea. Note that the prescription rates are roughly equal from 1997 to 2002, but there is a marked increase in the use of metformin and marked reduction in the use of sulfonylureas in the subsequent years that created the overall difference in the prescription rates (Figure 1).

STEP 1. SELECT AN INSTRUMENTAL VARIABLE

The most challenging task in an IV analysis is to identify a valid IV. IVs that are commonly used in pharmacoepidemiology include:

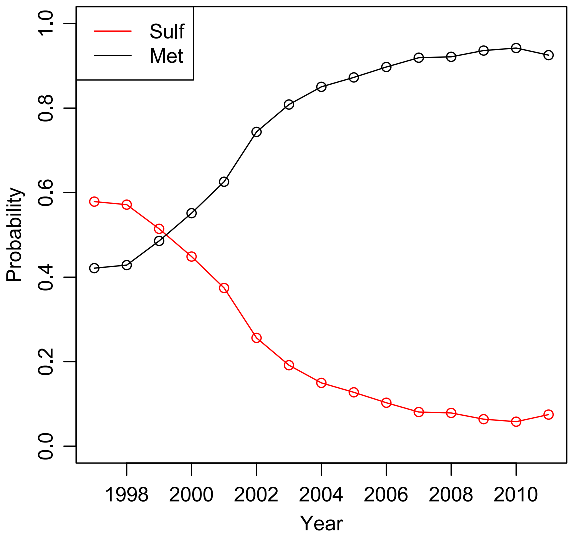

Figure 1. Rate of use of first line therapy over time [Colour figure can be viewed at wileyonlinelibrary.com]

i. *Randomized encouragement trials*. In an encouragement experiment, subjects selected at random are encouraged to take a treatment, and the remainder receive no encouragement.[14–16] These trials serve as prototype for the IV argument where the randomized encouragement assignment is a potential IV. By design, the independence of the IV and unmeasured confounders is guaranteed. The only concern is that the randomized encouragement might have a direct effect on the outcome. Note that randomized trials with nonadherence are special cases of randomized encouragement trials, and random assignment can be used as an IV.

ii. *Calendar time.* The popularity of one treatment versus another often changes over time.[17–20] Thus, calendar time can be considered as a potential IV. However, calendar time could violate the assumption that the IV is independent of unmeasured confounders if the study population's unmeasured characteristics (e.g., dietary habits in the diabetes example) change systematically during the study period. Also, calendar time might have a direct effect on the outcome rate because there are other changes in the way that patients are being treated over time (e.g., introduction of interventions to increase exercise).

iii. *Provider preference.* Naturally occurring variation in medical practice can be used to construct an IV.[10] Brookhart *et al.*[21] showed that under the assumption that systematic differences among patients in different practices are fully captured by the measured covariates, an IV can be defined as the treatment (e.g., metformin vs. sulfonylurea) that has higher chance of being prescribed by a particular provider or center. Different factors may contribute in providers' preference including different trainings and guidelines or marketing by pharmaceutical companies. However, provider preference cannot be directly observed and, thus, has to be inferred through a surrogate measurement.[21] Investigators have proposed different surrogates for provider preference based on the previous prescription pattern for a particular provider, such as the most recent prescription issued by the provider before the current patient, or the treatment that has been most frequently prescribed by the provider over a fixed period of time[22–24]. A major concern in the provider preference-based IVs is that providers who prescribe one treatment more often (e.g., metformin) may also provide better care in other ways, such as weight-reduction counseling. This could violate the assumption of no direct effect from the IV to the outcome. Another concern is that patients who go to providers who prefer treatment A may differ in characteristics than patients who go to providers who prefer treatment B. For example, it is conceivable that doctors in urban area are more likely to prescribe metformin because they see more patients with sedentary habits compared to doctors in rural area. This would violate the assumption that the IV is independent of unmeasured confounders if the living area is not recorded in the data.

iv. *Geographic distance to specialty care provider.* Sometimes, one of the treatments is only provided by specialty providers, for example, because it requires expensive equipment, such as designated trauma centers.[18,25–28] In these cases, the geographic distance that a patient has to travel to reach the nearest specialty provider can be used as an IV.

v. *Insurance plan*. Variations in insurance plan reimbursement or drug formulary policies can be used to construct an IV.[17,29] For example, a treatment option with lower co-payment amount can be an IV because it encourages the provider to prescribe the more affordable treatment. The validity of the insurance plan as an IV relies on the assumptions that there is no unmeasured confounding between patients in different types of plans.

Figure 1 presents time trends in the use of sulfonylureas and metformin from 1997 to 2011. During this period, the use of metformin rose very quickly, and the use of sulfonylureas declined. This variability led us to define our IV based on both provider preference and calendar time. In particular, we defined the IV within each general practitioner practice as the proportion of patients started on metformin versus a sulfonylurea over two year timeframes (i.e., 1997–1998, 1999–2000, 2001–2002, 2003–2004, 2005–2006, 2007–2008 or 2009–2011). We assigned $IV = met$ if the average was more than 50% and $IV = sulf$ otherwise. Thus, the value of IV was allowed to change over time for each practice. Note that while many IV analyses use the apparent preference of the individual prescriber as the IV, THIN data do not allow one to distinguish among individual prescribers within a practice. Therefore, we used practice preference to define the IV. Below, we refer to practice preference IV as PP IV.

Further identifying assumptions

Assumptions A1–A3 are the core assumptions for a valid IV.[7] In our example, A1 asserts that the IV based on provider preference is associated with the treatment choice of metformin versus sulfonylurea. Assumption A2 implies that after conditioning on the measured covariates, e.g., baseline BMI and HbA1c, patients who go to providers who prefer metformin do not have different characteristics than patients who go to providers who prefer sulfonylurea. Also, Assumption A3 means that the quality of care does not vary across providers with different preferences.

When the effect of treatment is constant among the population under study, i.e., there is no effect modification by any variable; IV methods can be used to estimate the treatment effect under the assumptions above. Even in the presence of effect modification, IV methods can still be used to estimate the treatment effect among a particular type of patients, so-called “compliers” or “marginal patients”.[30] These are patients who would receive the treatment indicated by the IV regardless of the value of the IV. In our example, compliers or marginal patients are those who would be always prescribed metformin if that was the practice preference, and would be always prescribed a sulfonylurea if that was the practice preference. The average treatment effect for this group of patients is called the complier average treatment effect, also known as the local average treatment effect (LATE). For IV methods to estimate the LATE, we need another assumption, namely monotonicity. In the context of the practice preference IV, this means that there are no “defier” patients who would be prescribed metformin if they were seen in a practice that preferred sulfonylureas but would be prescribed a sulfonylurea if they were seen in a practice that preferred metformin. Monotonicity is plausible when the IV does not provide any disincentive for the treatment indicated by the IV,[31,32] as seems to be the case in our example. This assumption is automatically satisfied for randomized encouragement trials in which only subjects encouraged to receive the treatment are able to receive it. Monotonicity seems to be plausible in our example, because there is no clinical reason that patients who have been prescribed metformin when the practice preference was sulfonylurea would have been prescribed sulfonylurea when the practice preference was metformin. In fact, it is very likely that patients who have been prescribed metformin when the practice preference was sulfonylurea be metformin always takers.

Under the monotonicity assumption, there are two other subgroups of patients besides compliers: “always takers” and “never takers”. For example, patients with high BMI might be always prescribed metformin regardless of whether that practice’s preference IV was a sulfonylurea or metformin, i.e., metformin always takers, because physicians might be concerned about the possibility of additional weight gain caused by sulfonylureas.[33,34] IV methods are not helpful for estimating the treatment effect among these patients because their treatment status does not change with the IV value, and thus their potential outcome is the same for both levels of the IV. For example, because of the potential weight gain caused by sulfonylureas, patients with high BMI are likely to be metformin always takers, and IV methods cannot be used to estimate the effect of metformin on this group of patients.

STEP 2. ASSESS THE IV ASSUMPTIONS AND STRENGTH OF THE IV

The validity of an IV estimator depends on whether assumptions A1–3 hold. It is therefore crucial to empirically assess whether the candidate IV satisfies all the required assumptions.

Assessing the association between the IV and the treatment and strength of the IV

The ability of an IV analysis to remove confounding depends on how strongly the treatment and the IV are associated, conditioning on measured covariates. Thus, an IV is defined as a weak IV if it is not a strong predictor of the treatment after controlling for the measured covariates. The strength of the IV can be assessed in at least two ways: (i) calculating the proportion of compliers by calculating the difference of the treatment assignment rate among subjects across the values of the IV; and (ii) calculating the F-statistic with 1 degree of freedom for the IV in a regression model that includes treatment as dependent variable and the IV and measured covariates as independent variables. The proportion of compliers represents the effective sample size in the IV analysis. So when the proportion is small, the IV analysis is making use of only a small proportion of the data and, thus, may have lower statistical power. In our example, the proportion of compliers is 0.43 and defined as a difference of sulfonylureas prescription rate among patients across providers with sulfonylureas versus metformin preferences. The F-statistic is 1601 when calculated in a model that includes antidiabetic treatment option indicator as the dependent variable and the PP IV, baseline BMI, baseline Hba1c, marital status and gender as independent variables. Stock *et al.*[35] suggest that an F-statistic less than 10 represents a weak IV. However, we should keep in mind that the F-statistic is an increasing function of the sample size. Thus, in large datasets, the F-statistic may be large even for weak IVs.[22] Another potential way to assess the strength of an IV is using the partial r^2 when adding the IV to a regression model that includes treatment as dependent variable and the IV and measured covariates as

DOI: 10.1002/pds

independent variables. The partial r^2 indicates the percentage of unexplained, i.e., not explained by the measured covariates included in stage 1 model, variance that can be explained by the IV. However, the partial r^2 can be small even for moderately strong IVs, which may make assessing the strength of IVs difficult. For example, if the IV increases probability of taking treatment from 0.2 to 0.4 (so 20% compliers) and half of the people have IV = 1, then partial $r^2 = 0.048$ and even if the IV increases probability of taking treatment from 0.25 to 0.75 (so 50% compliers) and half of the people have IV = 1, then partial $r^2 = 0.25$.

A weak IV can produce biased estimators,[36,37] magnify existing bias caused by small violations to the IV assumptions and introduce imprecision, manifest as wide confidence intervals.[38–40] In our example, the PP IV seems to be a reasonably strong IV according to the proportion of compliers and the F-statistic.

Assessing the independence of the IV and unmeasured confounders

The independence assumption of the candidate IV and unmeasured confounders (A2) cannot be completely tested using observed data. However, looking at the balance of measured covariates between values of the IV may provide insight about the validity of this assumption. This is similar to the insight that is provided by examining the balance of baseline covariates among exposure categories (the usual Table 1 of a cohort study) as a clue for potential unmeasured confounding in a conventional cohort study. Specifically, imbalance in measured confounders across categories of the IV makes assumption A2 less plausible because, for example, if the measured covariates are a proxy of the unmeasured confounders, an association between the measured confounders and the IV suggests that there will be an association between the IV and unmeasured confounders.

Table 1 presents the standardized difference as a measure of the imbalance between the measured covariates among patients with different treatments and PP IVs. For a given covariate, say baseline BMI, the standardized difference between the PP IV groups is the difference between the average of baseline BMI among patients who have seen a practice with sulfonylureas and metformin preference divided by the pooled standard deviation of the two groups. The standardized difference between the treatment groups is defined similarly. Between the treatment groups, the imbalance is considered potentially important if the standardized difference is greater than 0.2 [41], while between the IV groups the imbalance is potentially important if it is greater than 0.2 multiplied by the proportion of compliers, i.e., $0.2 \times 0.43 = 0.086$. This is because the bias in an IV analysis when there are differences between the IV groups is inflated by (1/proportion of compliers) compared to the bias in a propensity score or regression analysis when there are differences between the treatment groups.[9,42] Although the imbalance is reduced among the IV groups compared to the treatment group, the baseline BMI and HbA1c are still imbalanced. To further investigate this imbalance, we fit two separate models by regressing the baseline BMI and HbA1c on the IV and all the measured covariates. The p-values of the coefficients of the IV in these two models are 0.027 and 0.069, respectively, meaning that PP IV has stronger association with the baseline BMI than the baseline HbA1c, which is consistent with our results in Table 1. Thus, these variables must be controlled for by including them in the IV analysis. Further, this imbalance in measured variables reduces the plausibility that unmeasured variables are balanced. For example, the imbalance of the

Table 1. THIN Data. Covariate balance for treatment and IV groups. Bias_1 and Bias_2 are the denominator and numerator of the Bias ratio, respectively

	Treatment groups			
	Metformin	Sulfonylureas	Standardized difference	Bias_1
N	39 102	5415		
Sex (female)	17.5%	17.2%	0.06	0.03
Marital (married)	3.4%	2.8%	0.01	0.003
Baseline BMI, mean (SD)	32.38(6.43)	27.14(5.13)	0.91	5.24
Baseline HbA1c, mean (SD)	9.01(1.79)	9.50(2.07)	0.25	−0.49
	IV groups			
	Metformin	Sulfonylureas	Standardized difference	Bias_2
Sex (female)	41.9%	40.5%	0.03	−0.03
Marital (married)	17.5%	15.4%	0.06	0.05
Baseline BMI, mean (SD)	31.78(6.43)	30.12(6.51)	0.27	−3.90
Baseline HbA1c, mean (SD)	9.06(1.83)	9.29(1.78)	0.13	0.54

baseline BMI among the IV groups can be related to the practice location or region, which in that case, assuming that prescribing preferences differ systematically by region, would suggest that other unmeasured characteristics related to geographical locations might be imbalanced as well.

The bias caused by failing to adjust for a covariate in an IV analysis is approximately proportional to the difference of the average of the covariates across the values of the IV divided by the difference of the treatment assignment rate among subjects across the values of the IV. Similarly, the bias caused by failing to adjust for a covariate in a non-IV analysis is approximately proportional to the difference of the average of the covariate across the values of the treatment options. Brookhart and Schneeweiss[42] proposed the bias ratio to assess the association between the candidate IV and measured covariates. The bias ratio is the ratio of the bias between an IV analysis and a non-IV analysis that both fail to adjust for a covariate. In our example, the numerator of the bias ratio of the baseline BMI is the difference of the average baseline BMI across the values of the PP IV divided by the difference of the sulfonylureas assignment rate among subjects across the values of the PP IV. The denominator of the bias ratio is the difference of the average baseline BMI across the values of the treatment options (equation given in Appendix A). Bias ratios less than one indicate that the IV estimator is less biased than the non-IV estimator, while a bias ratio greater than one indicates that the IV estimator is more biased than the non-IV estimator.[9,43] The numerator and the denominator of the bias ratio correspond to the bias in the IV and non-IV analysis, respectively. In Table 1, the bias ratio calculated as $\text{Bias}_2/\text{Bias}_1$ shows that, for marital status ($0.05/0.003 > 1$) and baseline HbA1c ($0.54/0.49 > 1$), there would be more bias incurred from omitting these covariates from adjustment in the IV analysis than a non-IV analysis. This may signal potential IV outcome confounding by unmeasured covariates.

Assessing the exclusion restriction (ER) assumption

The exclusion restriction (ER) assumption (A3) is that the IV affects the outcome only through the treatment. In many clinical settings, where patients are taking other treatments concomitantly with the treatment under study, exploring the association between the IV and the concomitant treatments can help to assess the validity of the ER assumption. Specifically, assuming that the concomitant treatments affect the outcome, the ER assumption will be violated if an IV for the treatment of interest also affects prescribing of concomitant treatments.[10] For example, some antidiabetic treatments may be more likely to be prescribed to patients concomitantly with statins (lipid lowering treatments).[44] There is a clinical sense that taking statins may cause weight gain.[45] Assuming that these conjectures are true, if the IV (e.g., practice preference) for one antidiabetic treatment versus another is associated with prescribing statins, the IV analysis may be biased for the effect of antidiabetic treatments on BMI. In fact, our analysis shows that the proportion of statin prescription among providers with sulfonylureas and metformin preference are 0.65 and 0.80, respectively. We also regressed the statin use on the PP IV and the measured covariates. Because the coefficient of PP IV was significant (p-value$\,< 0.001$), if our conjecture about the association between statin and weight gain is true, our proposed PP IV may violate the ER assumption.

It is important to note that although it may be intuitively appealing to assess the validity of the ER assumption by looking at the p-value of the PP IV in a model that includes the BMI after two years of initiating the antidiabetic treatment as the dependent variable, and the PP IV, baseline BMI and baseline Hba1c as independent variables, such analysis could lead to falsely concluding that the ER assumption is violated. In fact, Baiocchi *et al.*[9] showed that the coefficient of the IV is usually nonzero even when the IV is valid.

STEP 3. ESTIMATE THE TREATMENT EFFECT

The two-stage least squares (2SLS) estimator is a commonly used approach to estimate the treatment effect in the form of a risk difference for a dichotomous outcome. In the first stage, exposure is the dependent variable, and the IV and measured covariates are the independent variables. This model gives the predicted probability of being exposed (e.g., being assigned to sulfonylurea) given the IV and measured covariates. In our diabetes example, the first stage model includes the treatment received (metformin vs. sulfonylurea) as the dependent variable and the practice preference IV, baseline BMI, baseline Hba1c, marital status and gender as independent variables. In the second stage, the outcome is the dependent variable, and the independent variables are the predicted value of the exposure variable given the IV and measured covariates (i.e., the fitted values from the first stage) and measured covariates. The coefficient for the association of the predicted value of the treatment with the outcome in the second stage regression is the IV estimator of treatment effect.[46] In our example, the second stage model includes BMI as the dependent variable and the predicted value of the first stage model (i.e., predicted probability of receiving a sulfonylurea vs. metformin), baseline BMI, baseline Hba1c, marital status and gender as independent variables. The coefficient for the association of the predicted value of the treatment with the outcome is the estimated causal effect of sulfonylurea (assuming that sulfonylurea is coded as 1 and metformin is coded as 0) on change in BMI among compliers, i.e., those who would be prescribed metformin if that was the practice preference and would be prescribed a sulfonylurea if that was the practice preference. Note that although the outcome is BMI after 2 years of treatment initiation, because we are adjusting for the baseline BMI, our analysis is conceptually equivalent to an analysis of the outcome of change in BMI.

The standard error from the second stage model systematically under-estimates the true standard error of the 2SLS estimator because the standard error from the second stage regression does not take into account for the sampling error in the first stage estimate of the probability of exposure.[47–49] One can obtain a 2SLS estimator with a valid standard error from “ivreg” in Stata, or the R package “ivmodel”. Alternatively, the standard errors of the 2SLS estimators can be estimated using bootstrap resampling, with both stages of 2SLS estimation replicated in each resample.

Example: Analysis of THIN data

Table 2 shows the treatment effect estimates obtained by the following estimators:

- 2SLS IV_1: the 2SLS IV estimator that does not include any covariates.
- 2SLS IV_2: the 2SLS IV estimator that only includes baseline BMI.
- 2SLS $IV_{3:}$ the 2SLS IV estimator that only includes baseline HbA1c.
- 2SLS IV_4: the 2SLS IV estimator that includes both baseline BMI and HbA1c.
- 2SLS IV_5: the 2SLS IV estimator that includes gender, marital status, baseline BMI and HbA1c.
- Regression_1: non-IV regression estimator that is obtained by regressing the outcome on the treatment and baseline HbA1c.
- Regression_2: non-IV regression estimator that is obtained by regressing the outcome on the treatment and baseline BMI.
- Regression_3: non-IV regression estimator that is obtained by regressing the outcome on the treatment, baseline BMI and baseline HbA1c.

Table 2. THIN Data. Estimating the effect of metformin versus sulfonylurea on BMI. Regression_1 represents the standard regression approach which estimates the treatment effect by including treatment and baseline HbA1c into a regression model; Regression_2 is similar to Regression_1 but includes baseline BMI instead of baseline HbA1c; Regression_3 includes gender, marital status, baseline BMI and HbA1c into a regression model. 2SLS IV_1 does not include any covariates; 2SLS IV_2 includes only baseline BMI; 2SLS IV_3 includes only baseline HbA1c; 2SLS IV_4 includes both baseline BMI and HbA1c; 2SLS IV_5 includes gender, marital status, baseline BMI and HbA1c

	Est.	S.D.	95%CI
Regression_1	−4.09	0.09	(−4.27,−3.91)
Regression_2	0.74	0.04	(0.66,0.82)
Regression_3	0.64	0.04	(0.56,0.72)
2SLS IV_1	−2.31	0.44	(−3.19,−1.43)
2SLS IV_2	1.27	0.19	(0.89,1.65)
2SLS IV_3	−2.35	0.46	(−3.27,−1.43)
2SLS IV_4	1.17	0.19	(0.79,1.55)
2SLS IV_5	1.17	0.19	(0.80,1.55)

The IV analyses 2SLS IV_1 and 2SLS IV_3 that did not include the baseline BMI, estimated that, contrary to their known effect,[50] sulfonylureas reduced BMI by 2.31 (95%CI: (−3.19,−1.43)) and 2.35 (95%CI: (−3.27,−1.43)) BMI units, respectively, compared with metformin. This is likely due to the association between the IV and baseline BMI (Table 1). In contrast, adjusting for the baseline BMI in the 2SLS IV analyses 2SLS IV_3, 2SLS IV_4 and 2SLS IV_5 estimated that sulfonylureas increased BMI by 1.27 (95%CI: (0.89,1.65)), 1.17 (95%CI: (0.79,1.55)) and 1.17 (95%CI: (0.80,1.55)) BMI units, respectively. Note that inclusion of baseline HbA1c in the analysis 2SLS IV_4 does not change the estimated effect by much compared with 2SLS IV_3 analysis. This is because the imbalance of baseline HbA1c among IV groups is small, as seen in Table 1. Similarly, because gender and marital status are balanced across IV groups (Table 1), the estimated effect by 2SLS IV_5 is almost identical to the one obtained by 2SLS IV_4. Comparing the results of the different IV models shows the importance of including the measured covariates in the IV analysis, particularly those that are imbalanced among the IV groups. While Regression_1 estimate shows sulfonylureas reduced BMI by 4.09 (95%CI: (−4.27,−3.91)), the other two regression estimates Regression_2 and Regression_3 that include baseline BMI in the models show that sulfonylureas increased BMI by 0.74 (95%CI: (0.66,0.82)) and 0.64 (95%CI: (0.56,0.72)), respectively. Note that the standard error of the latter two regression models is about five-fold lower than the standard error of 2SLS IV_2, 2SLS IV_4 and 2SLS IV_5 that also adjust for the baseline BMI. In general, IV-based estimators have higher variance compared to non-IV estimators, and the difference is bigger when the IV is not strong.[40]

We were concerned that the IV estimates in Table 1 may be affected by the difference in the prevalence of metformin and sulfonylurea in our dataset (88% metformin vs. 12% sulfonylurea). To examine the

DOI: 10.1002/pds

sensitivity of our estimate the imbalance in the rate of treatment prescriptions, we restricted the study period to 1997–2004 and estimated the treatment effect using a same model used in 2SLS IV^5. In the restricted period, the prescription rates of sulfonylureas and metformin are 27% and 73%, respectively. We used the 0.5 cut-point to construct the IV just as we did for unrestricted data, i.e., 1997–2011. The estimated treatment effect in the 2SLS model that includes gender, marital status, baseline BMI and HbA1c as covariates is 0.98 with standard error 0.13 (95%CI: 0.72–1.22). This point estimate is very close to the unrestricted result presented in the paper as 2SLS IV^5, which is 1.17 with a standard error 0.19. Note that despite the smaller sample size in the restricted analysis, the estimated standard error is smaller than the unrestricted analysis. This suggests that the dramatic treatment rate imbalance does not invalidate the effect estimate but it leads to an estimate with higher standard error.

STEP 4. PERFORM SENSITIVITY ANALYSIS TO ASSESS THE EFFECT OF VIOLATIONS OF ASSUMPTIONS

In general, IV assumptions cannot be completely tested, and the aforementioned assessment methods can provide incomplete insight about the validity of these assumptions. Thus, sensitivity analysis is warranted to quantify how sensitive the estimates are to the possible violations of the underlying assumptions.[7,26,38,42,51] The goal of sensitivity analysis is to assess how departures from certain assumptions may alter the study's conclusions. All of the model equations related to this section are presented in Appendix B, where we also provide a detailed discussion of how to perform sensitivity analysis in the context of 2SLS IV analysis.

Sensitivity to the violation of assumption A2

Suppose there is an unmeasured confounder that is associated with the IV. We can think of the unmeasured confounder as the component of confounding that is not captured by measured variables. That is, we can assume that the unmeasured confounder is unassociated with measured covariates, because if it were associated with measured covariates then controlling for them would also reduce the confounding associated with that unmeasured variable. For sake of the sensitivity analysis, we can also assume that the unmeasured confounder has been standardized to have a mean of zero and a variance of 1. We can further assume a range of plausible values for δ, the effect of a one standard deviation increase in the unmeasured confounder on the mean of the outcome, and τ, the effect of the IV on the unmeasured confounder, expressed in standard deviation units.[52] Then, the true treatment effect can be identified as a function of a two-dimensional sensitivity analysis (δ, τ) (see Appendix B.1 for model equations). If the treatment effect estimate does not change dramatically in a plausible range of sensitivity parameters (e.g., the sign remains the same or the confidence interval remains on the same side of the real line), we can conclude that the estimate is relatively insensitive to the plausible degrees of violation of A2 and the overall conclusion is robust.

In our example, the observed association between the IV and the baseline BMI in Table 1 suggested that the IV might be associated with unmeasured confounders. To assess the effect of this potential violation of the IV assumptions, we performed the sensitivity analysis discussed in this section. Table 3 shows that the effect of sulfonylureas on BMI would still be statistically significantly positive if there were an unmeasured confounder that increased follow-up BMI by 0.6 BMI units for a one standard deviation increase in the unmeasured confounder and was $\tau = 0.5$ standard deviation units higher in practices with a sulfonylurea preference than in practices with a metformin preference. A stronger effect of an unmeasured confounder on increasing follow-up BMI would result in a non-significant estimated effect of sulfonylureas on weight gain (e.g., see the row $\tau = 0.5$ and $\delta = 0.7$ in Table 3).

Sensitivity to the violation of assumption A3

The ER assumption is violated when there is a direct effect from the IV on the outcome. For example, considering practice preference as an IV, practices that issue sulfonylureas more often, i.e., IV = 1, may deliver sulfonylureas better by monitoring the BMI more carefully or better dosing the treatment than those that issue

Table 3. THIN data. Sensitivity analysis. δ is the effect of a one standard deviation increase in the unmeasured confounder on the mean of the outcome. τ is the effect of the IV on the unmeasured confounder, expressed in standard deviation units. $\boldsymbol{\beta}$ is the effect of sulfonylureas on BMI

τ	δ	β	95%CI
0.0	0.0	1.17	[0.79,1.55]
0.1	−3.0	1.91	[1.53,2.29]
0.5	−3.0	4.89	[4.41,5.37]
0.5	−0.5	1.79	[1.41,2.17]
0.5	0.6	0.43	[0.06,0.79]
0.5	0.7	0.30	[−0.07,0.67]
0.5	3.0	−2.55	[−3.00,−2.11]
0.1	3.0	0.43	[0.06,0.79]

DOI: 10.1002/pds

metformin more often, i.e., IV = 0. A sensitivity analysis that assesses the effects of such a violation employs a sensitivity parameter ρ that quantifies the amount of treatment effect modification across the different values of the IV.[9] In other words, the sensitivity parameter is the coefficient of the interaction term between the treatment options and the IV in the outcome model that includes as independent variables, in addition, the treatment indicator, measured covariates (see Appendix B.2 for model equations). We can estimate the treatment effect for plausible values of the sensitivity parameter ρ. If the estimate yields similar substantive conclusions across all plausible values of ρ, we can conclude that it is insensitive to the violation of assumption A3.

Our sensitivity analysis for the exclusion restriction assumption in THIN data suggests that sulfonylureas use is associated with weight gain as long as $\rho \leq 0.6$. In particular, $\rho \leq 0.6$ results in a treatment effect estimate of 0.37 with a 95% confidence interval [0.00,0.74]. In other words, if being prescribed sulfonylurea by a practice that prescribes sulfonylureas more often increases the BMI by more than 0.6, i.e., $\rho \leq 0.6$, compared to a practice that prescribes more metformin, then the effect of sulfonylureas on weight gain would not be significantly higher than metformin. This can happen if, for example, practices that prescribe metformin more often monitor sulfonylureas better than those who prescribe sulfonylureas more often.

STEP 5. SUMMARIZING THE IV ANALYSIS RESULTS

In our example, the PP IV seems to be reasonably strong with 43% compliance rate. We have assessed the plausibility of the independence assumption of the candidate IV and unmeasured confounders (A2) by looking at the balance of measured covariates between values of the IV. Although the imbalance is reduced among the IV groups compared to the treatment group, the baseline BMI and HbA1c are still imbalanced. This imbalance in measured variables reduces the plausibility that unmeasured variables are balanced. We were also concerned about the validity of the ER assumption. These concerns encouraged us to perform sensitivity analysis, and the results suggest that the effect of sulfonylurea on weight gain is significantly more than metformin for moderate violations of the assumptions. However, when the violations are severe, then the treatment effects are not significantly different.

IV ANALYSIS WITH BINARY OUTCOMES

Most of the IV analysis literature is focused on continuous outcomes. Having binary outcomes raises additional challenges. Bhattacharya *et al.*[53] showed that the two-stage procedures for binary outcomes that are analogous to 2SLS estimator do not, in general, result in an unbiased treatment effect estimate. This implies that, for example, replacing the second stage model in the 2SLS procedure with logit model does not result in a correct estimate of the odds ratio.[54] On the other hand, when the risk difference is the parameter of interest, the 2SLS does not constrain the probability of the binary outcome to be between 0 and 1. More information about issues inherent in using IV estimators for binary outcomes can be found elsewhere.[55–60]

DISCUSSION

IV analysis is a powerful technique in pharmacoepidemiologic studies where the investigator has good reason to suspect unmeasured confounding. However, IV analysis has to be used cautiously because the validity of IV estimates relies on assumptions that are, in general, untestable and impossible to be certain about. Thus, assessing the sensitivity of the estimate to violations of these assumptions is important and can strengthen the causal inferences that can be drawn from the study. We have introduced sensitivity analyses for the two untestable assumptions, i.e., independence of the IV and unmeasured confounders and the ER assumption. We first assume a range of plausible values for the sensitivity parameters and then estimate the treatment effect as a function of these parameters. If the treatment effect estimate does not change dramatically in a plausible range of sensitivity parameters, we can conclude that it is insensitive to the plausible degrees of violation of assumptions.

The ability of an IV analysis to remove confounding depends on how strongly the treatment and the IV are associated. Weak IVs result in unstable estimates with very wide confidence intervals that can significantly affect the power of the analysis. Also, weak IVs can lead to estimates that are sensitive to small departures from IV assumptions.[39,61] Specifically, weak IVs affect the power of the analysis and make the point estimates difficult to interpret as well as likely to diverge more from the true effect than the biased but much more stable conventional estimate, which accounts only for the measured confounders.[40]

The key advantage of IV methods is that they allow relaxation of the "no unmeasured[14] confounders"

DOI: 10.1002/pds

assumption, which is required by regression, matching, inverse probability weighting, and propensity score methods. However, this advantage comes at the cost of reliance on alternative assumptions and increased variance of the estimate of treatment effect. Investigators should consider IV analysis when unmeasured confounding is a major concern and IV assumptions are plausible. In addition to their use in primary analyses, IV methods can be considered for secondary or sensitivity analyses of conventionally analyzed studies.[9,10] Because IV methods rely on a different set of assumptions than the non-IV methods, e.g., regression or propensity score based methods, it is recommended to compare the estimates obtained by these methods. If both approaches yield similar conclusions, it is reassuring that the conclusion is correct. If the answers are very different, one should further investigate more on the plausibility of the required assumptions and perform sensitivity analysis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

KEY POINTS

- The key advantage of IV methods is that they allow relaxation of the “no unmeasured confounders” assumption. However, this advantage comes at the cost of reliance on alternative assumptions and increased variance of the estimate of treatment effect.
- The most challenging task in an IV analysis is to identify a valid IV.
- It is crucial to empirically assess whether the candidate IV satisfies all the required assumptions that are testable.
- Sensitivity analysis is warranted to quantify how sensitive the estimates are to the possible violations of the underlying assumptions, especially those untestable.

ETHICS STATEMENT

The University of Pennsylvania’s institutional review board (IRB) determined that this project met the criteria for IRB exemption.

ACKNOWLEDGEMENTS

This study was supported by the National Science Foundation: NSF grant SES-1260782.

REFERENCES

1. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996; **312**(7040): 1215–1218.
2. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1): 41–55.
3. D’Agostino R, Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**: 2265–2281.
4. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5): 550–560.
5. Guo S, Fraser MW. *Propensity Score Analysis: Statistical Methods and Applications: Statistical Methods and Applications*. Sage Publications: California, 2014.
6. Stel VS, Dekker FW, Zoccali C, Jager KJ. Instrumental variable analysis. *Nephrology Dialysis Transplantation* 2013; **28**(7): 1694–1699.
7. Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996; **91**(434): 444–455.
8. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *Jama* 2007; **297**(3): 278–285.
9. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Stat Med* 2014; **33**(13): 2297–2340.
10. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010; **19**(6): 537–554.
11. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol* 2003; **158**(9): 915–920.
12. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf* 2007; **16**(4): 393–401.
13. Bannister CA, Holden SE, Jenkins-Jones S, Morgan CL, Halcox JP, Schernthaner G, Mukherjee J, Currie CJ. Can people with type 2 diabetes live longer than those without? A comparison of mortality in people initiated with metformin or sulphonylurea monotherapy and matched, non-diabetic controls. *Diabetes Obes Metab* 2014; **16**(11): 1165–73.
14. Kramer MS, Chalmers B, Hodnett ED, Sevkovskaya Z, Dzikovich I, Shapiro S, Collet JP, Vanilovich I, Mezen I, Ducruet T, Shishko G. Promotion of Breastfeeding Intervention Trial (PROBIT): a randomized trial in the Republic of Belarus. *Jama* 2001; **285**(4): 413–420.
15. Hirano K, Imbens GW, Rubin DB, Zhou XH. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 2000; **1**(1): 69–88.
16. Connell AM. Employing complier average causal effect analytic methods to examine effects of randomized encouragement trials. *Am J Drug Alcohol Abuse* 2009; **35**(4): 253–259.
17. Ertefaie A, Small D, Flory J, Hennessy S. Application of a sensitivity analysis to assess bias in IV analyses due to selecting subjects based on treatment received. *Epidemiology* 2016; **27**(2).
18. McConnell KJ, Newgard CD, Mullins RJ, Arthur M, Hedges JR. Mortality benefit of transfer to level I versus level II trauma centers for head-injured patients. *Health Serv Res* 2005; **40**: 435–458.
19. Shetty KD, Vogt WB, Bhattacharya J. Hormone replacement therapy and cardiovascular health in the United States. *Med Care* 2009; **47**(5): 600–606.
20. Hampp C, Borders-Hemphill V, Moeny DG, Wysowski DK. Use of antidiabetic drugs in the US, 2003–2012. *Diabetes Care* 2014; **37**(5): 1367–1374.
21. Brookhart MA, Wang P, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006; **17**(3): 268.
22. Ionescu-Ittu R, Abrahamowicz M, Pilote L. Treatment effect estimates varied depending on the definition of the provider prescribing preference-based instrumental variables. *J Clin Epidemiol* 2012; **65**(2): 155–162.
23. Hennessy S, Leonard CE, Palumbo CM, Shi X, Ten Have TR. Instantaneous preference was a stronger instrumental variable than 3-and 6-month prescribing preference for NSAIDs. *J Clin Epidemiol* 2008; **61**(12): 1285–1288.
24. Abrahamowicz M, Beauchamp ME, Ionescu-Ittu R, Delaney JA, Pilote L. Reducing the variance of the prescribing preference-based instrumental variable estimates of the treatment effect. *Am J Epidemiol* 2011; **174**(4): 494–502.
25. Pracht EE, Tepas JJ, 3rd, Langland-Orban B, Simpson L, Pieper P, Flint LM. Do pediatric patients with trauma in Florida have reduced mortality rates when treated in designated trauma centers? *J Pediatr Surg* 2008; **43**(1): 212–221.
26. Baiocchi M, Small D, Lorch S, Rosenbaum P. Building a stronger instrument in an observational study of perinatal care for premature infants. *J Am Stat Assoc* 2010; **105**(492): 1285–1296.
27. Groenwold RHH, Hak E, Hoes AW. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. *J Clin Epidemiol* 2009; **62**(1): 22–28.
28. Harris KM, Remler DK. Who is the marginal patient? Understanding instrumental variables estimates of treatment effects. *Health Serv Res* 1998; **35**: 5.

29. Brookhart MA, Patrick AR, Dormuth C, Avorn J, Shrank W, Cadarette SM, Solomon DH. Adherence to lipid-lowering therapy and the use of preventive health services: an investigation of the healthy user effect. *Am J Epidemiol* 2007; **166**(3): 348–354.
30. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996; **91**(434): 444–455.
31. Balke A, Pearl J. Bounds on treatment effects for studies with imperfect compliance. *J Am Stat Assoc* 1997; **92**(439): 1171–1176.
32. Tan Z. Marginal and nested structural models using instrumental variables. *J Am Stat Assoc* 2010; **105**(489): 157–169.
33. Flory JH, Small DS, Cassano PA, Brillon DJ, Mushlin AI, Hennessy S. Comparative effectiveness of oral diabetes drug combinations in reducing glycosylated hemoglobin. *J Comp Eff Res* 2014; **3**: 29–39.
34. Forst T, Hanefeld M, Jacob S, Moeser G, Schwenk G, Pfützner A, Haupt A. Association of sulphonylurea treatment with all-cause and cardiovascular mortality: a systematic review and meta-analysis of observational studies. *Diab Vasc Dis Res* 2013.
35. Stock JH, Wright JH, Yogo M. A survey of weak instruments and weak identification in generalized method of moment. *J Bus Econ Stat* 2002; **20**(4): 518–529.
36. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc* 1995; **90**(430): 443–450.
37. Imbens GW, Rosenbaum PR. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *J R Stat Soc Series B* 2005; **168**(1): 109–126.
38. Small D, Rosenbaum P. War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *J Am Stat Assoc* 2008; **103**: 924–933.
39. Bound J, Jaeger D, Baker R. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. *J Am Stat Assoc* 1995; **90**: 443–450.
40. Ionescu-Ittu R, Delaney JA, Abrahamowicz M. Bias–variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: a simulation study. *Pharmacoepidemiol Drug Saf* 2009; **18**(7): 562–571.
41. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat-Simul Comm* 2009; **38**(6): 1228–1234.
42. Brookhart M, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *Int J Biostat* 2007; **14**: 1–23.
43. Jackson JK, Swanson SA. Toward a clearer portrayal of confounding bias in instrumental variable applications. *Epidemiology* 2015; **26**(4): 498–504.
44. Alsheikh-Ali AA, Abourjaily HM, Karas RH. Risk of adverse events with concomitant use of atorvastatin or simvastatin and glucose-lowering drugs (thiazolidinediones, metformin, sulfonylurea, insulin, and acarbose). *Am J Cardiol* 2002; **89**(11): 1308–1310.
45. Swerdlow DI, Preiss D, Kuchenbaecker KB, *et al.* HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. *The Lancet* 2015; **385**(9965): 351–361.
46. Wald A. The fitting of straight lines if both variables are subject to error. *Ann Math Stat* 1940; **11**: 284–300.
47. Imbens G, Angrist J. Identification and estimation of local average treatment effects. *Econometrica* 1994; **62**: 467–475.
48. Davidson R, MacKinnon J. *Estimation and Inference in Econometrics*. Oxford University Press: New York, 1993.
49. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2010.
50. Forst T, Hanefeld M, Jacob S, Moeser G, Schwenk G, Pfützner A, Haupt A. Association of sulphonylurea treatment with all-cause and cardiovascular mortality: a systematic review and meta-analysis of observational studies. *Diab Vasc Dis Res* 2013; **10**(4): 302–314.
51. Small D. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J Am Stat Assoc* 2007; **102**: 1049–1058.
52. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol* 1996; **25**(6): 1107–1116.
53. Bhattacharya J, Goldman D, McCaffrey D. Estimating probit models with self-selected treatments. *Stat Med* 2006; **25**(3): 389–413.
54. Vansteelandt S, Bowden J, Babnezhad M, Goetghebeur E. On instrumental variables estimation of causal odds ratios. *Stat Sci* 2011; **26**(3): 403–422.
55. Clarke PS, Windmeijer F. Instrumental variable estimators for binary outcomes. *J Am Stat Assoc* 2012; **50**: 1638–1652.
56. Blundell RW, Powell JL. Endogeneity in semiparametric binary response models. *Rev Econ Stud* 2004; **71**: 655–679.
57. Palmer TM, Thompson JR, Tobin MD, Sheehan NA, Burton PR. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *Int J Epidemiol* 2008; **37**: 1161–1168.
58. Keele, L., Small, D., & Grieve, R. Randomization based instrumental variables methods for binary outcomes with an application to the IMPROVE Trial (2014).
59. Hirano K, Imbens G, Rubin D, Zhou X. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 2000; **1**(1): 69–88.
60. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J Health Econ* 2008; **27**(3): 531–543.
61. Small DS, Rosenbaum PR. War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *J Am Stat Assoc* 2008; **103**(483): 924–933.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.