# The Missing Cause Approach to Unmeasured Confounding in Pharmacoepidemiology

Michal Abrahamowicz[1,2], Lise M. Bjerre[3,4,5], Marie-Eve Beauchamp[2], Jacques LeLorier[6,7], Rebecca Burne[1]

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada
[2] Division of Clinical Epidemiology, McGill University Health Centre, Montreal, QC, Canada
[3] Department of Family Medicine, University of Ottawa, Ottawa, ON, Canada
[4] School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, ON, Canada
[5] Bruyère Research Institute, Ottawa, ON, Canada
[6] Departments of Medicine & Pharmacology, University of Montreal, Montreal, QC, Canada
[7] Pharmacoepidemiology and pharmacoeconomics, University of Montreal Hospital Research Centre, Montreal, QC, Canada

*Correspondence to:* Michal Abrahamowicz, Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, QC, Canada.
michal.abrahamowicz@mcgill.ca

**Running head: Missing Cause Approach to Unmeasured Confounding**

**Abstract:** Unmeasured confounding is a major threat to the validity of pharmacoepidemiological studies of medication safety and effectiveness. We propose a new method for detecting and reducing the impact of unobserved confounding in large observational database studies. The method uses assumptions similar to the prescribing preferences-based instrumental variables (IV) approach. Our method relies on the new "missing cause" principle, according to which the impact of unmeasured confounding by (contra-)indication may be detected by assessing discrepancies between a) treatment actually received by individual patients and b) treatment they would be expected to receive based on the observed data. Specifically, we use the treatment-by-discrepancy interaction to test for the presence of unmeasured confounding and correct the treatment effect estimate for the resulting bias. Under standard IV assumptions, we first proved that unmeasured confounding induces a spurious treatment-by-discrepancy interaction in risk difference models for binary outcomes, and then simulated large pharmacoepidemiological studies with unmeasured confounding. In simulations, our estimates had four to six times smaller bias than conventional treatment effect estimates, adjusted only for measured confounders, and much smaller variance inflation than unbiased but very unstable IV estimates, resulting in uniformly lowest root mean square errors. The much lower variance of our estimates, relative to IV estimates, was also observed in an application comparing gastrointestinal safety of two classes of anti-inflammatory drugs. In conclusion, our missing cause-based method may complement other methods and enhance accuracy of analyses of large pharmacoepidemiological studies.

**Keywords**: pharmacoepidemiology, unobserved confounding, instrumental variables, bias, simulations.

## 1. Introduction

Accurate assessment of adverse effects of drugs is a methodologically challenging issue of major importance for public health. To ensure adequate sample size and follow-up duration, and avoid restrictive inclusion criteria of randomized trials, adverse drug effects are typically assessed using large administrative health databases [1, 2], which do not record lifestyle and clinical characteristics often affecting both the treatment choice and the adverse event risk [3]. The resulting unobserved confounding by indication is a major threat to the validity of observational pharmacoepidemiological studies of drug safety [4-6]. Therefore, it is important to develop and validate methods to control for, or at least reduce, bias due to unmeasured confounding [7].

A widely used approach to deal with unmeasured confounding involves instrumental variables (IV) [8-11]. Brookhart et al adapted this approach to pharmacoepidemiology by defining the IV as the physicians' subjective prescribing preferences [12, 13]. Such preferences are well documented in drug utilization studies [14-16]. The prescribing preference-based IV approach is increasingly employed in pharmacoepidemiology [14, 17, 18] and has stimulated further methodological investigations [19-23]. Under the standard IV assumptions [8, 24], the prescribing preference-based IV estimates remove even strong unmeasured confounding [22, 25]. However, the associated variance inflation implies that the mean square error (MSE) of IV estimates may be higher than the MSE of "conventional", biased but much more stable, estimates that adjust only for measured confounders [25]. Furthermore, even in very large database studies, it may be difficult to establish whether IV estimates are significantly different from the conventional estimates [20, 26-28], which may make researchers reluctant to base their conclusions on the less stable IV estimates [23, 29].

We propose a new method that may help detect and reduce the impact of unmeasured confounding. We adapt most of the standard IV assumptions [8] and, similar to IV applications in pharmacoepidemiology [12, 18], rely on prescribing preferences. However, in contrast to conventional IV analyses, in the final outcome model, we do *not* replace the actual treatment by the IV. Section 2 describes our conceptual framework and outlines an analytical proof of the underlying "missing cause" principle. Section 3 describes how our method is implemented. Simulations in section 4 assess the performance of our estimates and compare them to the prescribing preference-based IV estimates. A discussion of the implications and limitations of our work concludes the manuscript.

## 2. Conceptual framework

### *2.1. The missing cause principle*

Detection of unobserved confounding bias requires some analytical "detective" investigation. Detective work often focuses on apparently unexplained discrepancies between the observed facts and rational expectations to deduce unobserved causes or motives for the crime. We adapt this line of investigation to pharmacoepidemiology, and compare observed data on treatment actually received by individual patients with their expected treatment, to detect possible discrepancies, which may help unmask and correct for unobserved confounding.

Consider two hypothetical patients who have identical values for all measured covariates but received different treatments. In conventional multivariable analyses, which only match on or adjust for the observed variables, any difference between the outcomes of the two patients will be attributed to the difference in their respective treatment. However, there should be some reason why two apparently "identical" patients received different treatments. This may be due to different subjective prescribing preferences of their physicians and/or some objective differences in the patients' characteristics *not* recorded in the database, which would represent the "missing cause(s)" to justify the different treatment choices. A failure to account for such unobserved differences will induce unmeasured confounding bias if the missing cause is also associated with the outcome.

We now introduce the missing cause principle, which formalizes the above reasoning. In general, the observed treatment decisions depend on a combination of patients' characteristics, whether recorded or not in the study database, and physicians' subjective prescribing preferences. For individual patients, we cannot determine if some unrecorded variables did affect their treatment (e.g., choice of drug A vs. B). However, our missing cause principle postulates that the probability that a given patient's treatment has been partly chosen based on unobserved characteristics increases if the assigned treatment appears inconsistent with the treatment expected based on the patient's observed characteristics and his/her physician's prescribing preference. Physicians' preferences may be approximately quantified, based on the observed treatments prescribed to their patients [12, 15, 22]. Assuming that treatment decisions are rational, patients with higher discrepancy between observed and expected treatments are more likely to have some unobserved characteristics which may represent the missing cause(s) for their treatment assignment. Accordingly, the missing cause principle implies that the prevalence of unobserved characteristics associated with the choice of drug A increases for those users of drug A for whom this treatment choice appears more discrepant from the expectations based on observed characteristics and their physicians' preferences. In contrast, the presence of unobserved characteristics associated with drug A will be less likely among more discrepant users of drug B, as such characteristics would further decrease the probability of receiving drug B. Accordingly, the difference in the prevalence of unobserved determinants of treatment choice between the groups of patients prescribed drug A vs. drug B will tend to increase with increasing discrepancy between the observed and the expected treatments.

Notice that the difference in the distribution(s) of unobserved missing cause(s) of treatment choice between users of the two drugs will entail a confounding bias whenever these unobserved characteristics are also related to the outcome, and the resulting bias will increase with increasing difference in the prevalence of such umeasured confounder(s).

### *2.2. Proof of the missing cause principle*

We now provide a more formal proof of the missing cause principle for (linear) risk difference (RD) models, typically employed to implement IV methods in pharmacoepidemiology [12, 18, 22]. We adopt the classic IV assumptions [8, 13] and discuss them briefly below using the terminology of Angrist et al [8].

1) Standard unit treatment value assumption (SUTVA) implies that the potential outcome of a patient does not depend on the treatment assigned to any other patient, which seems rather incontestable in our context.

2) Exclusion restriction assumption implies that physicians' prescribing preferences are not independently associated with the outcome, except for their effects mediated through treatment assignment. The practical implications of this assumption are that prescribing preferences are 2a) independent of both unmeasured and measured patients' characteristics, including potential confounders of the treatment effect, and 2b) are not correlated with the quality of care provided by different physicians, which may affect the outcomes of their patients regardless of patients' characteristics.
3) Non-zero average causal effect assumption implies that the probability of receiving a given treatment varies among patients with identical characteristics but treated by different physicians. This is supported by well-documented subjective prescribing preferences and lies at the core of the IV approaches in pharmacoepidemiology [9, 12, 14, 15, 22, 26].

Furthermore, similar to IV applications in pharmacoepidemiology [12, 18], we also assume that unobserved confounders do not act as modifiers of the treatment effect.

Finally, in our proof and simulations in section 4, we do not rely on the monotonicity assumption [8], whose plausibility in the context of prescribing preferences has been questioned [30]. Indeed, in simulations, individual treatments are randomly generated from the patient-specific Bernoulli distribution with probability of treatment $A = 1$ that depends on both patients' characteristics and their physician's prescribing preference [22, 25]. Thus, while the expected probability of receiving treatment $A = 1$ is a monotone function of physicians' preferences, actual treatments assigned to individual patients may be inconsistent with the monotonicity assumption, due to sampling variance.

The impact of violating the exclusion restriction and the assumption that treatment effect is not modified by unobserved confounders [13], which both are difficult to verify empirically, is investigated in simulations reported in section 4.4.

Under the above assumptions, we outline a formal proof of the missing cause principle. Section A of Supplementary Materials provides a detailed proof. We consider a hypothetical study comparing binary outcomes $Y$ between users of two treatments ($A = 1$ vs. $A = 0$). Both $A$ and $Y$ arise from their respective RD models and both are affected by a binary unmeasured confounder $U$. Consistent with the non-zero average effects assumption [8], $A$ depends also on an observed continuous variable $Z$, which is not independently associated with either $Y$ or $U$ (exclusion restriction [8]), i.e. may serve as an instrument. In the analyses, both treatment and outcome models are correctly specified, except for (unobserved) $U$. We define the *treatment discrepancy D* as the probability, estimated conditional on $Z$, that a subject will receive the treatment opposite to his/her actual treatment (equation (2)). Then, in section A of Supplementary Materials, we prove that the difference in the prevalence of the unobserved confounder $U$ between subgroups $A = 1$ vs. $A = 0$, $P(U = 1|A = 1, D) - P(U = 1|A = 0, D)$, increases monotonically with increasing treatment discrepancy $D$. Thus, if $U = 1$ is associated with higher risk, the estimated RD also increases with increasing $D$, which validates the missing cause principle. In contrast, the RD will not change systematically with increasing $D$ if $U$ is not associated with either $A$ or $Y$.

## 3. Methods

We propose a new method for detecting and reducing the impact of unobserved confounding, which relies on the missing cause principle, for linear regression modeling of either the effect of a binary treatment on a continuous outcome or a risk difference for a binary outcome. Its implementation involves four steps.

*Step 1: Estimating the expected probability of treatment*

To estimate the expected probability of patient *j* ($j = 1,\ldots, n_i$) of physician *i* ($i = 1,\ldots, m$) receiving treatment $A_{ij} = 1$, we fit the multivariable linear RD model to data on all patients:

$$P\left(A_{ij} = 1 \middle| X_{ij,1}, \ldots, X_{ij,p,} M_{ij}\right) = \sum_{k=1}^{p} \beta_k X_{ij,k} + \gamma_i M_{ij} + \varepsilon_{ij} \qquad (1)$$

where $X_{ij,k}$ ($k = 1,\ldots, p$) is the value of covariate $X_k$ for patient *j* of physician *i*, $\varepsilon_{ij}$ is a Gaussian error term, and $M_{ij}$ ($i = 1,\ldots, m$) are *m* dummy indicators of individual physicians ($M_{ij} = 1$ for all patients of physician *i*). Accordingly, $\gamma_i$ estimates the preference of the $i^{\text{th}}$ physician for $A = 1$, independent of patients' characteristics. To avoid the violation of the positivity assumption, the estimated probabilities below 0.001 or above 0.999 are truncated and replaced by the respective boundary value.

*Step 2: Estimating treatment discrepancy*

Next, we estimate the discrepancy $D_{ij}$ between the treatment $A_{ij}$ actually received by a patient and his/her expected probability of receiving this treatment, estimated in (1):

$$D_{ij} = \begin{cases} 1 - \hat{P}\left(A_{ij} = 0\right) = \hat{P}(A_{ij} = 1) & for\ A_{ij} = 0 \\ \quad 1 - \hat{P}\left(A_{ij} = 1\right) & for\, A_{ij} = 1 \end{cases} \qquad (2)$$

Higher values of $D_{ij}$ in (2) indicate more "discrepant" patients.

*Step 3: Modeling and testing the treatment-by-discrepancy interaction in the outcome model*

Next, we expand the outcome model to account for potential effects of the treatment discrepancy $D_{ij}$ estimated in (2). Based on Lemma 2 and Theorem 1 (see section A of Supplementary Materials), we assume that, in the presence of unmeasured confounding, the risk of a binary outcome P($Y_{ij} = 1$) (or the expected value of a continuous outcome) changes monotonically with increasing $D_{ij}$, but in the *opposite* direction in the two treatment groups. This implies an interaction $f(D_{ij})A_{ij}$ between the treatment indicator $A_{ij}$ and a monotone function of discrepancy $f(D_{ij})$. For example, for binary outcomes, we fit the following multivariable risk difference model:

$$\begin{aligned} &P\left(Y_{ij} = 1 \middle| X_{ij,1}, \ldots, X_{ij,p}, A_{ij}, D_{ij}\right) \\ &\qquad = \alpha_0 + \sum_{k=1}^{p} \alpha_{X_k} X_{ij,k} + \alpha_A A_{ij} + \eta f\left(D_{ij}\right) + \theta f\left(D_{ij}\right) A_{ij} + \varepsilon_{ij} \end{aligned} \qquad (3)$$

where $Y_{ij}$ is the patient's outcome, conditional on his/her $p$ observed covariates ($X_{ij,1}, \ldots, X_{ij,p}$) and the received treatment $A_{ij}$, and $\varepsilon_{ij}$ is a Gaussian error. Equation (3) does not include the physician indicators $M_{ij}$ because prescribing preferences are assumed to have no independent effects on the outcome [12, 13]. A similar model is fitted to predict the expected value of a continuous outcome.

In real-life applications, the functional form $f(D_{ij})$ for the effect of increasing $D_{ij}$ on the outcome is analytically intractable, as it depends on (unknown) prevalence of unmeasured confounders and their impact on the treatment choice. Yet, section 2.2 demonstrates that $f(D_{ij})$ is monotone. In simulations in section 4, using $f(D_{ij}) = log(D_{ij}+1)$ improved the accuracy of the estimates relative to alternative simple monotone functions.

In equation (3), the coefficient $\eta$ for discrepancy $f(D_{ij})$ quantifies its impact on $P(Y_{ij} = 1)$ in group $A_{ij} = 0$, while the coefficient $\theta$ for treatment-by-discrepancy interaction $f(D_{ij})A_{ij}$ captures the differential effect of $D_{ij}$ for group $A_{ij} = 1$. According to section 2.2, under the assumptions stated therein, the effect of discrepancy $D_{ij}$ should vary across the two treatment groups if *both* the treatment choice and outcome are affected by unmeasured confounders. Therefore, we propose to test the $f(D_{ij})A_{ij}$ interaction in (3) with the model-based two-tailed t-test at $\alpha = 0.05$. In (3), adjusted RD for $A_{ij} = 1$ versus $A_{ij} = 0$ equals $\alpha_A A_{ij} + \theta f(D_{ij})A_{ij}$. Thus, the rejection of $H_0$: $\theta = 0$ implies that treatment effect does change systematically with increasing discrepancy $D_{ij}$ which, according to Theorem 1, indicates the presence of unobserved confounding.

*Step 4: Estimating the corrected treatment effect*

Equation (3) may also help correct the treatment effect estimate for unobserved confounding. Given the $f(D_{ij})A_{ij}$ interaction, the coefficient $\alpha_A$ for $A_{ij}$ in (3) estimates the treatment effect (adjusted for observed covariates) for hypothetical patients with $f(D_{ij}) = 0$. It is convenient to define $f$ so that $f(0) = 0$, which holds e.g. for $f(D_{ij}) = log(D_{ij}+1)$. Then, $\alpha_A$ estimates the adjusted treatment effect among patients who have (a) the same values for all observed covariates and (b) estimated $D_{ij} = 0$, i.e. the estimated probability of receiving their actual treatment is equal to 1.

Yet, according to equation (1), different patients with the *same* covariate vector will be assigned either $\hat{P}(A_{ij} = 0) = 1$ or $\hat{P}(A_{ij} = 1) = 1$ only if their respective treatments are entirely determined by their physicians' deterministic preferences, i.e. independent of individual characteristics, whether observed or not. Thus, assuming physicians' preferences are not independently associated with the outcome [12], $\alpha_A$ in (3) approximates the treatment effect estimated from a hypothetical cluster randomized trial, in which all patients of a given physician are randomized to either $A_{ij} = 1$ or $A_{ij} = 0$. In conclusion, $\alpha_A$ estimated in our interaction model in equation (3) should approximate the unbiased (causal) treatment effect, not affected by unmeasured confounders.

## 4. Simulation studies

*4.1 Simulation design and data generation*

To evaluate our method's performance, we simulated hypothetical studies comparing the risk of an adverse event between two drugs ($A = 1$ vs. $A = 0$) in large databases, in the presence of unmeasured confounding. We assumed *m* prescribing physicians (*m* = 400, 600, or 1,000), with 10 to 50 patients per physician, implying total N of about 12,000 to 30,000.

For each patient, we generated two observed confounders (continuous $X_1$ and binary $X_2$) and an unmeasured continuous confounder $X_3$. Consistent with stage 1 of the two-stage least squares (2SLS) IV estimation [12], the binary treatment was generated from a linear RD model. For patient $j$ $(j = 1,..., n_i)$ of physician $i$ $(i = 1,..., m)$, P($A_{ij}$ = 1) depended on $X_{ij,1}$-$X_{ij,3}$ and on the latent physician preference $\gamma_i$ (see section B.1.2 of the Supplementary Materials):

$$P\left(A_{ij} = 1 \middle| X_{ij,1}, X_{ij,2}, X_{ij,3}, \gamma_i\right) = \sum_{k=1}^{3} \beta_k X_{ij,k} + \gamma_i \tag{4}$$

In different simulation scenarios, either a continuous or a binary outcome was generated conditional on $X_{ij,1}$-$X_{ij,3}$ and on the received treatment $A_{ij}$, but independent of the physicians' preferences [12]. The binary outcome (adverse event) was generated from the RD model, consistent with stage 2 of the 2SLS estimation [12]:

$$P\left(Y_{ij} = 1 \middle| X_{ij,1}, X_{ij,2}, X_{ij,3}, A_{ij}\right) = \alpha_0 + \sum_{k=1}^{3} \alpha_k X_{ij,k} + \alpha_A A_{ij} \tag{5}$$

Continuous, normally distributed outcome $Y_{ij}$ was generated from the multivariable linear model:

$$Y_{ij} = \alpha_0 + \sum_{k=1}^{3} \alpha_k X_{ij,k} + \alpha_A A_{ij} + \varepsilon_{ij} \tag{6}$$

with normally distributed error $\varepsilon_{ij}$.

Across the simulated scenarios, we varied i) N, ii) assumptions regarding physicians' preferences ($\gamma_i$ in (4)), iii) true RD for treatment ($\alpha_A$ in (5) and (6)), and iv) the strength of the unmeasured confounding, by varying the parameters for unobserved $X_3$, i.e. $\beta_3$ in (4) and/or $\alpha_3$ in (5) or (6).

In main simulations, we assumed that i) classic IV assumptions hold [8], ii) physicians' preferences do not change over time, and iii) in both treatment and outcome models, respectively equations (1) and (3), the effects of measured covariates, physicians' preferences (for treatment model) and treatment (for outcome model) are correctly specified. Section 4.4 summarizes methods and results of sensitivity analyses which investigated the impact of violation of these assumptions.

Details of data generation methods and underlying assumptions are given in section B.1 of the Supplementary Materials, where Table A.1 shows all relevant parameters. For each simulated scenario, we generated 1000 independent random samples.

Simulations were performed with R version 3.1.1 [31]. The program is available in section D of Supplementary Materials.

### *4.2. Analyses of simulated data*

*Estimation models:* Each sample was analyzed using alternative multivariable linear regression models, including RD models for binary outcomes [12]. All models ignored the unmeasured confounder $X_3$ and adjusted for $X_1$ and $X_2$. Conventional Model 1 included only treatment indicator $A$, $X_1$ and $X_2$. Two prescribing preference-based IV models were estimated using the 2SLS approach [8]. The IV was defined as i) treatment received by the previous patient of the same physician [12] in binary IV Model 2, and ii) the proportion of all previous patients of the same physician prescribed treatment $A = 1$ [22] in continuous IV Model 3. Both Models 4 and 4A implemented our interaction-based model in equation (3), with $f(D_{ij}) = log(D_{ij}+1)$ in Model 4 vs. $f(D_{ij}) = D_{ij}$ in Model 4A.

*Power to detect unmeasured confounding* was estimated with the proportion of simulated samples in which the 95% confidence intervals (CIs) for the model-specific treatment effect excluded the naïve conventional Model 1 estimate. This avoided the analytical difficulties in using, e.g., the Durbin-Wu-Hausman test [32-34] for IV analyses of binary outcomes [23, 29]. In addition, for our Models 4/4A, we estimated how often the treatment-by-discrepancy interaction in (3) was significant at two-tailed $\alpha = 0.05$ (step 3 of section 3).

### *4.3. Simulation results*

Using $f(D_{ij}) = log(D_{ij}+1)$ in equation (3) systematically reduced bias and RMSE of Model 4 estimates, compared to Model 4A with $f(D_{ij}) = D_{ij}$ (data not shown). Therefore, we present results only for $f(D_{ij}) = log(D_{ij}+1)$. Across the simulated scenarios, the estimated treatment probabilities fell outside the [0.01, 0.99] interval and, thus, had to be truncated for only 0% to 4.4% of subjects.

Table 1 compares the power to detect unmeasured confounding for CI-based tests (described in section 4.2) for Models 2-4 and the interaction test for Model 4 across twelve simulation scenarios outlined in columns 2-4 (Table A.1 in Supplementary Materials provides details). With no unmeasured confounding, all type I error rates were close to nominal 5% (scenarios 1-2). In other scenarios, the power was systematically the lowest for the binary IV Model 2 (column 6) and the highest for our Model 4 (columns 8-9). Power increased with sample size (column 3) and strength of unmeasured confounding (column 4), but was only moderate even in very large datasets for both IV models (columns 6-7). The proposed test of the treatment-by-discrepancy interaction improved the power, over the CI-based tests, without inflating type I error (last column).

TABLE 1

Table 2 compares the point estimates of adjusted treatment effect. Both IV estimates were unbiased (data not shown), as expected given that data were generated in accordance with the standard IV assumptions [8]. Bias of Model 4 estimates was small, below or close to 10%, and four to six times lower than the bias of the conventional Model 1 estimates (columns 5 vs. 4, scenarios 3-12). Binary IV estimates were extremely unstable, with two to four times larger SDs than Model 4 (column 7), while continuous IV Model 3 yielded SDs about 60% higher than Model 4 (columns 8). The RMSE ratios for both IV models, relative to Model 4, were always much above 1 (columns 10-11), indicating that our estimates are, on average, substantially closer to the true treatment effect. In scenarios 3-12 with unmeasured confounding, our Model 4 yielded the lowest RMSE, at least 50% below the conventional Model 1 (column 9), but its relative advantages over alternative models reflect the bias/variance trade-off: biased but stable Model 1 estimates performed worst in large databases with strong unmeasured confounding, where unbiased IV estimates were less affected by variance inflation.

TABLE 2

Table 3 focuses on the accuracy of inference about the treatment effect. As expected, both unbiased IV Models 2 and 3 yielded uniformly correct coverage of the 95% CIs (data not shown), while conventional Model 1 had extremely low coverage in presence of unmeasured confounding (column 8, scenarios 3-12). Model 4 yielded coverage rates above 90% (column 9), except for the continuous outcome scenario 12, where a very narrow CI, due to N = 30,000, resulted in suboptimal 84% coverage. Yet, the CIs for Model 4 were substantially more precise than for IV-based Models 2-3, which yielded much larger SEs (columns 5-6 vs. 7). Our approach also offered excellent power to reject the null hypothesis of no treatment effect (column 13), in contrast to generally low power for binary IV Model 2 (column 11) and occasionally even for the continuous IV Model 3 (column 12, scenarios 2, 3, 5 and 8).

TABLE 3

### *4.4. Sensitivity analyses*

To assess the robustness of the above findings and conclusions, Table 4 reports the results of additional simulations. Each scenario shown in Table 4 modified specific assumption(s) or parameter(s) used to generate data for the original scenario 6 of Tables 1-3 (see column 2 of Table 4, with details in section B.2 in Supplementary Materials).

Scenarios 6a-6c show that even in the case of effect modification by either an observed or unobserved confounder, both IV Model 3 and our Model 4 yielded as unbiased estimates of the average (across the values of effect modifier) treatment effect as in scenario 6. Importantly, when unmeasured $X_3$ acted as effect modifier but not a confounder, our interaction-based test of confounding had a correct type I error rate (scenario 6c, column 3). Furthermore, the results were not affected by different misspecifications of the model used to estimate the expected treatment, including ignoring an interaction or a non-linear effect of an observed covariate or including variables that do not affect treatment (scenarios 6d-6f). In contrast, the results changed when estimating treatment probability with the (misspecified) linear RD model when treatment was in fact generated from the logistic model (scenario 6g), or assuming that physicians' prescribing preferences either were much weaker or changed over time (scenarios 6h-6i). The variance increased for Models 2-4, and the power of our interaction-based test for detecting unmeasured confounding decreased substantially (column 3), even if it was still higher than for CI-based tests (see section 4.2) for IV models (data not shown). The bias of Model 4 estimates also increased (column 6), but RMSEs of all models were at least 50% higher than for Model 4 (columns 7-9).

Scenario 6j shows that lowering the number of patients in physicians' practices from 10-50 to 2-50 had only a minor impact. However, reducing the maximum number of patients substantially, from 50 to 20 (scenarios 6k-6l), decreased the precision of physicians' preferences estimates in equation (1). This implied less accurate estimation of discrepancies in equation (2), resulting in increased bias of Model 4 estimates, even if this bias was systematically at least 50% smaller than for conventional Model 1 (columns 6 vs. 4 of Table 4). In scenario 6l, bias and coverage of our estimates improved if patients from smaller physicians' practices (<10 patients) were excluded from the analyses. In all scenarios 6j-6l, our Model 4 estimates always yielded RMSEs at least 39% lower than for all other models (columns 7-9).

Finally, scenarios 6m-6p demonstrated that in situations where the exclusion restriction assumption [8] was violated, both the IV Models 2 and 3 and our Model 4 yielded biased estimates, with no systematic difference between these three models. The bias varied from below 15% to above 50% depending if the distribution of unmeasured confounder $X_3$ varied slightly or substantially (ICC = 0.012 to ICC = 0.155) at random, across physicians' practices (scenarios 6m-6n). The bias became even more dramatic if the mean value of $X_3$ was correlated with physicians' preferences (scenarios 6o-6p). Thus, exclusion restriction violation affects our Model 4 estimates as strongly as IV estimates [8, 13, 30].

TABLE 4

## 5. Risk of gastrointestinal events in COX-2 vs. NSAID users

To illustrate the performance of Models 1-4 (section 4) with real data, we compared the risk of gastrointestinal (GI) side effects between users of more recently introduced cyclooxygenase-2 (COX-2) inhibitors vs. traditional non-steroidal anti-inflammatory drugs (NSAIDs). All models adjusted for the same *a priori* selected observed potential confounders. Section C of the Supplementary Materials provides details of study design, definitions of exposure, outcome and confounders, and detailed results.

During the 6-month follow-up, 33.8% (1,513/4,475) of COX-2 vs. 27.4% (1,184/4,318) of NSAID users had a GI event, yielding unadjusted RD = 6.4% (95% CI: 4.5%, 8.3%). However, when adjusted for systematically worse values of several observed GI risk factors among COX-2 users (Supplementary Materials Table A.2), the conventional multivariable Model 1 suggested lower GI risks for COX-2 users (adjusted RD = -1.9% (-3.5%, -0.3%)). Individual physicians' treatment preferences varied substantially (Supplementary Materials section C.3). Both IV Models 2 and 3 suggested GI risk *increases* for COX-2 users, but with very wide CIs (Table 5). Model 4 yielded a substantially more precise estimate and suggested a risk reduction (adjusted RD = -2.6% (-8.4, 3.2%)), similar to the modest risk reduction reported by meta-analyses of relevant trials [35, 36] and our conventional RD = -1.9%. Indeed, both inclusion of RD = -1.9% in all 95% CIs and p-value = 0.75 for our interaction test (Table 5) consistently indicated the absence of marked unmeasured confounding, suggesting that the extensive list of measured confounders, adjusted for in our multivariable analyses, might have accounted for possible confounding by indication.

TABLE 5

## 6. Discussion

Instrumental variables, based on prescribing preferences, represent one of the most promising approaches to deal with unmeasured confounding in pharmacoepidemiology [8, 13, 15]. In linear models, with a directly observable instrument, which meets standard assumptions, 2SLS IV estimates are asymptotically efficient [37]. However, the IV defined as prescribing preferences is latent and efficiency of alternative estimators may vary depending on how the available information is used. Rather than replacing the observed treatment by the instrument [12, 13], or conditioning it on the instrument, which could exacerbate the unmeasured confounding bias [38], our missing cause approach relies on treatment interaction with discrepancy between observed versus expected treatments. Thus, it may be considered an adaptation of the two-stage residual inclusion (2SRI) IV approach proposed by Nagelkerke [39, 40] to pharmacoepidemiological database studies of the effect of a binary treatment on a binary outcome, in which the instrument (prescribing preferences) is *not* directly observable and, thus, needs to be estimated. Terza et al provide analytical and simulation-based results suggesting that, in non-linear models, the 2SRI estimates, which rely on residuals from the model regressing the observed exposure on the instrument, may be more accurate than standard IV 2SLS estimates [40].

The main advantages of our missing-cause-based estimates are that they substantially reduce both the bias of conventional estimates and the variance of unstable but unbiased IV estimates. In almost all simulations, regardless of whether standard IV assumptions were met, our treatment effect estimates had the lowest RMSEs (Tables 2 and 4), but at the price of a small bias and reduced coverage (Tables 3 and 4). Such bias-variance trade-off is typical of methods proposed to reduce the variance of unbiased but unstable estimates, including inverse probability of treatment (IPT) weights stabilization or truncation in marginal structural models [41-44]. The lowest RMSE implies that in applications, where unmeasured confounding is expected [45, 46], our estimates will be, on average, closer to the true unknown treatment effect than IV or conventional estimates. In the Cox-2 inhibitors vs. NSAIDs example, our missing-cause estimate, relative to IV estimates, was both more consistent with clinical trials results [35, 36] and more precise.

Sensitivity analyses indicated that heterogeneity of treatment effect and misspecification of covariate effects in the model used to estimate the expected treatment had no material impact on our estimates (Table 4). Whereas weaker prescribing preferences and/or reduced number of patients per physician resulted in slightly increased bias of our estimates, the RMSE remained much lower than for IV estimates. However, in contrast to the binary IV model [8], the accuracy of our missing-cause-based estimates decreased if physicians' prescribing preferences changed over time, although they still yielded the best RMSE (Table 4, scenario 6i). In applications with frequent changes in preferences, the change-point approach proposed for IV analyses [22] may be adapted to the missing cause framework.

We recognize that our method cannot deal with some complex data structures. Simulated scenarios (6m-6p) show that the impact of exclusion restriction violation on our estimates is as strong (Table 4) as its well-documented impact on IV estimates [8, 13, 30]. This critical assumption is violated even if the distribution of unmeasured confounder(s) varies at random, but substantially, across physician practices. Physicians with higher mean values of unmeasured risk factors associated with higher probability of prescribing a given drug will be estimated, on average, to have stronger preferences for this drug. This induces a spurious association between estimated prescribing preferences and the outcome, independent of the treatment received by individual patients.

Whereas the exclusion restriction assumption is not directly verifiable in applications, it may be useful to assess if the distribution of measured risk factors, separately within patients prescribed each drug, shows important clustering by physician and/or correlates with estimated physicians' preferences [13].

In simulations, the test of treatment-by-discrepancy interaction, based on equation (3), had systematically better power to detect unobserved confounding than tests based on the IV models (Table 1). However, the power to detect moderate unmeasured confounding was adequate only in very large databases, partly reflecting generally low power of interaction tests [47, 48]. Furthermore, because the power of interaction tests depends on the variance of both variables, assuming weaker physicians' preferences reduced the power of all tests and increased the variance of our and the IV estimates [25]. In conclusion, a significant treatment-by-discrepancy interaction likely reflects important unobserved confounding, but a non-significant interaction does not exclude moderate bias. In the latter case, comparison of alternative estimates, together with substantive considerations regarding the plausibility and expected direction of confounding bias, may help interpret the results.

Clearly, further analyses of real-life data and additional simulations, with a wider range of underlying assumptions, are necessary to fully assess the practical usefulness and validity of the proposed missing-cause method. Future empirical studies should also assess the distributions of prescribing preferences and help better understand their determinants. In applications, confidence intervals for treatment effects estimated with our interaction model in equation (3) and IV models should be based on bootstrap resampling, to account for the dependence of the estimates on regressors estimated from the same data: $D_{ij}$ in equation (3) or IV.

The unobserved confounding problem is too complex to expect any *single* method to detect and remove, or even largely reduce, its impact in most real-life pharmacoepidemiologic analyses. Therefore, future studies may consider using our missing cause-based method, along with instrumental variables [12, 18, 22, 23] and other recent methods, including high-dimensional propensity scores, propensity score calibration, marginal structural models, or bias sensitivity analyses (e.g., [49-53]). Careful interpretation and comparison of both results and formal properties and limitations of the alternative methods, together with substantive insights, will then help derive more accurate conclusions and assess their robustness.

## Acknowledgements

This research was supported by the Canadian Institutes for Health Research (grant MOP-81275) and the Natural Sciences and Engineering Research Council of Canada (grant 228203). Michal Abrahamowicz is a James McGill Professor at McGill University. Rebecca Burne is supported by a doctoral studentship from the McGill University Health Centre. The authors thank Lawrence McCandless, Marie-Pierre Sylvestre, Yishu Wang, Willy Wynant and Yongling Xiao, as well as the two anonymous Reviewers and the Associate Editor, for useful comments and discussions.

## References

1. Abrahamowicz M, Tamblyn R. Drug Utilization Patterns. In *Encyclopedia of Biostatistics,* 2nd edition, Armitage P, Colton T (eds). John Wiley & Sons, Ltd: Chichester, 2005;1533-1553.

2. Skegg DCG. Evaluating the safety of medicines, with particular reference to contraception. *Statistics in Medicine* 2001; **20**(23):3557-3569.

3. Wolfe F, Flowers N, Burke TA, Arguelles LM, Pettitt D. Increase in lifetime adverse drug reactions, service utilization, and disease severity among patients who will start COX-2 specific inhibitors: quantitative assessment of channeling bias and confounding by indication in 6689 patients with rheumatoid arthritis and osteoarthritis. *The Journal of Rheumatology* 2002; **29**(5):1015-1022.

4. Slone D, Shapiro SH, Miettinen OS, Finkle WD, Stolley PD. Drug evaluation after marketing. *Annals of Internal Medicine* 1979; **90**(2):257-261.

5. Walker AM. Confounding by indication. *Epidemiology* 1996; **7**(4):335-336.

6. Shapiro SH. Confounding by indication? (Letter to the editor). *Epidemiology* 1997; **8**(1):110-111.

7. McMahon AD. Approaches to combat with confounding by indication in observational studies of intended drug effects. *Pharmacoepidemiology and Drug Safety* 2003; **12**(7):551-558.

8. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**(434):444-472.

9. Terza JV, Bradford WD, Dismuke CE. The use of linear instrumental variables methods in health services research and health economics: A cautionary note. *Health Services Research* 2008; **43**(3):1102-1120.

10. Staiger D, Stock JH. Instrumental variables regression with weak instruments. *Econometrica* 1997; **65**(3):557-586.

11. Johnston KM, Gustafson P, Levy AR, Grootendorst P. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine* 2008; **27**(9):1539-1556.

12. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006; **17**(3):268-275.

13. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *International Journal of Biostatistics* 2007; **3**(1):Article 14.

14. Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. *The New England Journal of Medicine* 2008; **358**(8):771-783.

15. 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. *Journal of Clinical Epidemiology* 2008; **61**(12):1285-1288.

16. Abrahamowicz M, Fortin PR, du Berger R, Nayak V, Neville C, Liang MH. The relationship between disease activity and expert physician's decision to start major treatment in active systemic lupus erythematosus: A decision aid for development of entry criteria for clinical trials. *The Journal of Rheumatology* 1998; **25**:277-284.

17. Schneeweiss S, Solomon DH, Wang PS, Rassen J, Brookhart MA. Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: an instrumental variable analysis. *Arthritis & Rheumatology* 2006; **54**(11):3390-3398.

18. Chen Y, Briesacher BA. Use of instrumental variable in prescription drug research with observational data: a systematic review. *Journal of Clinical Epidemiology* 2011; **64**(6):687-700.

19. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. *Journal of Clinical Epidemiology* 2009; **62**(12):1233-1241.

20. Ionescu-Ittu R, Abrahamowicz M, Pilote L. Treatment effect estimates varied depending on the definition of the provider prescribing preference-based instrumental variables. *Journal of Clinical Epidemiology* 2012; **65**(2):155-162.

21. Brookhart MA, Rassen JA, Wang PS, Dormuth C, Mogun H, Schneeweiss S. Evaluating the validity of an instrumental variable study of neuroleptics: can between-physician differences in prescribing patterns be used to estimate treatment effects? *Medical Care* 2007; **45**(10 Supl 2):S116-S122.

22. Abrahamowicz M, Beauchamp M-E, Ionescu-Ittu R, Pilote L, Delaney JAC. Reducing the Variance of the Prescribing Preference-Based Instrumental Variable Estimates of the Treatment Effect. *American Journal of Epidemiology* 2011; **174**(4):494-502.

23. Guo Z, Cheng J, Lorch SA, Small DS. Using an instrumental variable to test for unmeasured confounding. *Statistics in Medicine* 2014; **33**(20):3528-3546.

24. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Statistics in Medicine* 2014; **33**(13):2297-2340.

25. Ionescu-Ittu R, Delaney JA, Abrahamowicz M. Bias-variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental

variables: a simulation study. *Pharmacoepidemiology and Drug Safety* 2009; **18**(7):562-571.

26. Rassen JA, Mittleman MA, Glynn RJ, Brookhart MA, Schneeweiss S. Safety and effectiveness of bivalirudin in routine care of patients undergoing percutaneous coronary intervention. *European Heart Journal* 2010; **31**(5):561-572.

27. Schneeweiss S, Setoguchi S, Brookhart A, Dormuth C, Wang PS. Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. *Canadian Medical Association Journal* 2007; **176**(5):627-632.

28. Davies NM, Smith GD, Windmeijer F, Martin RM. COX-2 selective nonsteroidal anti-inflammatory drugs and risk of gastrointestinal tract complications and myocardial infarction: an instrumental variable analysis. *Epidemiology* 2013; **24**(3):352-362.

29. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety* 2010; **19**(6):537-554.

30. Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 2006; **17**(4):360-372.

31. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria, 2014.

32. Durbin J. Errors in variables. *Review of the International Statistical Institute* 1954; **22**:23-32.

33. Wu DM. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 1973; **41**:733-750.

34. Hausman J. Specification tests in econometrics. *Econometrica* 1978; **41**:1251-1271.

35. Mallen SR, Essex MN, Zhang R. Gastrointestinal tolerability of NSAIDs in elderly patients: a pooled analysis of 21 randomized clinical trials with celecoxib and nonselective NSAIDs. *Current Medical Research and Opinion* 2011; **27**(7):1359-1366.

36. Bhala N, Emberson J, Merhi A et al. Vascular and upper gastrointestinal effects of non-steroidal anti-inflammatory drugs: meta-analyses of individual participant data from randomised trials. *Lancet* 2013; **382**(9894):769-779.

37. Wooldridge JM. Instrumental Variables Estimation of Single-Equation Linear Models. In *Econometric Analysis of Cross-Section and Panel Data,* The MIT Press: Cambridge, MA, 2002;83-114.

38. Pearl J. Invited commentary: understanding bias amplification. *American Journal of Epidemiology* 2011; **174**(11):1223-1227.

39. Nagelkerke N, Fidler V, Bernsen R, Borgdorff M. Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine* 2000; **19**(14):1849-1864.

40. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics* 2008; **27**(3):531-543.

41. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; **168**(6):656-664.

42. Xiao Y, Abrahamowicz M, Moodie EEM, Weber R, Young J. Flexible Marginal Structural Models for Estimating the Cumulative Effect of a Time-Dependent Treatment on the Hazard: Reassessing the Cardiovascular Risks of Didanosine Treatment in the Swiss HIV Cohort Study. *Journal of the American Statistical Association* 2014; **109**(506):455-464.

43. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**(5):561-570.

44. Xiao Y, Moodie EEM, Abrahamowicz M. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods* 2013; **2**(1):1-20.

45. Solomon DH, Lunt M, Schneeweiss S. The risk of infection associated with tumor necrosis factor alpha antagonists: making sense of epidemiologic evidence. *Arthritis & Rheumatology* 2008; **58**(4):919-928.

46. Ramirez SP, Albert JM, Blayney MJ et al. Rosiglitazone is associated with mortality in chronic hemodialysis patients. *Journal of the American Society of Nephrology* 2009; **20**(5):1094-1101.

47. Abrahamowicz M, Beauchamp ME, Fournier P, Dumont A. Evidence of subgroup-specific treatment effect in the absence of an overall effect: is there really a contradiction? *Pharmacoepidemiology and Drug Safety* 2013; **22**(11):1178-1188.

48. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology* 2004; **57**(3):229-236.

49. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550-560.

50. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009; **20**(4):512-522.

51. Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration- A simulation study. *American Journal of Epidemiology* 2007; **165**(10):1110-1118.

52. McCandless LC, Richardson S, Best N. Adjustment for Missing Confounders Using External Validation Data and Propensity Scores. *Journal of the American Statistical Association* 2012; **107**(497):40-51.

53. Groenwold RH, Nelson DB, Nichol KL, Hoes AW, Hak E. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. *International Journal of Epidemiology* 2010; **39**(1):107-117.

**Table 1. Simulation results: Empirical power for detecting unmeasured confounding**

| Scenario [a] | Outcome | N (# events) [b] | Unmeasured confounding | Relative bias of conventional treatment estimate (Model 1) | Power (%) for detecting unmeasured confounding based on criteria: [c] | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | 95% CI of binary IV Model 2 [d] | 95% CI of continuous IV Model 3 [d] | 95% CI of interaction Model 4 [d] | P-value < 0.05 for two-tailed interaction test (Model 4) |
| Column 1 | Column 2 | Column 3 | Column 4 | Column 5 | Column 6 | Column 7 | Column 8 | Column 9 |
| 1 | Binary | 11,992 (1,370) | NO | 0.001 | 4.8 [c] | 4.7 [c] | 3.9 [c] | 4.9 [c] |
| 2 | Binary | 12,016 (2,964) | NO | 0.002 | 4.4 [c] | 3.4 [c] | 5.7 [c] | 5.7 [c] |
| 3 | Binary | 12,000 (3,170) | Weak | 0.349 | 5.2 | 11.1 | 15.7 | 18.5 |
| 4 | Binary | 29,975 (7,922) | Weak | 0.353 | 9.9 | 24.9 | 39.2 | 42.6 |
| 5 | Binary | 17,993 (6,556) | Moderate | 0.589 | 10.9 | 30.7 | 51.4 | 55.1 |
| 6 | Binary | 29,986 (10,918) | Moderate | 0.589 | 11.0 | 48.2 | 72.6 | 75.8 |
| 7 | Binary | 30,020 (9,459) | Moderate | NA [e] | 15.5 | 51.2 | 74.8 | 77.5 |
| 8 | Binary | 11,989 (3,156) | Weak | -0.335 | 6.4 | 11.4 | 16.4 | 18.7 |
| 9 | Continuous | 11,994 | Weak | -0.235 | 7.5 | 20.5 | 32.1 | 34.8 |
| 10 | Continuous | 29,996 | Weak | -0.234 | 11.2 | 42.8 | 64.0 | 67.2 |
| 11 | Continuous | 11,993 | Moderate | -0.468 | 13.7 | 51.6 | 73.3 | 77.0 |
| 12 | Continuous | 30,000 | Moderate | -0.470 | 30.4 | 89.7 | 99.0 | 99.1 |

[a] See Table A.1 in Supplementary Materials for details about each scenario.

[b] Mean total sample size N, across the 1,000 simulated samples, and mean number of events for the binary outcome in scenarios 1-8.

[c] For scenarios 1 and 2, the proportion of samples where the null hypothesis was rejected represents in fact the type I error because there is no unmeasured confounding.

[d] Percentage of samples where the 95% confidence interval for treatment effect of the given model excluded the conventional Model 1 estimate.

[e] True effect = 0.

**Table 2. Simulation results: Bias, standard deviation (SD), and root mean square error (RMSE) of treatment effect estimates**

| Scenario [a] | Outcome | True effect of *A* [b] | Bias | | SD ratio (relative to SD of interaction Model 4) | | | RMSE ratio (relative to RMSE of interaction Model 4) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Conventional Model 1 | Interaction Model 4 | Conventional Model 1 | Binary IV Model 2 | Continuous IV Model 3 | Conventional Model 1 | Binary IV Model 2 | Continuous IV Model 3 |
| Column 1 | Column 2 | Column 3 | Column 4 | Column 5 | Column 6 | Column 7 | Column 8 | Column 9 | Column 10 | Column 11 |
| 1 | Binary | 0.10 | 0.000 | 0.000 | 0.30 | 3.83 | 1.62 | 0.30 | 3.83 | 1.61 |
| 2 | Binary | 0.10 | 0.000 | 0.003 | 0.23 | 2.36 | 1.07 | 0.23 | 2.36 | 1.07 |
| 3 | Binary | 0.10 | 0.035 | 0.007 | 0.29 | 3.51 | 1.60 | 1.26 | 3.40 | 1.55 |
| 4 | Binary | 0.10 | 0.035 | 0.006 | 0.29 | 3.67 | 1.58 | 1.92 | 3.48 | 1.49 |
| 5 | Binary | 0.10 | 0.059 | 0.011 | 0.29 | 3.71 | 1.61 | 2.23 | 3.38 | 1.46 |
| 6 | Binary | 0.10 | 0.059 | 0.011 | 0.29 | 3.57 | 1.65 | 2.72 | 3.04 | 1.40 |
| 7 | Binary | 0.00 | 0.059 | 0.011 | 0.29 | 3.57 | 1.65 | 2.73 | 3.07 | 1.41 |
| 8 | Binary | 0.10 | -0.036 | -0.008 | 0.29 | 3.78 | 1.61 | 1.26 | 3.65 | 1.55 |
| 9 | Continuous | -5.00 | 1.173 | 0.191 | 0.28 | 3.59 | 1.63 | 1.70 | 3.45 | 1.57 |
| 10 | Continuous | -5.00 | 1.172 | 0.218 | 0.28 | 3.50 | 1.68 | 2.51 | 3.10 | 1.49 |
| 11 | Continuous | -5.00 | 2.340 | 0.462 | 0.30 | 3.89 | 1.65 | 2.80 | 3.24 | 1.38 |
| 12 | Continuous | -5.00 | 2.348 | 0.436 | 0.29 | 3.56 | 1.66 | 3.75 | 2.56 | 1.20 |

[a] Columns 2 to 4 in Table 1 describe the main features of each scenario. See Table A.1 in Supplementary Materials for details about each scenario.

[b] Risk difference for the binary outcome in scenarios 1 to 8, difference in the mean values of a continuous outcome for scenarios 9 to 12.

**Table 3. Simulation results: Inference about treatment effect**

| Scenario [a] | Outcome | True effect of *A* [b] | Mean SE [c] | | | | Coverage of 95% CI [c] | | Power (%) for testing $H_0$ of no effect of treatment [d] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Conventional Model 1 | Binary IV Model 2 | Continuous IV Model 3 | Interaction Model 4 | Conventional Model 1 | Interaction Model 4 | Conventional Model 1 | Binary IV Model 2 | Continuous IV Model 3 | Interaction Model 4 |
| Column 1 | Column 2 | Col. 3 | Column 4 | Column 5 | Column 6 | Column 7 | Column 8 | Column 9 | Column 10 | Column 11 | Column 12 | Column 13 |
| 1 | Binary | 0.10 | 0.006 | 0.072 | 0.032 | 0.020 | 94.1 | 95.3 | 100.0 | 30.2 | 89.1 | 99.8 |
| 2 | Binary | 0.10 | 0.008 | 0.090 | 0.042 | 0.036 | 93.7 | 93.6 | 100.0 | 22.3 | 68.5 | 79.5 |
| 3 | Binary | 0.10 | 0.008 | 0.100 | 0.045 | 0.027 | 0.7 | 93.6 | 100.0 | 17.1 | 61.9 | 96.6 |
| 4 | Binary | 0.10 | 0.005 | 0.062 | 0.028 | 0.017 | 0.0 | 92.5 | 100.0 | 34.6 | 94.8 | 100.0 |
| 5 | Binary | 0.10 | 0.007 | 0.088 | 0.004 | 0.024 | 0.0 | 92.5 | 100.0 | 20.0 | 71.1 | 99.6 |
| 6 | Binary | 0.10 | 0.005 | 0.068 | 0.031 | 0.019 | 0.0 | 91.6 | 100.0 | 31.5 | 90.8 | 100.0 |
| 7 | Binary | 0.00 | 0.005 | 0.066 | 0.030 | 0.018 | 0.0 | 90.3 | 100.0 [d] | 4.5 [d] | 5.2 [d] | 9.7 [d] |
| 8 | Binary | 0.10 | 0.008 | 0.100 | 0.045 | 0.027 | 0.4 | 93.8 | 100.0 | 16.6 | 62.7 | 91.6 |
| 9 | Continuous | -5.00 | 0.190 | 2.409 | 1.079 | 0.652 | 0.0 | 93.5 | 100.0 | 55.6 | 99.4 | 100.0 |
| 10 | Continuous | -5.00 | 0.120 | 1.504 | 0.678 | 0.411 | 0.0 | 91.7 | 100.0 | 92.3 | 100.0 | 100.0 |
| 11 | Continuous | -5.00 | 0.211 | 2.646 | 1.178 | 0.721 | 0.0 | 91.4 | 100.0 | 47.2 | 98.9 | 100.0 |
| 12 | Continuous | -5.00 | 0.133 | 1.639 | 0.740 | 0.454 | 0.0 | 84.0 | 100.0 | 87.5 | 100.0 | 100.0 |

[a] Columns 2 to 4 in Table 1 describe the main features of each scenario. See Table A.1 in Supplementary Materials for details about each scenario.

[b] Risk difference for the binary outcome in scenarios 1 to 8, difference in the mean values of a continuous outcome for scenarios 9 to 12.

[c] Standard errors (SEs) shown in column 7 and the resulting covariance-based 95% CIs, for which the coverage is reported in column 9, are only approximately accurate, as they rely on model-based SEs, which fail to account for the fact that *D* in equation (3) is a generated regressor that had to be estimated from the same data. However, in additional analyses for scenario 3, the alternative CI, based on 300 bootstrap resamples, had the same coverage rate as the corresponding covariance-based CIs (data not shown).

[d] For scenario 7, the numbers in columns 10 to 13 represent in fact the type I error, because in this scenario the treatment has no effect.

**Table 4. Simulation results: Sensitivity analyses with comparison of results to scenario 6 from the main simulations**

| Scenario [a] | Change with respect to scenario 6 of main simulations | Power (%): p-value < 0.05 for interaction test (Model 4) | Bias of treatment effect | | | RMSE ratio for treatment effect relative to interaction Model 4 | | | Coverage of 95% CI for treatment effect | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Conventional Model 1 | Continuous IV Model 3 [e] | Interaction Model 4 | Conventional Model 1 | Binary IV Model 2 | Continuous IV Model 3 | Conventional Model 1 | Continuous IV Model 3 [e] | Interaction Model 4 |
| Column 1 | Column 2 | Column 3 | Column 4 | Column 5 | Column 6 | Column 7 | Column 8 | Column 9 | Column 10 | Column 11 | Column 12 |
| 6 | - | 77.2 | 0.059 | 0.000 | 0.011 | 2.69 | 3.21 | 1.40 | 0.0 | 95.2 | 90.4 |
| Treatment ($A$) effect heterogeneity added in outcome model used for data generation (equation (5)) | | | | | | | | | | | |
| 6a | Interaction of $A$ with measured $X_1$ added in (5) | 73.9 | 0.059 | 0.000 | 0.012 | 2.68 | 3.06 | 1.35 | 0.0 | 95.0 | 89.7 |
| 6b | Interaction of $A$ with unmeasured $X_3$ added in (5) | 77.7 | 0.062 | 0.002 | 0.013 | 2.70 | 2.94 | 1.35 | 0.0 | 95.3 [1] | 88.9 |
| 6c | Interaction of $A$ with unmeasured $X_3$ and removal of main effect of $X_3$ in (5) | 4.5 | 0.003 | -0.001 | 0.000 | 0.37 | 3.69 | 1.59 | 87.3 | 95.1 | 95.0 |
| Misspecification of the model to estimate the expected treatment (with respect to model used to generate treatment (equation (4))) | | | | | | | | | | | |
| 6d | Interaction $X_1X_2$ added in (4) | 74.6 | 0.058 | 0.001 | 0.011 | 2.65 | 3.10 | 1.44 | 0.0 | 93.6 [2] | 91.4 |
| 6e | Non-linear (quadratic) effect of $X_1$ in (4) | 76.3 | 0.059 | 0.001 | 0.011 | 2.68 | 2.93 | 1.36 | 0.0 | 94.4 | 90.7 |
| 6f | Three spurious covariates, not associated with $A$ and $Y$, included in the analyses | 74.0 | 0.059 | -0.001 | 0.011 | 2.71 | 3.26 | 1.40 | 0.0 | 95.1 | 90.9 |
| 6g | Logistic regression, rather than RD model, used to generate treatment in (4) | 29.2 | 0.063 | 0.001 [3] | 0.028 | 1.68 | 4.91 | 1.95 | 0.0 | 94.7 | 79.5 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Change in physicians' prescribing preference used to generate treatment in equation (4) | | | | | | | | | | | |
| 6h | Weaker physicians' preferences | 25.1 | 0.059 | -0.005[4] | 0.025 | 1.63 | 5.47 | 2.05 | 0.0 | 95.6[5] | 83.9 |
| 6i | Change over time in preferences | 27.3 | 0.059 | 0.001[6] | 0.026 | 1.63 | 2.00 | 1.51 | 0.0 | 96.1[7] | 81.4 |
| Change in number of patients per physician (scenario 6 includes 10-50 patients/physician) | | | | | | | | | | | |
| 6j | 2-50 patients/physician | 66.8 | 0.059 | 0.001 | 0.014 | 2.48 | 3.13 | 1.39 | 0.0 | 96.3[8] | 89.5 |
| 6k | 2-20 patients/physician | 23.6 | 0.059 | -0.003 | 0.029 | 1.58 | 3.17 | 1.78 | 0.0 | 95.8[9] | 79.4 |
| 6l | 2-20 patients/physician, and exclude physicians with <10 patients | 23.8 | 0.059 | 0.001 | 0.024 | 1.51 | 3.25 | 1.75 | 0.0 | 95.7 | 85.9 |
| Violation of the exclusion restriction assumption at the data generation | | | | | | | | | | | |
| 6m | Distribution of unmeasured confounder $X_3$ varies slightly by physician (ICC = 0.012) [b] | 49.6 | 0.048 | 0.007[10] | 0.014 | 2.05 | 2.90 | 1.33 | 0.0 | 95.0 | 88.0 |
| 6n | Distribution of unmeasured confounder $X_3$ varies substantially by physician (ICC = 0.155) [b] | 28.4 | 0.034 | 0.061[11] | 0.058 | 0.56 | 1.44 | 1.12 | 0.0 | 45.9[12] | 15.9 |
| 6o | Unmeasured confounder $X_3$ moderately correlated with physicians' preferences for $A = 1$ [c] | 99.5 | 0.055 | 0.105 | 0.125 | 0.44 | 0.86 | 0.84 | 0.0 | 0.0 | 0.0 |
| 6p | Unmeasured confounder $X_3$ strongly correlated with physicians' preferences for $A = 1$ [d] | 100.0 | 0.093 | 0.260 | 0.265 | 0.35 | 0.99 | 0.98 | 0.0 | 0.0 | 0.0 |

[a] See Table A.1 and section B.2 of Supplementary Materials for details about each scenario. Scenario 6 has a binary outcome, average N = 29,986 (average number of events = 10,996), and moderate unmeasured confounding.

[b] Intra-class correlation (ICC) for clustering of unmeasured confounder $X_3$ within physicians' practices was calculated using one-way analysis of variance. Values reported are the mean of ICC coefficients across the 1,000 simulated samples for a scenario.

[c] Subgroup of physicians who strongly prefer $A = 0$ has $X_3 \sim U(0, 0.8)$ vs. $X_3 \sim U(0, 1)$ for the other subgroup who prefers $A = 1$.

[d] Subgroup of physicians who strongly prefer $A = 0$ has $X_3 \sim U(0, 0.5)$ vs. $X_3 \sim U(0, 1)$ for the other subgroup who prefers $A = 1$.

[e] Given that the continuous IV Model 3 generally performed similar or better than the binary IV Model 2, only the results for IV Model 3 are presented in Table 4. However, when the binary IV Model 2 obtained a smaller bias in absolute value or a coverage of the 95% CI closer to 95.0, its result is reported in footnote (with major improvements shown in bold): [1] 95.0; [2] 95.7; [3] 0.000; [4] -0.003; [5] 94.5; [6] 0.000; [7] 94.9; [8] 95.4; [9] 95.2; [10] 0.003; [11] 0.060; **[12] 85.4**.

**Table 5. Differences in risk of gastrointestinal events between COX-2 inhibitor vs. NSAID users**

| Model | Treatment RD (%) for COX-2 vs. NSAID users (95% CI) [a] | Log(*D*+1) (95% CI) [b] | Interaction log(*D*+1)*cox2 (95% CI) [c] | Interaction p-value |
|---|---|---|---|---|
| Column 1 | Column 2 | Column 3 | Column 4 | Column 5 |
| Unadjusted | 6.4 (4.5, 8.3) | - | - | - |
| Conventional Model 1 | -1.9 (-3.5, -0.3) | - | - | - |
| Binary IV Model 2 | 3.3 (-12.7, 19.3) | - | - | - |
| Continuous IV Model 3 | 2.5 (-6.2, 11.3) | - | - | - |
| Interaction-based Model 4 | -2.6 (-8.4, 3.2) | 6.4 (-2.9, 15.7) | 2.5 (-12.9, 17.8) | 0.751 |

[a] Estimated (adjusted) risk difference (positive values indicate higher risk for COX-2 inhibitor users).

[b] Effect of discrepancy f($D$) = log($D$+1) among NSAID users ($A$ = 0), see equation (3).

[c] Effect of interaction between f($D$) and the indicator of COX-2 inhibitor use ($A$ = 1), see equation (3).