Published in final edited form as:

Stat Med. 2015 February 20; 34(4): 664–684. doi:10.1002/sim.6369.

A Bayesian Approach for Instrumental Variable Analysis with Censored Time-to-Event Outcome

Gang Li^a and Xuyang Lu^{a,*}

^aDepartment of Biostatistics, UCLA School of Public Health, Los Angeles, California 90095-1772, U.S.A.

Abstract

Instrumental variable (IV) analysis has been widely used in economics, epidemiology, and other fields to estimate the causal effects of covariates on outcomes, in the presence of unobserved confounders and/or measurement errors in covariates. However, IV methods for time-to-event outcome with censored data remain underdeveloped. This paper proposes a Bayesian approach for IV analysis with censored time-to-event outcome by using a two-stage linear model. A Markov Chain Monte Carlo sampling method is developed for parameter estimation for both normal and non-normal linear models with elliptically contoured error distributions. Performance of our method is examined by simulation studies. Our method largely reduces bias and greatly improves coverage probability of the estimated causal effect, compared to the method that ignores the unobserved confounders and measurement errors. We illustrate our method on the Women's Health Initiative Observational Study and the Atherosclerosis Risk in Communities Study.

Keywords

instrumental variable analysis; time-to-event outcome; Mendelian Randomization; measurement error; Bayesian method; accelerated failure time model

1. Introduction

Instrumental variable (IV) analysis is a statistical tool that has been extensively used by economists, epidemiologists, and others. It provides a useful method to (1) estimate the causal effects of covariates on outcomes; (2) address the bias caused by unobserved confounders; (3) account for measurement errors in the covariates. It involves using an extra variable, the *instrumental variable* or *instrument*, that satisfies certain criteria. As shown in Figure 1, the primary variables in an IV analysis are: outcome variable Y ; true value of covariate W , which could be unobserved due to measurement errors; observed surrogate X of the covariate W ; observed confounders Z ; unobserved confounders U ; and instruments G . The primary aim is to estimate the causal effect of W on Y , indicated by the dotted line in Figure 1. Without the use of instruments G , it is usually difficult to infer causation from

association between W and Y , and estimation of this association could be biased by unobserved confounders U and/or measurement errors in W . The instrument vector G is selected based on three criteria: (1) G is independent of U and measurement errors in W ; (2) G is associated with W , conditional on Z ; (3) G is independent of Y given W and Z .

IV analysis has a long history dating back to the early 20th century. An extensive literature has been developed in the field of economics for IV analysis with continuous outcomes [1–10]. The observed covariate X is commonly referred to as an *endogenous variable* in economics, meaning that X is correlated with the error term in the model regressing outcome Y on X . The corresponding regression coefficient is called an *endogenous parameter*. IV analysis became popular in the field of epidemiology in the past two decades, especially with the special case of *Mendelian Randomization* (MR), where genetic markers are used as instruments [11–15]. The term was first used by Gray and Wheatley [16], though its meaning has evolved over time [17]. It refers to the similarities between random assortment of genes following Mendel's law of independent assortment and random treatment allocation in a randomized control trial [18]. In MR analysis, Y is usually a disease related outcome, X is usually a modifiable exposure or phenotype, and G is a vector of genetic instruments. Therefore, MR is also known as IV analysis using genetic instruments [15]. Genetic markers arise as a natural choice of instrument due to the fact that they are generally independent of typical confounders such as behavioral and environmental factors. The selection of genetic instruments is mostly based on biological and clinical information, since IV assumptions (1) and (3) are difficult to be validated directly. It is worth noting that for IV assumption (2), G does not have to be causal in the association between G and W to be an IV: The association could be due to a mediator variable that affects both G and W (e.g. when there is linkage disequilibrium between G and another genotype) [19].

The classic IV analysis estimates the endogenous parameter indirectly, by estimating the association between G and Y and the association between G and X . This is usually done by using a two-stage least squares (TSLS) estimation in a simultaneous equation model [1, 5, among others]. Variance estimation of IV estimates were derived based on the TSLS procedure, including the Murphy-Topel estimator [20] and the Huber/White/Sandwich estimator [21, 22]. Specific forms of the two variance estimators in IV analysis were given by Hardin [23] and Hardin and Carroll [24]. An alternative way to draw inference on variance is to use posterior variance and a credible interval from a Bayesian model. Bayesian IV methods were developed for continuous outcomes, based on a more general form of the IV model and the assumption that the error terms follow a bivariate normal distribution [25, 26]. Reviews of Bayesian approaches for normal linear IV models, as well as their advantages and disadvantages compared to the classical frequentist approaches, are given by [27, 28]. Conley *et al.* [29] further developed a semiparametric Bayesian approach using a Dirichlet process prior and showed that it is more efficient when the error terms are non-normal.

Causal inference, deconfounding, and measurement error correction are the three main objectives of IV analysis, which have been well-established for continuous outcomes. Angrist and Imbens [30] and Angrist *et al.* [31] showed that the TSLS can be used to estimate the average causal effect of the covariate, and that the IV estimand can be

embedded within the Rubin Causal Model, which is a well-established framework of causal inference with observational data [32–34]. Didelez and Sheehan [13] presented a formal framework for causal inference based on the MR approach. More comprehensive illustrations on causal inference based on IV analysis can be found in [35, 36] and the references therein. Didelez and Sheehan [13] also proved the consistency of causal parameter estimates in the presence of unobserved confounders, based on a classic IV model where outcome Y and covariate W are both continuous, and a threshold IV model for situations when covariate W and instrument G are both binary [7]. An extensive literature on IV analysis focusing on the aspect of measurement error correction can be found in [37–43] and the references therein.

For binary outcomes, such as the presence or absence of some disease, Y is usually modeled by a logistic or probit regression model. However, for non-linear models, it is not always clear how the causal parameter is defined and how it relates to the model parameters [13]. Moreover, in the presence of unobserved confounders U or measurement errors in W , when there is a non-null causal effect of covariate W on the binary outcome Y , the IV estimation of the endogenous parameter is no longer consistent; it can only serve as an approximation when W is restricted to a small range and its causal effect is small [44]. Palmer *et al.* [45] investigated the extent of bias in classic MR analysis by simulations and showed that the effect size of the endogenous parameter tends to be underestimated. McKeigue *et al.* [46] proposed a simplified model that ignores measurement errors in W (i.e. assuming $X = W$), ignores the random error term in W , and assumes that all variation in W given instrument G can be explained by a univariate unobserved confounder. This assumption is strong and unlikely to be realistic, and it was observed in our simulations that violation of the assumption could largely bias the parameter estimate. However, once the random error term in W is non-ignorable, the model is similar to a logistic regression with binary outcome and measurement error in the covariate. Kuchenhoff [47] proves the theoretical identifiability of this model, but also points out that it is not numerically identifiable.

It is often of interest in observational studies to investigate the effect of some covariate on time to the occurrence of events such as disease and death, while the same issues of causal inference, unobserved confounders and/or measurement errors need to be considered. The problem can be further complicated by censoring in the time-to-event outcome. By using instruments satisfying corresponding assumptions, the IV approach can be accommodated into survival analysis and provide a possible solution to this problem. The built-in Stata command “ivtobit” [48] and user-written Stata command “cmp” [49] can be used to fit a linear IV model with a tobit model for the censored time-to-event data. They use a maximum likelihood approach by assuming a bivariate normal error distribution. Bijwaard [50] proposed an IV Linear Rank estimator for right-censored time-to-event data, based on a Generalized Accelerated Failure Time model. In addition, generalized method of moments (GMM) estimators [51–53, among others] could potentially be used to draw parametric maximum likelihood inference for IV analysis with survival outcome. Nevertheless, IV methods for survival analysis remain underdeveloped, and yet to our knowledge no Bayesian IV model for survival outcome has been formally developed.

In this paper, we develop a Bayesian approach for IV analysis with censored time-to-event outcome, based on a two-stage linear model. A general form of the IV model is used so that all model parameters are identifiable. This is done by jointly modelling the random error term in the time-to-event outcome and the random error term in the continuous covariate by a bivariate normal or other parametric bivariate distribution such as an elliptically contoured distribution. This approach can be considered as an extension of the Bayesian model described in [25, 26], into analysis of arbitrarily censored survival data, with more general error distributions. Markov Chain Monte Carlo (MCMC) sampling methods are developed. Performance of the model is evaluated by simulation studies, under frequentist criteria of bias, standard deviation and coverage probability. By using a Bayesian approach, prior information (such as estimates and confidence intervals of the parameters from former studies) could be incorporated by using informative priors. We show that the causal effect of covariate W on time-to-event outcome Y (the endogenous parameter) can be consistently estimated in the presence of unobserved confounders U and measurement errors in W . We illustrate our method using two real data sets: the Women's Health Initiative Observational Study (WHI-OS) and the Atherosclerosis Risk in Communities (ARIC) Study. In the WHI-OS, we investigate the effect of high-sensitivity C-reactive protein (hsCRP) on time to occurrence of diabetes, using multiple selected single-nucleotide polymorphisms (SNPs) as genetic instruments. In the ARIC study, we examine the causal effect of systolic blood pressure (SBP) on time to diagnosis of coronary heart disease (CHD), using the IV model to correct for measurement errors in SBP. MCMC convergence diagnostics are conducted. Although we focus on right-censored time-to-event outcome for simplicity of illustration, our method can be easily extended to the case of arbitrary censoring (e.g. left censoring, interval censoring).

The rest of this article is organized as follows. Section 2 describes the model for IV analysis with time-to-event outcome, along with the MCMC procedure. Section 3 includes simulation studies in which the performance of the model is examined. Section 4 contains application of the model to two real studies and the MCMC convergence diagnostics. Section 5 contains some remarks and future directions.

2. The Model, Estimation and Inference Procedure

2.1. The Data and Model

For each subject i , let Y_i be the time-to-event outcome variable, and C_i be a corresponding right-censoring time. Only the smaller of the two, $T_i = \min(Y_i, C_i)$, can be observed. Let W_i be an unobserved continuous covariate of interest that is subject to measurement errors, X_i be the observed surrogate of W_i , Z_i be a vector of observed confounders, U_i be a vector of unobserved confounders, and G_i be a vector of instruments, $i = 1, \dots, n$. The primary aim is to estimate the causal effect of W_i on Y_i based on the observed right-censored data consisting of n independent and identically distributed observations $(T_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \dots, n$, where $\delta_i = I[Y_i \leq C_i]$ is the censoring indicator.

We consider the following two-stage linear model:

$$X_i = \alpha_0 + \alpha_1' G_i + \alpha_2' Z_i + \xi_{1i} \quad (1)$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2' Z_i + \xi_{2i} \quad (2)$$

where the random errors ξ_{1i} and ξ_{2i} jointly follow a bivariate normal distribution:

$$\begin{pmatrix} \xi_{1i} \\ \xi_{2i} \end{pmatrix} \stackrel{\text{i.i.d}}{\sim} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right) \quad (3)$$

The time-to-event outcome Y_i is usually a monotone transformed survival time. For example, the second-stage model (2) is an accelerated failure time model if Y_i is the log-transformed survival time [54]. The variable X_i in equation (2) is often referred to as an endogenous variable especially in economics, meaning that it is correlated with the error term ξ_{2i} . Our primary goal is to estimate and make inference on the endogenous parameter β_1 . Although we aim to estimate the endogenous parameter in a censored time-to-event context, the underlying model is the same as the two-stage linear model for continuous outcomes without censoring. Censoring only affects the estimation procedure, not the interpretation of the underlying causal model. In addition, the normality assumption in (3) is not essential for the development of our estimation and inferential method. In section 2.3, we note that our approach can be extended with minimal modifications to elliptically contoured models that include many useful non-normal models.

This model can be used to adjust for unobserved confounders and measurement errors simultaneously. This can be shown by the derivation of equations (1) and (2) as follows. With assumption of linear relationships among the variables Y_i , W_i , U_i , Z_i and G_i , the underlying structure in figure 1 can be modeled by

$$W_i = \alpha_0 + \alpha_1' G_i + \alpha_2' Z_i + \alpha_3' U_i + \varepsilon_{1i} \quad (4)$$

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2' Z_i + \beta_3' U_i + \varepsilon_{2i} \quad (5)$$

$$X_i = W_i + \varepsilon_{3i} \quad (6)$$

$i = 1, \dots, n$, where ε_{1i} , ε_{2i} , and ε_{3i} are independent random errors with mean zero and finite variances τ_1^2 , τ_2^2 , and τ_3^2 respectively. Note that ε_{3i} is the measurement error in the covariate. The unobserved confounder vector U_i is standardized to have mean 0 and covariance matrix Σ_U , where Σ_U is the correlation matrix of U_i . The variables ε_{1i} , ε_{2i} , ε_{3i} , U_i and G_i are assumed to be mutually independent. Since W_i is not observed, we can replace W_i in equations (4) and (5) by $W_i = X_i - \varepsilon_{3i}$ from equation (6) and have the following two-stage linear model:

$$X_i = \alpha_0 + \alpha_1' G_i + \alpha_2' Z_i + \alpha_3' U_i + \varepsilon_{1i} + \varepsilon_{3i} \quad (7)$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2' Z_i + \beta_3' U_i + \varepsilon_{2i} - \beta_1 \varepsilon_{3i} \quad (8)$$

Based on these two equations, parameters related to unobserved confounders U_i and random errors $(\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})$, which are $(\alpha_3, \beta_3, \tau_1^2, \tau_2^2, \tau_3^2)$, are not all identifiable. Since the parameter of primary interest is β_1 , we can reduce the two equations to the two-stage model with equations (1) and (2), by combining the unobserved confounders and random errors:

$$\xi_{1i} = \alpha_3' U_i + \varepsilon_{1i} + \varepsilon_{3i} \quad \text{and} \quad \xi_{2i} = \beta_3' U_i + \varepsilon_{2i} - \beta_1 \varepsilon_{3i}$$

With normality assumptions on U_i , ε_{1i} , ε_{2i} and ε_{3i} , the random errors ξ_{1i} and ξ_{2i} jointly follow a bivariate normal distribution as shown in equation (3), where

$$\sigma_1^2 = \alpha_3' \Sigma_U \alpha_3 + \tau_1^2 + \tau_3^2, \sigma_2^2 = \beta_3' \Sigma_U \beta_3 + \tau_2^2 + \beta_1^2 \tau_3^2, \text{ and } \rho \sigma_1 \sigma_2 = \alpha_3' \Sigma_U \beta_3 - \beta_1 \tau_3^2.$$

2.2. Estimation and Inference

We develop a Bayesian approach with MCMC techniques to draw inferences on the endogenous parameter β_1 . For the two-stage IV model (1)–(3), denote

$$\theta = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \rho), \vec{T} = (T_1, \dots, T_n), \vec{\delta} = (\delta_1, \dots, \delta_n), \vec{X} = (X_1, \dots, X_n), \\ \vec{Z} = (Z_1, \dots, Z_n) \text{ and } \vec{G} = (G_1, \dots, G_n). \text{ The likelihood function of observing } (\vec{T}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) \text{ is:}$$

$$\begin{aligned} \mathcal{L}(\theta | \vec{T}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) &= P(\vec{T}, \vec{\delta} | \vec{X}, \vec{Z}, \vec{G}, \theta) \cdot P(\vec{X}, \vec{Z}, \vec{G} | \theta) \\ &= \prod_{i=1}^n f_1(T_i | X_i, Z_i, G_i)^{\delta_i} S(T_i | X_i, Z_i, G_i)^{1-\delta_i} \cdot f_2(X_i, Z_i, G_i) \end{aligned} \quad (9)$$

where

$$\begin{aligned} S(T | X, Z, G) &= 1 - \Phi \left(\frac{T - \beta_0 - \beta_1 X - \beta_2' Z - \frac{\sigma_2^2}{\sigma_1^2} \rho (X - \alpha_0 - \alpha_1' G - \alpha_2' Z)}{\sqrt{(1-\rho^2)\sigma_2^2}} \right) \\ f_1(T | X, Z, G) &= \phi \left(\frac{T - \beta_0 - \beta_1 X - \beta_2' Z - \frac{\sigma_2^2}{\sigma_1^2} \rho (X - \alpha_0 - \alpha_1' G - \alpha_2' Z)}{\sqrt{(1-\rho^2)\sigma_2^2}} \right) \\ f_2(X, Z, G) &= \phi \left(\frac{X - \alpha_0 - \alpha_1' G - \alpha_2' Z}{\sqrt{\sigma_1^2}} \right) \end{aligned}$$

$\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and the probability density function of standard normal distribution, respectively. The detailed derivation of the likelihood is given in the appendix (subsection A.1).

We use independent vague priors for the parameters: a normal distribution $N(0, \zeta^2)$ with large variance ζ^2 for each element in $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1$ and β_2 , an inverse-gamma distribution $\text{Inv-Gamma}(\gamma_1, \gamma_2)$ with small shape parameter γ_1 and small scale parameter γ_2 for σ_1^2 and σ_2^2 , and a uniform distribution $\text{Unif}(-1, 1)$ for ρ . The MCMC method is used to generate

samples from the posterior distributions of the parameters: In each iteration, a random walk Metropolis-Hastings algorithm [55, 56] is used to update the parameters one by one, while other parameters are fixed at their current states. Highly correlated parameters are updated simultaneously using a multiple-block Metropolis-Hastings algorithm [57] in order to have faster convergence for the Markov chains. Uniform proposal distributions are used for the random walk in our simulations and real data analysis, with widths chosen to obtain appropriate acceptance rates. The detailed MCMC algorithm is provided in the appendix (subsection A.2). A sufficiently large amount of MCMC samples are generated from the posterior distribution. Sample mean of a parameter can be used to approximate the posterior mean and serve as an estimation of the parameter. Credible intervals of the parameters can be constructed by using the empirical quartiles of the simulated samples. We implemented the method in R [58]. Our program is available online at http://www.biostat.ucla.edu/people/gangli/IV_MH.R, along with an example with simulated data at http://www.biostat.ucla.edu/people/gangli/IV_example.R.

Convergence of the MCMC algorithm can be examined visually by graphical methods including trace plots, histograms and autocorrelation plots, and quantitatively by using the Brooks-Gelman-Rubin diagnostics [59]. Detailed convergence diagnostics are presented in section 4.3 for the real data examples.

2.3. Extension to Non-Normal Models

The method developed earlier for the normal IV model (1)–(3) can be easily extended to a more general IV model with elliptically contoured distributed errors. Specifically, the normality assumption (3) can be replaced by the following assumption

$$\begin{pmatrix} \xi_{1i} \\ \xi_{2i} \end{pmatrix} \stackrel{\text{i.i.d}}{\sim} EC_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, g \right), \quad (10)$$

where a k -dimensional random vector ξ is said to have an elliptically contoured distribution (ECD), denoted by $\xi \sim EC_n(\mu, \Sigma, g)$, if it has the following density function

$$|\Sigma|^{-1/2} g \left[(x - \mu)' \Sigma^{-1} (x - \mu) \right]$$

for a given function g . Clearly the k -dimensional multivariate normal distribution is an ECD with $g(x) = (2\pi)^{-\frac{k}{2}} e^{-\frac{x}{2}}$. The ECD also includes many non-normal multivariate distributions such as the multivariate t , the multivariate Cauchy, the multivariate Laplace, the multivariate uniform, scale mixtures of normal distributions, and the multivariate stable distributions. See Fang and Anderson [60], Chmielewski [61], and the references therein for some nice survey of ECD and its applications in various areas, including robust regression [62], risk measure [63], hyperspectral imaging data modeling [64], etc.

The family of ECDs share many nice properties of the multivariate normal, including that all marginal distributions and all conditional distributions of an ECD are ECD. With a bivariate ECD assumption on random errors $(\xi_{1i}, \xi_{2i})'$ other than the bivariate normal, we will simply modify the functions $S(T/X, Z, G)$, $f_1(T/X, Z, G)$ and $f_2(X, Z, G)$ in the likelihood function

(9) accordingly: $S(T/X, Z, G)$ and $f_1(T/X, Z, G)$ are derived from the conditional distribution of ξ_{2i} given ξ_{1i} ; $f_2(X, Z, G)$ is derived from the marginal distribution of ξ_{1i} . Both are univariate ECDs in explicit form determined by the given function g .

3. Simulation Studies

A simulation study is conducted to assess the performance of our Bayesian IV model under frequentist criteria of bias, standard deviation (SD), and coverage probability (CP).

Synthetic data is generated following the underlying model with unobserved confounders and measurement errors given by equations (7) and (8), with a variety of sample sizes, instrument strengths, censoring rates, and effects of covariate. Specifically, we use a simulation setting that is motivated by the WHIOS real data example in section 4.1.

For the first-stage equation (7): A univariate instrument G_i and a univariate unobserved confounder U_i are used. For simplicity of illustration, we do not include observed confounder vector Z_i in our simulations, i.e. $\alpha_2 = \beta_2 = 0$. G_i and U_i both follow a standard normal distribution $N(0, 1)$; ε_{1i} and ε_{3i} follow normal distributions $N(0, 0.04)$ and $N(0, 0.015)$, respectively. The regression parameters are set as $\alpha_0 = 0.5$, $\alpha_1 = \sqrt{0.005}$, $\alpha_3 = \sqrt{0.04}$. This gives $E(X) = 0.5$ and $V ar(X) = 0.1$, while $Cor^2(X, G) = 5\%$, similar to the mean and variance of the covariate and its correlation with the genetic instrument in the WHI-OS example. We also consider a stronger instrument with $Cor^2(X, G) = 10\%$ by setting $\alpha_1 = \sqrt{0.01}$ and $\varepsilon_{1i} \sim N(0, 0.035)$ while other parameters remain the same. For the second-stage equation (8): Different values are used for the endogenous parameter: $\beta_1 = (0, -0.5, -1)$, representing none, small, or moderate causal effects, respectively, of underlying true covariate W on outcome Y . These values of β_1 correspond to acceleration factors of 1, 1.6 and 2.7 when the outcome Y is a log-transformed survival time and an accelerated failure time model is used. Other regression parameters are set as $\beta_0 = 5 - \alpha_0\beta_1$ and $\beta_3 = \sqrt{0.05}$, and the random error ε_{2i} follows a normal distribution $N(0, 0.45)$. This gives $E(Y) = 5$ and $V ar(Y) = 0.5$ when $\beta_1 = 0$, similar to the time-to-event outcome in the WHI-OS example. The underlying right-censoring time is set to be conditionally independent of Y_i : $C_i = \beta_0 + \beta_1 X_i + \beta_3 U_i + \varepsilon_{ci}$, where $\varepsilon_{ci} \sim N(\mu_c, 2)$ and μ_c is adjusted to give censoring rates 25%, 50% and 75%. The observed time T_i and censoring indicator δ_i are derived by $T_i = \min(Y_i, C_i)$ and $\delta_i = I[Y_i \leq C_i]$. We use sample size $n = (300, 500, 800)$ for each of the parameter settings.

We first use a regular linear regression survival model (without using IV) to evaluate the bias caused by the unobserved confounder U_i and the measurement error ε_{3i} . We generate 5000 synthetic data sets for each of the simulation settings. The following linear model is applied to each set of observed data (T_i, δ_i, X_i) , $i = 1, \dots, n$:

$$Y_i = \beta_1 X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(\mu, \sigma^2) \quad (11)$$

We denote this as the ‘simple method’. Estimates and 95% confidence intervals of the parameters $(\beta_1, \mu \text{ and } \sigma^2)$ are derived based on maximum likelihood estimation and asymptotic normal approximation [54]. This can be easily implemented by using the *survreg* function in R package ‘survival’ [65] or the LIFEREG procedure [66] in the SAS®

software. Expectation and SD of the parameter estimates are approximated by the sample mean and sample SD of the 5000 estimates from independent simulated data sets. Coverage probability is approximated by the proportion of confidence intervals that cover the true value of the parameter.

We then apply our proposed IV model (1)–(3) to the simulated data. We generate 2000 synthetic data sets for each of the simulation settings described earlier. For each data set, the likelihood is constructed using the observed data $(T_i, \delta_i, X_i, G_i)$, $i = 1, \dots, n$ and likelihood function (9). Independent vague priors are used: $N(0, 100^2)$ for each of α_0 , α_1 , β_0 , and β_1 ; Inv-Gamma(0.001, 0.001) for σ_1^2 and σ_2^2 , and Unif(−1, 1) for ρ . MCMC samples are generated from the posterior distribution, which is the product of the likelihood and the priors. The detailed MCMC algorithm is described in the appendix (subsection A.2). Specifically, a width of the proposal distribution that gives an acceptance rate between 0.3 and 0.4 is chosen for each parameter. We run 110,000 iterations and use the first 10,000 iterations as burn-in to achieve a state of convergence. We thin the resulting chain by taking every 5th sample, in order to reduce autocorrelation. Posterior means and credible intervals are derived for each parameter based on the resulting 20,000 posterior samples. Similar to simulations using the simple method described earlier, expectation and SD of the parameter estimate are approximated by the sample mean and sample SD of the 2000 posterior means from the independent simulated data sets. Coverage probability is approximated by the proportion of credible intervals that cover the true value of the parameter.

Results of the parameter of primary interest, β_1 , from the two analysis models are summarized in Table 1. The β_1 estimates from the simple method have substantial bias in all simulation settings: bias $\approx 0.45, 0.37$ and 0.29 for $\beta_1 = 0, -0.5$ and -1 , respectively, insensitive to different sample sizes and different censoring rates. The coverage probabilities are poor, especially when sample size is large or censoring rate is low (e.g. CP = 0.36 when $n = 800$; CP = 0.43 when censoring rate = 25%). This is because the SD is smaller in these situations, while the bias remains large. On the other hand, the bias in estimating β_1 from our IV method is much smaller. For example, with instrument strength $R^2(X, G) = 0.05$ and $n = 800$, the bias of the IV estimate is 0.04. With a stronger instrument of $R^2(X, G) = 0.1$ and $n = 800$, the bias is further reduced to 0.015. Coverage probabilities of the 95% credible intervals from the IV method are very close to the nominal level. Moreover, the bias from the IV method tends to reduce as sample size increases, under different scenarios of censoring rates, instrument strengths and true values of β_1 , suggesting consistency of the Bayesian IV estimation. The SD and coverage probability seem to be insensitive to different values of β_1 . Unsurprisingly, the IV method tends to perform better in terms of bias and SD with stronger instrument and/or smaller censoring rate, and the SD decreases as sample size increases.

Using the IV method will result in larger variation in the estimates compared to the simple method, due to the uncertainty from the first-stage model (1). Our simulation results show that the β_1 estimates from our IV method have a four-to-five-fold increase in SD with instrument strength $R^2(X, G) = 0.05$ and a roughly threefold increase in SD with instrument strength $R^2(X, G) = 0.1$ compared to the simple method. This is an inevitable trade-off between accuracy and precision. The gains in bias reduction come at the cost of lower

statistical power. It is worthwhile to apply the IV method when the major concern lies with bias in parameter estimation, and instruments with reasonable strength are available.

We further investigate the performance of our method by examining all the parameter estimates, with extensive simulations by varying the true values of all parameters. Our method performs well for estimation of all parameters in the IV analysis model: α_0 , α_1 , β_0 , β_1 , σ_1^2 , σ_2^2 , and ρ , in terms of bias and coverage probability (partial results reported in the appendix (subsection A.3)). In addition, we have conducted simulations with some vague dependent priors. The conclusions are consistent with the ones with vague independent priors.

In order to investigate the robustness of our proposed model against deviation of the normality assumption, we conducted another simulation study with different true underlying distributions for the time-to-event outcome Y . The simulation setting is the same as the previous one with $R^2(X, G) = 0.05$, except for the random error terms ε_{2i} in equation (2) and ε_{ci} in censoring time C_i . ε_{2i} follows distributions of four different forms: normal, exponential, weibull, and a mixture of normal distributions, while fixing $E(\varepsilon_{2i}) = 0$ and $V ar(\varepsilon_{2i}) = 0.45$. ε_{ci} follows the same distribution as ε_{2i} , fixing the censoring rate at 50%. We apply the proposed IV method to each of the data sets, similar to described earlier. Results of estimating β_1 are summarized in Table 2. Again, when the distribution of ε_{2i} is normal, the IV estimation of β_1 has bias becoming close to 0 as sample size increases, and the coverage probability is close to the nominal level of 95%. These properties also hold for IV estimation of β_1 when ε_{2i} follows the three non-normal distributions. This suggests that the IV method based on normality assumption is quite robust against the deviation from normality.

4. Real Data Examples

4.1. Women's Health Initiative Observational Study

We illustrate the proposed IV method using two real data examples. The first one is a prospective case-control study nested within the Women's Health Initiative Observational Study (WHI-OS). In this study, we want to investigate the effect of high-sensitivity C-reactive protein (hsCRP) on development of diabetes. hsCRP is an inflammatory marker that has been positively associated with diabetes [67–69]. However, whether lowering hsCRP level will result in diabetes prevention is uncertain. Therefore, we apply the Mendelian Randomization (MR) method (i.e. IV analysis with genetic instruments) with time-to-event outcome to make inference about the causal effect of hsCRP on diabetes and account for potential impact of unobserved confounders and measurement errors.

In the WHI-OS, 82069 postmenopausal women (50–59 years of age) with no history of diabetes were followed-up for a mean of 5.5 years. 1584 cases of diabetes were identified and matched with 2198 controls (by age, ethnicity, clinical center, time of blood draw, and length of follow-up). We focus on the subgroup of whites (954 cases and 968 controls) to avoid the potential problem of population stratification. Time to diabetes diagnosis from baseline was recorded for each case, and time to last visit from baseline for each control. Plasma concentration of hsCRP is measured for each subject. Descriptive statistics of

baseline characteristics are summarized by case-control status in Table 3. More detailed descriptions of the study are given in Liu *et al.* [69] and Chan *et al.* [70]. 13 haplotype-tagging single-nucleotide polymorphisms (tSNPs) across 2.3 kb of the *CRP* (C-reactive protein, pentraxin-related) genes that had been shown to account for most of the genetic variation within the *CRP* locus are used as instruments. Details of selection of the tSNPs are given in Lee *et al.* [71]. None of the 13 tSNPs show indication of Hardy-Weinberg disequilibrium (all p-values from a Hardy-Weinberg equilibrium test > 0.05 after Bonferroni correction).

We first apply a simple method similar to what has been described in section 3 to estimate the association between hsCRP and time to diabetes diagnosis:

$$Y_i = \beta_1 X_i + \beta_2' Z_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(\mu, \sigma^2) \quad (12)$$

$i = 1, \dots, n$. For each subject i , Y_i is the log-transformed time to diabetes diagnosis (in days); X_i is the log-transformed hsCRP level; Z_i is a vector of observed potential confounders including age, body mass index, cigarette smoking, alcohol intake, hormone-replacement therapy, family history of diabetes and physical activity. The censoring indicator δ_i is 1 for cases and 0 for controls. The observed time variable T_i is log-transformed time to diabetes diagnosis for cases and log-transformed time to last visit for controls. Since the outcome Y_i is a log-transformed survival time, this model is a log-normal accelerated failure time model. By using the SAS® procedure LIFEREG [66], β_1 has an estimate (SE) of -0.446 (0.089) with a p-value $< .001$. This correspond to a 95% confidence interval of $(-0.621, -0.272)$ as summarized in Table 4. This significant negative association is consistent with the previous finding by Liu *et al.* [69], who analyzed the same data set and found that hsCRP was significantly associated with increased diabetes risk (odds ratio = 1.16 with 95% confidence interval $(1.03, 1.30)$ and a p-value $< .001$).

We then apply the proposed IV method with two-stage model (1)–(3) to estimate the causal effect of hsCRP on time to diabetes diagnosis. For each subject i , instrument G_i is a vector of the 13 tSNPs, where each tSNP is coded as an additive effect model, i.e. coded as either 0, 1, or 2 depending on the number of minor alleles. Y_i , X_i , Z_i , T_i and δ_i are defined as earlier. The likelihood is constructed using observed data $(T_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \dots, n$ and likelihood function given by equation (9). Independent vague priors are used: $N(0, 100^2)$ for each element of a_0 , a_1 , a_2 , β_0 , β_1 and β_2 ; Inv-Gamma(0.001, 0.001) for σ_1^2 and σ_2^2 , and Unif($-1, 1$) for ρ . In the MCMC sampling, a width of the proposal distribution that gives an acceptance rate between 0.35 and 0.4 is chosen for each parameter. We generate 30 chains from different initial values, with 1, 100, 000 iterations (100, 000 burn-ins) in each chain. We thin the chains by taking every 10th sample to reduce autocorrelation. A detailed discussion of convergence is given in section 4.3. Posterior mean, posterior standard deviation and a credible interval are derived for each parameter based on the resulting 3, 000, 000 combined samples. Figure 2(a) is a histogram of the resulting MCMC samples of β_1 . The brackets on the horizontal axis denote the 95% credible interval. The posterior distribution of β_1 is fairly normal, with mean -0.162 , standard deviation 0.426 and 95% credible interval $(-0.987, 0.685)$, as summarized in Table 4.

Based on these results, we see that although hsCRP is significantly associated with development of diabetes, there is no sufficient evidence of causal effect of hsCRP on time to diabetes diagnosis among white postmenopausal women. This is consistent with the previous finding by Brunner *et al.* [72], who applied the MR approach in a case-control study and found that the associations between C-reactive protein (CRP) and diabetes incidence are likely to be noncausal. One possible explanation for the association is that hsCRP level is affected by causal factors of diabetes, such as obesity [73]. On the other hand, the instruments in this IV analysis may not be strong enough to provide sufficient statistical power to detect small effect sizes (partial R-squared = 0.028), even though the sample size is reasonably large.

4.2. Atherosclerosis Risk in Communities Study

The second data example is a subset in the Atherosclerosis Risk in Communities (ARIC) Study. The ARIC study is a multi-center prospective cohort study of cardiovascular disease and its risk factors. A total of 15,792 subjects aged 45-64 years were recruited from four US communities in 1987-89. They received 4 clinical examinations at 3-year intervals (Visits 1-4). Medical, social and demographic data were collected at each visit. Hospitalization information was obtained by annual telephone follow-up and active surveillance in the communities. A more detailed description of the study is reported elsewhere [74]. In this study, we are interested in estimating the association between systolic blood pressure (SBP) and development of coronary heart disease (CHD), after correcting for potential bias due to measurement error through IV analysis. For each of Visits 1-4, a subject's SBP level is an average of three measurements. We use Visit 2 (1990-92) as baseline, and use the SBP level at Visit 1 (1987-89) as an instrument of the baseline SBP level. We exclude subjects that (1) have missing baseline information, (2) do not have information after baseline, and/or (3) have developed their first CHD event prior to baseline. After the exclusion, our data consists of 12,782 subjects, 768 of which have CHD events during the follow-up. Descriptive statistics of baseline characteristics are summarized in Table 5.

For each subject i , outcome Y_i = time to the first CHD event from baseline (Visit 2) in years; censoring indicator $\delta_i = 1$ if the subject has at least one CHD event during the follow-up, and 0 otherwise; observed time $T_i = Y_i$ if $\delta_i = 1$, and time to the last visit from baseline in years otherwise; covariate of interest X_i = standardized log-transformed SBP level at baseline; instrument G_i = standardized log-transformed SBP level at Visit 1; both X_i and G_i are standardized to have standard deviation 1; Z_i is a vector of observed potential confounders at baseline, including ethnicity (black vs. non-black) and other potential risk factors of CHD developed by the Framingham Heart Study: gender, age, total cholesterol level, high-density lipoprotein cholesterol level, smoking behavior, and diabetes status [75].

Similar to the previous example in Section 4.1, we first apply the simple method with equation (12) to estimate the association between SBP and time to CHD. β_1 has an estimate (SE) of -0.779 (0.087) with a p-value $<.001$, corresponding to a 95% confidence interval of $(-0.950, -0.608)$ as summarized in Table 6. We then apply the proposed IV method with two-stage model (1)–(3), primarily aiming to correct for potential measurement error bias. Similarly, the likelihood is constructed using observed data $(T_i, \delta_i, X_i, Z_i, G_i)$, $i = 1, \dots, n$ and

likelihood function (9). Similar independent vague priors are used as described in Section 4.1. In the MCMC sampling, a width of the proposal distribution that gives an acceptance rate between 0.2 and 0.25 is chosen for each parameter. Due to the high correlations between some of the parameters ($\text{cor}(\alpha_0, \alpha_1) \simeq -0.95$, $\text{cor}(\beta_0, \beta_1) \simeq -0.97$ in the posterior samples), we used a multiple-block Metropolis-Hastings algorithm [57] to update the parameters: α_0 and α_1 are updated simultaneously; β_0 and β_1 are updated simultaneously. This dramatically improves convergence and results in greatly reduced autocorrelations. We generate 40 chains from different initial values, with 2, 100, 000 iterations (100, 000 burn-ins) in each chain. The chains are thinned to reduce autocorrelation by taking every 20th sample. A detailed discussion of convergence is given in section 4.3. Posterior mean, posterior standard deviation and a credible interval are derived for each parameter based on the resulting 4, 000, 000 combined samples. Figure 2(b) is a histogram of the resulting MCMC samples of β_1 . The brackets on the horizontal axis denote the 95% credible interval. The posterior distribution of β_1 appears to be normal. It has mean -1.180 , standard deviation 0.141 and 95% credible interval $(-1.460, -0.907)$, as summarized in Table 6. A standard deviation increase in log-transformed SBP level is associated with an acceleration of 1.18 years in time to the first CHD event. We observe a larger effect size of SBP on CHD development compared to the simple analysis. This result suggests that the effect size of β_1 calculated by the simple method is possibly attenuated by measurement errors in X_i .

In the IV analysis, we assume that the SBP measurement at an earlier visit is an instrument of the SBP measurement at a later visit. This assumption is weaker than the assumption that both the earlier and later measurements are replicates of noisy surrogate [42,43]. Since a subject's SBP level at a certain time point is naturally predictive of his/her SBP level three years later, G_i is a strong instrument of X_i (partial R-squared = 0.35). Furthermore, measurement of the instrument is not necessary to be accurate: the SBP at Visit 1 can still serve as an instrument if it is also subject to measurement errors.

4.3. MCMC Convergence Diagnostics

We assess the convergence of MCMC sampling in the two real data examples. Trace plots of parallel chains with diverse initial values are monitored for all the parameters. Figure 3 shows the trace plots of β_1 : (a) for the WHI-OS data and (b) for the ARIC data. Different chains are marked with different colors. The chains seem to be mixing well and stable over the whole period. We further visually compare the histograms of individual chains to the histogram of combined samples. No obvious difference is observed. Autocorrelation plots of individual chains are monitored for each parameter. Figure 4 shows the autocorrelation plots of two representative individual chains for β_1 : (a) for the WHI-OS data and (b) for the ARIC data. The autocorrelation is relatively low in the individual chains (generally less than 0.1 after a lag of 1000 for the WHI-OS data and less than 0.1 after a lag of 500 for the ARIC data). Finally, we use the Brooks-Gelman-Rubin diagnostics [59] to quantitatively measure convergence. The 'potential scale reduction factor' (PSRF) is calculated for each parameter, together with its 95% confidence interval. Approximate convergence is diagnosed when the upper limit of PSRF is close to 1. The 95% upper confidence limit of PSRF for β_1 is <1.005 for both real data examples, indicating good convergence properties of the method.

5. Discussion

We have developed a Bayesian approach for IV analysis to examine the causal effect of a covariate on a censored time-to-event outcome, in the presence of unobserved confounders and/or measurement errors in the covariate. We show by simulations that the proposed method largely reduces bias in estimation and it greatly improves coverage probability of the endogenous parameter, compared to the ‘simple method’ where the unobserved confounders and measurement errors are ignored. The method works well in a variety of settings, provided that the instrumental variable assumptions described early in introduction are satisfied. To test the instrumental variable assumptions, one could use the Instrumental Variable Bayesian Model Averaging (IVBMA) method proposed by Eicher *et al.* [76]. The IVBMA estimate accounts for model uncertainty by taking the weighted average of IV estimates from different potential models, weighted by both the first and second stage posterior model probabilities. The Bayesian Sargan test based on the IVBMA framework can be used to detect violation of the IV assumptions.

For simplicity, we assume the random errors ξ_{1i} and ξ_{2i} jointly follow a bivariate normal distribution or other parametric distribution such as a bivariate elliptically contoured distribution. This allows a natural extension of the linear two-stage model for uncensored continuous outcomes into an IV model for censored time-to-event outcomes with MCMC procedures. However, the idea used in this paper can be extended to more flexible settings with different error distributional assumptions. In a sequel, we will extend our model to a semiparametric Bayesian model, in order to deal with more flexible semiparametric error distributions in the presence of arbitrarily censored data which is difficult to handle using the classic frequentist approach.

Acknowledgements

We thank the associate editor and two referees for their insightful suggestions, which led to significant improvements of the paper. We also thank Dr. Donatello Telesca for his helpful comments and suggestions. Moreover, we are grateful to Drs. Simin Liu and Katie Kei-hang Chan of the Department of Epidemiology at UCLA for permission to use their data. Gang Li's research was supported in part by NIH grant CA016042, NIH U54 RR031268-01 and NIH grant P01AT003960.

Appendix

A1. Likelihood Derivation

We use notations defined in Section 2.1 and Section 2.2. For time-to-event data subject to right-censoring, the likelihood of observing $(\vec{T}, \vec{\delta})$ is:

$$\mathcal{L} = \prod_{i=1}^n f(T_i)^{\delta_i} S(T_i)^{1-\delta_i} \quad (13)$$

where $S(y) = P(Y > y)$ is the survival distribution function and $f(y) = -\frac{d}{dy}S(y)$ is the survival time density function.

Based on the two-stage IV model (1) and (2), the likelihood function of observing

$(\vec{T}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G})$ can be written as:

$$\mathcal{L}(\theta | \vec{T}, \vec{\delta}, \vec{X}, \vec{Z}, \vec{G}) = P(\vec{T}, \vec{\delta} | \vec{X}, \vec{Z}, \vec{G}, \theta) \cdot P(\vec{X}, \vec{Z}, \vec{G} | \theta) \quad (14)$$

where the second part is the marginal likelihood of the first-stage model (1), and the first part is the conditional likelihood of the second-stage model (2).

For the first part: from the bivariate normality assumption of ξ_1 and ξ_2 following equation (3), the conditional distribution of ξ_{2i} given ξ_{1i} is:

$$\xi_{2i} | \xi_{1i} \sim N\left(\frac{\sigma_2}{\sigma_1} \rho \xi_{1i}, (1 - \rho^2) \sigma_2^2\right)$$

$i = 1, \dots, n$. Since $\xi_{1i} = X_i - \alpha_0 - \alpha_1' G_i - \alpha_2' Z_i$ from the first-stage model (1), the conditional distribution becomes:

$$\xi_{2i} | X_i, Z_i, G_i \sim N\left(\frac{\sigma_2}{\sigma_1} \rho (X_i - \alpha_0 - \alpha_1' G_i - \alpha_2' Z_i), (1 - \rho^2) \sigma_2^2\right)$$

Therefore, given $\vec{X}, \vec{Z}, \vec{G}, \alpha_0, \alpha_1$ and α_2 , the second-stage model (2) has conditional survival function

$$\begin{aligned} S(T | X, Z, G) &= P(Y > T | X, Z, G) \\ &= P(\beta_0 + \beta_1 X + \beta_2' Z + \xi_2 > T) \\ &= P(\xi_2 > T - \beta_0 - \beta_1 X - \beta_2' Z) \\ &= 1 - \Phi\left(\frac{T - \beta_0 - \beta_1 X - \beta_2' Z - \frac{\sigma_2}{\sigma_1} \rho (X - \alpha_0 - \alpha_1' G - \alpha_2' Z)}{\sqrt{(1 - \rho^2) \sigma_2^2}}\right) \end{aligned}$$

and conditional density function

$$\begin{aligned} f_1(T | X, Z, G) &= -\frac{\partial}{\partial T} S(T | X, Z, G) \\ &= \phi\left(\frac{T - \beta_0 - \beta_1 X - \beta_2' Z - \frac{\sigma_2}{\sigma_1} \rho (X - \alpha_0 - \alpha_1' G - \alpha_2' Z)}{\sqrt{(1 - \rho^2) \sigma_2^2}}\right) \end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative density function and the probability density function of standard normal distribution, respectively. From (13), we have

$$P(\vec{T}, \vec{\delta} | \vec{X}, \vec{Z}, \vec{G}, \theta) = \prod_{i=1}^n f_1(T_i | X_i, Z_i, G_i, \theta)^{\delta_i} S(T_i | X_i, Z_i, G_i, \theta)^{1 - \delta_i} \quad (15)$$

For the second part: the marginal distribution of ξ_{1i} is:

$$\xi_{1i} \sim N(0, \sigma_1^2)$$

which gives the marginal density function for the first-stage model (1):

$$f_2(X, Z, G) = \phi\left(\frac{X - \alpha_0 - \alpha_1'G - \alpha_2'Z}{\sqrt{\sigma_1^2}}\right)$$

Therefore, the likelihood of observing $\vec{X}$, $\vec{Z}$ and $\vec{G}$ is:

$$P(\vec{X}, \vec{Z}, \vec{G} | \theta) = \prod_{i=1}^n f_2(X_i, Z_i, G_i) \quad (16)$$

From (14), (15) and (16), we have the joint likelihood function (9).

A2. MCMC algorithm

We use the MCMC sampling method to generate samples from the posterior distribution of the parameters. In each iteration, a random walk Metropolis-Hasting algorithm is used to update the parameters in $\theta = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \rho)$ one by one, while other parameters are fixed at their current state. It updates the current state of the parameter, denoted as z , by candidate z^* generated from a proposal distribution $h(z^*/z)$, with acceptance probability:

$$a(z, z^*) = \min\left(1, \frac{h(z|z^*)\pi(z^*)}{h(z^*|z)\pi(z)}\right)$$

where $\pi(\cdot)$ is a probability density function, in our case, the unstandardized posterior distribution function (i.e. the product of prior distribution and likelihood function) of the parameter. The following is the detailed procedure to update z :

1. Generate a candidate sample z^* from proposal distribution $h(z^*/z)$.
2. Calculate

$$\begin{aligned} \log(a(z, z^*)) &= \log\left(\frac{h(z|z^*)\pi(z^*)}{h(z^*|z)\pi(z)}\right) \\ &= [\log(h(z|z^*)) - \log(h(z^*|z))] + [\log(\mathcal{P}(z)) - \log(\mathcal{P}(z^*))] + [\log(\mathcal{L}(z)) - \log(\mathcal{L}(z^*))] \end{aligned}$$

where $\mathcal{P}(\cdot)$ is the prior distribution density function and $\mathcal{L}(\cdot)$ is the likelihood function. In our case, $\mathcal{L}(\cdot)$ is the joint likelihood function (9).

3. Generate a random number r from $\text{Unif}(0, 1)$.
4. If $\log(a(z, z^*)) > \log(r)$, we accept z^* as the next sample of the parameter; else we keep z as the next sample.

5. Repeat from step 1.

Independent diffuse priors are used for the parameters: a normal distribution $N(\mu, \zeta^2)$ with large variance (e.g. $\mu = 0, \zeta^2 = 100^2$) for each element in $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1$ and β_2 ; an inverse-gamma distribution $\text{Inv-Gamma}(\gamma_1, \beta_2)$ with small shape parameter and small scale parameter (e.g. $\gamma_1 = \beta_2 = 0.001$) for σ_1^2 and σ_2^2 ; and a uniform distribution $\text{Unif}(-1, 1)$ for ρ . Uniform proposal distributions are used for the random walk: $\text{Unif}(z - \omega, z + \omega)$ for each element in $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1$ and β_2 ; $\text{Unif}(\max(z - \omega, 0), z + \omega)$ for σ_1^2 and σ_2^2 ; and $\text{Unif}(\max(z - \omega, -1), \min(z + \omega, 1))$ for ρ . Different positive ω is chosen for each parameter to obtain an appropriate acceptance rate (e.g. 20% ~ 40% depending on the sample size).

The detailed derivation of the log of acceptance probability for parameters in θ is as follows:

- α_0 : Denote the current state and candidate sample as α_0 and α_0^* , respectively. With prior distribution $N(\mu, \zeta^2)$ and proposal distribution $\text{Unif}(\alpha_0 - \omega, \alpha_0 + \omega)$, the log of acceptance probability:

$$\begin{aligned} \log(a(\alpha_0, \alpha_0^*)) &= \frac{1}{2\zeta^2} \left((\alpha_0 - \mu)^2 - (\alpha_0^* - \mu)^2 \right) \\ &+ \sum_{i=1}^n \left[\delta_i \left(\frac{1}{2} (q_i^2 - q_i^{*2}) \right) + (1 - \delta_i) (\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))) + \frac{1}{2} (\nu_i^2 - \nu_i^{*2}) \right] \end{aligned} \quad (17)$$

where

$$q_i = \frac{1}{\sqrt{(1 - \rho^2) \sigma_2^2}} \left[\log(T_i) - (\beta_0 + \beta_1 X_i + \beta_2' Z_i) - \frac{\sigma_2}{\sigma_1} \rho (X_i - \alpha_0 - \alpha_1' G_i - \alpha_2' Z_i) \right] \quad (18)$$

$$\nu_i = \frac{1}{\sqrt{\sigma_1^2}} (X_i - \alpha_0 - \alpha_1' G_i - \alpha_2' Z_i) \quad (19)$$

q_i^* and ν_i^* are similar to q_i and ν_i , respectively, both equations with all α_0 replaced by α_0^* .

- α_1 : We update α_1 by updating its elements one-by-one. To update the j -th element α_{1j} with candidate sample α_{1j}^* , and with prior distribution $N(\mu, \zeta^2)$ and proposal distribution $\text{Unif}(\alpha_{1j} - \omega, \alpha_{1j} + \omega)$, the log of acceptance probability is similar to (17), with the first term replaced by $\frac{1}{2\zeta^2} \left((\alpha_{1j} - \mu)^2 - (\alpha_{1j}^* - \mu)^2 \right)$. q_i and ν_i stay the same as (18) and (19). q_i^* and ν_i^* are similar to q_i and ν_i , respectively. Both equations have all α_1 replaced by α_1^* , where α_1^* is α_1 with the j -th element replaced by α_{1j}^* .
- α_2 : We update α_2 by updating its elements one-by-one. To update the j -th element α_{2j} with candidate sample α_{2j}^* , and with prior distribution $N(\mu, \zeta^2)$ and proposal distribution $\text{Unif}(\alpha_{2j} - \omega, \alpha_{2j} + \omega)$, the log of acceptance probability is similar to

(17), with the first term replaced by $\frac{1}{2\zeta^2} \left((\alpha_{2j} - \mu)^2 - (\alpha_{2j}^* - \mu)^2 \right)$. q_i and v_i stay the same as (18) and (19). q_i^* and ν_i^* are similar to q_i and v_i , respectively: Both equations have all α_2 replaced by α_{2j}^* , where α_{2j}^* is α_2 with the j -th element replaced by α_{2j}^* .

- β_0 : We update the current state β_0 with candidate sample β_0^* . With prior distribution $N(\mu, \zeta^2)$ and proposal distribution $\text{Unif}(\beta_0 - \omega, \beta_0 + \omega)$, the log of acceptance probability:

$$\log(a(\beta_0, \beta_0^*)) = \frac{2}{2\zeta^2} \left((\beta_0 - \mu)^2 \right) + \sum_{i=1}^n \left[\delta_i \left(\frac{1}{2} (q_i^2 - q_i^{*2}) \right) + (1 - \delta_i) (\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))) \right] \quad (20)$$

where q_i stays the same as (18). q_i^* is similar to q_i , with β_0 replaced by β_0^* .

- β_1 : We update the current state β_1 with candidate sample β_1^* . With prior distribution $N(\mu, \zeta^2)$ and proposal distribution $\text{Unif}(\beta_1 - \omega, \beta_1 + \omega)$, the log of acceptance probability is similar to (20), with the first term replaced

$\frac{1}{2\zeta^2} \left((\beta_1 - \mu)^2 - (\beta_1^* - \mu)^2 \right)$. q_i stays the same as (18). q_i^* is similar to q_i , with β_1 replaced by β_1^* .

- β_2 : We update β_2 by updating its elements one-by-one. To update the j -th element β_{2j} with candidate sample β_{2j}^* and with prior distribution $N(\mu, \zeta^2)$ and proposal distribution $\text{Unif}(\beta_{2j} - \omega, \beta_{2j} + \omega)$, the log of acceptance probability is similar to

(20), with the first term replaced by $\frac{1}{2\zeta^2} \left((\beta_{2j} - \mu)^2 - (\beta_{2j}^* - \mu)^2 \right)$. q_i stays the same as (18). q_i^* is similar to q_i , with β_2 replaced by β_{2j}^* , where β_{2j}^* is β_2 with the j -th element replaced by β_{2j}^* .

- σ_1^2 : We update the current state σ_1^2 with candidate sample σ_1^{2*} . With prior distribution $\text{Inv-Gamma}(\gamma_1, \gamma_2)$ and proposal distribution $\text{Unif}(\sigma_1^2, \sigma_1^{2*})$, the log of acceptance probability:

$$\begin{aligned} \log(a(\sigma_1^2, \sigma_1^{2*})) = & [\log(\sigma_1^2 + \omega - \max(0, \sigma_1^2 - \omega)) - \log(\sigma_1^{2*} + \omega - \max(0, \sigma_1^{2*} - \omega))] \\ & + \left[(\gamma_1 + 1) (\log \sigma_1^2 - \log \sigma_1^{2*}) + \gamma_2 \zeta \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_1^{2*}} \right) \right] \\ & + \sum_{i=1}^n \left[\delta_i \left(\frac{1}{2} (q_i^2 - q_i^{*2}) \right) + (1 - \delta_i) (\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))) \right] \\ & + \frac{1}{2} ((\log \sigma_1^2 - \log \sigma_1^{2*}) + (\nu_i^2 - \nu_i^{*2})) \end{aligned}$$

where q_i and v_i stay the same as (18) and (19). q_i^* and ν_i^* are similar to q_i and v_i , respectively: Both equations have all σ_1^2 replaced by σ_1^{2*} .

- σ_2^2 : We update the current state σ_2^2 with candidate sample σ_2^{2*} . With prior distribution $\text{Inv-Gamma}(\gamma_1, \gamma_2)$ and proposal distribution Unif

$(\max(\sigma_2^2 - \omega, 0), \sigma_2^2 + \omega)$, the log of acceptance probability:

$$\begin{aligned} \log(a(\sigma_2^2, \sigma_2^{2*})) = & [\log(\sigma_2^2 \omega - \max(0, \sigma_2^2 - \omega)) - \log(\sigma_2^{2*} \omega - \max(0, \sigma_2^{2*} - \omega))] \\ & + [(\gamma_1 + 1)(\log \sigma_2^2 - \log \sigma_2^{2*}) + \gamma_2 \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_2^{2*}} \right)] \\ & + \sum_{i=1}^n \left[\delta_i \left(\frac{1}{2} (\log \sigma_2^2 - \log \sigma_2^{2*} + q_i^2 - q_i^{*2}) \right) + (1 - \delta_i) (\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))) \right] \end{aligned}$$

where q_i stays the same as (18). q_i^* is similar to q_i , with all σ_2^2 replaced by σ_2^{2*} .

- ρ : We update the current state ρ with candidate sample ρ^* . With prior distribution $\text{Unif}(-1, 1)$ and proposal distribution $\text{Unif}(\max(\rho - \omega, -1), \min(\rho + \omega, 1))$, the log of acceptance probability:

$$\begin{aligned} \log(a(\rho^2, \rho^*)) = & [\log(\min(\rho + \omega, 1) - \max(\rho - \omega, -1)) - \log(\min(\rho^* + \omega, 1) - \max(\rho^* - \omega, 1))] \\ & + \sum_{i=1}^n \left[\delta_i \left(\frac{1}{2} (\log(1 - \rho^2) - \log(1 - \rho^{*2}) + q_i^2 - q_i^{*2}) \right) \right. \\ & \quad \left. + (1 - \delta_i) (\log(1 - \Phi(q_i^*)) - \log(1 - \Phi(q_i))) \right] \end{aligned}$$

where q_i stays the same as (18). q_i^* is similar to q_i , with all ρ replaced by ρ^* .

A3. Supplementary simulation results

Estimation of other parameters besides β_1 in the IV model on simulated data with normal random errors

parameter	n	$R^2(X, G) = 0.05$			$R^2(X, G) = 0.1$		
		Bias	SD	CP	Bias	SD	CP
α_0	300	0.001	0.011	0.951	0.002	0.018	0.937
	500	0.002	0.014	0.949	0.002	0.013	0.949
	800	0.001	0.011	0.951	0.001	0.011	0.950
α_1	300	0.002	0.011	0.941	0.002	0.018	0.950
	500	0.002	0.014	0.943	0.001	0.013	0.955
	800	0.002	0.011	0.941	0.001	0.011	0.943
β_0	300	0.014	0.213	0.959	0.030	0.247	0.961
	500	0.034	0.265	0.964	0.015	0.192	0.960
	800	0.014	0.213	0.959	0.012	0.154	0.952
σ_1^2	300	0.001	0.005	0.942	0.001	0.007	0.945
	500	0.001	0.006	0.943	0.001	0.006	0.954
	800	0.001	0.005	0.942	0.001	0.005	0.951
σ_2^2	300	0.046	0.065	0.949	0.060	0.090	0.951
	500	0.071	0.087	0.941	0.039	0.062	0.946
	800	0.046	0.065	0.949	0.025	0.047	0.951
ρ	300	0.022	0.163	0.960	0.040	0.181	0.964

parameter	<i>n</i>	$R^2(X, G) = 0.05$			$R^2(X, G) = 0.1$		
		Bias	SD	CP	Bias	SD	CP
	500	0.046	0.196	0.959	0.019	0.149	0.965
	800	0.022	0.163	0.960	0.013	0.124	0.950

Simulation settings are the same as those for Table 1, with $\beta_1 = 0$ and censoring rate 50%.

Results are based on 2000 simulations.

Bias is calculated as the absolute difference between the sample mean of the parameter estimates and the true value of the parameter.

Standard deviation (SD) is calculated as the sample standard deviation of the parameter estimates.

Coverage probability (CP) is the proportion of 95% credible intervals that cover the true value of the parameter.

References

1. Wright S. Correlation and Causation. *Journal of Agricultural Research*. 1921; 20(7):557–585.
2. Wright, S. Appendix to The Tariff on Animal and Vegetable Oils. MacMillan; New York: 1928.
3. Haavelmo T. The Statistical Implications of a System of Simultaneous Equations. *Econometrica*. 1943; 11:1–12.
4. Haavelmo T. The Probability Approach in Econometrics. *Econometrica*. 1944; 12(Supplement):iii–115.
5. Theil, H. Economic forecasts and policy. North-Holland; Amsterdam: 1958.
6. Goldberger AS. Structural equation Methods in the Social Sciences. *Econometrica*. 1972; 40:979–1001.
7. Bowden, RJ.; Turkington, DA. Instrumental Variables. Cambridge University Press; Cambridge: 1984.
8. Heckman JJ, Robb R. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*. 1985; 30(1-2):239–267.
9. Heckman JJ, Hotz VJ. Choosing Among Alternative Non-experimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association*. 1989; 84:862–880.
10. Morgan, MS. The History of Econometric Ideas. Cambridge University Press; 1991.
11. Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*. 2004; 33(1):30–42. DOI: 10.1093/ije/dyh132. [PubMed: 15075143]
12. Thomas D, Conti D. Commentary: the concept of 'Mendelian Randomization'. *International Journal of Epidemiology*. 2004; 33(1):21–25. [PubMed: 15075141]
13. Didelez V, Sheehan NA. Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*. 2007; 16:309–330. DOI: 10.1177/0962280206077743. [PubMed: 17715159]
14. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*. 2008; 27:1133–1328. [PubMed: 17886233]
15. Wehby GL, Ohsfeldt RL, Murray JC. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Statistics in Medicine*. 2008; 27:2745–2749. DOI: 10.1002/sim.3255. [PubMed: 18344186]
16. Gray R, Wheatley K. How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplant*. 1991; 7(Supplement 3):9–12. [PubMed: 1855097]
17. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*. 2003; 32(1):1–22. [PubMed: 12689998]

18. Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *American journal of epidemiology*. 2006; 163(5):397–403. [PubMed: 16410347]
19. Ogbuanu IU, Zhang H, Karmaus W. Can we apply the Mendelian randomization methodology without considering epigenetic effects? *Emerging Themes in Epidemiology*. 2009; 6:3. DOI: 10.1186/1742-7622-6-3. [PubMed: 19432981]
20. Murphy KM, Topel RH. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*. 1985; 3(4):370–379.
21. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967; 1: 221–233.
22. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980; 48:817–838.
23. Hardin JW. The robust variance estimator for two-stage models. *The Stata Journal*. 2002; 2(3): 253–266.
24. Hardin JW, Carroll RJ. Variance estimation for the instrumental variables approach to measurement error in generalized linear models. *The Stata Journal*. 2003; 3(4):342–350.
25. Kleibergen F, Van Dijk HK. Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory*. 1998; 14:701–743.
26. Hoogerheide L, Kleibergen F, Van Dijk HK. Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data. *Journal of Econometrics*. 2007; 138:63–103. DOI: 10.1016/j.jeconom.2006.05.015.
27. Kleibergen F, Zivot E. Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics*. 2003; 114(1):29–72.
28. Lancaster, T. *An introduction to modern Bayesian econometrics*. Wiley-Blackwell; 2004. Chapter 8, Instrumental Variables
29. Conley T, Hansen C, McCulloch R, Rossi P. A semi-parametric Bayesian approach to the instrumental variables problem. *Journal of Econometrics*. 2008; 144:276–305. DOI: 10.2139/ssrn.917432.
30. Angrist JD, Imbens GW. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association*. 1995; 90(430):431–442.
31. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*. 1996; 91(434):444–455.
32. Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*. 1974; 66(5):688–701.
33. Rubin DB. Bayesian Inference for Causal Effects. *The Annals of Statistics*. 1978; 6:34–58.
34. Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association*. 1986; 81(396):945–960.
35. Pearl, J. *Causality: models, reasoning, and inference*. Cambridge University Press; New York: 2000. Chapter 7, Structural and Counterfactual Models. ISBN 0-521-77362-8
36. Heckman, JJ. *Econometric causality*. Working Paper 13934. National Bureau of Economic Research; Apr. 2008
37. Durbin J. Errors in Variables. *Review of the International Statistical Institute*. 1954; 22:23–32.
38. Fuller, WA. *Measurement Error Models*. Wiley; New York: 1987.
39. Carroll RJ, Stefanski LA. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine*. 1994; 13(12):1265–1282. DOI: 10.1002/sim.4780131208. [PubMed: 7973207]
40. Buzas JS, Stefanski LA. Instrumental variables estimation in generalized linear measurement error models. *Journal of the American Statistical Association*. 1996; 91:999–1006.
41. Goetghebuer E, Vansteelandt S. Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research*. 2005; 14:397–415. [PubMed: 16178139]

42. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement error in nonlinear models: a modern perspective. 2nd edition.. Chapman and Hall CRC Press; Boca Raton, FL: 2006. Chapter 6, Instrumental variables
43. Gustafson P. Measurement error modelling with an approximate instrumental variable. *Journal of the Royal Statistical Society: Series B.* 2007; 69(5):797–815.
44. Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. *Statistical Science.* 2010; 25(1):22–40. DOI: 10.1214/09-STS316.
45. Palmer TM, Thompson JR, Tobin MD, Sheehan NA, Burton PR. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *International journal of epidemiology.* 2008; 37(5):1161–1168. DOI: 10.1093/ije/dyn080. [PubMed: 18463132]
46. McKeigue PM, Campbell H, Wild S, Vitart V, Hayward C, Rudan I, Wright AF, Wilson JF. Bayesian methods for instrumental variables analysis with genetic instruments (“Mendelian randomization”): example with urate transporter SLC2A9 as instrumental variable for effect of urate levels on metabolic syndrome. *International journal of epidemiology.* 2010; 39:907–918. DOI: 10.1093/ije/dyp397. [PubMed: 20348110]
47. Kuchenhoff H. The identification of logistic regression models with errors in the variables. *Statistical Papers.* 1995; 36:41–48. DOI: 10.1007/BF02926017.
48. StataCorp. Stata 12 Base Reference Manual. Stata Press; College Station, TX: 2011.
49. Roodman D. Fitting fully observed recursive mixed-process models with cmp. *The Stata Journal.* 2011; 11(2):159–206.
50. Bijwaard, GE. Tinbergen Institute Discussion Papers. Tinbergen Institute; 2008. Instrumental Variable Estimation for Duration Data.. 08-032/4 DOI: 10.2139/ssrn.1114023
51. Hansen LP. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica.* 1982; 50:1029–1054.
52. Hall, AR. Generalized Method of Moments. Oxford University Press; New York: 2005.
53. Yin G, Ma Y, Liang F, Yuan Y. Stochastic Generalized Method of Moments. *Journal of Computational and Graphical Statistics.* 2011; 20(3):714–727. DOI: 10.1198/jcgs.2011.09210. [PubMed: 22375093]
54. Klein, J.; Moeschberger, M. Survival Analysis: Techniques for Censored and Truncated Data. 2nd edition. Springer-Verlag; 2003. ISBN: 038795399X
55. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. *Journal of Chemical Physics.* 1953; 21:1087–1091.
56. Hastings WK. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika.* 1970; 57:97–109.
57. Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. *American Statistician.* 1995; 49:327–335. DOI: 10.1080/00031305.1995.10476177.
58. R Development Core Team. R. A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2010. ISBN 3-900051-07-0, URL <http://www.R-project.org>
59. Brooks SP, Gelman A. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics.* 1998; 9:262–285. DOI: 10.1080/10618600.1998.10474787.
60. Fang, KT.; Anderson, TW. Statistical Inference in Elliptically Contoured and Related Distributions. Allerton Press; New York: 1990.
61. Chmielewski MA. Elliptically symmetric distributions: a review and bibliography. *International Statistical Review.* 1981; 49:67–74. DOI: 10.2307/1403038.
62. Lange KL, Little RJA, Taylor JMG. Robust Statistical Modeling Using the t Distribution. *Journal of the American Statistical Association.* 1989; 84(408):881–896.
63. Landsman Z, Valdez EA. Tail conditional expectations for elliptical distributions. *North American Actuarial Journal.* 2003; 7:55–71. DOI: 10.1080/10920277.2003.10596118.
64. Marden, DB.; Manolakis, DG. Using elliptically contoured distributions to model hyperspectral imaging data and generate statistically similar synthetic data.. *Algorithms and Technologies for*

- Multispectral, Hyperspectral, and Ultraspectral Imagery X; Proceedings of SPIE; 2004. p. 558-572. DOI: 10.1117/12.542077
65. Therneau, T. A Package for Survival Analysis in S. R package version 2.37-4. 2013. <http://CRAN.R-project.org/package=survival>
 66. SAS Institute Inc.. SAS/STAT® 9.2 User's Guide. SAS Institute Inc.; Cary, NC: 2008.
 67. Han TS, Sattar N, Williams K, Gonzalez-Villalpando C, Lean ME, Haffner SM. Prospective study of C-reactive protein in relation to the development of diabetes and metabolic syndrome in the Mexico City Diabetes Study. *Diabetes Care*. 2002; 25:2016–2021. [PubMed: 12401749]
 68. Freeman DJ, Norrie J, Caslake MJ, Gaw A, Ford I, Lowe GD, O'Reilly DS, Packard CJ, Sattar NC. C-reactive protein is an independent predictor of risk for the development of diabetes in the West of Scotland Coronary Prevention Study. *Diabetes*. 2002; 51:1596–1600. [PubMed: 11978661]
 69. Liu S, Tinker L, Song Y, Rifai N, Bonds DE, Cook NR, Heiss G, Howard BV, Hotamisligil GS, Hu FB, Kuller LH, Manson JE. A prospective study of inflammatory cytokines and diabetes mellitus in a multiethnic cohort of postmenopausal women. *Archives of Internal Medicine*. 2007; 167:1676–1685. [PubMed: 17698692]
 70. Chan KH, Brennan K, You NC, Lu X, Song Y, Hsu YH, Chaudhuri G, Nathan L, Tinker L, Liu S. Common variations in the genes encoding C-reactive protein, tumor necrosis factor- α , and interleukin-6, and the risk of clinical diabetes in the Women's Health Initiative Observational Study. *Clinical Chemistry*. 2011; 57(2):317–325. DOI: 10.1373/clinchem.2010.154526. [PubMed: 21149504]
 71. Lee CC, You NC, Song Y, Hsu YH, Manson J, Nathan L, Tinker L, Liu S. Relation of genetic variation in the gene coding for C-reactive protein with its plasma protein concentrations: findings from the Womens Health Initiative Observational Cohort. *Clinical Chemistry*. 2009; 55:351–360. DOI: 10.1373/clinchem.2008.117176. [PubMed: 19095725]
 72. Brunner EJ, Kivimaki M, Witte DR, Lawlor DA, Davey Smith G, Cooper JA, Miller M, Lowe GD, Rumley A, Casas JP, Shah T, Humphries SE, Hingorani AD, Marmot MG, Timpson NJ, Kumari M. Inflammation, insulin resistance, and diabetes: Mendelian randomization using CRP haplotypes points upstream. *PLoS Med*. 2008; 5(8):e155. DOI: 10.1371/journal.pmed.0050155. [PubMed: 18700811]
 73. Keavney B. More evidence against a causal association between C-reactive protein and diabetes. *PLoS Med*. 2008; 5:e174. DOI: 10.1371/journal.pmed.0050174. [PubMed: 18700815]
 74. The ARIC investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *American Journal of Epidemiology*. 1989; 129:687–702. [PubMed: 2646917]
 75. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998; 97(18):1837–1847. DOI: 10.1161/01.CIR.97.18.1837. [PubMed: 9603539]
 76. Eicher, TS.; Lenkoski, A.; Raftery, AE. Bayesian model averaging and endogeneity under model uncertainty: an application to development determinants. University of Washington; 2009. Working paper

Figure 1.

Directed acyclic graph for instrumental variable model in the presence of unobserved confounders and measurement errors in the intermediate covariate. G is a vector of instruments that satisfy the instrumental variable assumptions; W is an unobserved intermediate covariate that is subject to measurement errors; X is the observed surrogate of W ; Z is a vector of observed confounders; U is a vector of unobserved confounders; and Y is the outcome variable. The causal effect of W on Y , i.e. parameter β_1 , is of primary interest.

Figure 2. Histograms of the posterior samples for β_1 : (a) WHI-OS example; (b) ARIC example. The brackets denote the limits of the 95% credible interval.

Figure 3. Trace plots of the posterior samples for β_1 : (a) WHI-OS example; (b) ARIC example.

Figure 4. Autocorrelation plots of individual chains for β_1 : (a) WHI-OS example; (b) ARIC example.

Table 1
 β_1 Estimation with and without Instrumental Variable Analysis on Simulated Data with Normal Random Errors

Performance of the Instrumental Variable (IV) method and the simple method under various sample sizes (n), censoring rates (CR), true values of the endogenous parameter (β_1), and instrument strengths ($R^2(X, G)$).

n	CR	β_1	Simple estimate without IV				IV estimate			
							$R^2(X, G) = 0.05$			
			Bias	SD	CP	Bias	SD	CP	Bias	SD
300	25%	0	0.444	0.139	0.108	0.029	0.594	0.971	0.017	0.453
		-0.5	0.378	0.141	0.228	0.047	0.582	0.969	0.017	0.445
		-1	0.297	0.141	0.432	0.005	0.602	0.969	0.003	0.444
		50%	0	0.449	0.157	0.186	0.048	0.658	0.972	0.486
	50%	-0.5	0.372	0.158	0.351	0.046	0.657	0.972	0.030	0.512
		-1	0.294	0.163	0.551	0.022	0.641	0.981	0.015	0.492
		75%	0	0.447	0.206	0.402	0.075	0.748	0.975	0.602
		-0.5	0.377	0.207	0.551	0.088	0.759	0.973	0.056	0.601
		-1	0.300	0.210	0.705	0.097	0.761	0.981	0.043	0.619
	75%	0	0.447	0.107	0.013	0.002	0.479	0.953	0.006	0.348
		-0.5	0.372	0.107	0.067	0.004	0.483	0.962	0.005	0.338
		-1	0.297	0.110	0.229	0.006	0.490	0.964	0.008	0.346
		50%	0	0.449	0.122	0.044	0.020	0.532	0.962	0.011
500	25%	-0.5	0.371	0.125	0.153	0.023	0.538	0.962	0.007	0.387
		-1	0.296	0.127	0.348	0.011	0.539	0.964	0.001	0.396
		75%	0	0.438	0.156	0.210	0.066	0.631	0.970	0.462
		-0.5	0.371	0.164	0.363	0.041	0.626	0.974	0.037	0.484
	50%	-1	0.296	0.166	0.552	0.043	0.660	0.967	0.007	0.497
		0	0.449	0.083	<0.001	0.008	0.366	0.963	0.002	0.269
		-0.5	0.374	0.086	0.008	0.011	0.379	0.956	0.005	0.274
		-1	0.297	0.087	0.071	0.001	0.388	0.958	0.010	0.277
		50%	0	0.447	0.096	0.003	0.027	0.428	0.961	0.300
	75%	-0.5	0.371	0.164	0.363	0.041	0.626	0.974	0.037	0.484
		-1	0.296	0.166	0.552	0.043	0.660	0.967	0.007	0.497
		0	0.449	0.083	<0.001	0.008	0.366	0.963	0.002	0.269
		-0.5	0.374	0.086	0.008	0.011	0.379	0.956	0.005	0.274
800	25%	-1	0.297	0.087	0.071	0.001	0.388	0.958	0.010	0.277
		50%	0	0.447	0.096	0.003	0.027	0.428	0.961	0.300
	50%	-0.5	0.371	0.164	0.363	0.041	0.626	0.974	0.037	0.484
		-1	0.296	0.166	0.552	0.043	0.660	0.967	0.007	0.497
		0	0.449	0.083	<0.001	0.008	0.366	0.963	0.002	0.269
		-0.5	0.374	0.086	0.008	0.011	0.379	0.956	0.005	0.274
	75%	-1	0.297	0.087	0.071	0.001	0.388	0.958	0.010	0.277
		50%	0	0.447	0.096	0.003	0.027	0.428	0.961	0.300
		-0.5	0.371	0.164	0.363	0.041	0.626	0.974	0.037	0.484
		-1	0.296	0.166	0.552	0.043	0.660	0.967	0.007	0.497
		0	0.449	0.083	<0.001	0.008	0.366	0.963	0.002	0.269
		-0.5	0.374	0.086	0.008	0.011	0.379	0.956	0.005	0.274

Simple estimate without IV						IV estimate					
						$R^2(X, G) = 0.05$			$R^2(X, G) = 0.1$		
n	CR	β_1	Bias	SD	CP	Bias	SD	CP	Bias	SD	CP
75%		-0.5	0.372	0.097	0.032	0.040	0.425	0.969	0.001	0.316	0.953
		-1	0.299	0.099	0.145	0.017	0.434	0.965	0.005	0.307	0.956
		0	0.449	0.125	0.055	0.027	0.518	0.961	0.012	0.387	0.951
		-0.5	0.372	0.126	0.163	0.026	0.527	0.963	0.015	0.387	0.953
	-1	0.296	0.128	0.358	0.037	0.535	0.954	0.002	0.400	0.960	

Results are based on 2000 simulations using the IV method and 5000 simulations using the simple method. Bias is calculated as the absolute difference between the sample mean of the β_1 estimates and the true value of β_1 . Standard deviation (SD) is calculated as the sample standard deviation of the β_1 estimates. Coverage probability (CP) is the proportion of 95% credible intervals (for IV estimates) or confidence intervals (for simple estimates) that cover β_1 .

Table 2
 β_1 Estimation with Instrumental Variable Analysis on Simulated Data with Normal and Non-Normal Random Errors

Performance of the Instrumental Variable (IV) method with normality assumption under various underlying random error distributions, sample sizes (n), and true values of the endogenous parameter (β_1).

Error Distribution	n	β_1	IV estimate		
			Bias	SD	CP
Normal	300	0	0.066	0.621	0.973
		-0.5	0.008	0.642	0.975
		-1	0.032	0.634	0.975
	500	0	0.007	0.501	0.970
		-0.5	0.024	0.522	0.964
		-1	0.006	0.523	0.965
	800	0	0.012	0.415	0.959
		-0.5	0.007	0.412	0.956
		-1	0.002	0.412	0.963
Exponential	300	0	0.045	0.453	0.960
		-0.5	0.046	0.444	0.972
		-1	0.028	0.466	0.969
	500	0	0.017	0.347	0.972
		-0.5	0.006	0.366	0.955
		-1	0.004	0.378	0.964
	800	0	0.014	0.277	0.959
		-0.5	0.010	0.293	0.954
		-1	0.015	0.298	0.956
Weibull	300	0	0.059	0.704	0.974
		-0.5	0.054	0.743	0.963
		-1	0.015	0.740	0.968
	500	0	0.004	0.602	0.958
		-0.5	0.001	0.570	0.976
		-1	0.029	0.596	0.966
	800	0	0.021	0.476	0.954
		-0.5	0.005	0.479	0.958
		-1	0.008	0.493	0.949
Normal Mixture	300	0	0.024	0.644	0.969
		-0.5	0.029	0.638	0.974
		-1	0.021	0.695	0.969
	500	0	0.031	0.526	0.960
		-0.5	0.019	0.531	0.964
		-1	0.015	0.546	0.957

IV estimate					
Error Distribution	n	β_1	Bias	SD	CP
	800	0	0.003	0.423	0.962
		-0.5	0.008	0.438	0.952
		-1	0.004	0.442	0.958

'Error Distribution' refers to distribution of ε_{2i} in equation (2) and ε_{Ci} in censoring time C_i .

Distribution 'Normal Mixture' is a mixture of two normal distributions $N\left(-\sqrt{8/9}, 1/9\right) \cdot 0.5 + N\left(\sqrt{8/9}, 1/9\right) \cdot 0.5$.

Results are based on 2000 simulations.

Censoring rate is 50% and instrument strength is $R^2(X, G) = 0.05$.

Bias is calculated as the absolute difference between the sample mean of the β_1 estimates and the true value of β_1 .

Standard deviation (SD) is calculated as the sample standard deviation of the β_1 estimates.

Coverage probability (CP) is the proportion of 95% credible intervals that cover β_1 .

Table 3

Baseline characteristics of a white subgroup within the Women's Health Initiative Observational Study (WHI-OS)

Characteristic	Controls (n=968)	Case (n=954)
Age, mean $\pm$ SD, year	63.9 $\pm$ 6.9	63.9 $\pm$ 6.9
BMI, mean $\pm$ SD, kg/m ²	26.5 $\pm$ 5.1	32.5 $\pm$ 6.8
Physical activity, MET-h/wk		
Median	10.5	5.8
Interquartile range	3.8 – 20.6	0.9 – 14.0
Smoking status, %		
Nonsmoker	52.3	49.7
Past smoker	42.6	43.7
Current smoker	5.1	6.6
Alcohol intake, %		
Nondrinker	9.8	13.4
Past drinker	16.5	22.3
Current drinker, <1 drink/week	31.9	40.1
Current drinker, $\geq$ 1 drink/week	41.8	24.2
Hormone-replacement therapy, %		
Never	37.1	48.5
Past	14.3	15.3
Current	48.6	36.2
Family history of diabetes, %		
Yes	30.2	51.5
No	69.8	48.5
HsCRP, mg/L		
Median	2.05	4.06
Interquartile range	0.91 – 4.24	2.16 – 7.55

BMI = body mass index; hsCRP = high-sensitivity C-reactive protein.

Family history of diabetes is defined as self-reported diabetes in a first-degree relative.

Medians and interquartile ranges are provided for continuous variables with skewed distributions.

Table 4
Instrumental Variable (IV) analysis versus simple method in a subgroup analysis of whites within the Women's Health Initiative Observational Study (WHI-OS)

The simple method uses a log-normal accelerated failure time model to estimate the association between high-sensitivity C-reactive protein (hsCRP) and time to diabetes diagnosis. The IV analysis uses our proposed Bayesian IV method to estimate the causal effect of hsCRP on time to diabetes diagnosis, using 13 selected tSNPs as genetic instruments. Both models adjust for observed potential confounders including age, body mass index, cigarette smoking, alcohol intake, hormone-replacement therapy, family history of diabetes and physical activity.

	estimate of β_1	SE	95% CI
Simple method	-0.446	0.089	(-0.621, -0.272)
IV analysis	-0.162	0.426	(-0.987, 0.685)

CI stands for confidence interval for the simple method and credible interval for the IV analysis.

SE in the IV analysis is estimated by the posterior standard deviation of β_1 .

The 13 selected tSNPs are: rs4275453, rs2808634, rs3093059, rs2794521, rs1417938, rs1800947, rs1130864, rs1205, rs3093075, rs3093068, rs2808629, rs2369146, and rs1470515.

Table 5

Baseline characteristics of a subset of the Atherosclerosis Risk in Communities (ARIC) Study

Characteristic	Baseline (Visit 2) (<i>n</i> = 12, 782)
Age, mean $\pm$ SD, year	56.9 $\pm$ 5.7
Gender, % of Female	57
Ethnicity, % of Black	24.6
Current smoker, %	21.6
Diabetes, %	11
Total cholesterol level, mean $\pm$ SD, mmol/L	5.42 $\pm$ 1.01
HDL cholesterol level, mean $\pm$ SD, mmol/L	1.29 $\pm$ 0.43
Systolic blood pressure, mean $\pm$ SD, mmHg	145 $\pm$ 17.7

HDL = high-density lipoprotein.

Table 6
Instrumental Variable (IV) analysis versus simple method in a subset of the
Atherosclerosis Risk in Communities (ARIC) Study

The simple method uses a linear regression survival model with normally distributed residuals to estimate the association between standardized log-transformed systolic blood pressure (SBP) level at baseline and time to the first CHD event. The IV analysis uses our proposed Bayesian IV method to estimate this association, by using standardized log-transformed SBP level at Visit 1 as an instrument to correct for potential bias due to measurement errors in baseline SBP. Both models adjust for observed potential confounders including gender, age, total cholesterol level, high-density lipoprotein cholesterol level, smoking behavior, and diabetes status.

	estimate of β_1	SE	95% CI
Simple method	-0.779	0.087	(-0.950, -0.608)
IV analysis	-1.180	0.141	(-1.460, -0.907)

CI stands for confidence interval for the simple method and credible interval for the IV analysis.

SE in the IV analysis is estimated by the posterior standard deviation of β_1 .