Harvard University

Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 179

Instrumental Variable Estimation in a Survival Context

Eric J. Tchetgen Tchetgen* Stefan Walter† Stijn Vansteelandt‡

Torben Martinussen** Maria Glymour††

*Harvard University, etchetge@hsph.harvard.edu

†University of California - San Francisco, swalter@psg.ucsf.edu

‡University of Ghent, stijn.vansteelandt@ugent.be

**University of Copenhagen, tma@sund.ku.dk

††University of California - San Francisco, mglymour@psg.ucsf.edu

http://biostats.bepress.com/harvardbiostat/paper179

Instrumental variable estimation in a survival context

Eric J. Tchetgen Tchetgen[1], Stefan Walter[2],

Stijn Vansteelandt[3], Torben Martinussen[4], Maria Glymour[2]

[1]Departments of Biostatistics and Epidemiology,

Harvard University;

[2]Department of Epidemiology & Biostatistics,

University of California, San Francisco;

[3]Department Applied Mathematics, Computer Science and Statistics,

Ghent University;

[4]Department of Biostatistics, University of Copenhagen.

Correspondence: Eric J. Tchetgen Tchetgen, Departments of Biostatistics and Epidemiology, Harvard School of Public Health 677 Huntington Avenue, Boston, MA 02115.

Abstract

Bias due to unobserved confounding can seldom be ruled out with certainty when using non-experimental data to draw inferences about causal effects. The instrumental variable (IV) design offers under certain assumptions, the opportunity to tame confounding bias, without directly observing all confounders. The IV approach is very well developed in the context of linear regression but also for certain generalized linear models with non-linear link function. However, IV methods are not as well developed for censored survival outcomes. In this paper, the authors develop the instrumental variable approach in a survival context, under an additive hazards model, and they describe two simple strategies for estimating causal effects for this context. The first strategy entails a straightforward two stage regression approach analogous to two stage least squares commonly used for IV estimation in linear models, whereby the fitted value from a first stage regression of the exposure on the IV is entered in place of the exposure in the second stage hazard model to recover the causal effect in view. The second strategy is a so-called control function approach, which entails adding as confounding control covariate to the additive hazards regression model for the exposure effect, the residual from a first stage regression of the exposure on the IV. Formal conditions are given justifying each strategy, and the methods are illustrated in a novel application to a Mendelian randomization study of the causal association between diabetes status at start of follow-up and mortality using data from the Health and Retirement Study. It is also shown that analogous estimation strategies can also be used under a proportional hazards model specification provided the outcome is rare over the entire follow-up period under consideration.

Confounding bias remains to date a major potential source of bias in observational studies conducted in epidemiology. In recent years, epidemiologists have slowly expanded their analytic toolbox to account for unobserved confounding, by adopting the instrumental variable design, an approach for analyzing non-experimental studies, historically favored by economists and other social scientists.[1,2] The IV design entails selecting an observed pre-exposure variable which is known to only be associated with the outcome to the extent that the latter is causally affected by the exposure of interest, and the IV is directly related to the exposure. Thus, the IV is selected such that it does not directly affect the outcome (known as the exclusion restriction), and, although correlated with the exposure in view, it is independent of confounders of the exposure-outcome relation. A valid IV may be hard to find in practice, but when successfully selected to meet these criteria, an IV can sometimes be used to account for unobserved confounding bias.[1–4]

Instrumental variable methods are particularly well developed in the context of linear models,[4,5] and similar methods are likewise well developed for regression analysis with certain nonlinear link functions (e.g. log, logit, probit).[5–7] Right censored survival outcomes are of common occurrence in epidemiologic practice, and regression analysis on the hazard scale is typically used. The Cox proportional hazards model is perhaps the most popular regression framework for survival data,[8] and recently, Aalen's additive hazards model has also gained popularity as an alternative framework.[9] An important appeal of the additive hazards framework is that, unlike proportional hazards models, the class of additive hazards models is closed under marginalization over regressors. Specifically, collapsing over a continuous regressor in an additive hazards model can under fairly reasonable assumptions, produce a marginal additive hazards model. This property is known not to generally hold for the Cox proportional hazards model, except perhaps for a disease that remains rare over the follow-up period of the study. In this paper, the authors exploit the collapsibility of additive hazards to develop two straightforward strategies for IV estimation with a

censored survival outcome. The first strategy entails a straightforward two stage regression approach analogous to two stage least squares commonly used for IV estimation in linear models.[5] The current setting differs from that of standard two stage least squares in that, here, the fitted value from the first stage regression of the exposure on the IV is entered in place of the exposure in a second stage additive hazards model, instead of a standard linear regression model to recover the causal effect in view. The second proposed strategy is a so-called control function approach,[5] which entails adding to the additive hazards regression model for the exposure effect, the residual from a first stage regression of the exposure on the IV also known as a control function. Formal conditions are given justifying each strategy, and the methods are illustrated in a novel application to a Mendelian randomization study of the causal association between diabetes diagnosis and mortality using data from the Health and Retirement Study.[10] Finally, it is shown that when the disease outcome is rare, in the sense that its cumulative incidence remains low over the follow-up period under consideration, then analogous two stage regression and control function strategies may be used under a Cox proportional hazards model.

Two-stage regression approach

Suppose one has observed independent and identically distributed data on (T^*, A, Z) for a sample of size n, where A is an exposure of interest, Z is the candidate IV, T is the time to event outcome and $T^* = \min(T, Y)$ with Y the potential censoring time. Unless stated otherwise, we assume that Y is independent of (T, A) conditional on Z. To introduce the causal model of interest, suppose that the effect of the IV Z on the outcome T is unconfounded, however, we will suppose that the effect of A on T remains confounded whether one conditions on Z or not. Let U denote the unobserved confounder of the effect of A on T, so that conditioning on U recovers the causal effect of A on T. To further ground ideas, we will suppose the data are generated under the Aalen

additive hazards model

$$h\left(t|A,U,Z\right)=b_0\left(t\right)+b_a\left(t\right)A+b_u(U,t) \tag{1}$$

where $h\left(t|A,U,Z\right)$ is the hazard function of T evaluated at t, conditional on A,U and Z, and the functions $\left(b_0\left(\cdot\right),b_a\left(\cdot\right),b_u\left(\cdot,\cdot\right)\right)$ are unrestricted. The model states that conditional on U, the effect of A on T encoded on the additive hazards scale is linear in A for each t, however, the effect size $b_a\left(t\right)$ may vary with t. The model is quite flexible in the unobserved confounder association with the outcome $b_u\left(\cdot,\cdot\right)$, which is allowed to remain unrestricted at each time point t and across time points. In the Mendelian randomization study we will consider below, A represents diabetes status measured at start of follow-up coded as a binary variable (1 if diabetic and 0 otherwise), T is time to death, and Z is defined as a genetic risk score for diabetes, which combines several genetic variants previously established to predict diabetes risk. The approach is described in additional detail below. More generally, A could be continuous, such as say body mass index (BMI), in which case the above Aalen model assumes linearity of the conditional hazards difference at each t. The null hypothesis of no causal effect of A (BMI or diabetes status) on T (mortality) is encoded by $b_a\left(t\right)=0$ for all t. An important sub-model to consider is the constant hazards difference model obtained by setting

$$b_a\left(t\right)=b_a \tag{2}$$

where b_a is an unknown constant. Crucially, the model assumes no interaction between A and U. Collapsibility over U makes $b_a\left(t\right)$ interpretable as a marginal causal hazards difference (upon standardization with respect to the population distribution of U), which is an appealing feature of the model since it would indeed be uncomfortable to come up with an effect size that is only interpretable conditional on the unobserved U. The baseline hazard function $b_0\left(t\right)$ is a priori unrestricted. Finally, the right hand side of equation (1) does not depend on Z, even though the

left hand side of the equation conditions on Z, so that the model encodes explicitly the assumption that Z and T are conditionally independent given (U, A), i.e. the exclusion restriction condition.[11] In practice, additional pre-exposure covariates X (e.g. age, gender, education, ...etc) may be observed, and one may wish to account for such covariates in an IV analysis. In order to ease the presentation, we will first describe the proposed methodology without covariates, so as to more easily focus on key ideas, however, later, we will describe how the methods can be modified to incorporate such covariates.

Until otherwise stated, suppose that A is continuous, e.g. body mass index (BMI). Then, in addition to equation (1) , one may specify a standard linear model for A :

$$A = c_0 + c_z Z + \Delta, \text{ where } \Delta \text{ is mean zero residual error independent of } Z \tag{3}$$

We do not further specify the distribution of Δ, and we allow for U and Δ to be conditionally associated given Z, i.e. $COV(\Delta,U|Z) \neq 0$, inducing confounding by U. Throughout, we will assume that $c_z \neq 0$ so that there is a non-null association between Z and A. However, c_z may not have a causal interpretation, in the event of unobserved confounding of the effect of Z on A. We will however assume that any unobserved common cause of Z and A must be independent of U. Let $M = m(Z) = E(A|Z) = c_0 + c_z Z$.

The proposed two-stage approach is based on the following result, which provides an analytic expression for the conditional hazard model $\widetilde{h}(t|Z)$, of T evaluated at t, conditional on Z, under model restrictions (1) and (3) .

RESULT 1: Under assumptions (1) *and* (3) , *and assuming that* U *is independent of* Z, *one has that*

$$\widetilde{h}(t|Z) = \widetilde{b}_0(t) + b_a(t) M \tag{4}$$

for $\widetilde{b}_0(t)$ *a baseline hazard function.*

Result 1 states that under assumptions (1), (3) and the assumption of independence of U and Z, the hazard function of T at t conditional on Z is linear in $M = m(Z)$. Suppose for a moment that, contrary to fact, $M = E(A|Z)$ were observed, thus rendering model (4) a standard Aalen additive hazards model with covariate M. Inference about $B(t) = \left(\widetilde{b}_0(t), b_a(t)\right)^T$ for such a model has been well studied and can be obtained using the R package TIMEREG.[12] Let $B^*(t)$ denote Aalen's least squares estimator of $B(t)$ under model (4) which we provide in the Appendix for completeness and which can be computed using TIMEREG. The proposed two-stage approach entails in the first stage, estimating M with $\widehat{M}$, the fitted value of the ordinary least squares regression of A on Z, i.e. $\widehat{M} = \widehat{c}_0 + \widehat{c}_z Z$, where $(\widehat{c}_0, \widehat{c}_z)$ is the ordinary least squares estimator of (c_0, c_z). The second stage then involves obtaining Aalen's least squares estimator $\widehat{B}^*(t)$ of $B(t)$, defined similarly to $B^*(t)$, with $\widehat{M}$ substituted for M. Estimation under assumption (2) is also easily accommodated in TIMEREG.[12,13] However, some care is generally required to obtain valid inferences about the regression parameter $B(t)$, because one must acknowledge in computing standard errors and confidence intervals, the additional uncertainty due to the first stage estimation of M. Standard errors obtained in R will fail to appropriately account for this extra variability and thus will tend to understate uncertainty. A simple remedy is to perform either the jackknife or the nonparametric bootstrap, either of which will produce more accurate estimates of standard errors.[14] For completeness, we also provide in the Appendix, an analytic expression of a consistent estimator of the corrected standard error of $\widehat{B}^*(t)$.

Occasionally, the first stage OLS estimate $(\widehat{c}_0, \widehat{c}_z)$ may be obtained from an independent sample to that used for the second stage estimation of (4). This type of IV design, known as a split IV design,[15] has become quite common in Mendelian randomization studies in which genetic variants define the IV, so that the first stage regression may be obtained in a sample that only includes data

on the genetic variants and the phenotype defining the exposure of interest, but no information on the outcome. The approach then entails obtaining from an independent sample, an estimate of the causal hazards difference, by combining $(\widehat{c}_0, \widehat{c}_z)$ with genotype information in the second sample to construct $\widehat{M}$, which is then used to fit the hazard model. Note that information on phenotype A need not be available in the second sample in which the outcome is available. Also, the split sample IV design is known in linear models to confer robustness to the weak instrumental variable problem, in that, in the event of a (nearly) null relation between the IV and the exposure of interest, the approach is guaranteed to deliver an effect estimate that is biased towards the null.[15] A similar robustness property is expected to hold for the above two-stage linear-additive hazards regression analysis. An additional appeal of the split IV design is that, uncertainty in the first stage estimation can essentially be ignored, if as in the case in the empirical example presented below, the sample size for the first stage is considerably larger than that of the second stage.

Control function approach

In this section, we will consider an alternative approach to two-stage regression for estimation. Consider the sub-model of (1) which further specifies

$$b_u(U,t) = \rho_0(t)\,\Delta + \varepsilon(t) \tag{5}$$

where Δ is the residual error defined in (3), $\varepsilon(t)$ is a random error independent of (Δ, Z), which may not have mean zero, and the unknown function $\rho_0(t)$ is a priori unrestricted. The model makes explicit the dependence between Δ and U, encoded in a nonnull value of $\rho_0(t) \neq 0$, and induces confounding bias. The residual error $\varepsilon(t)$ introduces additional variability to ensure that the relation between U and Δ is not assumed deterministic, however, beside for independence

with (Δ, Z), the distribution of $\varepsilon(t)$ is otherwise unrestricted (up to certain regularity conditions provided in the Appendix). Let $\overline{h}(t|A,Z)$ denote the observed hazard function of T given (A,Z), evaluated at t. Then, we have the following result:

RESULT 2: Under assumptions (1), (3) *and* (5) *one has that*

$$\overline{h}(t|A,Z) = \overline{b}_0(t) + b_a(t)A + \rho_0(t)\Delta \tag{6}$$

for $\overline{b}_0(t)$ *a baseline hazard function.*

Result 2 provides an explicit parametrization of the hazard function of T conditional on A and Z, under assumptions (1), (3) and (5) . This result shows that an appropriate model specification of $\overline{h}(t|A,Z)$ is essentially obtained upon replacing $b_u(U,t)$ with $\rho_0(t)\Delta$, and by allowing the baseline hazard function $\overline{b}_0(t)$ to differ from $b_0(t)$. Intuitively, the residual Δ captures any variation due to unobserved correlates of A, not accounted for in M, that may also be associated with T, and thus serves as a proxy measure of unobserved confounding. For this reason, "$\rho_0(t)\Delta$" is referred to as a control function, akin to the control function sometimes used in IV estimation of linear and nonlinear models.[5] For estimation, we propose to use $\widehat{\Delta} = A - \widehat{M}$ as an estimate of the unobserved residual Δ which we use to fit an additive hazards model, with regressors $\left(A, \widehat{\Delta}\right)$ under (8) . Such an additive hazards model can be estimated using the methods and statistical software described in the previous section and the nonparametric bootstrap equally applies as an approach to appropriately account for uncertainty due to in-sample estimation of $\widehat{\Delta}$. In situations where a split sample IV design is adopted, the first sample estimation can essentially be ignored, when the size of the sample is much greater than that used in the second stage. Furthermore, as in the previous section, the first stage sample does not need to include outcome data, however, unlike

in the previous section, the second stage sample must have data collected on the IV, the exposure and the outcome for all observations. The model is easily modified to incorporate heterogeneity with respect to Z in the degree of confounding bias, by simply extending model (5) with the additional covariate ΔZ, such that

$$b_u(U,t) = (\rho_0(t) + \rho_1(t) Z)\Delta + \varepsilon(t),$$

where $\varepsilon(t)$ is a random error independent of (Δ, Z). This in turn yields the conditional hazard model

$$\overline{h}(t|A,Z) = \overline{b}_0(t) + b_a(t) A + \rho_0(t)\Delta + \rho_1(t)\Delta Z$$

such that the null hypothesis of no confounding bias, now corresponds to the joint null hypothesis

$$\rho_0(t) = \rho_1(t) = 0 \text{ for all } t.$$

Finally, one may note that it suffices for the control function approach, that censoring is independent of T conditional on (A, Z), a somewhat weaker independent censoring assumption than needed for the stage regression approach.

Binary exposure

The control function approach can also be used in the context of a binary or more general discrete exposure. In the simple case of a binary exposure, the methods described in the previous section apply upon estimating M with an appropriate regression model for binary data, e.g. $\text{logit}M = \text{logit}m(Z) = \text{logit}\Pr(A = 1|Z) = c_0 + c_z Z$. The approach can be motivated under a modified set of assumptions to account for the binary nature of the treatment. Specifically, suppose

that

$$b_u(U,t) = E\left\{b_u(U,t)|A,Z\right\} + \varepsilon(t) \tag{7}$$

where $\varepsilon(t)$ is an independent error, and both A and Z are binary. The assumption is best understood if $b_u(U,t) = b_u^*(t)U$ is linear in U, in which case the assumption amounts to a location shift model for the density of U conditional on A and Z, i.e. (A,Z) are associated with U only on the mean scale. The assumption is certain to hold say if U were normal with constant variance, but the model also allows for a more flexible distribution. Then, we have the following result.

RESULT 3: Assuming Z is a valid binary IV and both assumptions (1) *and* (5) *hold, one has that*

$$\overline{h}(t|A,Z) = \widetilde{b}_0(t) + b_a(t)A + \left\{\rho_0(t) + \rho_1(t)Z\right\}\Delta, \tag{8}$$

for $\widetilde{b}_0(t)$ a baseline hazard function, and

$$\Delta = A - \Pr(A=1|Z).$$

The model of equation (8) is again an Aalen additive hazards model which can be estimated in a manner analogous to the control function approach described in the previous section for a continuous exposure. Although the result assumes binary Z, we may nonetheless use model (8) with continuous Z, under an additional assumption that $E\left\{b_u(U,s)|A,Z\right\}$ is linear in Z.

Sensitivity analysis

We briefly describe a sensitivity analysis technique that may be used to assess the extent to which a violation of the exclusion restriction might impact inference. The approach is motivated by noting

that a violation of the assumption can be encoded by modifying the Aalen model (1) as followed

$$h\left(t|A,U,Z\right)=b_0\left(t\right)+b_a\left(t\right)A+b_z\left(t\right)Z+b_u(U,t)$$

whereby $b_z\left(t\right)$ encodes the causal effect of Z on T on the additive hazards scale, so that $b_z\left(t\right)=0$ for all t recovers model $\left(1\right)$, while $b_z\left(t\right)\neq 0$ for some t implies violation of the exclusion restriction. The function $b_z\left(t\right)$ cannot be identified from the model, and entails a sensitivity parameter. To fix ideas, it is convenient to take $b_z\left(t\right)$=b_z independent of t. A sensitivity analysis is then obtained by setting b_z to a specific value $b_z^*\neq 0$ and to obtain inferences under the assumed departure from the exclusion restriction. This can be achieved by simply including the term "b_z^*Z" as a known offset to the additive hazards model which may then be estimated via either two stage regression or the control function approach. A sensitivity analysis is obtained by varying the assumed value b_z^* and subsequently obtaining inferences about the effect of A for each assumed value.

Covariate Adjustment

Suppose that one has collected a vector of pre-exposure confounders X of the effects of (Z,A) on Y. In this section, we show how the proposed IV methods are easily modified to incorporate X. Formal justification for the approach is relegated to the Appendix. The first stage regression model can be formulated as followed to make explicit the dependence on X,

$$A=c_0+c_zZ+c_x^TX+\Delta \tag{9}$$

where c_x encodes the regression association of X with A conditional on Z, and Δ is assumed to be independent of Z given X. In the Appendix, we show that under certain assumptions, the second

stage regression obtained in Result 1 can be modified to account for X using the more general Aalen model

$$\widetilde{h}(t|Z,X) = \widetilde{b}_0(t) + b_a(t)\,M + b_x^T(t)\,X \tag{10}$$

with $M = m(X,Z) = c_0 + c_z Z + c_x^T X$ and $b_x^T(t)$ encoding the effect of X on the hazard of T at t, conditional on M on the additive hazards scale. Two stage estimation using the above regression models can be implemented in R using the same procedure as previously described without additional difficulty. The control function approach can also be modified along the same lines, by fitting the regression model

$$\overline{h}(t|A,Z) = \overline{b}_0(t) + b_a(t)\,A + b_x^T(t)\,X + \rho_0(t)\,\Delta \tag{11}$$

instead of (8). Formal justification for this modification can be obtained for continuous A by replacing assumption (5) with

$$b_u(U,X,t) = \rho(t)\,\Delta + b_x^T(t)\,X + \varepsilon(t) \tag{12}$$

where $\varepsilon(t)$ is a random error independent of (Δ, Z, X). We briefly note that previous state assumptions about censoring will need to be modified by also conditioning on X.

IV for Cox proportional hazards model

Suppose that the underlying failure time outcome follows the proportional hazards model

$$h(t|A,U,Z) = h_0(t)\exp(b_a A + b_u(U,t)) \tag{13}$$

where b_a is the log-hazards ratio encoding the effect of exposure, and $b_u(U,t)$ denotes the effect of an

unmeasured confounder U at time t, so that $COV(\Delta,U|Z) \neq 0$, where Δ is given by the exposure model (3), and Z is independent of U. Equation (13) also encodes the exclusion restriction by defining the left hand-side conditional on (A, Z, U), while the right-hand side does not depend on Z. In the following, we focus on the fairly common setting, of a rare outcome, which we encode by near unity conditional survival curve throughout follow-up:

$$\begin{aligned} S(t|A,U,Z) &= P(T \geq t|A,U,Z) \\ &\approx 1 \end{aligned}$$

for all t during follow-up. Let $f(T|A,U,Z)$ denote the density of T given (A,U,Z), and recall that the conditional hazard function is defined as $h(t|A,U,Z) = f(t|A,U,Z)/S(t|A,U,Z)$. Then, we have that:

$$\begin{aligned} f(t|A,U,Z) &= h(t|A,U,Z)\, S(t|A,U,Z) \\ &\approx h(t|A,U,Z) \end{aligned}$$

Likewise, by the rare disease assumption,

$$\begin{aligned} f(t|Z) &= E[f(t|A,U,Z)|Z] \\ &= h(t|Z)\, S(t|Z) \\ &\approx h(t|Z) \end{aligned}$$

therefore under the exposure model (3) , one obtains:

$$\begin{aligned} h(t|Z) &\approx E\left[h(t|A,U,Z)\,|Z\right] \\ &= E\left[h_0(t)\exp\left(b_a A + b_u(U,t)\right)|Z\right] \\ &= E\left[h_0(t)\exp\left(b_a M + b_a\Delta + b_u(U,t)\right)|Z\right] \text{ by equation (3)} \\ &= h_0^*(t)\exp\left(b_a M\right) \text{ by independence of } (U,\Delta) \text{ with } Z. \end{aligned} \tag{14}$$

where $h_0^*(t) = E\left[\exp\left(b_a\Delta + b_u(U,t)\right)\right] h_0(t)$. The above derivation formally justifies the use of two stage regression analysis, analogous to the two stage procedure previously described for the Aalen model, whereby the fitted value $\widehat{M}$ obtained via standard OLS, is used as a regressor in place of M in the standard Cox proportional hazards regression defined with equation (14) , which can then be estimated via standard maximum partial likelihood. A similar argument provides justification for the use of a control function approach for Cox regression under rare disease. In this vein, suppose that $b_u(U,t)$ follows equation (5) , so that

$$\begin{aligned} h(t|A,Z) &\approx E\left[h(t|A,U,Z)\,|A,Z\right] \\ &= E\left[h_0(t)\exp\left(b_a A + b_u(U,t)\right)|A,Z\right] \\ &= E\left[h_0(t)\exp\left(b_a A + \rho(t)\,\Delta + \varepsilon(t)\right)|A,Z\right] \\ &= h_0^{**}(t)\exp\left(b_a A + \rho(t)\,\Delta\right) \end{aligned}$$

where $h_0^{**}(t) = h_0(t)\,E\left[\exp\left(\varepsilon(t)\right)\right]$. One immediately recovers a standard Cox proportional hazards model by letting $\rho(t) = \rho$, so that (ρ, b_a) can be estimated via standard maximum partial likelihood, upon substituting $\widehat{\Delta}$ for Δ. For inference, we recommend either the jackknife or the nonparametric bootstrap. It is also fairly straightforward to incorporate covariate adjustment in an analogous

manner to the additive hazards setting, details are omitted.

Intuition for the above results for Cox regression can be gained upon noting that a Cox regression analysis is essentially a loglinear regression for the risk of the outcome performed repeatedly over the follow-up period, among persons that remain at risk for the outcome. For a rare outcome, the joint distribution of the instrumental variable, the unobserved confounder and the exposure in view, is nearly stable across risk sets, so that the IV assumptions are ensured to hold within each risk set, and the exclusion restriction is satisfied within each risk set. The framework then essentially reduces to IV for loglinear regression analysis, for which two stage regression has previously been shown to apply under assumptions analogous to those considered here.[16]

Empirical illustration.

The prevalence of type 2 diabetes mellitus (T2DM) is increasing across all age groups in the United States possibly as a consequence of the obesity epidemic.[17,18] In addition, no decline has been observed in the excess mortality among persons suffering from T2DM relative to persons without T2DM.[19] Obtaining an unbiased estimate of the mortality risk associated with T2DM is key to predicting the future health burden in the population and to evaluate the effectiveness of possible public health interventions.

In order to illustrate the proposed instrumental variable approach for survival analysis, we used data from the Health and Retirement Survey (HRS), a public survey with repeated assessments every 2 years initiated in 1992, to investigate the mortality risk associated with T2DM being instrumented by externally validated, genetic predictors of T2DM. HRS is a well-documented nationally representative sample of individuals 50 years of age or older and their spouses.[10] Genotype data was collected on a subset of HRS respondents in 2006 and 2008. Genotyping was completed on the Illumina Omni-2.5 chip platform and imputed using the 1000G phase 1 reference panel and filed with

the Database for Genotypes and Phenotypes (dbGaP, study accession number: phs000428.v1.p1) in April 2012. Exact information on the process performed for quality control is available via HRS and dbGaP21.[20] From the 12,123 HRS participants for whom genotype data was available, we restricted the sample to 8,446 Non-Hispanic Whites with valid self-reported diabetes status at baseline. For deaths occurring between 1998-2008 the date of death was confirmed through the National Death Index. Mortality status for 2008-2010 was obtained by interviewing surviving relatives. Follow-up was determined as years since sampling of DNA (2006 or 2008 respectively). The current analysis was exempt by the Institutional Review Board at the Harvard School of Public Health.

We used the control function approach discussed previously to estimate the relationship between diabetes status (coded 1 for diabetic and 0 otherwise) on Mortality. As genetic instrument we used 39 independent single nucleotide polymorphisms previously established to be significantly associated with T2DM.[21] In addition we created a polygenic risk score calculated on the basis of these 39 established genetic variants and their "external" effect size based on the meta-analysis of 34,840 T2DM cases and 114,981 controls from Morris et al.[21] The polygenic risk score was calculated by multiplying each log odds ratio coefficient by the corresponding number of risk alleles and summing across the products.

For comparison, we first performed an observational analysis, which entailed fitting a standard Aalen additive hazards model for T2DM. Next, we implemented the proposed control function instrumental variable approach, which is appropriate for binary endogenous variable. In addition to the first stage residual, we also adjusted for possible effect heterogeneity of the degree of selection bias by including an interaction between the first stage residual and the first stage risk score. All regression models further adjusted for: age, sex and the top four genomewide principal components to account for possible population stratification. Inferences were based on 5000 nonparametric

Table 1. Observational and IV analysis of HRS data to estimate the effects of baseline diabetes status on Mortality under an Aalen additive hazards model.

	Beta*	95% CI
Observational Analysis**		
Diabetes Status (Yes vs. No)	0.031	(0.027,0.035)
IV Survival Models**		
Diabetes Status (Yes vs. No)	0.082	(0.075,0.089)
First Stage Residual	-0.023	(-0.028,-0.017)
First Stage Residual by Estimated Diabetes Risk status Interaction	-0.024	(-0.052,0.001)

*risk difference of mortality rate per additional person years.

**All models adjust for age , gender and first four genome wide principal components to control for possible population stratification

bootstrap samples.

Participants were on average 68.5 years old (Standard Deviation (SD): 10.4) at baseline and 1,891 self-reported diabetics (22.4%). The average follow-up time was 4.10 years (SD = 1.10). In total we observed 644 deaths over 34035 person-years. The 39 SNPs jointly included in a first stage logistic regression model to predict diabetes status explained 3.4 % of the variation in diabetes in the study sample (Nagelkerke R^2).

Table 1 shows results from both observational and IV analyses. In the observational analysis, being diabetic was associated with an increase in the hazard rate of beta=0.03 (95%-Confidence Interval (95%-CI): 0.025 – 0.035) per person-year. This means approximately 3 additional deaths occured for each year of follow-up for every hundred diabetic persons, than for every hundred diabetic-free persons. The genetic IV approach produced a notably larger effect associated with T2DM, with a T2DM increase in the mortality rate of beta = 0.08 (95%CI: 0.075 – 0.090) per person-year, nearly three times the rate estimated by the observational additive hazards model. We obtained further evidence of significant negative confounding bias reflected by an association between the first stage residual and mortality rate, beta=-0.023 (95%CI:-0.028,-0.017), as well as marginal evidence of confounding bias heterogeneity beta=-0.024 (95%CI:-0.052,0.001).

Using the external genetic risk score as a single IV gave essentially the same results, which are

not reported.

Closing remarks

In this paper, the authors describe an instrumental variable framework of inference about causal effects of a continuous or binary exposure on a right censored failure time under an additive hazards regression model. Two strategies for estimation are described, a two stage regression approach for continuous exposure, similar to two stage least squares commonly used in linear regression IV settings, and a control function approach that equally applies for binary and continuous exposure. A notable distinction from the classical linear regression setting is that, in the current context unlike in the latter, the first stage regression model for the exposure must be correctly specified in order for the proposed methods to generally remain valid. An important exception holds for the two stage regression approach under the null hypothesis of no causal effect of A, in which case the first stage regression can be mis-specified without altering the nominal type 1 error rate of the corresponding test statistic. A variety of extensions of the methods are also discussed, including a simple strategy for covariate adjustment, and a sensitivity analysis approach for assessing the extent to which a violation of the exclusion restriction assumption could impact inference. In certain settings, both the exposure of interest and the instrumental variable may be time-updated, in which case, the methods described above may not directly apply. Instrumental variable estimation of the joint effects of time-updated exposures present several challenges beyond the scope of the current manuscript, and to the best of our knowledge, methods are currently not available to address these challenges for either the additive hazards model, or the Cox proportional hazards model studied herein, although IV methods for joint effects are available under alternative modeling assumptions, such as the semiparametric accelerated failure time model.[22] However, available methods for the semiparametric accelerated failure time model, whether for point or time varying exposures, may be

computationally challenging in practice, because they sometimes require censoring a set of subjects (defined according to the causal parameter) whose event time was observed for unbiasedness. In the point exposure scenario, alternative IV methods have been proposed under a structural proportional hazards model,[23,24] which do not require artificial censoring and which do not rely on a rare disease assumption, however, in contrast with the methods we have developed herein for a proportional hazards model, earlier proposals were limited to either binary instrumental variable or binary exposure variable.[23,24] In a more recent proposal, MacKenzie and colleagues use an instrumental variable to estimate a Cox proportional hazards model subject to additive unobserved confounding.[25] Specifically, they specify a so-called additive multiplicative hazards model,[26,27]

$$h\left(t|A,U,Z\right)=h_0\left(t\right)\exp\left(b_aA\right)+b_u(U,t)) \tag{15}$$

with the key restriction

$$E\left\{b_u(U,t)|T(a)\geq t\right\}=0 \tag{16}$$

where $T(a)$ is the potential outcome of T under treatment a. The model combines features of models (1) and (13) since it incorporates both an additive effect of U and a multiplicative effect of A. The restriction (16) ensures that the marginal hazard model of $T(a)$ follows a Cox proportional hazards model. The authors note that this restriction is generally satisfied if $b_u(U,t)$ can be written $d(t)U+\ln \mathrm{mgf}_U\left\{d(t)\right\}$, where mgf_U stands for the moment generating function of U. Under such unobserved confounding, the authors show that a valid IV can be used to recover a consistent estimator of b_a. Although interesting this model may be more contrived than initially meets the eye, because supposing that U were observed, assumptions (15) and (16) imply that the conditional hazard function of $T(a)$ at time t given U does not only depend on the value of U, but further depends on the underlying distribution of the unobserved confounder. For instance, if U were

normally distributed $N(\mu, \sigma_U^2)$, we would have $\ln \text{mgf}_U \{d(t)\} = \mu d(t) + \sigma_U^2 d(t)^2/2$. The model would then imply that the density of T conditional on (A, U) is made to depend explicitly on the parameters (μ, σ_U^2) of the density of the covariate U. Such a parametrization is nonstandard and somewhat artificial in the sense that it would not naturally be entertained by an analyst if U were in fact observed.

Finally, the control function approach described in this paper may also be seen as an extension of the two stage residual inclusion (2SRI) approach of Terza et al.[28] In order to ease a comparison between the two methods, it is helpful to restate the key assumption underlying 2SRI within our context using our notation. This is easiest achieved by simply replacing equation (5) with the more restrictive model:

$$b_u(U, t) = \rho_0(t)\,\Delta \tag{17}$$

obtained by setting $\varepsilon(t) \equiv 0$ for all t, thus essentially assuming the relationship between U and Δ is deterministic. This assumption may be unrealistic in most health related applications, since it essentially rules out the existence of any other (unobserved) cause of A, that like Z may not be directly related to the outcome, a possibility that cannot generally be ruled out with certainty. By allowing for $\varepsilon(t)$ in equation (5) avoids this type of restriction. It is also notable that assumption (17) may be overly restrictive for binary (or discrete) A, since the distribution of the residual Δ is completely determined by the mean $M(Z)$ of A, and therefore the IV assumption that Z is independent of U is not compatible with the model. In this paper, we have provided an alternative formulation of the control function approach for binary A, which circumvents this difficulty.

Acknowledgement 1 *Eric Tchetgen Tchetgen's work is funded by NIH grant R01AI104459. Torben Martinussen's work is part of the Dynamical Systems Interdisciplinary Network, University of Copenhagen.*

APPENDIX

Aalen's least squares estimator: Consider the Aalen additive hazards model:

$$h\left(t|C\right) = {B'}^{T}\left(s\right)C$$

with C a vector of covariates with first entry equal to 1, to encode the baseline hazard function. Let $R(t)$ denote the at risk process at time t. The least squares estimator of $B'\left(t\right)$ due to Aalen can be written:

$$\widehat{B}'\left(t\right) = \left\{\sum_i R_i(t)C_iC_i^T\right\}^{-1}\sum_i C_i dN_i(t)$$

where G^{-1} denotes the inverse of the matrix G, and $dN_i(t)$ is the counting process associated with T_i, at time t.

A consistent estimator of the asymptotic variance:We use notation $V = c_0 + c_z Z$ and let $M(t)$ be the martingale with respect to the filtration where V is also observed. Let also $c = (c_0, c_z)$, we then have $n^{1/2}(\hat{c} - c) = n^{-1/2}\sum_i \epsilon_i^c + o_p\left(1\right)$, where the ϵ_i^c's are zero-mean iid terms.

Let $\hat{V}_i = \hat{c}(1, Z_i)^T$ and $\hat{V}(t)$ be the $n \times 2$-matrix with ith row $(R_i(t), R_i(t)\hat{V}_i)$. The estimator of $B(t) = (\tilde{B}_0(t), B_a(t))^T$ is therefore

$$\hat{B}(t) = \int_0^t \{\hat{V}(t)^T\hat{V}(t)\}^{-1}\hat{V}(t)^T dN(t),$$

where $N(t)$ is the vector counting processes. We have

$$R_i(t)\hat{V}_i = R_i(t)\ V_i + R_i(t)(\hat{c} - c)(1, Z_i)^T,$$

and

$$n^{-1}\hat{V}(t)^T\hat{V}(t), \quad n^{-1}V(t)^TV(t)$$

are therefore asymptotically equivalent since $\hat{c}$ is a consistent estimator of c. Throughout, we freely assume without formally stating the necessary regularity conditions hold to establish asymptotic stochastic convergence results.[29] Then, we can further write

$$\hat{V}(t)^TdN(t) = V(t)^TdN(t) + \{0, (\hat{c}-c)\sum_i R_i(t)(1,Z_i)^TdN_i(t)\}^T.$$

But then

$$\begin{aligned} n^{1/2}\hat{B}(t) &= \int_0^t \{n^{-1}\hat{V}(t)^T\hat{V}(t)\}^{-1}n^{-1/2}\hat{V}(t)^TdN(t) \\ &= \int_0^t \{n^{-1}V(t)^TV(t)\}^{-1}n^{-1/2}V(t)^TdN(t) \\ &\quad + \int_0^t \{n^{-1}V(t)^TV(t)\}^{-1}\{0, n^{1/2}(\hat{c}-c)n^{-1}\sum_i R_i(t)(1,Z_i)^TdN_i(t)\}^T, \end{aligned}$$

and

$$\int_0^t \{n^{-1}V(t)^TV(t)\}^{-1}n^{-1/2}V(t)^TdN(t) = n^{1/2}B(t) + \int_0^t \{n^{-1}V(t)^TV(t)\}^{-1}n^{-1/2}V(t)^TdM(t).$$

Hence,

$$\begin{aligned} n^{1/2}\{\hat{B}(t) - B(t)\} &= \int_0^t \{n^{-1}V(t)^TV(t)\}^{-1}n^{-1/2}V(t)^TdM(t) \\ &\quad + \int_0^t \{n^{-1}V(t)^TV(t)\}^{-1}\{0, n^{1/2}(\hat{c}-c)n^{-1}\sum_i R_i(t)(1,Z_i)^TdN_i(t)\}^T \\ &\quad + o_P(1). \end{aligned}$$

This gives us that

$$n^{1/2}\{\hat{B}(t) - B(t)\} = n^{-1/2} \sum_i \epsilon_i^B(t),$$

where the $\epsilon_i^B(t)$ are iid zero-mean processes that can be estimated by

$$\hat{\epsilon}_i^B(t) = \int_0^t \{n^{-1}\hat{V}(s)^T\hat{V}(s)\}^{-1} Y_i(s)(1,\hat{V}_i)^T d\hat{M}_i(s)$$
$$+ \int_0^t \{n^{-1}\hat{V}(t)^T\hat{V}(t)\}^{-1}\{0, \hat{\epsilon}_i^c n^{-1} \sum_i R_i(t)(1,Z_i)^T dN_i(t)\}^T.$$

with $d\widehat{M}_i(t) = dN_i(t) - R_i(t)(1,\hat{V}_i)d\hat{B}(t)$ and

$$\hat{\epsilon}_i^c = \{n^{-1}(1,Z_i)^T(1,Z_i)\}^{-1}(1,Z_i)^T A_i.$$

Hence the variance of $\hat{B}(t)$ is consistently estimated by $\sum_i \hat{\epsilon}_i^B(t)\{\hat{\epsilon}_i^B(t)\}^T$. We can likewise construct uniform confidence bands and do testing investigating for instance the hypothesis $H_0: \ b_a(t) = b_a$. The first term of $\hat{\epsilon}_i^B(t)$ is the usual martingale term pretending that c is known while the second term gives the needed extra variation due to that c is estimated.

Proof of Result 1: *We consider a more general model which allows for covariates X. In this vein, suppose that conditional on (U,X,Z), the hazard function of T follows the semiparametric Aalen model*

$$h\left(t|A,U,X,Z\right) = b_0\left(s\right) + b_a\left(s\right)A + b_{x,u}(X,U,s)$$

The corresponding survival function is given by

$$S\left(t|A,U,X,Z\right)=\exp\left\{-\int_0^t\left[b_0\left(s\right)+b_a\left(s\right)A+b_{x,u}(X,U,s)\right]ds\right\}$$
$$=\exp\left\{-\int_0^t b_0\left(s\right)+b_a\left(s\right)Mds\right\}\times\exp\left\{-\int_0^t b_a\left(s\right)\Delta+b_{x,u}(X,U,s)ds\right\}$$

which induces the following survival function at time t conditional on (X,Z) upon marginalization with respect to (A,U):

$$S\left(t|X,Z\right)=E\left[S\left(t|A,U,X,Z\right)|X,Z\right]$$
$$=\exp\left\{-\int_0^t\left[b_0\left(s\right)+b_a\left(s\right)M\right]ds\right\}\times E\left[\exp\left\{-\int_0^t\left[b_a\left(s\right)\Delta+b_{x,u}(X,U,s)\right]ds\right\}|X,Z\right]$$
$$=\exp\left\{-\int_0^t\left[b_0\left(s\right)+b_a\left(s\right)M\right]ds+Q(t,X)\right\}$$

where

$$Q(t,X)=\log E\left[\exp\left\{-\int_0^t\left[b_a\left(s\right)\Delta+b_{x,u}(X,U,s)\right]ds\right\}|X,Z\right]$$
$$=\log E\left[\exp\left\{-\int_0^t\left[b_a\left(s\right)\Delta+b_{x,u}(X,U,s)\right]ds\right\}|X\right]$$

In the absence of covariates, one recovers the result given in the text, where

$$\widetilde{b}_0\left(t\right)=b_0\left(t\right)-\frac{\partial Q(t)}{\partial t}=b_0\left(t\right)-\frac{\partial\log E\left[\exp\left\{-\int_0^t\left[b_a\left(s\right)\Delta+b_{x,u}(U,s)\right]ds\right\}\right]}{\partial t}$$

More generally, in the presence of covariates, one obtains the additive hazard function:

$$\widetilde{h}(t|Z,X)=b_0\left(t\right)+b_a\left(t\right)M-\frac{\partial\log E\left[\exp\left\{-\int_0^t\left[b_a\left(s\right)\Delta+b_{x,u}(U,s)\right]ds\right\}|X\right]}{\partial t}$$

which reduces to equation (10) *under linear specification of the above function,i.e.*

$$b_0(t) - \frac{\partial \log E\left[\exp\left\{-\int_0^t \left[b_a(s)\Delta + b_{x,u}(U,s)\right] ds\right\} |X\right]}{\partial t} = \widetilde{b}_0(t) + b_x^T(t) X$$

Proof of Result 2: *To allow for covariates, suppose the following Aalen additive hazards model holds:*

$$h(t|A,U,Z) = b_0(t) + b_a(t) A + b_u(U,t)$$

and further assume (9) *and* (12) *hold, then:*

$$\begin{aligned} h(t|A,U,X,Z) &= b_0(s) + b_a(s) A + b_{x,u}(X,U,s) \\ &= b_0(s) + b_a(s) A + \rho(t)\Delta + b_x^T(t) X + \varepsilon(t) \end{aligned}$$

The corresponding survival function is then given by

$$\begin{aligned} S(t|A,U,X,Z) &= \exp\left\{-\int_0^t \left[b_0(s) + b_a(s) A + \rho(s)\Delta + b_x^T(s) X + \varepsilon(s)\right] ds\right\} \\ &= \exp\left\{-\int_0^t \left[b_0(s) + b_a(s) A + \rho(s)\Delta + b_x^T(s) X\right] ds\right\} \times \exp\left\{-\int_0^t \varepsilon(s) ds\right\} \end{aligned}$$

This in turn induces the conditional survival curve given (A,X,Z)

$$\begin{aligned} S(t|A,X,Z) &= \exp\left\{-\int_0^t \left[b_0(s) + b_a(s) A + \rho(s)\Delta + b_x^T(s) X\right] ds\right\} \times E\left[\exp\left\{-\int_0^t \varepsilon(s) ds\right\}\right] \\ &= \exp\left\{-\int_0^t \left[b_0(s) + b_a(s) A + \rho(s)\Delta + b_x^T(s) X\right] ds + \log E\left[\exp\left\{-\int_0^t \varepsilon(s) ds\right\}\right]\right\} \end{aligned}$$

with corresponding hazard function

$$\bar{b}_0(t) + b_a(t) A + \rho(t) \Delta + b_x^T(t) X$$

where

$$\bar{b}_0(t) = b_0(t) - \frac{\partial \log E\left[\exp\left\{-\int_0^t \varepsilon(s) ds\right\}\right]}{\partial t}$$

Proof of Result 3: *Under assumptions* (1) *and* (5) *one has that*

$$\begin{aligned}
S(t|A,U,X,Z) &= \exp\left\{-\int_0^t \left[b_0(s) + b_a(s) A + E\left\{b_u(U,s)|A,Z\right\} + \varepsilon(s)\right] ds\right\} \\
&= \exp\left\{-\int_0^t \left[b_0(s) + b_a(s) A + E\left\{b_u(U,s)|A,Z\right\} - E\left\{b_u(U,s)|Z\right\}\right] ds\right\} \\
&\exp\left\{-\int_0^t \left(\varepsilon(s) - E\left\{b_u(U,s)|Z\right\}\right) ds\right\} \\
&= \exp\left\{-\int_0^t \left[b_0(s) + b_a(s) A + \left(\rho_0(s) + \rho_1(s) Z\right)(A - m(Z))\right] ds\right\} \\
&\exp\left\{-\int_0^t \left(\varepsilon(s) - E\left\{b_u(U,s)\right\}\right) ds\right\}
\end{aligned}$$

where

$$\rho_0(s) = E\left\{b_u(U,s)|A=1,Z=0\right\} - E\left\{b_u(U,s)|A=0,Z=0\right\}$$

$$\begin{aligned}
\rho_1(s) = {} & E\left\{b_u(U,s)|A=1,Z=1\right\} - E\left\{b_u(U,s)|A=0,Z=1\right\} \\
& - E\left\{b_u(U,s)|A=1,Z=0\right\} + E\left\{b_u(U,s)|A=0,Z=0\right\}
\end{aligned}$$

and we use the fact that for binary A and Z,

$$\begin{aligned}
E\left\{b_u(U,s)|A,Z\right\} &= \left[E\left\{b_u(U,s)|A=1,Z\right\} - E\left\{b_u(U,s)|0,Z\right\}\right]A \\
&\quad + E\left\{b_u(U,s)|0,Z\right\} \\
&= \left\{E\left\{b_u(U,s)|A=1,Z=1\right\} - E\left\{b_u(U,s)|0,Z=1\right\}\right. \\
&\quad \left. -E\left\{b_u(U,s)|A=1,Z=0\right\} + E\left\{b_u(U,s)|0,Z=0\right\}\right\}ZA \\
&\quad + E\left\{b_u(U,s)|A=1,Z=0\right\} - E\left\{b_u(U,s)|0,Z=0\right\}A \\
&\quad + \left[E\left\{b_u(U,s)|0,Z=1\right\} - E\left\{b_u(U,s)|0,Z=0\right\}\right]Z \\
&\quad + E\left\{b_u(U,s)|0,Z=0\right\}
\end{aligned}$$

and

$$E\left\{b_u(U,s)|Z\right\} = E\left\{b_u(U,s)\right\}$$

by the independence property of the IV with U. We may conclude that

$$\begin{aligned}
S\left(t|A,X,Z\right) &= E\left\{S\left(t|A,U,X,Z\right)|A,X,Z\right\} \\
&= \exp\left\{-\int_0^t \left[b_0\left(s\right) + b_a\left(s\right)A + \left(\rho_0\left(s\right) + \rho_1\left(s\right)Z\right)\left(A - m(Z)\right)\right]ds\right\} \\
&\quad E\left[\exp\left\{-\int_0^t \left(\varepsilon(s) - E\left\{b_u(U,s)\right\}\right)ds\right\}\right] \\
&= \exp\left\{-\int_0^t \left[\widetilde{b}_0\left(t\right) + b_a\left(s\right)A + \left(\rho_0\left(s\right) + \rho_1\left(s\right)Z\right)\Delta\right]ds\right\}
\end{aligned}$$

with

$$\widetilde{b}_0\left(t\right) = b_0\left(t\right) - \frac{\partial \log E\left[\exp\left\{-\int_0^t \left[\varepsilon(s) - E\left\{b_u(U,s)\right\}\right]ds\right\}\right]}{\partial t}$$

proving the result.

References

[1] Angrist, J and Krueger, A (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. Journal of Economic Perspectives 15, 69-85.

[2] Heckman, JJ (1997), Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations, Journal of Human Resources, 32, 441–462.

[3] Hernan MA and Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology. 2006;17:360–372.

[4] Wright, S. (1928),The Tariff on Animal and Vegetable Oils, NewYork:MacMillan, the Appendix.

[5] Wooldridge, J. M. (2002),Econometric Analysis of Cross Section and Panel Data, Cambridge, MA: MIT Press.

[6] Vansteelandt, S and Goetghebeur, E (2003), Causal Inference With Generalized Structural Mean Models, Journal of the Royal Statistical Society, Ser. B, 65, 817–835.

[7] Robins, JM and Rotnitzky, A (2004), Estimation of Treatment Effects in Randomized Trails With Noncompliance and a Dichotomous Outcome Using Structural Mean Models, Biometrika, 91, 763–783.

[8] Cox, D.R. (1972). Regression Models and Life Tables, Journal of the Royal StatisticalSociety: Series B, 34: 187-220.

[9] Aalen, O.O. (1989). A Linear Regression Model for the Analysis of Life Times, Statistic in Medicine, 8: 907-925.

[10] Juster FT and Suzman R. An Overview of the Health and Retirement Study. The Journal of human resources 1995;30 (Special Issue on the Health and Retirement Study: Data Quality and Early Results):S7-S56.

[11] Angrist, JD, Imbens, GW, and Rubin, DB. (1996). Identification of Causal Effects Using Instrumental Variables, Journal of the American Statistical Association, 91, 444–472.

[12] Scheike T, Martinussen T and Silver J (2012). The "timereg" R package. Available at http://cran.r-project.org/web/packages/timereg/timereg.pdf.

[13] Huffer, FW and McKeague, IW (1987). Weighted Least Squares Estimation for Aalen's Additive Risk Model Journal of the American Statistical Association, 86, 114-129 (1991).

[14] Efron, B and Tibshirani, RJ. (1993). An Introduction to the Bootstrap, Chapman & Hall, New York.

[15] Angrist, JD and Krueger AB. (1995) Split-Sample Instrumental Variables Estimates of the Return to Schooling. Journal of Business and Economic Statistics 13, 225-235.

[16] Didelez V, Meng S and Sheehan N. Assumptions of IV methods for observational epidemiology, Statistical Science, 25, 22-40, 2010.

[17] Mozumdar A and Liguori G. Persistent increase of prevalence of metabolic syndrome among US adults: NHANES III to NHANES 1999–2006. Diabetes Care 2011;34(1):216-219.

[18] Demmer RT, Zuk AM, Rosenbaum M and Desvarieux M. Prevalence of diagnosed and undiagnosed type 2 diabetes mellitus among US adolescents: results from the continuous NHANES, 1999–2010. American Journal of Epidemiology 2013;178(7):1106-1113.

[19] Stokes A and Mehta NK. Mortality and excess risk in US adults with pre-diabetes and diabetes: a comparison of two nationally representative cohorts, 1988–2006. Population health metrics 2013;11(1):3.

[20] The NCBI dbGaP Database of Genotypes and Phenotypes.Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L,Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST.Nat Genet. 2007 Oct; 39(10):1181-6.

[21] Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics 2012;44(9):981.

[22] Robins JM. (1993).Analytic methods for estimating HIV treatment and cofactor effects. Methodological Issues of AIDS Mental Health Research. Eds: Ostrow DG, Kessler R. New York: Plenum Publishing. pp. 213-290.

[23] Loeys, T, Goetghebeur, E and Vandebosch, A (2005). Causal proportional hazards models and time-constant exposure in randomized clinical trials. Lifetime Data Analysis,11, 435-449.

[24] Cuzick, J, Sasieni, P, Myles, J, Tyrer J (2007). Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. Journal of the Royal Statistical Society - Series B 69, 565-588.

[25] MacKenzie T, Tosteson T, Morden N, Stukel, T, O'Malley J. Using instrumental variables to estimate a Cox's proportional hazards regression subject to additive confounding.Health Serv Outcomes Res Method (2014) 14:54–68.

[26] Lin, DY and Ying, Z (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. The annals of Statistics, 1712-1734.

[27] Martinussen, T and Scheike, TH. (2002). A flexible additive multiplicative hazard model. Biometrika, 89(2), 283-298.

[28] Terza, JV, Basu, A, and Rathouz, PJ (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. Journal of health economics, 27(3), 531-543.

[29] Martinussen T, Scheike T. Dynamic regression models for survival data. (2006), Springer-Verlag.