Statistica Sinica **19** (2009), 1223-1246

CORRECTING INSTRUMENTAL VARIABLES ESTIMATORS FOR SYSTEMATIC MEASUREMENT ERROR

Stijn Vansteelandt[1], Manoochehr Babanezhad[1] and Els Goetghebeur[1,2]

[1] *Ghent University and* [2] *Harvard School of Public Health*

Abstract: Instrumental variables (IV) estimators are well established in a broad range of fields to correct for measurement error on exposure. In a distinct prominent stream of research, IV's are becoming increasingly popular for estimating causal effects of exposure on outcome since they allow for unmeasured confounders which are hard to avoid. Because many causal questions emerge from data which suffer severe measurement error problems, we combine both IV approaches in this article to correct IV-based causal effect estimators in linear (structural mean) models for possibly systematic measurement error on the exposure. The estimators rely on the presence of a baseline measurement that is associated with the observed exposure and known not to modify the target effect. Simulation studies and the analysis of a small blood pressure reduction trial ($n = 105$) with treatment noncompliance confirm the adequate performance of our estimators in finite samples. Our results also demonstrate that incorporating limited prior knowledge about a weakly identified parameter (such as the error mean) in a frequentist analysis can yield substantial improvements.

Key words and phrases: Causal inference, instrumental variables, measurement error, noncompliance, prior information, two-stage least squares estimators, weak identifiability.

1. Introduction

Instrumental variables (IV) methods have a long tradition in economics and econometrics, where they are used in connection with structural equation models. They have more recently entered the medical, epidemiological, and biostatistical literature (for reviews, see e.g., Greenland (2000), and Martens, Pestman, de Boer, Belitser and Klungel (2006)). To estimate the average causal effect of an exposure on an outcome in the presence of unmeasured confounders, these methods rely on so-called IV's. These are variables which (i) are associated with the exposure; (ii) have no direct effect on the outcome; and (iii) do not share common causes with the outcome (Hernán and Robins (2006)). IV's arise naturally in double-blind randomized trials with treatment noncompliance because randomization (i.e., the instrument) is associated with received treatment

(i.e., exposure), often does not affect the outcome other than through received treatment, and shares no common causes with the outcome by virtue of randomization. They are hence frequently used to adjust for treatment noncompliance in randomized experiments (see e.g., Goetghebeur and Vansteelandt (2005), for a review) and for the analysis of randomized encouragement designs (Ten Have, Elliott, Joffe, Zanutto and Datto (2004)). At the same time, they are becoming increasingly popular in observational settings where the conditions for an IV are harder to justify. In genetics, for instance, the random assortment of genes transferred from parents to offspring - called 'Mendelian randomization' - resembles the use of randomization in experiments, and is therefore a natural IV for estimating the effect of genetically affected exposures on a given trait (Didelez and Sheehan (2007), and Lawlor, Harbord, Sterne, Timpson and Smith (2008)). Casas, Bautista, Smeeth, Sharma and Hingorani (2005) use this idea to assess the influence of plasma homocysteine level on the risk of stroke with homozygosity at a specific allele as an IV. In most observational studies no real or natural randomization is present, in which case the availability of an IV must be assessed on theoretical grounds. For instance, Leigh and Schembri (2004) use the cigarette price per region as an IV to estimate the effect of smoking on health, assuming that the price of cigarettes may only impact health by mediating exposure to cigarette smoke.

With the increasing popularity of IV methods in causal inference comes the growing concern for their performance under common complications, such as misclassification or measurement error on exposure. In the context of noncompliance adjustment in clinical trials (Dunn (1999), and Goetghebeur and Vansteelandt (2005)) for instance, simple measures of compliance with drug therapy, such as pill counts, are notorious for overestimating the amount of drug actually taken (Urquhart and De Klerk (1998)). HIV prevention studies tend to rely on self reported measures of sexual activity and accompanying preventive action, including use of condoms or microbicide gels, which are subject to 'pleasing bias' (Van Damme et al. (2002)). Many other exposure measures are popular even though they are bias prone.

Random measurement error on exposure is not alarming for IV estimators in linear (structural mean) models (Goetghebeur and Vansteelandt (2005)). These estimators continue to be asymptotically unbiased when random measurement error is ignored, with at most a slight loss of efficiency. When measurement error is systematic, tests of the causal null hypothesis of no effect remain valid, but effect estimates may become biased. Because systematic error is a real concern in many practical settings (e.g., overreporting of drug compliance, underreporting of alcohol use, ...), our goal in this article is to investigate how IV estimators for the parameters in linear (structural mean) models may be adjusted for systematic measurement error. Goetghebeur and Vansteelandt (2005) show how this can be done when the average size of the error is known. This allows for

sensitivity analyses, but leaves open the question of how to estimate the average size of the measurement error and subsequently correct for it. Because of identifiability problems, the latter can only be realized when extraneous information is available. One common source of information is an IV for the measurement error (Buzas and Stefanski (1996), Carroll, Ruppert, Crainiceanu, Tosteson and Karagas (2004) and Carroll, Ruppert, Stefanski and Crainiceanu (2006)). In contrast to the original IV used for confounder adjustment, we define this to be a (pre-exposure) surrogate for the observed exposure (in the sense that it is correlated with exposure), which is assumed not to modify the exposure effect of interest. The general interest in such variables stems from the fact that we can identify settings where such variables exist (see later) and that other sources of information on the measurement error, such as repeated measurements or validation samples, are typically not available in large classes of problems (e.g., noncompliance adjustment).

In the next section, we build on ideas from linear regression models with error in the covariates (Carroll et al. (2006)) to show how an IV for the measurement error can help correct IV-based causal effect estimators for systematic error under linear structural mean models (Goetghebeur and Lapp (1997), and Robins (1994)). In Section 2.3, we diagnose poor performance of the error-adjusted estimator in small to moderate sample sizes as compared to the standard estimator which ignores measurement error. We show in Section 3 that this is due to the average magnitude of the error being weakly identified at causal effects close to zero. In Section 3, we accommodate this by imposing liberal bounds on the magnitude of the average error. This leads to reliable estimators for the causal effect of observed exposure with good performance in finite samples. The latter is confirmed through the analysis of a placebo-controlled hypertension trial in Section 4, and through simulation studies in Section 5. Our results reveal how the incorporation of prior information (in the form of bounds on weakly identified nuisance parameters) in a frequentist analysis can recover considerable precision for the target parameter.

2. Adjusting for Measurement Error

2.1. Assumptions

We consider data on a scalar exposure Z_i, a scalar outcome Y_i, and possibly a vector of baseline (i.e., pre-exposure) covariates $\mathbf{X}_i$ drawn from independent subjects $i = 1, \ldots, n$, to study the average effect of exposure Z_i on outcome Y_i. We define this effect as an expected contrast

$$\mathrm{E}(Y_i - Y_{i0}|Z_i, \mathbf{X}_i), \tag{2.1}$$

between observed outcomes Y_i and potential exposure-free outcomes Y_{i0} (Rubin (1978)). The latter indicates a reference response which would have been measured for subject i if all conditions had been the same as in the current study, but

no exposure had been received; that is, if the assigned experimental treatment contained no active dose. Because true exposure Z_i is imprecisely measured, the observed exposure level W_i for subject i may differ from the actual exposure level Z_i, which is unobserved.

Since Y_{i0} and Z_i are not generally observed, identification of the causal effect (2.1) requires assumptions.

Assumption A1 (Causal IV assumption). R_i is a *causal instrumental variable* ($IV-C$) for inferring the causal effect of Z_i on Y_i; that is, R_i is conditionally dependent on Z_i, given $\mathbf{X}_i$, and satisfies the following:

1. Exclusion restriction (Angrist, Imbens and Rubin (1996)): R_i has no direct effect on the outcome (only an indirect effect via the exposure is possible). That is, with Y_{i0r} the potential outcome that we would have observed for subject i if (R_i, Z_i) were set to $(r, 0)$, we assume that $Y_{i0r} = Y_{i0}$ for all values of r in the support of R_i.
2. Randomization assumption: within strata of baseline covariates $\mathbf{X}_i$, $\mathrm{E}(Y_{i0}|\mathbf{X}_i, R_i) = \mathrm{E}(Y_{i0}|\mathbf{X}_i)$.

In double-blind randomized trials of an asymptomatic disease, one expects these assumptions to hold for randomization R_i since patients and physicians are unaware of the assigned treatment (Robins (1994)).

Assumption A2 (Consistency assumption). To link exposure-free outcomes to observed outcomes, $Y_i = Y_{i0}$ for subjects with $Z_i = 0$.

Assumption A3 (Model assumption). The expected causal effect (2.1) follows the linear structural mean model (Robins (1994), and Goetghebeur and Lapp (1997))

$$\mathrm{E}(Y_i - Y_{i0}|Z_i, \mathbf{X}_i, R_i) = \gamma(\mathbf{X}_i; \boldsymbol{\psi}^*)Z_i, \tag{2.2}$$

where $\gamma(\mathbf{X}_i; \boldsymbol{\psi})$ is a known function smooth in $\boldsymbol{\psi}$, satisfying $\gamma(\mathbf{X}_i; \mathbf{0}) = 0$, and where $\boldsymbol{\psi}^*$ is an unknown finite-dimensional parameter.

For instance, in placebo-controlled randomized experiments with $R_i = 1$ for subjects randomized to the experimental arm and $R_i = 0$ for placebo control, and with Z_i denoting exposure to the experimental treatment, we may choose

$$\mathrm{E}(Y_i - Y_{i0}|Z_i, \mathbf{X}_i, R_i) = \psi^* Z_i. \tag{2.3}$$

Here, ψ^* expresses the expected change in outcome when those exposed to $Z_i = 1$ would have their exposure set to zero. When treatment effects are potentially modified by pre-treatment covariates, one may add covariate-exposure interactions, as in

$$\mathrm{E}(Y_i - Y_{i0}|Z_i, \mathbf{X}_i, R_i) = (\psi_1^* + \boldsymbol{\psi}_2^{*'}\mathbf{X}_i)Z_i.$$

Here, $\boldsymbol{\psi}_2^*$ defines the change in the average effect of unit exposure per unit increase in $\mathbf{X}_i$.

Note that we restrict our development to models (2.2) which postulate the causal effect to be linear in the exposure. This is a standard restriction in the literature on IV-estimation and on two-stage-least-squares estimation of causal effects (Hernán and Robins (2006)) because linear structural mean models with nonlinear exposure effects suffer from identification problems, even in the absence of measurement error (Vansteelandt and Goetghebeur (2005)). For similar reasons, no effect modification by the IV-C is allowed. Note, however, that the linear model can be seen as a first order approximation and that model (2.2) will therefore often give a reasonable approximation, even for nonlinear causal effects.

Assumption A4 (Measurement error IV assumption). Given the difficulty in obtaining information about measurement error characteristics, we introduce an *instrumental variable for the measurement error* ($IV{-}M$). In contrast to an IV-C which satisfies Assumption A1, we define this to be a surrogate $\mathbf{T}_i \subseteq \mathbf{X}_i$ for the observed exposure (in the sense that is it is conditionally associated with W_i, given $(\mathbf{S}_i, R_i)$, where $\mathbf{S}_i$ is such that $\mathbf{X}_i \equiv (\mathbf{S}_i, \mathbf{T}_i)$), which is measured prior to exposure and is such that it does not modify the causal effect of received exposure on the outcome, i.e., such that

$$\mathrm{E}(Y_i - Y_{i0}|Z_i, \mathbf{X}_i, R_i) = \mathrm{E}(Y_i - Y_{i0}|Z_i, \mathbf{S}_i, R_i). \tag{2.4}$$

We thus assume that $\gamma(\mathbf{X}_i; \boldsymbol{\psi})$ in (2.2) does not involve $\mathbf{T}_i$. With a slight abuse of notation, we denote it by $\gamma(\mathbf{S}_i; \boldsymbol{\psi})$. Importantly, note that the IV-M $\mathbf{T}_i$ differs from and satisfies different assumptions than the IV-C R_i, which satisfies Assumption A1. The former IV will be used to correct for systematic measurement error, the latter to infer a causal effect of Z on Y.

The use of a no-interaction assumption such as (2.4) is increasingly common in causal inference, in particular in the context of IV-estimation. For instance, Ten Have, Joffe, Lynch, Brown, Maisto and Beck (2007), Joffe, Small and Hsu (2007), and Albert (2008) use similar no-interaction assumptions to infer direct causal effects. Vansteelandt and Goetghebeur (2004), and Fischer and Goetghebeur (2004) rely on no-interaction assumptions for assessing effect modification by treatment-free responses. In this study, the interest in Assumption A4 is motivated by the fact that other sources of information on the measurement error, such as repeated measurements or validation samples, are typically not available in large classes of problems (e.g., noncompliance adjustment), and by the fact that we can identify settings where the assumption is reasonable. For instance, in randomized clinical trials, one source of an IV for the measurement error on treatment noncompliance would be a measurement of placebo compliance during

a run-in period of the study. Indeed, run-in placebo compliance is associated with treatment compliance and likely not further related to the treatment effect, given the actual compliance during the active study period (unless in the presence of side effects, where large differences between treatment and placebo compliance may be suggestive of side effects and thus of treatment activity). More generally, one can use a second causal IV as an IV for the measurement error. Indeed, an IV-C is associated with the considered exposure by Assumption A1, and does not modify the target causal effect by Assumption A3. It thus satisfies the conditions for an IV-M. The use of multiple IV-C's turns out feasible in practice as it is commonly considered in econometrics and, more recently, also in Mendelian randomization studies (Didelez and Sheehan (2007)). For instance, to assess the effect of C-reactive protein on insulin resistance, one may use the CRP-gene as an IV-C and the interleukin-6 gene - which is known to be associated with C-reactive protein through other pathways than the CRP-gene and which thus applies as a second IV-C - as an IV-M. Note furthermore that the restrictions for an IV-M are much weaker than those for an IV-C as an IV-M need not satisfy the exclusion restriction, nor the randomization assumption (see Assumption A1). Note also that (2.4) is weaker than the typical IV-assumption encountered in measurement error models (Carroll et al. (2006)) as it does allow for the IV to be associated with the outcome, conditional on the exposure.

Assumption A5 (Constant average measurement error). For simplicity and because information about the average error is weak, we develop our approach below for constant (but unknown) average error $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta^*$. This assumption is standard in the measurement error literature, but is straightforwardly relaxed (e.g., by postulating $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta_0^* + \delta_1^{*'}\mathbf{X}_i$).

2.2. Inference

Our goal is to estimate the parameter $\boldsymbol{\psi}^*$ indexing (2.2) under model $\mathcal{A}$, which is the model for the observed data $(Y_i, W_i, R_i, \mathbf{X}_i)$ defined by assumptions A1−A5 with the conditional density

$$f(R_i|\mathbf{X}_i) \text{ known.} \tag{2.5}$$

The latter assumption holds in a randomized trial when R_i indicates randomized assignment, because treatment allocation is then under the control of the investigator. If (2.5) fails, then all further results remain valid upon replacing $f(R_i|\mathbf{X}_i)$ with a consistent estimator.

It will follow from our Proposition 1 (whose proof is given in the Supplementary Materials on `http://www.stat.sinica.edu.tw/statistica`) that the average measurement error δ^* is all that must be known for identifying $\boldsymbol{\psi}^*$ under model $\mathcal{A}$.

Proposition 1. *Model $\mathcal{A}$ is the same model for the observed data as the conditional mean independence model $\mathcal{B}$ for the observed data model, defined by* (2.5) *and*

$$E\Big[Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi}^*)(W_i - \delta^*)|\mathbf{X}_i, R_i\Big] = E\Big[Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi}^*)(W_i - \delta^*)|\mathbf{X}_i\Big]. \quad (2.6)$$

Note the essential difference between models $\mathcal{A}$ and $\mathcal{B}$. Model $\mathcal{A}$ is expressed in terms of counterfactuals and therefore has parameters with a causal interpretation. Model $\mathcal{B}$ imposes the same restrictions on the observed data as model $\mathcal{A}$, but is not expressed in terms of counterfactuals. This makes the parameters in this model harder to interpret, but simplifies inference as the model is expressed in terms of observed data only. Note also that model $\mathcal{A}$ imposes only weak restrictions on the error distribution. First, it allows the error to be associated with both the true exposure Z_i and observed exposure W_i. It thus encompasses both the classical and Berkson error model (Carroll et al. (2006)). In addition, by avoiding assumptions about the conditional association between W_i and Y_i given Z_i, it allows for so-called differential error, which is associated with outcome conditional on exposure (see the proof of Proposition 1 for a more explicit argument). This can be important. For instance, in a clinical trial, patients may be more reluctant to 'confess' to noncompliance when their outcome stayed below target. Finally, model $\mathcal{A}$ makes no assumptions on the measurement error distribution other than Assumption 5. This is useful because the error distribution can be quite complex. For instance, with low level exposures, negative errors become constrained by the fact that negative exposures are never reported.

By Proposition 1 and the fact that $\boldsymbol{\psi}^*$ is the same functional of the observed data under models $\mathcal{A}$ and $\mathcal{B}$, inference for $\boldsymbol{\psi}^*$ is the same under both models. It follows that the set of all consistent and asymptotically normal (CAN) estimators for $\boldsymbol{\psi}^*$ is the same under models $\mathcal{A}$ and $\mathcal{B}$, where the latter can be obtained as in Robins (1994) by solving the mean independence estimating equations

$$\sum_{i=1}^{n} \mathbf{d}(R_i, \mathbf{X}_i)\Big[Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i - \delta) - q(\mathbf{X}_i)\Big] = 0 \quad (2.7)$$

jointly for $\boldsymbol{\theta} = (\boldsymbol{\psi}', \delta)'$, with $\mathbf{d}(R_i, \mathbf{X}_i) = \mathbf{g}(R_i, \mathbf{X}_i) - \mathrm{E}\{\mathbf{g}(R_i, \mathbf{X}_i)|\mathbf{X}_i\}$ and with $\mathbf{g}(R_i, \mathbf{X}_i)$ and $q(\mathbf{X}_i)$ arbitrary (non-trivial) index functions of the dimension of $\boldsymbol{\theta}$. Note that (2.7) is designed to make the predicted exposure-free outcomes $Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i - \delta)$ mean independent of R_i, conditional on $\mathbf{X}_i$, in order to satisfy Assumption A1. The index functions $\mathbf{g}(R_i, \mathbf{X}_i)$ and $q(\mathbf{X}_i)$ can be arbitrarily chosen without affecting the consistency of the resulting estimators of $\boldsymbol{\psi}^*$. In particular, they can be chosen in view of efficiency. Under the homoscedasticity assumption that the conditional variance of $Y_i - \gamma(\mathbf{X}_i;\boldsymbol{\psi})(W_i - \delta)$, given $(R_i, \mathbf{X}_i)$,

is constant, semi-parametric efficiency (Robins (1994)) is for instance obtained by setting $q(\mathbf{X}_i)$ equal to

$$q_{opt}(\mathbf{X}_i) = E\Big\{Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i-\delta)|\mathbf{X}_i, R_i\Big\}$$

and $\mathbf{d}(R_i,\mathbf{X}_i)$ equal to $\mathbf{d}_{opt}(R_i,\mathbf{X}_i) = \mathbf{g}_{opt}(R_i,\mathbf{X}_i) - \mathrm{E}\{\mathbf{g}_{opt}(R_i,\mathbf{X}_i)|\mathbf{X}_i\}$, with

$$\mathbf{g}_{opt}(R_i,\mathbf{X}_i) = E\left\{\frac{\partial\gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i-\delta)}{\partial\boldsymbol{\theta}}\Big|\mathbf{X}_i,R_i\right\}.$$

These choices will be used later in the data analysis and simulation study.

Theorem 1.

1. *Under weak regularity conditions, the solution* $\hat{\boldsymbol{\psi}}(\mathbf{d},q)$ *to* (2.7) *satisfies* $\sqrt{n}(\hat{\boldsymbol{\psi}}(\mathbf{d},q)-\boldsymbol{\psi}^*) \to N(0,\Gamma(\mathbf{d},q))$ *in distribution, where*

$$\Gamma(\mathbf{d},q) = E^{-1}\left\{\frac{\partial\mathbf{U}_i(\mathbf{d},q;\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}}\right\}\mathrm{Var}\left\{\mathbf{U}_i(\mathbf{d},q;\boldsymbol{\psi}^*)\right\}E^{-1\prime}\left\{\frac{\partial\mathbf{U}_i(\mathbf{d},q;\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}}\right\}, \quad (2.8)$$

   with $\mathbf{d}(R_i,\mathbf{X}_i) = (\mathbf{d}_\psi(R_i,\mathbf{X}_i), d_\delta(R_i,\mathbf{X}_i))$ *and*

$$\mathbf{U}_i(\mathbf{d},q;\boldsymbol{\psi}) = \left[\mathbf{d}_\psi(R_i,\mathbf{X}_i) - \frac{E\{\mathbf{d}_\psi(R_i,\mathbf{X}_i)\gamma(\mathbf{S}_i;\boldsymbol{\psi})\}}{E\{d_\delta(R_i,\mathbf{X}_i)\gamma(\mathbf{S}_i;\boldsymbol{\psi})\}}d_\delta(R_i,\mathbf{X}_i)\right] \times\Big[Y_i-\gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i-\delta)-q(\mathbf{X}_i)\Big].$$

2. *The average error* δ^* *is not root-n estimable at* $\boldsymbol{\psi}^*=\mathbf{0}$.
3. *For arbitrary* $(\mathbf{d},q)$, $\Gamma(\mathbf{d}_{\mathrm{opt}},q_{\mathrm{opt}}) \le \Gamma(\mathbf{d},q)$ *where* $A \le B$ *is defined as* $A-B$ *being semi-positive definite.*

Part 1 of Theorem 1 (whose proof is given in the Supplementary Materials) confirms that the solution $\hat{\boldsymbol{\psi}}(\mathbf{d},q)$ to (2.7) is a root-n CAN estimator of $\boldsymbol{\psi}^*$. This is even so at $\boldsymbol{\psi}^*=\mathbf{0}$ where δ^* is not root-n estimable. Theorem 1 also shows how to calculate the efficient score $\mathbf{U}_i(\mathbf{d}_{\mathrm{opt}},q_{\mathrm{opt}};\boldsymbol{\psi})$ for $\boldsymbol{\psi}^*$ in model $\mathcal{A}$. For example, in Section 4, we consider the analysis of a placebo-controlled randomized trial with Z_i denoting compliance to the experimental treatment. Because the placebo arm ($R_i=0$) is unexposed, $Z_i = Z_iR_i$ and there is no measurement error in that arm so that we modify Assumption A5 to $E(W_i-Z_i|\mathbf{X}_i,R_i)=\delta^*R_i$. With $\mathbf{X}_i=\mathbf{T}_i$, $\gamma(\mathbf{S}_i;\psi)=\psi$ and, assuming homoscedasticity and constant randomization probabilities $\pi = P(R_i=1) = P(R_i=1|\mathbf{X}_i)$, the semi-parametric efficient score for ψ^* is

$$(R_i-\pi)\Big[E(W_i|R_i=1,\mathbf{X}_i)-E\{E(W_i|R_i=1,\mathbf{X}_i)\}\Big]\Big\{Y_i-\psi(W_i-\delta)R_i-q_{opt}(\mathbf{X}_i)\Big\}.$$

This score differs from the efficient score in the absence of biased measurement error (i.e., assuming that $\delta^*=0$) in that it carries the additional term

$E\{E(W_i|R_i = 1, \mathbf{X}_i)\}$, which corrects for estimation of the error mean. This term reduces the variance of the estimating functions and, as such, encodes efficiency loss. Specifically, note that the efficient score becomes 0 when the IV-M, T, is uncorrelated with the observed exposure, and hence that ψ^* is not root-n estimable in that case. By the same token, instruments for the measurement error that are weakly correlated with observed exposure may yield unstable effect estimates.

2.3. Bias-variance trade-off

The anticipated loss of efficiency of the error-adjusted estimator raises the question of whether the bias correction developed so far is useful. To this end, we investigate the bias-variance trade-off for the error-adjusted and the standard unadjusted estimator for the causal effect ψ^*, in a specific case. Tractable expressions for the mean-squared error of both estimators, are obtained when $Z \sim N(\mu_z, \sigma_z^2), T|Z \sim N(\nu_0 + \nu_1 Z, \sigma_{t|z}^2), Y_0|Z, T \sim N(\alpha_0 + \alpha_1 Z + \alpha_2 T, \sigma_0^2)$ and $Y = Y_0 + (\psi + \epsilon)RZ$ with $\epsilon|Y_0, Z, T \sim N(0, \sigma^2)$.

Under the working assumption of no systematic measurement error (i.e., fixing $\delta^* = 0$ in equation (2.7) and not estimating it), the efficient score for ψ^* is $U_u(\psi) = (0.5 - R)\mathrm{E}(W|T, R = 1)\{Y - \psi RW - E(Y|R = 0, T)\}$ in model $\mathcal{A}$ with $\mathbf{X}_i = \mathbf{T}_i$ under the above data-generating mechanism. It follows after some algebra that the solution $\hat{\psi}_u$ to $\sum_{i=1}^n U_{ui}(\psi) = 0$ has bias which can be approximated with

$$\mathrm{E}^{-1}\left(\frac{\partial U_u(\psi)}{\partial \psi}\right)\mathrm{E}\{U_u(\psi)\} = \frac{\psi\delta(\mu_z + \delta)}{\sigma_z^2 - \sigma_{z|t}^2 + (\mu_z + \delta)^2},$$

where $\sigma_{z|t}^2 = \sigma_z^2\sigma_{t|z}^2/(\nu_1^2\sigma_z^2 + \sigma_{t|z}^2)$ is the conditional variance of Z given T, and asymptotic variance given by

$$\frac{1}{n}\left[\frac{4\sigma_0^2 + 4\alpha_1^2\sigma_{z|t}^2 + 2\psi^2\sigma_u^2 + \psi^2\delta^2}{\sigma_z^2 - \sigma_{z|t}^2 + (\mu_z + \delta)^2} + \frac{\psi^2\delta^2(\sigma_z^2 - \sigma_{z|t}^2)}{\{\sigma_z^2 - \sigma_{z|t}^2 + (\mu_z + \delta)^2\}^2}\right].$$

Allowing for systematic measurement error, the efficient estimator $\hat{\psi}_c$ for ψ^* under model $\mathcal{A}$ has no asymptotic bias and asymptotic variance

$$\frac{1}{n}\,\frac{\sigma_0^2 + \alpha_1^2\sigma_{z|t}^2 + 0.5\psi^2\sigma_u^2}{0.5^2(\sigma_z^2 - \sigma_{z|t}^2)}.$$

Note that the bias of the unadjusted estimator and the asymptotic variance of both estimators is inversely proportional to the multiple correlation coefficient for the regression of Z on T. The variance of the error-adjusted estimator becomes infinite when Z and T are uncorrelated.

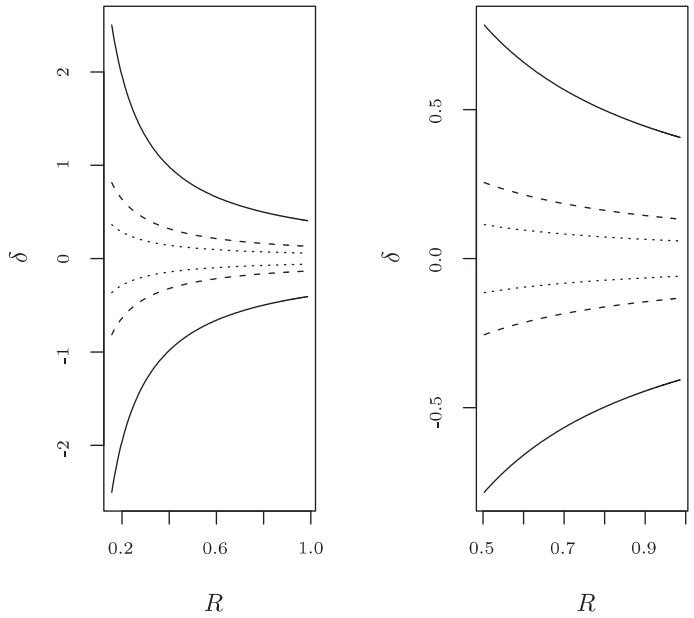

Figure 1. Curves indicating the tuples (R, δ) where the standard SMM estimator and the error-adjusted instrumental variable estimator have the same mean squared error, for R equalling the correlation between Z and T, for different sample sizes $n = 105, 1,000$ and 5,000 and with $\mu_z = 0.85$, $\sigma_z^2 = 0.11$, $\nu_0 = 0.75$, $\nu_1 = 0.12$, $\sigma_{t|z}^2 = 0.012$, $\alpha_0 = -4.4$, $\alpha_1 = 6.8$, $\alpha_2 = -13.7$, $\sigma_0^2 = 53.2$, $\sigma_u^2 = 0$, $\psi = -7.5$ and $\sigma^2 = 0$. Left: for R from 0 to 1; Right: for R from 0.5 to 1.

Figure 1 shows the range of values δ for the average error under which the standard estimator, which ignores measurement error, has smaller mean squared error than the error-adjusted estimator. This is displayed as a function of the sample size and the correlation between Z and T. Specifically, the values of δ between the solid lines indicate data-generating mechanisms under which the standard estimator outperforms the error-adjusted estimator in terms of mean squared error. The figure was constructed using parameter values which are reflective of the hypertension study that we analyze in Section 4. It shows that at small sample sizes ($n = 105$), correction for systematic measurement error leads to smaller mean squared error, but only when the systematic error component is substantial (i.e., of about the size of the average exposure μ_z) and, at the same time, the IV-M, T, is strongly correlated with Z. Figure 1 indicates further that bias correction using the error-adjusted estimator is only helpful at moderate degrees of error and moderate correlations between T and Z when sample sizes are very large.

3. Incorporating Prior Information

The previous results demonstrate the poor performance of the error-adjusted estimator, even in settings where the sample size is moderate and good (pre-exposure) predictors of exposure are available. In particular, tests of the causal null hypothesis can lose substantial power by using this approach rather than the standard test of the causal null (i.e., that R and Y are conditionally independent, given $\mathbf{X}_i$), which is immune to measurement error on the exposure (Goetghebeur and Vansteelandt (2005)). This is surprising, considering that the score test of $\boldsymbol{\psi}^* = \mathbf{0}$ under model $\mathcal{A}$ does not involve δ^* and, hence, does not need to correct for measurement error when testing the causal null hypothesis. Curiously, it follows that one can validly and efficiently test the causal null hypothesis without correction for measurement error, but that a score test of $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$ with $\boldsymbol{\psi}_0$ arbitrarily close to (but different from) 0, would require correcting for measurement error and hence could imply a serious and sudden loss of power.

The root cause of this apparent discontinuity is the fact that, as shown in Part 2 of Theorem 1, δ^* is not root-n estimable at $\boldsymbol{\psi}^* = \mathbf{0}$, so that estimation of δ^* affects the distribution of the score test statistic even though it gets multiplied by $\boldsymbol{\psi}^* = \mathbf{0}$ in the test statistic (i.e., even at the causal null hypothesis). In particular, it follows from the proof of Theorem 1 that $\sqrt{n}\{\hat{\delta}(\mathbf{d}, q) - \delta^*\}\boldsymbol{\psi}^*$, with $\hat{\delta}(\mathbf{d}, q)$ the solution for δ to (2.7), is bounded in probability with strictly positive variance for each value of $\boldsymbol{\psi}^*$, suggesting that $\hat{\delta}(\mathbf{d}, q)\boldsymbol{\psi}^*$ varies around 0, even when $\boldsymbol{\psi}^* = \mathbf{0}$. This happens with decreasing variance as the sample size increases.

Similar problems of inestimability at a local point in the parameter space have been noted in other measurement error problems (Gustafson (2005)). More general problems of inferring a parameter $\boldsymbol{\psi}^*$ when a nuisance parameter δ^* disappears under the null ($\boldsymbol{\psi}^* = \mathbf{0}$) have been discussed mainly in the econometrics literature (Davies (1977, 1987), Hansen (1992), and Andrews and Ploberger (1994)). To the best of our knowledge, attention has only been given to testing problems in which the test statistic involves a nuisance parameter which is unidentified at the null. Some of these approaches assume that the nuisance parameter lies within a known open set, and base inference on the supremum of a score or likelihood ratio test statistic taken over all values of the nuisance parameters in the chosen set (Davies (1977, 1987)). Andrews and Ploberger (1994) postulate a prior distribution for the nuisance parameter and base inference on the average of a score or likelihood ratio test statistic over the chosen prior distribution. Our problem is different in that our main focus is on estimation rather than testing, and that a score test for the causal null hypothesis does not involve the nuisance parameter. Nonetheless, inspired by the work of Davies (1977, 1987) and by sensitivity analyses for IV-estimators with measurement error (Goetghebeur and Vansteelandt (2005)), we proceed by considering estimation under the

assumption that the average error δ^* lies within a known open set Δ. This strategy is further motivated by the fact that (a) subject-matter experts often have a good sense of the extent of expected mismeasurement (Gustafson (2005)); (b) it forces the estimate for δ^* to have bounded variation around the truth, contrary to what happens under the approach of Section 2.2.

3.1. Improved error adjustment

Our first approach under the assumption that $\delta^* \in \Delta =]\Delta_l, \Delta_u[$ is to solve equations (2.7) with δ replaced by $\{I(\lambda < 0)\Delta_l + I(\lambda > 0)\Delta_u\}\lambda/(1 + |\lambda|)$ and λ unknown. This guarantees estimates $\delta^*(\hat{\lambda})$ within the set Δ and thus greatly improves the stability of estimators for the causal effect $\boldsymbol{\psi}^*$. A drawback that becomes apparent in the simulation study of Section 5, is that tests of the causal null hypothesis may still lose substantial power under this approach due to the fact that also λ is not root-n estimable at $\boldsymbol{\psi}^* = \mathbf{0}$. To accommodate this, we develop a second, recommended approach which trades bias for precision by solving a weighted average of the estimating functions for the standard SMM estimator and for the error-adjusted estimator of Section 2.3. Estimating functions for the standard estimator are weighted proportionally to the estimated probability that δ^* falls outside the chosen set Δ. The philosophy behind this choice is that estimates for δ^* will not likely fall within the set Δ in situations where little information on the error mean is available. Hence more weight will be given to the standard unadjusted estimator in those cases.

For pedagogic purposes, we explain our proposal for the case $\gamma(\mathbf{X}_i; \psi) = \psi$, and with Assumption A5 modified to $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta^* R_i$, so as to represent the setting of our application in Section 4. We further delete reference to the index functions (d, q) in the estimators. For each value ψ in a chosen grid, we calculate an estimator $\hat{\delta}(\psi)$ for δ^* that solves (2.7) for the given ψ with $d_\delta(R_i, \mathbf{X}_i)$ in place of $d(R_i, \mathbf{X}_i)$. Next, we consider a weighted average of the estimating function $U_{\psi i}(\psi, \delta)$ for ψ^* (as defined in (2.7) with $d_\psi(R_i, \mathbf{X}_i)$ in place of $d(R_i, \mathbf{X}_i)$), evaluated at the profile estimator $\delta = \hat{\delta}(\psi)$ and at $\delta = 0$, respectively,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{U}_i(\psi) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{P}\{\hat{\delta}(\psi) \in \Delta\} U_{\psi i}\{\psi, \hat{\delta}(\psi)\} + \hat{P}\{\hat{\delta}(\psi) \notin \Delta\} U_{\psi i}(\psi, 0). \tag{3.1}$$

In this expression, the weights involve the estimated probability $\hat{P}\{\hat{\delta}(\psi) \notin \Delta\}$ that $\hat{\delta}(\psi)$ falls outside the chosen interval $\Delta =]\Delta_l, \Delta_u[$. Using a similar development as in the proof of Theorem 1, this probability can be approximated by

$$P\{\hat{\delta}(\psi) \notin \Delta\} = 1 + \Phi\left(\frac{\Delta_l - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right) - \Phi\left(\frac{\Delta_u - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right), \tag{3.2}$$

with $\Phi(.)$ the cumulative standard normal distribution function, and where δ may be replaced by a consistent estimator (for instance, $\min[\Delta_u, \max\{\Delta_l, \hat{\delta}(\psi)\}]$) and $\sigma(\psi)$ by a consistent estimator $\hat{\sigma}(\psi)$ for the standard deviation of the scaled estimating function $E^{-1}[d_\delta(R,T,X)R]U_{i\delta}(\psi,\delta)$ for δ^*. We define the improved error-adjusted estimator $\tilde{\psi}$ for ψ^* as the value of ψ at which the score test (3.1) becomes zero. Curiously, this estimator assigns much weight to the standard estimating equations (which do not adjust for measurement error) when the error mean is estimated to be large. This is (a) because the philosophy behind the estimator is that such large values for the error mean are indicative of imprecision; and (b) because the estimating functions are designed to equal the unadjusted estimating functions at the causal null hypothesis (see further).

Theorem 2. *Suppose that* $\gamma(\mathbf{X}_i;\psi) = \psi$, $Z_i = Z_i R_i$ *and* $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta^* R_i$. *Then, under regularity conditions stated in the Appendix and for any fixed* ψ, $(1/\sqrt{n})\sum_{i=1}^n \tilde{U}_i(\psi) \to N(0,\Sigma(\psi))$ *in distribution, where* $\Sigma(\psi)$ *is the variance of*

$$
\begin{aligned}
&P\{\hat{\delta}(\psi) \in \Delta\}U_{i\psi}(\psi,\delta) + P\{\hat{\delta}(\psi) \notin \Delta\}U_{i\psi}(\psi,0) \\
&\quad - \Bigg[P\{\hat{\delta}(\psi) \in \Delta\} + \Bigg\{\varphi\bigg(\frac{\Delta_l - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\bigg) \\
&\quad - \varphi\bigg(\frac{\Delta_u - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\bigg)\Bigg\}\frac{\sqrt{n}|\psi|\delta}{\sigma(\psi)}\Bigg]\frac{E\{d_\psi(R,\mathbf{X})R\}}{E\{d_\delta(R,\mathbf{X})R\}}U_{i\delta}(\psi,\delta)
\end{aligned}
$$

with $\varphi(.)$ *the standard normal density function.*

Theorem 2 (whose proof is given in the Supplementary Materials) can be used to construct $(1-\alpha)100\%$ confidence intervals for ψ^* as the range of values ψ_0 for ψ such that the two-sided score test based on (3.1) does not reject the null hypothesis $H_0 : \psi^* = \psi_0$ at the $\alpha 100\%$ significance level. To evaluate this score test, one may replace the variance of the score test statistic by the sample variance with $P\{\hat{\delta}(\psi) \in \Delta\}$ replaced by $\hat{P}\{\hat{\delta}(\psi) \in \Delta\}, \delta$ by $\hat{\delta}(\psi)$, and $\sigma(\psi)$ by $\hat{\sigma}(\psi)$. The resulting confidence intervals have the desirable feature that they exclude 0 if and only if the standard test of the causal null hypothesis (i.e., that $Y \perp\!\!\!\perp R|\mathbf{X}$) rejects. Indeed, at the null hypothesis $\hat{P}\{\hat{\delta}(0) \notin \Delta\} = 1$, and hence the score test statistic becomes

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^n U_{\psi i}(\psi,0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n d_\psi(R_i,\mathbf{X}_i)\{Y_i - q(\mathbf{X}_i)\}
$$

for an arbitrary function $d_\psi(R_i,\mathbf{X}_i)$ with conditional mean zero, given $\mathbf{X}_i$. When $\hat{\delta}(\psi)$ in $U_{\psi i}(\psi,\hat{\delta}(\psi))$ is restricted to Δ as described in the first paragraph of this

section, this statistic is a score test statistic of the causal null hypothesis under the observed data model defined by restriction (2.5).

Unfortunately, the suggested confidence intervals are not uniform asymptotic confidence intervals. The reason is that, at each sample size, there exists a ψ^* depending on n which is sufficiently close to zero that the score test statistic (3.1) is significantly biased as a result of bias in the estimating functions of the standard unadjusted SMM estimator. Specifically, it follows from the proof of Theorem 2 that the improved error-adjusted estimator $\tilde{\psi}$ is asymptotically biased within root-n shrinking neighbourhoods of zero (i.e., when $\psi^* = k/\sqrt{n}$ for some constant k) and may not converge to a normal distribution along such sequences. Curiously, $\tilde{\psi}$ is asymptotically unbiased and normally distributed along faster converging sequences (i.e., when $\psi^* = kn^{-a}$ for some constant k and $a > 1/2$) and, in particular, at $\psi^* = 0$. The reason is that, although the probability that $\hat{\delta}(\psi) \in \Delta$ now converges to 0 and hence $\tilde{\psi}$ is asymptotically equivalent to the standard unadjusted SMM estimator, ψ^* is sufficiently close to zero to make any bias in the estimator negligible. Likewise, $\tilde{\psi}$ is asymptotically unbiased and normally distributed along slower converging sequences (i.e., when $\psi^* = kn^{-a}$ for some constant k and $0 \leq a < 1/2$). The reason is that the estimated probability of $\hat{\delta}(\psi) \in \Delta$ now converges to 1 so that the improved error-adjusted estimator is asymptotically equivalent to the error-adjusted estimator of Section 2.2, which is asymptotically unbiased.

The practical implication of the foregoing discussion is that the improved error-adjusted estimator $\tilde{\psi}$ and confidence intervals have no guaranteed performance in finite samples in the sense that, for each sample size, one can find a causal effect ψ^* which is close, but not too close, to zero so that $\tilde{\psi}$ is significantly biased and that confidence intervals for ψ^* do not cover ψ^* at the nominal level. This local bias is the price we pay for estimators with smaller variability and limited loss of power for testing the causal null hypothesis. Because this problem only appears within $n^{-1/2}$ distances from zero and not within larger or shorter distances, we expect adequate performance in many practical situations. However, to be conservative we develop uniform asymptotic confidence intervals in the next section.

3.2. Uniform asymptotic confidence intervals

Uniform asymptotic $(1-\alpha)100\%$ confidence intervals are expected to have better finite-sample properties than the intervals of the previous section because they guarantee the existence of a minimal sample size such that, at larger sample sizes, they cover ψ^* with at least $(1-\alpha)100\%$ chance regardless of the value of ψ^*. Following ideas in Robins (2005), we construct such intervals by first constructing, for each ψ, an asymptotic uniform $(1-\epsilon)100\%$ confidence interval

$C(\psi)$ for δ^*, where the choice of $\epsilon < \alpha$ will be discussed later. Because we assume the parameter space for δ^* to be Δ, a conservative asymptotic interval $C(\psi)$ may be obtained as

$$\left\{\hat{\delta}(\psi) \pm z_{\epsilon/2}\frac{\hat{\sigma}(\psi)}{|\psi|\sqrt{n}}\right\} \cap \Delta.$$

Using Theorem 5.1 in Robins (2005), an asymptotic uniform $(1-\alpha)100\%$ confidence interval for ψ^* may be obtained as the set of ψ-values for which

$$\inf_{\delta \in C(\psi)} \left| \text{Var}^{-1/2}\{U_{\psi i}(\psi, \delta)\} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_{\psi i}(\psi, \delta) \right| < z_{(\alpha-\epsilon)/2}.$$

The optimal choice of ϵ that leads to confidence intervals of minimum length is difficult to determine (Robins (2005)). We propose to choose ϵ in function of ψ as $0.5\alpha|\psi|/(1+|\psi|)$. This choice guarantees that $C(\psi)$ will equal Δ for $\psi^* = 0$, and be a $(1-\alpha/2)100\%$ confidence interval for δ^* at causal effects ψ^* far from 0. The philosophy behind this choice is that estimates for δ^* will be highly imprecise at causal effects close to zero and hence, given that the parameter space for δ^* is bounded, we expect no difference between 100% confidence intervals and $(1-\alpha)100\%$ confidence intervals for δ^* at $\psi^* = 0$. As such, we need not offer the significance level for ψ^* at small causal effects and will thus get narrower intervals in return. Specifically, the proposed confidence intervals have the feature that they involve no correction for measurement error at $\psi^* = 0$; this is desirable because there is no bias due to measurement error at $\psi^* = 0$.

4. Application

We analyze data from a placebo-controlled randomized hypertension trial which enrolled some 300 hypertensive patients (Goetghebeur and Lapp (1997)). After a run-in period of four weeks where all patients received placebo tablets, they were randomized to four weeks of one of two active treatments (A or B) or placebo. All treatments were prescribed at one tablet per day. Here we analyze the subset of 105 patients randomized to A or placebo, for whom treatment compliance was electronically measured, ignoring 5 patients with missing diastolic blood pressure or compliance.

An intent-to-treat analysis reveals an average difference in blood pressure reduction over the active four week study period of 7.5 mmHg (95% CI 4.0; 11.0) without adjustment. This reveals the effect of assignment to treatment A (instead of placebo) on expected diastolic blood pressure reduction from baseline (i.e., the time of randomization). Primary interest lies however in the effect of *received* treatment on average blood pressure reduction. We therefore fit model (2.3) with Y_i the blood pressure reduction over the active study period, R_i the randomization indicator as the IV-C (which is 1 if assigned to experimental treatment and 0

if assigned placebo), Z_i the average number of prescribed pills taken, and $\mathbf{X}_i$ the age of patient i. Assuming that compliance measurements are free of systematic error, we estimate that the average blood pressure reduction would have been 9.6 mmHg (95% CI 3.5; 11.8) smaller over the study period among those who chose to take on average one pill per day, had they not taken the exposure. Note that this estimand averages the effect over patients with different compliance patterns, but with the same average pill intake. Distinguishing between these patients would require more detailed compliance measures, but would suffer from identifiability problems (Vansteelandt and Goetghebeur (2005)).

In reality, there are concerns that electronic compliance measurements carry systematic errors and thus that the above estimate may be biased. Because this study was not designed to correct for measurement error, no natural IV's for the measurement error have been recorded. Our analysis is hence for illustrative purposes only, and will use age as an IV-M (i.e., $\mathbf{T}_i = \mathbf{X}_i$ equals age). Age was chosen because effect modification through age is not anticipated (nor observed) in this study population, which consists of middle aged hypertensive patients (5th, 95th percentiles: 41 and 69 years), and thus Assumption A4 is anticipated to be approximately true. A more adequate analysis would use placebo compliance during the run-in period (where no electronic adherence measures were taken) as an IV-M. Using the error-adjusted estimator of Section 2.2, we estimate a larger treatment effect of 27.0 mmHg (95% CI -91.2; 145.2). To improve this imprecise result, we impose the weak assumption that the average error is smaller than 0.25. We believe this assumption to be reasonable, given that the observed percentage of assigned dose taken (i.e., the observed exposure) is 0.85 (i.e., 85%) on average. Choosing $\Delta = [-0.25, 0.25]$ thus allows for 30% of the observed average exposure to be due to systematic error. Using the improved error-adjusted estimator for inference, we estimate a slightly smaller effect of 9.0 mmHg (95% CI 4.4; 17.4) as compared to the standard analysis. As predicted by the theory, the estimate is less precise than the unadjusted estimator, but still significantly different from 0 at the 5% significance level. The uniform asymptotic 95% confidence interval (2.7; 16.8) has a more guaranteed performance in finite samples.

To investigate the sensitivity of our result to the choice of Δ, Figure 2 shows the improved error-adjusted estimate, along with uniform 95% confidence intervals as a function of the assumed maximum error mean Δ_u, with $\Delta = [-\Delta_u, \Delta_u]$. It reveals reasonable stability. Comparison with the sensitivity analysis results of Goetghebeur and Vansteelandt (2005) shows that the error-adjustment described in this article reduces uncertainty.

Note that our analysis is limited to a linear dose-response relationship. Because this linearity assumption is untestable, sensitivity analyses can be undertaken, as illustrated for these data in Vansteelandt and Goetghebeur (2005).

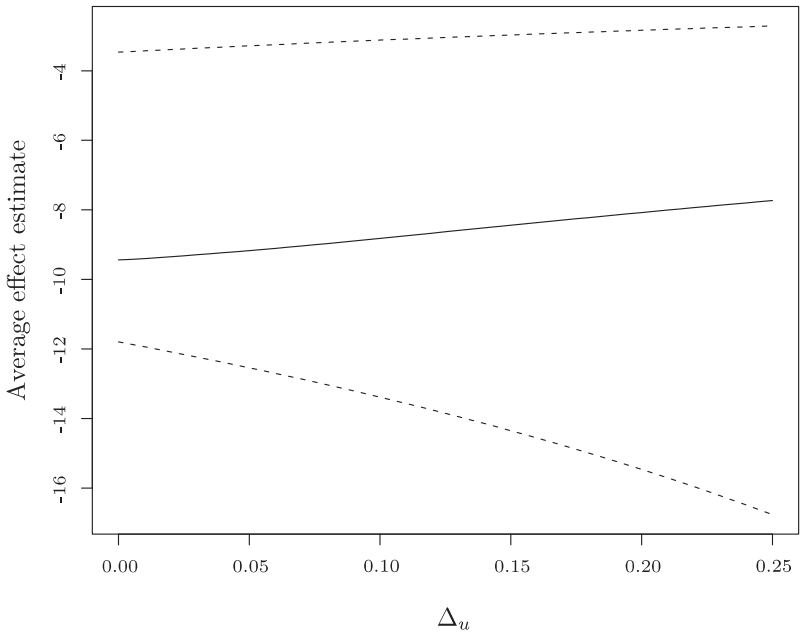

Figure 2. Improved error-adjusted estimate, along with uniform 95% confidence intervals as functions of the maximum error mean Δ_u, with $\Delta = [-\Delta_u, \Delta_u]$.

Note also that we a priori assume age to be a valid IV-M because tests for effect modification by age are underpowered when age is used as an IV for estimating the error mean (see Section 4 in Vansteelandt and Goetghebeur (2004) for related remarks).

5. Simulation Study

To investigate the behaviour of the error-adjusted estimators in finite samples with ψ^* possibly close to zero, we conducted simulation experiments. Each experiment was based on 5,000 replications of random samples of size 105 (i.e., the sample size of the blood pressure study) or 1,000, generated as follows. In each experiment, the instrument T for the measurement error was normal with mean 0.83 and standard deviation 0.14, and R was independently generated from a Bernoulli distribution with success probability 0.5. The true exposure Z and exposure-free response were generated as $Z = T + 0.32\epsilon_Z$ and $Y_0 = -4.4 + 6.8Z - 7.3T + 7.3\epsilon_0$ for independent standard normal variates ϵ_Z, ϵ_0. Finally, we generated Y as $Y_0 + \psi RZ$, and the observed exposure W as $W = (Z+U)R$ where $U \sim N(\delta, 0.01)$.

Table 1 summarizes the results for estimation of ψ using (i) the standard IV estimator which ignores systematic measurement error (STD); (ii) the error-adjusted estimator of Section 3.1 (IV1); (iii) the error-adjusted estimator of

Table 1. Bias of the different effect estimators and coverage and average length of corresponding 95% confidence intervals.

Δ	n	ψ	δ	Bias				Coverage					Average length CI				
				STD	IV1	IV2	IV3	STD	IV1	IV2	IV3	UI	STD	IV1	IV2	IV3	UI
0.5	105	-7.5	0.15	1.11	-3.77	-2.65	0.68	86.8	96.5	99.8	97.7	99.8	5.87	3039	56.2	13.3	18.8
0.5	105	-7.5	0	-0.020	-3.77	-2.31	-0.046	93.7	96.5	99.9	98.7	100	6.96	3039	55.7	10.8	29.2
0.5	105	0	0	-0.015	-3.63	-0.019	-0.0096	93.5	96.5	100	94.1	96.2	6.94	3027	63.8	8.88	21.1
0.5	1000	-7.5	0.15	1.13	-0.15	-0.28	0.81	36.4	95.1	99.9	97.8	100	1.90	14.0	14.0	4.83	11.4
0.5	1000	-7.5	0	-0.0048	-0.15	-0.52	-0.62	95.0	95.1	98.1	95.1	100	2.25	14.0	14.0	12.4	16.7
0.5	1000	0	0	-0.0042	-0.14	0.0032	-0.0036	94.9	95.2	100	95.0	96.2	2.24	14.0	13.9	2.78	5.86
0.25	105	-7.5	0.15	1.11	-3.77	0.59	1.06	86.8	96.5	100	97.0	99.5	5.87	3039	56.4	8.42	10.9
0.25	105	-7.5	0	-0.020	-3.77	-0.63	-0.062	93.7	96.5	100	98.8	99.9	6.96	3039	56.8	10.7	14.2
0.25	105	0	0	-0.015	-3.63	-0.013	-0.010	93.5	96.5	100	94.2	94.9	6.94	3027	67.8	8.88	11.6
0.25	1000	-7.5	0.15	1.13	-0.15	0.42	0.78	36.4	95.1	100	98.0	99.9	1.90	14.0	13.9	4.35	5.69
0.25	1000	-7.5	0	-0.0048	-0.15	-0.37	-0.19	95.0	95.1	100	95.8	99.6	2.25	14.0	14.0	5.83	7.61
0.25	1000	0	0	-0.0042	-0.14	-0.0012	-0.0036	94.9	95.2	100	95.0	95.5	2.24	14.0	13.9	2.78	3.43
0.05	105	-7.5	0.15	1.11	-3.77	1.07	1.11	86.8	96.5	100	91.3	94.8	5.87	3039	63.9	6.54	7.47
0.05	105	-7.5	0	-0.020	-3.77	-0.052	-0.015	93.7	96.5	100	96.3	98.2	6.96	3039	64.4	7.87	9.07
0.05	105	0	0	-0.015	-3.63	-0.014	-0.016	93.5	96.5	100	94.1	94.3	6.94	3027	67.9	7.53	8.51
0.05	1000	-7.5	0.15	1.13	-0.15	1.02	1.11	36.4	95.1	100	54.2	74.9	1.90	14.0	13.9	2.32	2.78
0.05	1000	-7.5	0	-0.0048	-0.15	-0.031	-0.0063	95.0	95.1	100	98.3	99.7	2.25	14.0	13.9	2.83	3.43
0.05	1000	0	0	-0.0042	-0.14	-0.0037	-0.0042	94.9	95.2	100	95.0	95.1	2.24	14.0	13.9	2.36	2.55

Section 3.3 which guarantees estimates for δ to stay within $\Delta = [\Delta_l, \Delta_u]$ with $\Delta_u = -\Delta_l$ equal to 0.5, 0.25 or 0.05, by defining $\delta = \{I(\lambda < 0)\Delta_l + I(\lambda > 0)\Delta_u\}\lambda/(1+|\lambda|)$ for unknown λ (IV2); the improved error-adjusted estimator of Section 3.3 with the same choices for Δ (IV3). In addition, the table shows uniform asymptotic 95% confidence intervals (UI) corresponding to these choices. The results for the different estimators were as predicted by the theory. The error-adjusted estimator (IV1) was extremely variable at small sample sizes, but performed adequately at larger sample sizes, even at $\psi = 0$. Estimator (IV2) was less variable, although still substantially less precise than the standard unadjusted estimator. Figures 3 and 4 show that estimator (IV1) was normally distributed in moderate sample sizes, even at $\psi = 0$, but not in small samples. It also shows that the improved error-adjusted estimator (IV3) was much less variable than the error-adjusted estimator (IV1). While the former followed a normal distribution in small samples, deviations from normality appeared in larger sample sizes as a result of convergence to a normal distribution not being uniform in ψ. By the same token, the improved error-adjusted estimator was more biased than the error-adjusted estimator in larger samples, and even than the standard IV estimator in some scenarios. Informally, this happened because data sets which carry evidence for causal effects close to zero, yielded estimated probabilities of $\hat{\delta}(\psi) \in \Delta$ close to zero. The bias then arose because the small estimated causal effects in such data sets were more attracted toward the estimates obtained from a standard structural mean analysis (which ignores measurement error) than large estimated causal effects. Additional simulations (not displayed) have shown that, as predicted by the theory, this bias and deviation from normality disappears again in larger sample sizes. Furthermore, note that the confidence intervals for the improved error-adjusted estimator retained their coverage despite these deviations, although there was a tendency for the approach to be conservative. Finally, as predicted by the theory, the uniform confidence intervals were conservative and also wider on average than those obtained via the improved error-adjusted estimator.

The impact of narrower intervals $\Delta = [-0.25, 0.25]$ was large at small sample sizes, but moderate at large sample sizes. For instance, confidence intervals based on the improved error-adjusted estimator had an average length of 8.42 (instead of 13.3) and coverage of 97.0% (instead of 97.7%) in the small samples, and 4.35 (instead of 4.83) and 98.0% (instead of 97.8%), respectively, in the large samples. The impact of $\Delta = [-0.05, 0.05]$ not including the error mean was to induce bias of the order of magnitude of the standard unadjusted estimator. The 95% confidence intervals based on the improved error-adjusted estimator and uniform 95% confidence intervals then no longer cover at the nominal rate. Coverage of those intervals was still better than the coverage of 95% confidence

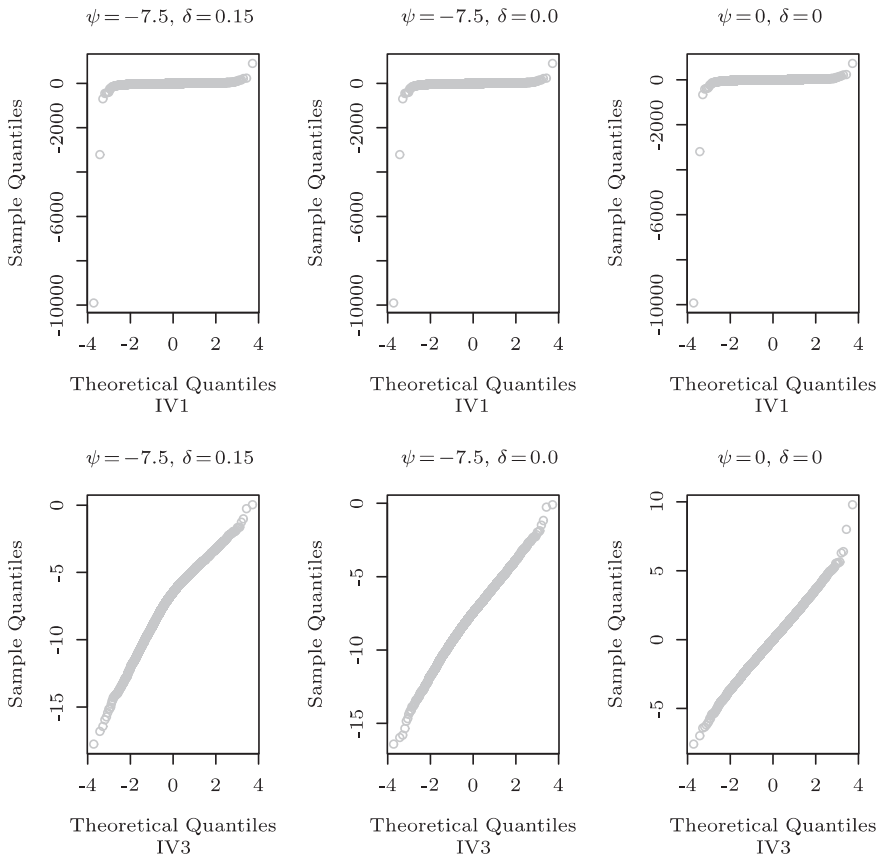

Figure 3. QQ-plots for $n = 105$ and $\Delta = [-0.5, 0.5]$. Row 1: error-adjusted estimator IV1; Row 2: improved error-adjusted estimator IV3.

intervals based on the standard unadjusted estimator, but at the expense of being wider.

6. Conclusions

We have proposed a general procedure to correct IV estimators for systematic error in the exposure when an additional IV for the measurement error is available. This procedure complements the sensitivity analysis approach of Goetghebeur and Vansteelandt (2005) and is especially attractive when the IV-M assumption (A4) is likely to be met. This is the case in placebo-controlled randomized trials with noncompliance, where measurements $\mathbf{T}_i$ on run-in placebo compliance may very well meet assumption (A4). With concern for compliance mismeasurement, recording run-in compliance may thus be favourable. More generally, causal IV's can be used as IV's for the measurement error.

On theoretical grounds and on the basis of simulation experiments, we rec-

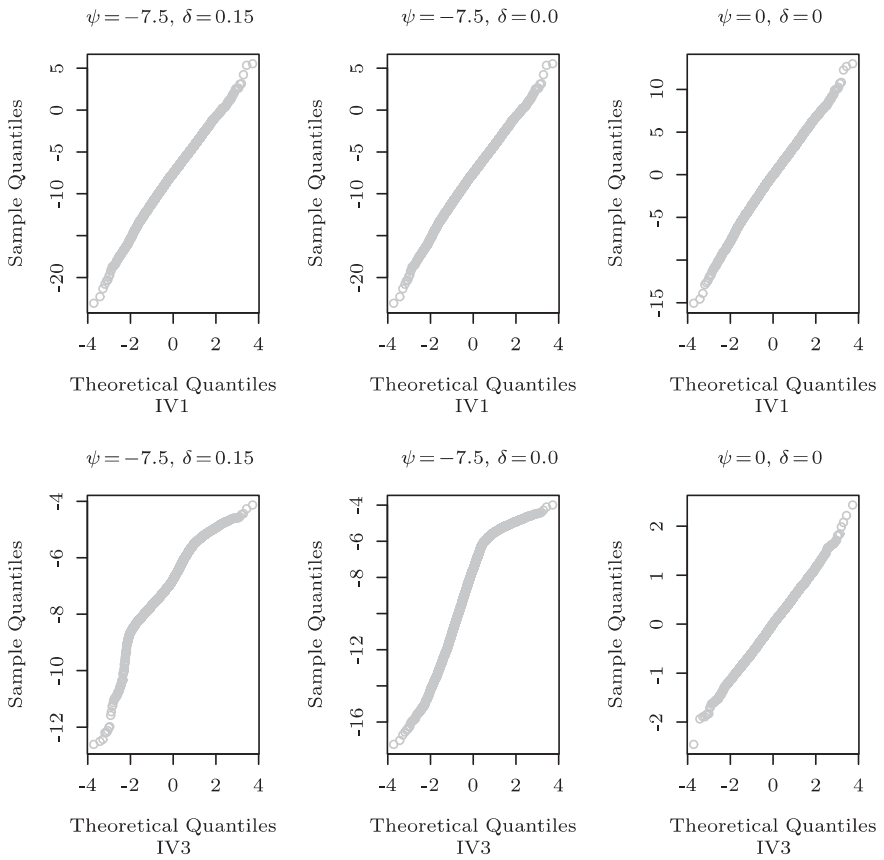

Figure 4. QQ-plots for $n = 1,000$ and $\Delta = [-0.5, 0.5]$. Row 1: error-adjusted estimator IV1; Row 2: improved error-adjusted estimator IV3.

ommend the improved error-adjusted estimator of Section 3.1. This estimator was designed so that adjustment for measurement error does not compromise the power of tests of the causal null. This is attractive, knowing that standard tests of the causal null hypothesis (i.e., that the causal instrument R is independent of outcome) ignore exposure measurements and are thus valid in the presence of measurement error. Because the proposed estimator does not converge uniformly to a normal distribution, we recommend the uniform confidence intervals of Section 2.3.

For illustrative purposes, we have developed this work under structural mean models that assume linear exposure effects that are not modified by pre-exposure covariates. Extensions to linear structural mean models that allow for effect modification by baseline covariates are methodologically straightforward, but computationally more demanding. Finally, we believe our results to be more broadly useful from a theoretical perspective as they suggest, in line with Gustafson

(2005), that incorporating some prior information on a weakly identified nuisance parameter may yield substantial efficiency improvements for the target parameter. Similar ideas may therefore prove useful in related settings (Vansteelandt and Goetghebeur (2004), Fischer and Goetghebeur (2004), and Ten Have et al. (2007)) with weak identification. In addition, our results indicate how such prior information may be adopted in a frequentist analysis.

Acknowledgements

The authors acknowledge support from IAP research network grant nr. P06/03 from the Belgian government (Belgian Science Policy). The second author would like to thank Iran's Minister of Science for financially supporting his PhD study at Ghent University. The third author acknowledges partial support from NIH grant AI24643.

References

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statist. Medicine* **27**, 1282-1304.

Andrews, D. W. K. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under alternative. *Econometrica* **62**, 1383-1414.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Amer. Statist. Assoc.* **91**, 444-455.

Buzas, J. S. and Stefanski, L. A. (1996). Instrumental variable estimation in generalized linear measurement error models. *J. Amer. Statist. Assoc.* **91**, 999-1006.

Casas, J. P., Bautista, L. E., Smeeth, L., Sharma, P. and Hingorani, A. D. (2005). Homocysteine and stroke: evidence on a causal link from Mendelian randomisation. *Lancet* **365**, 224-232.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Tosteson, T. D. and Karagas, M. R. (2004). Nonlinear and nonparametric regression and instrumental variables. *J. Amer. Statist. Assoc.* **99**, 736-750.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. CRC Press.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under alternative. *Biometrika* **64**, 247-254.

Davies, R. B. (1987). Hypothesis-testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33-43.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statist. Meth. Medical Res.* **16**, 309-330.

Dunn, G. (1999). The problem of measurement error in modelling the effect of compliance in a randomized trial. *Statist. Medicine* **18**, 2863-2877.

Fischer, K. and Goetghebeur, E. (2004). Structural mean effects of noncompliance: Estimating interaction with baseline prognosis and selection effects. *J. Amer. Statist. Assoc.* **99**, 918-928.

Goetghebeur, E. and Lapp, K. (1997). The effect of treatment compliance in a placebo-controlled trial: regression with unpaired data. *Appl. Statist.* **46**, 351-364.

Goetghebeur, E. and Vansteelandt, S. (2005). Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statist. Meth. Medical Res.* **14**, 397-415.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *Int. J. Epidem.* **29**, 722-729.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statist. Sci.* **20**, 111-129.

Hansen, B. E. (1992). The likelihood ratio test under nonstandard conditions - testing the Markov switching model of GNP. *J. Appl. Econometrics* **7**, S61-S82.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference - An epidemiologist's dream? *Epidemiology* **17**, 360-372.

Joffe, M. M., Small, D. and Hsu, C. Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statist. Sci.* **22**, 74-97.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. and Smith, G. D. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statist. Medicine* **27**, 1133-1163.

Leigh, J. P. and Schembri, M. (2004). Instrumental variables technique: cigarette price provided better estimate of effects of smoking on SF-12. *J. Clin. Epidem.* **57**, 284-293.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V. and Klungel, O. H. (2006). Instrumental variables application and limitations. *Epidemiology* **17**, 260-267.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23**, 2379-2412.

Robins, J. M. (2005). Optimal structural nested models for optimal sequential decisions. In Proceedings of the Second Seattle Symposium in Biostatistics (Edited by D. Y. Lin and P. J. Haegerty), pages 189-326. Springer Verlag, New York.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6**, 34-58.

Ten Have, T. R., Elliott, M. R., Joffe, M., Zanutto, E. and Datto, C. (2004). Causal models for randomized physician encouragement trials in treating primary care depression. *J. Amer. Statist. Assoc.* **99**, 16-25.

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A. and Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics* **63**, 926-934.

Urquhart, J. and De Klerk, E. (1998). Contending paradigms for the interpretation of data on patient compliance with therapeutic drug regimens. *Statist. Medicine* **17**, 251-267.

Van Damme, L. Ramjee, G., Alary, M, Vuylsteke, B., Chandeying, V., Rees, H., Sirivongrangson, P., Mukenge-Tshibaka, L., Ettiègne-Traoré, V., Uaheowitchai, C., Karim, S. S., Mâsse, B., Perriëns, J., Laga, M. and COL-1492 Study Group. (2002). Effectiveness of COL-1492, a nonoxynol-9 vaginal gel, on HIV-1 transmission in female sex workers: a randomised controlled trial. *The Lancet* **360**, 971-977.

Vansteelandt, S. and Goetghebeur, E. (2004). Using potential outcomes as predictors of treatment activity via strong structural mean models. *Statist. Sinica* **14**, 907-925.

Vansteelandt, S. and Goetghebeur, E. (2005). Sense and sensitivity when correcting for observed exposures in randomized clinical trials. *Statist. Medicine* **24**, 191-210.

Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9), 9000 Gent, Belgium.

E-mail: Stijn.Vansteelandt@ugent.be

Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9), 9000 Gent, Belgium.

E-mail: Manoochehr.Babanezhad@ugent.be

Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9), 9000 Gent, Belgium.

E-mail: Els.Goetghebeur@ugent.be

(Received October 2007; accepted September 2008)