American Journal of Epidemiology
© The Author 2008. Published by the Johns Hopkins Bloomberg School of Public Health.
All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

Vol. 169, No. 3
DOI: 10.1093/aje/kwn299
Advance Access publication November 25, 2008

Practice of Epidemiology

Instrumental Variable Analysis for Estimation of Treatment Effects With Dichotomous Outcomes

Jeremy A. Rassen, Sebastian Schneeweiss, Robert J. Glynn, Murray A. Mittleman, and M. Alan Brookhart

Initially submitted May 5, 2008; accepted for publication August 26, 2008.

Instrumental variable analyses are increasingly used in epidemiologic studies. For dichotomous exposures and outcomes, the typical 2-stage least squares approach produces risk difference estimates rather than relative risk estimates and is criticized for assuming normally distributed errors. Using 2 example drug safety studies evaluated in 3 cohorts from Pennsylvania (1994–2003) and British Columbia, Canada (1996–2004), the authors compared instrumental variable techniques that yield relative risk and risk difference estimates and that are appropriate for dichotomous exposures and outcomes. Methods considered include probit structural equation models, 2-stage logistic models, and generalized method of moments estimators. Employing these methods, in the first study the authors observed relative risks ranging from 0.41 to 0.58 and risk differences ranging from −1.41 per 100 to −1.28 per 100; in the second, they observed relative risks of 1.38–2.07 and risk differences of 7.53–8.94; and in the third, they observed relative risks of 1.45–1.59 and risk differences of 3.88–4.84. The 2-stage logistic models showed standard errors up to 40% larger than those of the instrumental variable probit model. Generalized method of moments estimation produced substantially the same results as the 2-stage logistic method. Few substantive differences among the methods were observed, despite their reliance on distinct assumptions.

antipsychotic agents; confounding factors (epidemiology); instrumental variable; pharmacoepidemiology

Abbreviations: APM, antipsychotic medication; COX-2, cyclooxygenase 2; GMM, generalized method of moments; NSAID, nonsteroidal antiinflammatory drug; PACE, Pharmaceutical Assistance Contract for the Elderly.

Instrumental variable analysis is a technique for the control of unmeasured confounding in nonrandomized data, which is becoming more common in epidemiology and health services research (1–5). The technique rests on the idea that an instrument—a variable that is related to treatment but neither directly nor indirectly related to outcome, except through the effect of the treatment itself—can be identified in observed data and then, in the simplest case, substituted for actual exposure (6, 7). If such a variable can be found one can estimate or place bounds on the causal effect of treatment, provided that all necessary assumptions are met (8).

The most commonly used technique for instrumental variable analysis is the 2-stage least squares method (6, 9). In the setting of dichotomous exposures and outcomes, 2-stage least squares produces a risk difference estimate but a relative measure of effect may be desired. Further, there is a statistical issue of fitting dichotomous outcomes and exposures with the 2-stage least squares approach: 2-stage least squares relies on linear models, which can lead to model misspecification when predicting dichotomous exposure or outcome as a function of many covariates (10). This misspecification can yield predicted values of treatment or exposure outside the 0–1 range, as well as inconsistent estimates of treatment effect.

In this paper, we review a range of instrumental variable methods that one can use for dichotomous treatments and outcomes, including "2-stage" estimation techniques based on linear and probit models, a 3-stage estimator, and a generalized method of moments (GMM) approach. We empirically compare the performance of the methods in

Correspondence to Dr. Jeremy A. Rassen, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (e-mail: jrassen@post.harvard.edu).

a reanalysis of data from 3 published pharmacoepidemiologic studies (2, 3, 11). In each example, the physician prescribing preference instrument is used to determine the safety of medications. We consider the risk of severe gastrointestinal complications among users of nonsteroidal antiinflammatory drugs (NSAIDs) in 1 cohort, and in 2 other cohorts we consider the risk of mortality related to use of antipsychotic medications (APMs).

MATERIALS AND METHODS

Instrumental variable methods

Instrumental variable estimators can be formally defined in a number of ways. Informally, an instrument is an observed variable Z that predicts treatment X but is unassociated with outcome Y, either directly or indirectly through unmeasured confounders U, except via the effect of treatment (3, 6, 8, 12). In the presence of measured confounders C, these assumptions are generally assumed to hold within strata of the covariates. Alternatively, Pearl (13) approaches instrumental variables using graphical techniques.

While this simple definition provides a basis for thinking about instrumental variables, it is not necessarily universal across all instrumental variable approaches; in particular, the implied assumption of no treatment effect heterogeneity varies (14). Below we outline a series of methods for instrumental variable analysis along with their respective assumptions. In all cases, we simplify the discussion by assuming no treatment effect heterogeneity, though approaches to dealing with heterogeneity do exist (7). The techniques we used for estimation in our example studies are described following the analytic approaches.

Structural equation models

Instrumental variable estimators as used in economics have been specified as systems of simultaneous equations, often termed *structural equation models.* Typically, the first equation models the treatment assignment mechanism: It predicts treatment as a function of observed confounders and variables that are related to the treatment but unrelated to the outcome (i.e., instruments). The second equation is a model for the outcome that includes the treatment and the observed confounders.

Linear structural equation models and 2-stage least squares

The most common of these structural equation approaches involves 2 simultaneous linear models. As before, let X be treatment, Y be outcome, C be 1 or more measured confounders, and Z be the instrument. Let α_i and β_i be coefficients and ε_1 and ε_2 be errors:

$$X = \alpha_0 + \alpha_1 Z + \alpha_2 C + \varepsilon_1. \qquad (1)$$

$$Y = \beta_0 + \beta_1 X + \beta_2 C + \varepsilon_2. \qquad (2)$$

In this model, ε_1 and ε_2 are often assumed to have a bivariate normal distribution, an extension of the assumption of normally distributed error for a single linear model. In the presence of unmeasured factors related to both treatment and outcome, the correlation between ε_1 and ε_2 will be nonzero. In this case, X will be correlated with ε_2 such that ordinary least squares estimation of β_1 (often intended to be interpretable as a causal risk difference) will be biased because of an association between X and Y via unmeasured risk factors U.

However, if Z is uncorrelated with U after control for C, then β_1 can be consistently estimated using 2-stage least squares. This procedure works by sequential application of 2 ordinary least squares regressions in which predicted values of treatment X from the first stage are entered into the second stage as a replacement for actual treatment. By replacing the confounded treatment variable with an unconfounded *prediction* of treatment, the bias due to unmeasured confounding can be avoided (6, 12). Note that 2-stage least squares is a method of moments approach (discussed below) that, unlike maximum likelihood estimation, does not make distributional assumptions about the error terms.

Two-stage least squares estimation may be problematic in the context of dichotomous exposures and outcomes. The linear models of treatment and outcome may produce predicted values outside of the 0–1 range; plus, with dichotomous outcomes, the assumption of a bivariate normal distribution of errors will be violated. However, it has been suggested that these are theoretical rather than practical problems and that 2-stage least squares estimates are unbiased, though their standard errors are possibly incorrect (10, 15). One alternative, the use of a generalized linear model with an identity link and binomially distributed errors, can fail to fit (16).

A more fundamental issue is that 2-stage least squares is based on an additive model even in instances where a multiplicative model may be more appropriate. Several alternative models for dichotomous exposures and outcomes have been proposed, most of them estimating relative measures of risk (odds ratios, risk ratios) rather than absolute measures like 2-stage least squares' estimate of risk difference.

Probit structural equation models

Economists frequently use probit structural models in places where 2-stage least squares may not be suited to the question at hand (6, 12). Unlike ordinary least squares and 2-stage least squares, probit models explicitly model probabilities and, as such, constrain the predicted values of treatment and outcome to the 0–1 range.

The specification of a system of probit equations is given by

$$X = I[(\alpha_0 + \alpha_1 Z + \alpha_2 C) > \varepsilon_1]. \qquad (3)$$

$$Y = I[(\beta_0 + \beta_1 X + \beta_2 C) > \varepsilon_2]. \qquad (4)$$

ε_1 and ε_2 are thresholds that are assumed to have a bivariate normal distribution. $I(x)$ is the indicator function, returning 0 if the condition is not met and 1 if it is.

Unlike logistic models in which the β_i's are interpretable as logarithms of odds ratios, the β_i's in probit models have no such natural interpretation. However, it has been shown that by multiplying a probit coefficient by approximately 1.6, probit coefficients can be made to approximate logistic coefficients (17, 18).

2-stage logistic model

If one wants to use instrumental variables to estimate odds ratios in a study with dichotomous treatment and outcome, a natural choice might be to create 2 sequential logistic regressions in a manner parallel to 2-stage least squares. Using maximum likelihood estimation, the first stage would predict treatment as a function of an instrument and covariates, and the second stage would predict outcome based on the first stage's predicted treatment and the covariates. Such an approach is problematic, however, as the first-stage logistic model must be specified correctly in order for the second stage to be unbiased (10); further, to our knowledge there is no definition for a bivariate logistic distribution of errors. Therefore, even with a valid instrumental variable, the 2-stage logistic approach is not guaranteed to yield unbiased estimates (19, 20). In practice, this bias may be small; the technique has been successfully employed for the closely related problem of measurement error correction (19, 21).

3-stage model

If one is concerned with misspecification of the first stage and interested in an estimate of risk difference, then replacing the 2-stage least squares method's first ordinary least squares model with a logistic model may be reasonable. However, Angrist (10) cautions against this: As above, if the first-stage logistic model is incorrect, the resulting second-stage estimates will be inconsistent, whereas 2-stage least squares can be consistent even with first-stage misspecification. As a solution, Angrist demonstrated a 3-stage approach: A logistic model is used as a precursor to 2-stage least squares; the continuous probability value predicted by the logistic estimation is used as the instrument in 2-stage least squares estimation.

GMM instrumental variable approaches

Instrumental variable estimators can also be constructed by making assumptions about the moments of the error term in a regression model (15); 2-stage least squares is a specific example of this approach. These moment assumptions are expressed as equations whose solution yields an estimate of the treatment effect and of the other parameters in the model. Depending on the assumptions expressed, the equations may be solvable in closed form, as with 2-stage least squares, or may need to be solved with iterative techniques. Since these so-called GMM estimators rely on moment assumptions rather than the distributional assumptions of maximum likelihood, they may in certain cases be more robust and less sensitive to parametric requirements, though possibly also less efficient.

As an example, suppose that one is interested in estimating a usual mean-model regression:

$$Y = \mu(X, C, \beta) + \varepsilon. \qquad (5)$$

In this framework, $\mu(X, C, \beta)$ could be any function, but we will use the logit:

$$\mu(X, C, \beta) = \frac{1}{1 + e^{\beta_0 + \beta_1 X + \beta_2 C}}. \qquad (6)$$

Imposing the structure of the instrumental variable framework yields the following moment assumptions. First, we must express the fact that, as in any mean model, the residuals should sum to 0:

$$\begin{aligned} 0 &= \tfrac{1}{n}\sum \left[Y_i - \mu(X_i)\right] \\ &= \tfrac{1}{n}\sum \left[Y_i - \tfrac{1}{1 + e^{\beta_0 + \beta_1 X + \beta_2 C}}\right]. \end{aligned} \qquad (7)$$

Second, the errors must be uncorrelated with the confounders:

$$\begin{aligned} 0 &= \tfrac{1}{n}\sum \{C[Y_i - \mu(X_i)]\} \\ &= \tfrac{1}{n}\sum \left\{C\left[Y_i - \tfrac{1}{1 + e^{\beta_0 + \beta_1 X + \beta_2 C}}\right]\right\}. \end{aligned} \qquad (8)$$

Finally, the errors must be uncorrelated with the instrument:

$$\begin{aligned} 0 &= \tfrac{1}{n}\sum \{Z[Y_i - \mu(X_i)]\} \\ &= \tfrac{1}{n}\sum \left\{Z\left[Y_i - \tfrac{1}{1 + e^{\beta_0 + \beta_1 X + \beta_2 C}}\right]\right\}. \end{aligned} \qquad (9)$$

Using iterative techniques like the Newton-Raphson method, one can solve for the combination of β_i's that will yield a 0 for each of equations 7–9. In our example, β_0 and β_2 are nuisance parameters, but the estimate of β_1 will be an estimate of the causal log odds ratio of interest.

Calculating risk differences from nonlinear models

If one considers linear models for risk differences estimated by ordinary least squares to be undesirable in the setting of dichotomous outcomes, an alternative approach is available. We used marginal-effects models to calculate risk differences from probit and logistic models. (Note that the term ''marginal effect'' as used here refers to ''change in the slope of the probability function'' rather than ''effect on the marginal patient''.) With marginal effects models, investigators look at the underlying cumulative density function and use the derivatives at a chosen point (or average of the derivatives at a number of points) to estimate a risk difference. See the Appendix for details on this approach.

Estimation of risk differences and relative risks

We considered various modeling approaches in our analysis, starting with 2-stage least squares and its estimate of a risk difference. For other estimates of risk difference, we considered a logistic first stage with an ordinary least squares second stage and an instrumental variable probit

Table 1. Characteristics of 3 Cohorts of Adults Aged 65 Years or Older, by Type of Treatment Received, Pennsylvania (1994–2003) and British Columbia, Canada (1996–2004)

Treatment	Pennsylvania								British Columbia APM Cohort			
	COX-2 Inhibitor Cohort				APM Cohort							
	COX-2 Inhibitor		Nonselective NSAID		Conventional APM Treatment		Cohort Atypical APM Treatment		Conventional APM Treatment		Atypical APM Treatment	
	No. or Mean	%	No. or Mean	%	No. or Mean	%	No. or Mean	%	No. or Mean	%	No. or Mean	%
No. of patients	32,074		17,637		8,056		12,031		12,756		23,785	
Mean age, years	79.75		77.79		83.30		83.58		79.89		80.32	
Male sex		14.1		18.9		20.1		15.1		39.7		35.1
Medical history												
Cerebrovascular disease		—[a]		—		28.3		30.2		10.8		9.9
Congestive heart failure		30.3		24.6		31.8		30.4		8.4		6.0
Myocardial infarction		1.84		1.64		3.4		3.3		2.7		2.3
Ischemic heart disease		—		—		28.3		23.8		3.8		2.7
Other cardiovascular disease		16.4		14.8		55.4		57.7		20.2		16.6
Hypertension		72.8		70.1		57.2		64.2		22.3		24.1
Osteoarthritis		48.5		33.4		—		—		—		—
Rheumatoid arthritis		5.0		2.7		—		—		—		—
Diabetes mellitus		—		—		25.5		26.3		15.0		13.8
Warfarin use		13.3		6.5		—		—		—		—
Corticosteroid use		8.7		7.8		—		—		—		—
Delirium		—		—		11.7		15.2		7.4		8.4
Mood disorders		—		—		21.8		35.5		15.6		25.3
Psychotic disorders		—		—		21.7		24.4		11.2		16.7
Other psychiatric disorders		—		—		5.7		7.9		3.1		4.5
Nursing home stay		7.6		7.1		15.5		20.2		31.0		26.8
No. of generic drugs[b]		7.82		6.65		6.65		7.82		7.36		7.34

Abbreviations: APM, antipsychotic medication; COX-2, cyclooxygenase 2; NSAID, nonsteroidal antiinflammatory drug.

[a] Variable was not measured in this cohort.

[b] Number of generic drugs for which the participant had filled a pharmacy prescription in the previous 180 days.

marginal effects model evaluated both at the mean of all covariates and averaged over all observations (6). With regard to relative risks and odds ratios, we considered 2-stage logistic models, a GMM instrumental variable logistic model, and an instrumental variable probit model.

For the probit/probit model, we wished to report odds ratios, so we used the scaling factor proposed by Amemiya et al. and others (16–18) to transform the coefficients of the probit regression into coefficients that would approximate the log odds ratio coefficients of a logistic regression. Because this is an ad hoc approach (though with theoretical roots), we calculated several values of the scaling factor, ranging from 1.4 to 1.8, including Amemiya et al.'s suggested value of 1.6. We also used an empirically derived transformation figure calculated as the ratio of the unadjusted probit effect to the unadjusted logistic effect (log odds ratio).

These combinations were chosen on the basis of past practice (1–3) and discussion in the literature (8, 10, 22, 23).

For the GMM models, we implemented an estimation procedure based on the estimator defining equations 7–9 above. The GMM approach has been implemented in several other instances (15, 24).

For reference, we also fitted each model's non-instrumental–variable analog: crude and adjusted ordinary least squares, logistic, and probit models. All analyses were carried out using Stata, version 9 (25, 26), with the exception of the GMM models, for which analyses were carried out in R (27). R programming code is available from the corresponding author (J. A. R.).

We bootstrapped standard errors for all models for which there were not analytic standard error estimates (28). We used cluster sampling and conducted 1,000 iterations.

Physician prescribing preference as an instrumental variable

We used physician prescribing preference as the instrumental variable for use in studies comparing 2 treatment regimens.

Physician prescribing preference, as proposed by Brookhart et al. (3), posits that a doctor's prescribing decision depends on both characteristics of the patient at hand and the physician's preference for a specific drug or class of drugs. Her preference should be largely independent of patient characteristics and outcome and may therefore serve as an instrumental variable (and, as such, a predictor of treatment among marginal patients), given that certain assumptions are fulfilled (8).

Because true preference is unmeasured, we used a simple binary surrogate measure of preference Z: the treatment given by the doctor to the patient she saw most recently prior to the current patient and who received either of the study treatments (3, 4).

$$Z_i = X_{i-1};\; i > 1. \tag{10}$$

Example study 1: NSAID initiation and risk of severe gastrointestinal complications

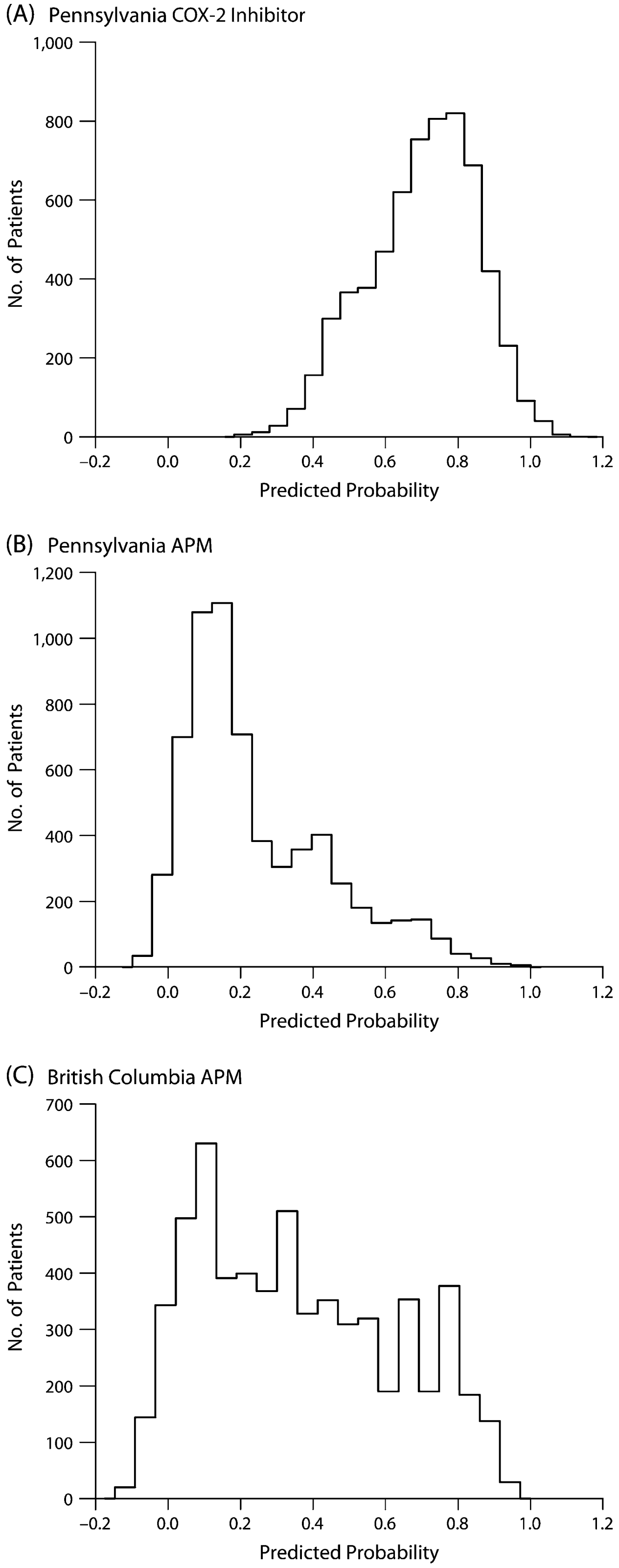
We performed a study of initiation of NSAID therapy and its relation to severe gastrointestinal complications (4). A dichotomous exposure variable was coded by class of NSAID; nonselective NSAIDs (ibuprofen, naproxen, diclofenac) were the referent category, and they were compared with cyclooxygenase 2 (COX-2) inhibitors (celecoxib, rofecoxib, valdecoxib). Outcome was defined as the cumulative risk of a gastrointestinal complication (hospitalization for gastrointestinal hemorrhage or peptic ulcer disease, or a medical insurance claim for associated services) within 180 days of treatment initiation. The study was performed in the Pennsylvania population described below.

Example study 2: APM initiation and risk of short-term mortality

As a second example, we performed a study of initiation of APMs and the associated risk of short-term mortality (29). APMs are categorized into 2 groups: conventional (older) agents and atypical (newer) agents (30). Previous studies have raised the question of increased death rates among users of atypical antipsychotic agents as compared with placebo (31–33). Outcome was defined as the cumulative risk of death from any cause within 180 days of treatment initiation. The study was performed separately in the British Columbia and Pennsylvania populations described below.

Cohorts of patients

We drew 3 study cohorts from the 2 populations (British Columbia and Pennsylvania). Our study populations each comprised patients aged 65 years or older who initiated treatment with the study drugs. For the COX-2 inhibitor study and the Pennsylvania APM study, we drew cohorts from Pennsylvania's Pharmaceutical Assistance Contract for the Elderly (PACE) program, a drug assistance program for the state's low-income seniors (2, 3). We received claims from 1994–2003 for those PACE participants also enrolled in Medicare. For the APM study, we separately drew


(A) Pennsylvania COX-2 Inhibitor

(B) Pennsylvania APM

(C) British Columbia APM

Figure 1. Probability of receiving treatment as predicted by a first-stage ordinary least squares model. Most probabilities fell within the acceptable 0–1 range. (A) Pennsylvania cyclooxygenase 2 (COX-2) inhibitor cohort (1994–2003); (B) Pennsylvania antipsychotic medication (APM) cohort; (C) British Columbia, Canada, APM cohort (1996–2004).

Table 2. Basic Measures of Association Observed for Drug Safety in 3 Cohorts of Patients, Pennsylvania (1994–2003) and British Columbia, Canada (1996–2004)

Exposure:	COX-2 Inhibitor				Conventional APM				Conventional APM			
Referent:	Nonselective NSAID				Atypical APM				Atypical APM			
Outcome:	Severe Gastrointestinal Complications				Death				Death			
Population:	Pennsylvania				Pennsylvania				British Columbia			
	%	No. of Events	Risk Measure	95% CI	%	No. of Events	Risk Measure	95% CI	%	No. of Events	Risk Measure	95% CI
Risk of outcome in exposed group	1.57	503			16.22	1,307			04.16	1,806		
Risk of outcome in referent group	1.38	243			13.53	1,628			9.70	2,307		
Crude risk difference (×100)			0.19	−0.03, 0.41			2.69	1.68, 3.70			4.46	3.75, 5.17
Crude risk ratio			1.14	0.98, 1.33			1.20	1.12, 1.28			1.46	1.38, 1.55
Crude odds ratio			1.14	0.98, 1.33			1.24	1.14, 1.34			1.54	1.44, 1.64

Abbreviations: APM, antipsychotic medication; COX-2, cyclooxygenase 2; NSAID, nonsteroidal antiinflammatory drug.

a cohort of patients from all residents of British Columbia, Canada, aged ≥65 years who initiated therapy between 1996 and 2004 (11).

All identifying information was transformed into anonymous identifiers. Covariates were each measured at baseline and included diagnoses and drug therapies occurring in the 180 days prior to treatment initiation. Further details on both populations and the covariates assembled are given elsewhere (2, 11). The studies were approved by the institutional review board of Brigham and Women's Hospital (Boston, Massachusetts), and we had data-use agreements in place with the Centers for Medicare and Medicaid Services (Baltimore, Maryland) and the British Columbia Ministry of Health (Victoria, British Columbia, Canada).

Table 3. Comparison of Risk Difference Models in 3 Cohorts of Patients, Pennsylvania (1994–2003) and British Columbia, Canada (1996–2004)[a]

Exposure:	COX-2 Inhibitor		Conventional APM		Conventional APM	
Referent:	Nonselective NSAID		Atypical APM		Atypical APM	
Outcome:	Severe Gastrointestinal Complications		Death		Death	
Population:	Pennsylvania		Pennsylvania		British Columbia	
	RD × 100	95% CI	RD × 100	95% CI	RD × 100	95% CI
Crude OLS model	0.19	−0.03, 0.41	2.69	1.65, 3.73	4.46	3.69, 5.23
Adjusted OLS model	−0.07	−0.30, 0.16	3.91	2.68, 5.13	3.55	2.74, 4.37
2-stage least squares model[b]	−1.28	−2.56, 0.01	7.69	1.26, 14.12	4.00	0.94, 7.06
Logistic/OLS model[b,c]	−1.36	−2.58, -0.15	7.64	1.55, 13.74	4.84	1.80, 7.88
3-stage model[b,c]	−1.35	−2.53, −0.17	7.53	1.83, 13.24	4.76	1.81, 7.72
Crude probit marginal effects[d] model	0.19	−0.03, 0.41	2.69	1.68, 3.70	4.46	3.69, 5.23
Adjusted probit marginal effects[d] model	−0.05	−0.26, 0.17	3.51	2.32, 4.69	3.20	2.42, 3.97
IV-based probit marginal effects[d] model[b,c]	−1.41	−3.14, 0.32	8.94	1.64, 16.24	3.88	0.67, 7.08

Abbreviations: APM, antipsychotic medication; CI, confidence interval; COX-2, cyclooxygenase 2; IV, instrumental variable; NSAID, nonsteroidal antiinflammatory drug; OLS, ordinary least squares; RD, risk difference.

[a] All risk differences are scaled by 100.

[b] Instrumental variable analysis.

[c] Confidence intervals are based on bootstrapped standard errors.

[d] Marginal effects are the difference in predicted probability due to change of treatment status, evaluated at the means of all covariates and scaled by 100.

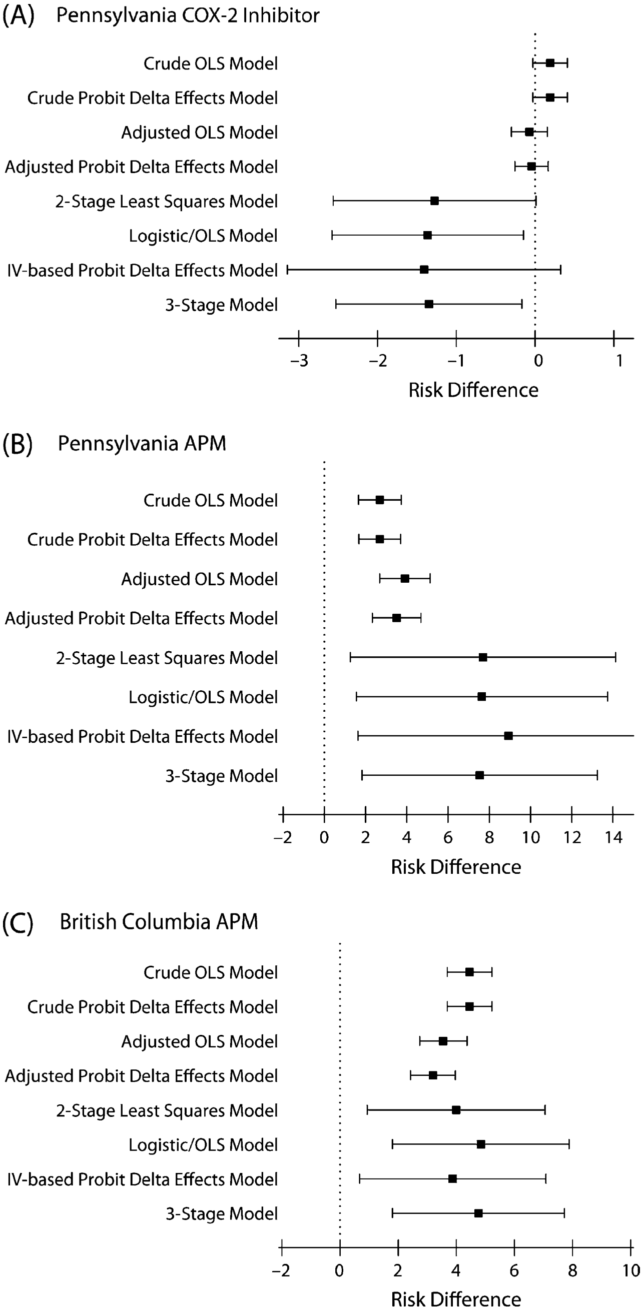


Figure 2. Point estimates for 8 models of risk difference in 3 cohorts of patients. All risk difference estimates are scaled by 100. (A) Pennsylvania cyclooxygenase 2 (COX-2) inhibitor cohort (1994–2003); (B) Pennsylvania antipsychotic medication (APM) cohort; (C) British Columbia, Canada, APM cohort (1996–2004). OLS, ordinary least squares; IV, instrumental variable. Bars, standard error.

RESULTS

Characteristics of participants in the 3 cohorts of elderly drug initiators are summarized in Table 1; each of these covariates was controlled for in the adjusted analyses. Generally speaking, patients in Pennsylvania were older and sicker than those in British Columbia.

Figure 1 shows histograms of the predicted probabilities from first-stage ordinary least squares models for each cohort. The vast majority of predicted values fell within the appropriate 0–1 range, though the APM analyses produced some predicted probabilities below 0.

Table 2 presents basic measures of absolute and relative risk for outcomes in each of the 3 cohorts. The outcome in the COX-2 inhibitor study was relatively rare, while the APM outcomes were much more frequent. In keeping with the observation that the Pennsylvania patients were generally sicker than their counterparts in British Columbia, the baseline risk was higher in the Pennsylvania population than in British Columbia, though British Columbia's crude risk difference was higher.

Table 3 and Figure 2 present results from the estimates of risk difference. Most of the multistage models had similar results for both the point estimate and the standard error, though the British Columbia APM cohort exhibited differences between the 2-stage least squares model and the logistic first stage. The probit marginal-effects model's results were similar to the risk differences estimated by the 2-stage models, and the marginal effects in our data were not sensitive to whether the effect was calculated at the mean of the covariates or averaged over all observations. We display the data at the mean of all covariates.

Similarly, Table 4 and Figure 3 present results from estimates of the odds ratio. The various approaches to instrumental variable regression all yielded comparable point estimates. As with the risk differences, the standard errors in the instrumental variable models were far higher than those in the crude and adjusted models. The GMM and 2-stage logistic estimates agreed in 2 of the 3 cohorts but disagreed markedly in the Pennsylvania APM cohort. With regard to the probit scaling factor, scaling by 1.6 or 1.8 generally made little difference. Scaling by the ratio of the crude estimates produced larger variation.

When addressing the issue of incorrect specification of the error distribution in 2-stage least squares, one suggestion is the use of a generalized linear model with a noncanonical error structure (16). For our dichotomous outcome, we attempted to use generalized linear models with binomial error structures and identity and log links, with and without covariates. In almost all cases, the models without covariates converged as expected, but as the covariates were added, the models failed to fit, either because predicted values were outside the acceptable range or because of nonconvergence. Because of these models' unpredictable behavior, we did not include them in our final analysis of modeling techniques.

DISCUSSION

Instrumental variable analysis has traditionally been performed using 2-stage least squares models that predict risk differences. In this paper, we sought to empirically compare the performance of 2-stage least squares with alternative instrumental variable approaches appropriate for dichotomous

Table 4. Comparison of Relative Risk Models in 3 Cohorts of Patients, Pennsylvania (1994–2003) and British Columbia, Canada (1996–2004)

Exposure:	COX-2 Inhibitor		Conventional APM		Conventional APM	
Referent:	Nonselective NSAID		Atypical APM		Atypical APM	
Outcome:	Severe Gastrointestinal Complications		Death		Death	
Population:	Pennsylvania		Pennsylvania		British Columbia	
	RR	95% CI	RR	95% CI	RR	95% CI
Logistic model						
Crude	1.14	0.98, 1.33	1.24	1.14, 1.34	1.54	1.43, 1.65
Adjusted	0.97	0.82, 1.13	1.39	1.25, 1.54	1.42	1.31, 1.54
2-stage logistic model[a,b]						
Crude	0.54	0.25, 1.19	1.37	1.06, 1.77	1.70	1.39, 2.08
Adjusted	0.41	0.18, 0.94	2.01	1.21, 3.35	1.59	1.18, 2.14
Logistic GMM model[a,b]						
Crude	0.56	0.28, 1.16	1.36	1.08, 1.71	1.67	1.41, 1.97
Adjusted	0.47	0.25, 0.86	1.38	1.03, 1.85	1.45	1.17, 1.80
Probit model[c]						
Crude						
×1.6	1.09	1.01, 1.17	1.20	1.14, 1.27	1.43	1.37, 1.50
×1.8	1.10	1.01, 1.19	1.23	1.16, 1.31	1.50	1.43, 1.58
Adjusted						
×1.6	0.98	0.90, 1.06	1.31	1.22, 1.41	1.36	1.28, 1.44
×1.8	0.98	0.90, 1.06	1.36	1.26, 1.47	1.41	1.33, 1.50
×CR[d]	0.97	0.87, 1.07	1.36	1.26, 1.47	1.44	1.35, 1.53
IV probit model[a,c]						
×1.6	0.58	0.39, 0.87	1.88	1.29, 2.74	1.44	1.16, 1.79
×1.8	0.54	0.35, 0.84	2.04	1.37, 3.04	1.51	1.20, 1.90
×CR[d]	0.42	0.25, 0.70	2.07	1.38, 3.09	1.54	1.22, 1.96

Abbreviations: APM, antipsychotic medication; B, bootstrap; CI, confidence interval; COX-2, cyclooxygenase 2; CR, crude ratio; GMM, generalized method of moments; IV, instrumental variable; NSAID, nonsteroidal antiinflammatory drug; RR, relative risk.

[a] Instrumental variable analysis.

[b] Confidence intervals are based on bootstrapped standard errors.

[c] Probit models' coefficients are scaled by the indicated amounts.

[d] Scaled by the ratio of the crude logit estimate to the crude probit estimate.

outcomes, in a reanalysis of 3 data sets. With regard to risk differences, we observed little difference in point estimate or precision between estimates from 2-stage least squares models and models using logistic first stages and ordinary least squares second stages; in our data, this equivalence held with both rare and frequent outcomes. With regard to relative measures of risk, 2-stage logistic regression was satisfactory in our data, though GMM models or 2-stage probit models are considered by some to be more theoretically sound. As expected, standard errors for all instrumental variable approaches were substantially larger than those of ordinary regression.

In each of the examples, we anticipated the presence of strong unmeasured confounding. In the APM example, we hypothesized that the strongest unmeasured confounder would be frailty, whereby frail patients would be less likely to receive conventional treatment and more likely to die such that the crude association would underestimate the true effect. We also hypothesized that this effect would be stronger in Pennsylvania than in British Columbia, given that the Pennsylvania patients were older and sicker than the British Columbia patients. In the COX-2 inhibitor case, we expected the strongest unmeasured confounders to be gastrointestinal risk factors that the physician considered for the treatment choice but were not recorded in claims data such that the crude association would overestimate the risk.

In the Pennsylvania studies, the data bore these hypotheses out: In both the risk difference and relative risk models, we saw the expected movement of point estimates. In the British Columbia cohort, we saw a subtler movement that could be explained by chance; the instrumental variable analyses suggested that there was relatively little unmeasured confounding

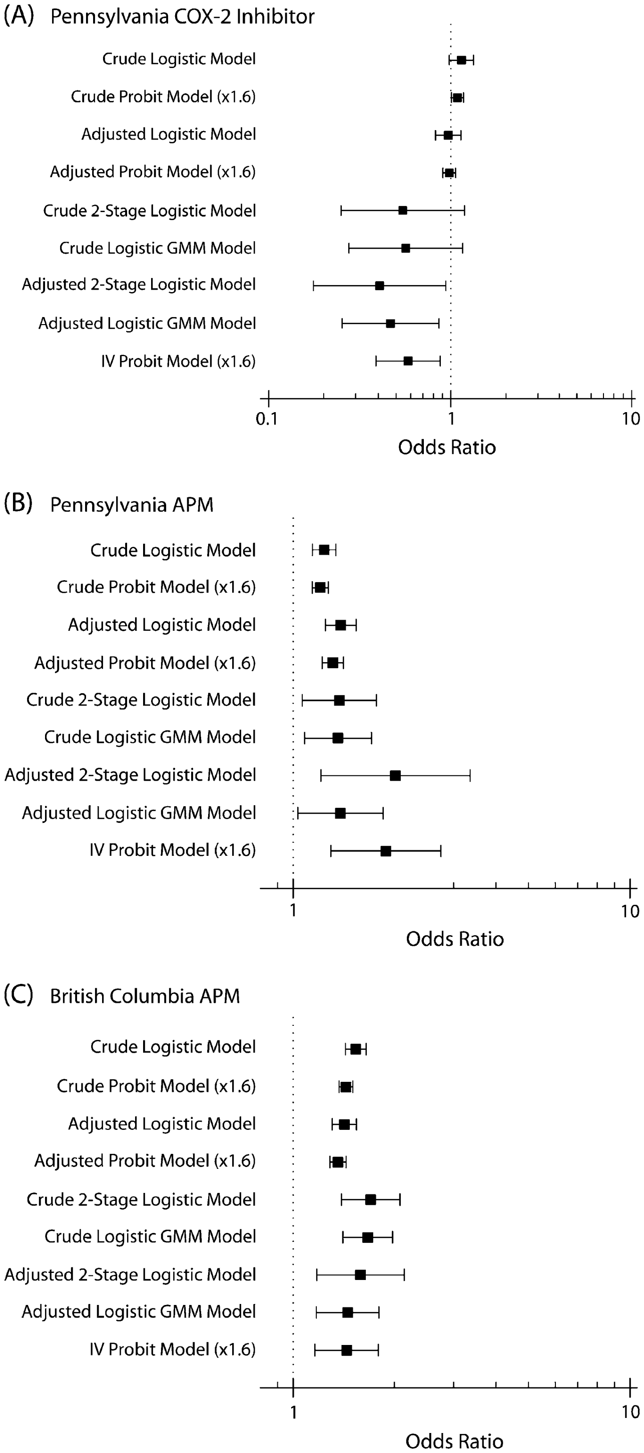


Figure 3. Point estimates for 10 models of relative risk or odds ratio in 3 cohorts of patients. (A) Pennsylvania cyclooxygenase 2 (COX-2) inhibitor cohort (1994–2003); (B) Pennsylvania antipsychotic medication (APM) cohort; (C) British Columbia, Canada, APM cohort (1996–2004). GMM, generalized method of moments; IV, instrumental variable. Bars, standard error.

or, alternatively, that the instrumental variable did not adjust for the unmeasured confounding that was present.

In the analysis of our data, we did not see great benefit to the marginal effects models that predicted risk differences; the marginal effects estimate and the 2-stage least squares estimate were very similar. While the marginal effects were derived from a 2-stage probit model which is appropriate for dichotomous data, the lack of analytic standard errors and the need to pick a point or set of points at which to estimate the marginal effect appeared to us to be drawbacks that outweighed the potential benefits. We did consider the marginal effect estimates to be helpful in validating the 2-stage least squares results that we observed.

We also did not see a large benefit of the 3-stage model as compared with the logistic/ordinary least squares model or 2-stage least squares. Again, while the 3-stage model is theoretically justified, the complexity did not seem to provide benefit in our data. In the 1 case where the approaches differed (British Columbia APM), the variation was well within the margin of error.

In the realm of the odds ratio, the 2-stage logistic model performed similarly to the GMM model in the COX-2 inhibitor cohort and in the British Columbia APM cohort; there was a difference in the Pennsylvania APM group. The 2-stage logistic model was faster and more straightforward to compute than GMM since we could implement it in a standard software package. While time and software may be practical concerns, when choosing a model we would place more weight on whether GMM's moments-based approach is more appropriate than the 2-stage logistic model's parametric requirements in cases where there is concern about model misspecification.

Many of our results were estimates of the risk ratio or odds ratio rather than the risk difference. Though the choice of preferred measure will be motivated by the investigators' needs in a particular study, we did see 1 overall benefit of the relative measures. Instrumental variable methods generally yield estimates with larger variance than do their conventional counterparts, and confidence intervals can be consequently large. As such, the confidence interval for a risk difference estimate may move outside reasonable bounds, when the confidence interval of a relative measure of the same effect could remain more plausible. Consider our COX-2 inhibitor example: The lower bound of the 2-stage least squares confidence interval for COX-2 inhibitors in Pennsylvania (left columns of Table 3) is −2.56 per 100, with a baseline risk in the unexposed of 1.38 per 100 (Table 2). This risk difference seems implausibly large, even allowing for the possibility of substantial treatment effect heterogeneity. On the other hand, from Table 4, the associated odds ratio predicted by the 2-stage logit model is 0.41, with a 95% confidence interval of (0.18, 0.94). The confidence interval is still wide, but to our eyes it does not present the same challenge for interpretability.

There are alternatives to the approaches presented here. In particular, a causal parameter using a structural mean model (34, 35) on the multiplicative scale can be computed. Following the equation shown in the appendix of the paper by Hernán et al. (8), we computed unadjusted causal risk ratios of 0.44 for the Pennsylvania COX-2 inhibitor cohort, 1.32

for the Pennsylvania APM cohort, and 1.61 for the British Columbia APM cohort; these figures were similar to the results obtained using 2-stage logistic regression. Bootstrapped standard errors were in line with other methods for the APM examples but much wider for the COX-2 inhibitor example. The width of the COX-2 inhibitor confidence interval may be due to the rarity of the lower frequency of events: 1.38% risk in the referent group in the COX-2 inhibitor study versus 13.53% in the Pennsylvania APM study. Beyond structural mean models, there are also approaches suggested by Abadie (36) and Mullahy (37).

We evaluated a number of models for estimating causal risk differences and relative risks in settings of expected strong unmeasured confounding. In our data, we saw relatively little difference between the various instrumental variable approaches, despite their reliance on different assumptions. The methods may yield differences that are more substantively relevant when many continuous covariates need to be included in the model and thus when modeling assumptions are likely to be more important. While we attempted solely to explore the issues empirically, a theoretical treatment of the comparative effectiveness of these models may be in order.

ACKNOWLEDGMENTS

Author affiliations: Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Boston, Massachusetts (Jeremy A. Rassen, Sebastian Schneeweiss, Robert J. Glynn, M. Alan Brookhart); Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts (Sebastian Schneeweiss, Murray A. Mittleman); Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts (Robert J. Glynn); Cardiovascular Epidemiology Research Unit, Beth Israel Deaconess Medical Center, Boston, Massachusetts (Murray A. Mittleman); and Harvard Medical School, Boston, Massachusetts (Jeremy A. Rassen, Sebastian Schneeweiss, Robert J. Glynn, Murray A. Mittleman, M. Alan Brookhart).

Funding for this study was provided by Brigham and Women's Hospital, Division of Pharmacoepidemiology and Pharmacoeconomics.

Conflict of interest: none declared.

REFERENCES

1. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*. 1994;272(11):859–866.
2. Wang PS, Schneeweiss S, Avorn J, et al. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. *N Engl J Med*. 2005;353(22):2335–2341.
3. Brookhart MA, Wang PS, Solomon DH, et al. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*. 2006; 17(3):268–275.
4. Schneeweiss S, Solomon DH, Wang PS, et al. 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 Rheum*. 2006;54(11): 3390–3398.
5. Stukel TA, Fisher ES, Wennberg DE, et al. 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.
6. Greene WH. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall; 2003.
7. Angrist JD, Imbens G, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996; 94(434):444–455.
8. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*. 2006;17(4):360–372.
9. Kennedy P. *A Guide to Econometrics*. 5th ed. Cambridge, MA: MIT Press; 2003.
10. Angrist JD. Estimations of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. *J Bus Econ Stat*. 2001;19(1):2–16.
11. Schneeweiss S, Setoguchi S, Brookhart MA, et al. Mortality in users of conventional and atypical antipsychotic medications in British Columbia seniors. *Can Med Assoc J*. 2007;126(5): 627–632.
12. Wooldridge JM. *Introductory Econometrics: A Modern Approach*. 3rd ed. Mason, OH: Thomson/South-Western; 2006.
13. Pearl J. *Causality: Models, Reasoning, and Inference*. New York, NY: Cambridge University Press; 2000.
14. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results [electronic article]. *Int J Biostat*. 2007;3(1):article 14.
15. Johnston KM, Gustafson P, Levy AR, et al. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Stat Med*. 2008;27(9):1539–1556.
16. Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. *Am J Epidemiol*. 2005;162(3): 199–200.
17. Amemiya T. Qualitative response models: a survey. *J Econ Lit*. 1981;19(4):1483–1536.
18. Agresti A. *Categorical Data Analysis*. New York, NY: John Wiley & Sons, Inc; 1990.
19. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models*. Boca Raton, FL: Chapman & Hall, Inc; 1998.
20. Henneman TA, van der Laan MJ, Hubbard AE. *Estimating Causal Parameters in Marginal Structural Models With Unmeasured Confounders Using Instrumental Variables.* (U.C. Berkeley Division of Biostatistics Working Paper Series, paper 104). Berkeley, CA: The Berkeley Electronic Press; 2002.
21. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol*. 1990;132(4):734–745.
22. Amemiya T. The non-linear two-stage least-squares estimator. *J Econom*. 1974;2(7):105–110.
23. Didelez V, Sheehan N. *Mendelian Randomisation and Instrumental Variables: What Can and What Can't Be Done.* (Technical report 05-02). London, United Kingdom: Department of Health Sciences, University of Leicester; 2005. (http://www.homepages.ucl.ac.uk/~ucakvdi/tech_rep_mendel.pdf).

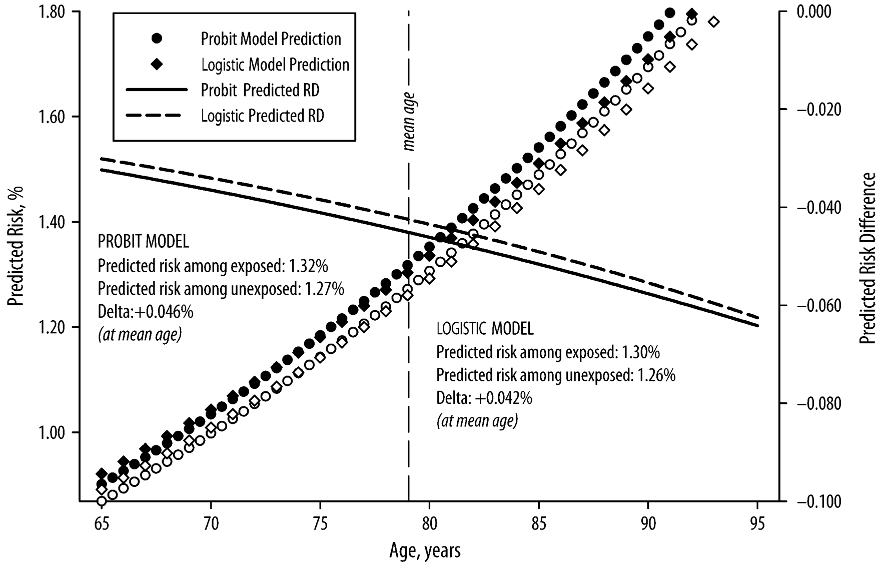
24. Foster EM. Instrumental variables for logistic regression: an illustration. *Soc Sci Res.* 1997;26(4):487–504.
25. Stata Corporation. *Stata, Version 9* [computer program]. College Station, TX: Stata Corporation; 2007.
26. Department of Economics, Boston College. *ivreg2: Stata Module to Extended Instrumental Variables/2-SLS, GMM and AC/HAC, LIML and k-Class Regression* [computer program]. Boston, MA: Department of Economics, Boston College; 2006.
27. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2007.
28. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall, Inc; 1993.
29. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003;158(9): 915–920.
30. Salzman C. *Clinical Geriatric Psychopharmacology.* 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2005.
31. Ray WA, Meredith S, Thapa PB, et al. Antipsychotics and the risk of sudden cardiac death. *Arch Gen Psychiatry.* 2001; 58(12):1161–1167.
32. Kuehn BM. FDA warns antipsychotic drugs may be risky for elderly. *JAMA*. 2005;293(20):2462.
33. Janssen, L.P. Risperdal® (risperidone) tablets/oral solution. Risperdal® M-Tab® (risperidone) orally disintegrating tablets [modified package insert]. Titusville, NJ: Janssen, L.P.; 2006. (http://www.fda.gov/cder/foi/label/2006/021444s008s015, 020588s024s028s029, 020272s036s041lbl.pdf). (Accessed November 6, 2006).
34. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun Stat Theory Methods*. 1994;23(8):2379–2412.
35. Goetghebeur E, Stijn V. Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Stat Methods Med Res.* 2005;14(4):397–415.
36. Abadie A. *Semiparametric Estimation of Instrumental Variable Models for Causal Effects*. Cambridge, MA: National Bureau of Economic Research, Inc; 2000.
37. Mullahy J. Instrumental-variable estimation of count data models: applications to models of cigarette smoking behavior. *Rev Econ Stat*. 1997;79(4):586–593.

APPENDIX

Estimating Risk Differences With Nonlinear Models

Both probit and logistic models can be used to estimate relative risks by examining the part of the slope of the distribution function that is ‘‘contributed’’ by the effect of the treatment. Since the function tracks a probability, the slope is the difference in probability of outcome associated with changing nothing but treatment status: a risk difference.

Mathematically, the risk difference can be calculated by taking the partial derivative of the logistic or probit cumulative distribution function with respect to the treatment. This is often called the ‘‘marginal effect’’ in software and in the literature, but note that the word ‘‘marginal’’ in this case is not referring to the marginal patient. This slope must be evaluated at a particular point and will not be the same throughout since the underlying function is not a straight line. Two approaches are conventional: 1) to evaluate the slope at the center of all measured variables (i.e., where age is at its mean, male gender is at its mean prevalence, etc.) and 2) to evaluate the slope at each point defined by the



Appendix Figure. Comparison of risks of severe gastrointestinal complications as predicted by the logit and 1.6× scaled probit models, plotted across the range of ages, using data on use of cyclooxygenase 2 inhibitors from Pennsylvania (1994–2003). Circles show the predicted risks from the probit model; diamonds show the predicted risks from the logit model. Filled symbols indicate predicted risk in exposed patients; unfilled symbols indicate predicted risk in unexposed patients. The difference between the exposed and the unexposed, interpretable as an age-adjusted risk difference (RD), is plotted on the right-hand vertical axis.

observations in the data (1,000 observations will define up to 1,000 points) and then take the mean of the observed slopes (6).

The Appendix Figure shows an example from the Pennsylvania cyclooxygenase 2 inhibitor study. In this figure, the probit- and logistic-predicted probabilities are plotted; we have ‘‘deconstructed’’ the slope (risk difference) by showing the exposed and unexposed groups separately. To demonstrate that the choice of evaluation point will affect the estimate, variation in 1 covariate (age) is plotted. The marginal effect observed at the mean age, interpretable as an age-adjusted risk difference, is 0.046 as estimated by the probit model and 0.042 as estimated by the logit model. The probit figure falls in between the observed crude risk difference of 0.19 and the fully adjusted risk difference of −0.05 (Table 3). Risk differences vary from approximately −0.03 to approximately −0.06 across the age range.