HHS Public Access

Author manuscript

Epidemiology. Author manuscript; available in PMC 2016 May 01.

Published in final edited form as:

Epidemiology. 2015 May ; 26(3): 402–410. doi:10.1097/EDE.0000000000000262.

Instrumental variable estimation in a survival context

Eric J. Tchetgen Tchetgen[1], **Stefan Walter**[2], **Stijn Vansteelandt**[3], **Torben Martinussen**[4], and **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

Abstract

Bias due to unobserved confounding can seldom be ruled out with certainty when estimating the causal effect of a nonradomized treatment. 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 and also for certain generalized linear models with a non-linear link function. However, IV methods are not as well developed for regression analysis with a censored survival outcome. In this paper, we develop the instrumental variable approach for regression analysis in a survival context, primarily under an additive hazards model, for which we describe two simple methods for estimating causal effects. The first method is a straightforward two-stage regression approach analogous to two-stage least squares commonly used for IV analysis in linear regression. In this approach, 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 a valid estimate of the treatment effect of interest. The second method is a so-called control function approach, which entails adding to the additive hazards outcome model, 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 to evaluate the effect of diabetes on mortality using data from the Health and Retirement Study. We also establish that analogous strategies can also be used under a proportional hazards model specification provided the outcome is rare over the entire follow-up.

Keywords

Instrumental variable; Unobserved confounding; Aalen additive hazards model; Two stage regression; Control function; Cox proportional hazards model

Unmeasured confounding is an important possible source of bias when estimating the effect of a nonrandomized intervention, treatment or exposure, such as in epidemiologic

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

observational studies. In recent years, epidemiologists have slowly expanded their analytic toolbox to address unobserved confounding, by adopting the instrumental variable (IV) design, an approach for analyzing non- experimental data. The IV approach has historically been favored by economists but has until recently received less attention in epidemiology.[1,2] A valid IV is a pre-exposure variable associated with the exposure of interest, and only associated with the outcome through its association with the exposure. Thus, a valid IV must not influence the outcome through a pathway other than through the exposure (the exclusion-restriction assumption), and, although correlated with the exposure, the IV must be independent of unobserved confounders of the exposure-outcome association. 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 common in epidemiologic practice, and hazard regression is typically used to model such outcomes. The Cox proportional hazards model is perhaps the most popular regression framework for survival data.[8] Aalen's additive hazards model offers a flexible alternative for modeling associations on the hazard scale.[9] An important appeal of additive hazards models is that, unlike proportional hazards, a hazards difference is a collapsible effect measure. Specifically, collapsing over a continuous regressor in an additive hazards model can under fairly reasonable assumptions, recover a marginal additive hazards model (see eAppendix for a detailed discussion of collapsibility of the Aalen model). In this paper, we exploit the collapsibility of additive hazards to develop two straightforward approaches for IV estimation with a censored survival outcome. The first approach is a straightforward two-stage regression approach analogous to two-stage least squares commonly used for IV estimation in linear models.[5] The current context is somewhat different from that of standard two-stage least squares in that, to estimate the treatment effect, here the fitted value from the first-stage regression of the exposure on the IV is substituted for the exposure in a second-stage additive hazards model, rather than in a standard linear regression model. The second proposed approach 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 several extensions of the methods are also given. The new methods are illustrated below 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] Several additional results are relegated to the Supplemental Appendix. There, we establish 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. We also describe a straightforward sensitivity analysis technique to assess the extent to which a violation of the exclusion restriction might affect the proposed methods to make IV inferences under an additive hazards model, and the methods are easily extended to the Cox model when appropriate.

Two-stage regression approach

Suppose that one has observed independent and identically distributed data on (T^*, A, Z) for n persons, where A is a treatment, Z is the 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, but 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

$$h(t|A,U,Z)=b_0(t)+b_a(t)A+b_u(U,t) \quad (1)$$

where $h(t|A, U, Z)$ is the hazard function of T evaluated at t, conditional on A, U and Z, and the functions ($b_0(\cdot)$, $b_a(\cdot)$, $b_u(\cdot,\cdot)$) 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, although, the effect size $b_a(t)$ may vary with t. The model is quite flexible in the unobserved confounder association with the outcome $b_u(\cdot,\cdot)$, 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 binary diabetes status measured at baseline (1 if diabetic and 0 otherwise), T is time to death, and Z is 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 $b_a(t) = 0$ for all t. An important sub-model to consider is the constant hazards difference model obtained by setting

$$b_a(t)=b_a \quad (2)$$

where b_a is an unknown constant. Note that the model assumes no interaction between A and U. Collapsibility over U makes $b_a(t)$ 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 interpretable only conditional on the unobserved U. The baseline hazard function $b_0(t)$ 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, necessary for a valid IV analysis.[11] In practice, additional pre-exposure covariates X (e.g. age, sex, 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; 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 :

$$A=c_0+c_zZ+\Delta, \text{where } \Delta \text{ is mean zero residual error independent of } Z \quad (3)$$

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, just as with the usual IV analyses, c_z may not have a causal interpretation, in the event of unobserved confounding of the effect of Z on A. We must, however, assume that any unobserved common cause of Z and A is independent of U. Let $M = m(Z) = E(A|Z) = c_0 + c_zZ$. In words, M is the predicted mean value of the treatment variable as a function of the IV, the usual first-stage of two-stage least-squares IV analyses.

The proposed two-stage approach for IV in a survival context is based on the following result, which provides an analytic expression for the conditional hazard model $h(t|Z)$, 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 obtains*

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

with $b_0(t)$ *a baseline hazard function.*

Result 1 states that under assumptions (1) and (3), as well as 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 $B(t) = (b_0(t), b_a(t))^T$ 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 $\hat{M}$, the fitted value of the ordinary least-squares regression of A on Z, i.e. $\hat{M} = \hat{c}_0 + \hat{c}_zZ$, where $(\hat{c}_0, \hat{c}_z)$ is the ordinary least-squares estimator of (c_0, c_z). The second-stage then involves obtaining Aalen's least-squares estimator $\hat{B}^*(t)$ of $B(t)$, defined similarly to $B^*(t)$, with $\hat{M}$ 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 eAppendix) an analytic expression of a consistent estimator of the corrected standard error of $\hat{B}^*(t)$.

Occasionally, the first-stage ordinary least-squares estimate ($\hat{c}_0$, $\hat{c}_z$) may be obtained from a sample that is independent of that used for the second-stage estimation of (4). In this type of split-sample IV design,[15] uncertainty in the first-stage estimation can essentially be ignored, but inferences must be interpreted conditional on the external sample. A major issue in IV estimation appears when the association between the IV and the treatment is weak, the so-called problem of weak instruments.[16] When the IV is weak, standard confidence intervals may not have adequate coverage and estimates may be sensitive to small violations in the exclusion restriction.[17,18] For this reason, it is important in practical situations to assess the strength of the first-stage association between *Z* and *X*. In case of a weak IV, split sample IV is well known to be robust to weak IV bias in the context of linear models in the sense that the bias is guaranteed to be towards the null of no causal effect.[15] The split sample IV may also be as effective to address weak IV bias in the present context.

Control function approach

In this section, we consider an alternative approach to two-stage regression. Consider the sub-model of (1) that further specifies the influence of the unobserved confounder *U* on the hazard function:

$$b_u(U,t)=\rho_0(t)\Delta+\varepsilon(t) \quad (5)$$

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; other than 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*

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

for $\bar{b}_0(t)$ *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 $\bar{b}_0(t)$ to differ from $b_0(t)$. Intuitively, the residual Δ captures any variation in the hazard function due to unobserved correlates of *A*, not accounted for in *M*. These unobserved correlates must include any confounders of the $A-Y$ association, and so Δ can be used as a proxy measure of unobserved confounders. 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 $\hat{\Delta}=$

$A - M\hat{}$ as an estimate of the unobserved residual Δ that we use to fit an additive hazards model, with regressors $(A, \Delta\hat{})$ 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 applies equally as an approach to appropriately account for uncertainty due to in-sample estimation of Δ. In situations where a split-sample IV design is adopted, the first sample estimation uncertainty can essentially be ignored. 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, because it is necessary in the control function approach to calculate the residual $A - M\hat{}$.

Binary exposure

The control function approach can also be used in the context of a binary or discrete exposure. In the simple case of binary *A*, the methods described in the previous section apply upon estimating *M* using binary regression, e.g. logit*M* =logit*m* (*Z*) =logitPr(*A* = 1|*Z*) = $c_0 + c_z Z$. The approach can be motivated under a modified set of assumptions to account for binary *A*. Suppose that

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

where $\varepsilon(t)$ is an independent error, and *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.

RESULT 3

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

$$\overline{h}(t|A,Z) = \tilde{b}_0(t) + b_a(t)A + \{\rho_0(t) + \rho_1(t)Z\}\Delta, \quad (8)$$

for $b_0\tilde{}(t)$ *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 $E\{b_u(U,s)|A,Z\}$ is linear in *Z*.

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*,

$$A=c_0+c_zZ+c_x^TX+\Delta \quad (9)$$

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

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

with $M=m(X,Z)=c_0+c_zZ+c_x^TX$ 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

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

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

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

where $\varepsilon(t)$ is a random error independent of (Δ, *Z*, *X*). Our previous assumption that censoring is independent of *T* conditional on *A* and *Z* will need to be extended to require that censoring is independent of *T* conditional on *A*, *Z* and *X*.

In the eAppendix, we formally establish that, under a rare disease condition, analogous two-stage regression and control function methods likewise apply in the context of Cox proportional hazards model. Intuitively, when the outcome is rare, 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. Then, a Cox regression analysis is essentially equivalent to 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. The framework then essentially reduces to IV for loglinear regression analysis, for which two-stage regression has previously been shown to apply under analogous assumptions as considered here.[17]

Empirical illustration

The prevalence of type 2 diabetes mellitus is increasing across all age groups in the United States possibly as a consequence of the obesity epidemic.[18,19] In addition, no decline has been observed in the excess mortality among persons suffering from diabetes relative to persons without diabetes.[20] Obtaining an unbiased estimate of the mortality risk associated

with diabetes is key to predicting the future health burden in the population and to evaluating 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 Study, a cohort initiated in 1992 with repeated assessments every 2 years. We used externally validated genetic predictors of type 2 diabetes as IVs to estimate effects on mortality among HRS participants. The Health and Retirement Study is a well-documented nationally representative sample of persons aged 50 years or older and their spouses.[10] Genotype data were collected on a subset of 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 Health and Retirement Study and dbGaP21.[21] From the 12,123 participants for whom genotype data was available, we restricted the sample to 8,446 non-hispanic white persons with valid self-reported diabetes status at baseline. Self-reported diabetes in the Health and Retirement Study has been shown to have 87% sensitivity and 97% specificity for Hemoglobin A1c defined diabetes among non-Hispanic white HRS participants.[22] For deaths occurring between 1998 and 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 determined 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) and mortality. As genetic instruments, we used 39 independent single nucleotide polymorphisms previously established to be significantly associated with diabetes.[23]

For comparison, we first performed an observational analysis, which entailed fitting a standard Aalen additive hazards model for diabetes. Next, we implemented the proposed control-function instrumental variable approach, which is appropriate for binary endogenous variables, while the two-stage approach is strictly justified only for continuous 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 bootstrap samples.

Participants were, on average, 68.5 years old (standard deviation [SD]=10.4 years old) at baseline and 1,891 self-reported that they had diabetes (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% (Nagelkerke R^2) of the variation in diabetes in the study sample, and were strongly associated as a set with the endogenous variable (Likelihood ratio test Chi-square

statistic = 176.75 with 39 degrees of freedom, which corresponds to a significance value $<10^{-6}$).

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 [CI]=0.025 to 0.035) per person-year. This means that, over the course of the follow-up, an average of 3 additional deaths occurred for each year of follow-up in each 100 persons with diabetes alive at the start of the year, compared with each 100 diabetes-free persons alive at the start of the year, conditional on age and sex. The genetic IV approach produced a notably larger effect associated with diabetes, with a diabetes-associated increase in the mortality rate of beta = 0.08 [95%CI=0.075 to 0.090] per person-year, nearly three times the rate estimated by the observational additive hazards model. We obtained further evidence of negative confounding bias reflected in the observed association between the first-stage residual and mortality rate (beta=−0.023 [95%CI=−0.028 to −0.017], as well as with marginal evidence of confounding bias heterogeneity, beta=−0.024 [−0.052,0.001]. We also note that a split-sample IV analysis was obtained using external first-stage regression coefficients,[23] the results were essentially identical and are therefore omitted.

The assumption that all 39 SNPs that define the IV affect a person's time from baseline to death only through baseline diabetes status may not be entirely credible, even if all 39 SNPs only affect mortality through diabetes. This is because there is likely to be a nonnegligible direct effect from one of the SNPs to diabetes incidence among persons who are diabetes-free at baseline. This would constitute a violation of the so-called exclusion restriction and therefore would invalidate our genetic IV for assessing the mortality effects of baseline diabetes. Nonetheless, although possibly positively biased under the alternative hypothesis, the two-stage regression estimator could still be interpreted as a valid test of the null hypothesis of no association between diabetes disease (whether baseline or time-updated) and mortality. In addition, there may be further pleiotropic effects of at least one of the SNPs through a pathway not involving diabetes, which would constitute an even more serious violation, as it would also invalidate our IV analysis as a valid test of a causal association between diabetes and death. In light of these possible limitations, the reader should interpret these analyses with care and only as an empirical illustration of the methods. In the eAppendix, we provide R code used to implement the HRS analysis.

Discussion

A well-known result about standard 2SLS for IV estimation in linear models is that the approach is completely robust to a mis-specified model for the first-stage regression.[5] In contrast, in principle, the first-stage regression in the current setting must be correctly specified in order for either of the proposed methods to be valid. This highlights the importance of performing routine regression diagnostics for the first-stage model. An important exception occurs for two-stage regression under the null hypothesis of no causal effect of *A*, in which case model mis-specification of the first-stage regression does not directly impact the validity of IV inferences about the causal effect of *A*. This local robustness property, however, does not apply to the control function approach. In the

eAppendix, we further describe a straightforward modification of the two-stage regression approach, which is guaranteed to be consistent even when the first-stage regression is mis-specified and whether the null hypothesis of no causal effect of A holds or not. We briefly describe the approach, which entails substituting an estimate of the risk set-specific first-stage regression $M(t) = m(Z, t) = E(A|Z, T^* > t)$ for M in the second-stage Aalen regression model. Note that under independent censoring $E(A|Z, T^* > t) = E(A|Z, T > t)$. In the eAppendix, we formally establish that, after making this substitution, the second-stage Aalen estimator will be consistent for the causal effect of A whether the first-stage model is correct or not, provided a separate linear model is used to estimate $m(Z, t)$, using standard ordinary least squares analysis among person at risk at time t. Furthermore, if the cumulative risk for the failure-time outcome remains low over the follow-up, one may pool the first-stage regression across all person-time contributions to estimate a common regression model independent of t. This is because the rare-disease assumption would ensure that the regression fit remains relatively stable over the follow-up.

In certain settings, both A and Z 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 presents several challenges, and methods to appropriately handle such challenges are beyond the scope of the paper. To the best of our knowledge, methodology is currently lacking for such settings under either an additive hazards model or a Cox proportional hazards model. However, IV methods to evaluate the joint effects of a time-updated exposure are available for the semi-parametric accelerated failure time model of Robins.[24] Unfortunately estimation of semiparametric accelerated-failure-time models can be computationally burdensome in practice, because they often require artificial censoring of subjects with observed event time. In the point-exposure case, alternative IV methods have been proposed under a structural proportional hazards model,[25,26] which do not require artificial censoring and which do not rely on a rare disease assumption. However, in contrast with the methods developed herein for a proportional hazards model, earlier proposals are limited to either binary instrumental variable or binary exposure variable.[25,26] Furthermore, existing methods for Cox regression are primarily aimed at estimating a so-called complier causal effect and rely for identification on a monotonicity assumption about the effect of the IV on the exposure, an assumption we do not make in the current paper (also see Nie et al[27] for a related approach for estimating the complier survival curve.)

The methods proposed here address an important gap in the IV literature. Previous, motivations of IV for hazard regression appropriate for binary or continuous exposures and IV have relied on less plausible assumptions or placed overly stringent restrictions, and thus may be less relevant for routine application. For example, in a recent proposal, MacKenzie and colleagues[28] use an instrumental variable to estimate a Cox proportional hazards model subject to additive unobserved confounding. Specifically, they focus on a so-called additive multiplicative hazards model,[29,30]

$$h(t|A,U,Z)=h_0(t)\exp(b_aA)+b_u(U,t) \quad (13)$$

with the key restriction

$$E\{b_u(U,t)|T(a) \geq t\}=0 \quad (14)$$

where $T(a)$ is the potential outcome of T under treatment a. The model combines features of both the Cox model and the Aalen model, since it includes both an additive effect of U and a multiplicative effect of A. The restriction (14) ensures that the marginal hazard model of $T(a)$ follows a Cox proportional hazards model. MacKenzie and colleagues note that this restriction is generally satisfied if $b_u(U, t)$ can be written $d(t)U + \ln \text{mgf}_U\{d(t)\}$, for some function $d(t)$, where mgf_U stands for the moment-generating function of U. Under such unobserved confounding, MacKenzie and colleagues 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 it initially seems, because, supposing that U were observed, assumptions (13) and (14) imply that the conditional hazard function of $T(a)$ at time t given U depends only 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.

The control-function approach described in this paper may also be seen as an extension of the two-stage residual inclusion approach of Terza et al.[31] In order to ease a comparison between the two methods, it is helpful to restate the key assumption underlying their approach using our notation. This is best achieved by simply replacing equation (5) with the more restrictive model:

$$b_u(U,t)=\rho_0(t)\Delta \quad (15)$$

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, i.e. that Z includes all existing IVs. Allowing for $\varepsilon(t)$ in equation (5) avoids this type of restriction. It is also notable that assumption (15) may be overly restrictive for binary (or discrete) A, since the distribution of the residual Δ is completely determined by the mean $M = m(Z)$ of A given Z, 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.

Very recently, Li et al[32] published a paper in which they independently derive the special case of Result 1, under the additional assumption of a constant hazards difference (2). Therefore, the current paper provides a number of generalizations beyond the result of Li et al, who not only considered a more restrictive additive hazards model, but also formally addressed only the situation of a continuous endogenous variable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

Eric Tchetgen Tchetgen's work is funded by NIH grant R01AI104459. Maria Glymour and Stefan Walter were funded by NIH grant RC4MH092707. Torben Martinussen's work is part of the Dynamical Systems Interdisciplinary Network, University of Copenhagen.

APPENDIX

Proof of Result 1

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

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

The corresponding survival function is given by

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

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

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

where

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

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

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

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

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

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

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

Proof of Result 2

To allow for covariates, suppose the following Aalen additive hazards model holds:

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

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

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

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*

$$S(t|A,U,Z)=\exp\left\{-\int_0^t[b_0(s)+b_a(s)A+E\{b_u(U,s)|A,Z\}+\varepsilon(s)]ds\right\}$$
$$=\exp\left\{-\int_0^t[b_0(s)+b_a(s)A+E\{b_u(U,s)|A,Z\}-E\{b_u(U,s)|Z\}]ds\right\}$$

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

where

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

$$\rho_1(s)=E\{b_u(U,s)|A=1,Z=1\}-E\{b_u(U,s)|A=0,Z=1\}-E\{b_u(U,s)|A=1,Z=0\}+E\{b_u(U,s)|A=0,Z=0\}$$

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

$$E\{b_u(U,s)|A,Z\}=[E\{b_u(U,s)|A=1,Z\}-E\{b_u(U,s)|A=0,Z\}]A$$
$$+E\{b_u(U,s)|A=0,Z\}$$
$$=\{E\{b_u(U,s)|A=1,Z=1\}-E\{b_u(U,s)|A=0,Z=1\}$$
$$-E\{b_u(U,s)|A=1,Z=0\}+E\{b_u(U,s)|A=0,Z=0\}\}ZA$$
$$+E\{b_u(U,s)|A=1,Z=0\}-E\{b_u(U,s)|A=0,Z=0\}A$$
$$+[E\{b_u(U,s)|A=0,Z=1\}-E\{b_u(U,s)|A=0,Z=0\}]Z$$
$$+E\{b_u(U,s)|A=0,Z=0\}$$

and

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

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

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

$$E\left[\exp\left\{-\int_0^t(\varepsilon(s)-E\{b_u(U,s)\})ds\right\}\right] = \exp\left\{-\int_0^t\left[\tilde{b}_0(t)+b_a(s)A+(\rho_0(s)+\rho_1(s)Z)\Delta\right]ds\right\}$$

with

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

proving the result.

References

1. Angrist J, Krueger A. Instrumental variables and the search for identification: From supply and demand to natural experiments. Journal of Economic Perspectives. 2001; 15:69–85.
2. Heckman JJ. Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. Journal of Human Resources. 1997; 32:441–462.
3. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology. 2006; 17:360–372. [PubMed: 16755261]
4. Wright, S. The Tariff on Animal and Vegetable Oils. New York: MacMillan; 1928. the Appendix.
5. Wooldridge, JM. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press; 2002.
6. Vansteelandt S, Goetghebeur E. Causal Inference With Generalized Structural Mean Models. Journal of the Royal Statistical Society, Ser. B. 2003; 65:817–835.
7. Robins JM, Rotnitzky A. Estimation of Treatment Effects in Randomized Trails With Noncompliance and a Dichotomous Outcome Using Structural Mean Models. Biometrika. 2004; 91:763–783.
8. Cox DR. Regression Models and Life Tables. Journal of the Royal Statistical Society: Series B. 1972; 34:187–220.
9. Aalen OO. A Linear Regression Model for the Analysis of Life Times. Statistic in Medicine. 1989; 8:907–925.
10. Juster FT, Suzman R. An Overview of the Health and Retirement Study. The Journal of human resources. 1995; 30:S7–S56. (Special Issue on the Health and Retirement Study: Data Quality and Early Results).
11. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association. 1996; 91:444–472.
12. Scheike T, Martinussen T, Silver J. The "timereg" R package. 2012 Available at http://cran.r-project.org/web/packages/timereg/timereg.pdf.
13. Huffer FW, McKeague IW. Weighted Least Squares Estimation for Aalen's Additive Risk Model. Journal of the American Statistical Association. 1987; 86:114–129. (1991).
14. Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.
15. Angrist JD, Krueger AB. Split-Sample Instrumental Variables Estimates of the Return to Schooling. Journal of Business and Economic Statistics. 1995; 13:225–235.
16. Bound J, Jaeger DA, Baker RM. Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. Journal of the American Statistical Association. 1995; 90:443–450.
17. Imbens GW, Rosenbaum PR. Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education. Journal of the Royal Statistical Society: Series A (Statistics in Society. 2005; 168:109–126.

18. Small DS, Rosenbaum PR. War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases. Journal of the American Statistical Association. 2008; 103:924–933.
19. Didelez V, Meng S, Sheehan N. Assumptions of IV methods for observational epidemiology. Statistical Science. 2010; 25:22–40.
20. Mozumdar A, 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. [PubMed: 20889854]
21. Demmer RT, Zuk AM, Rosenbaum M, 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. [PubMed: 23887044]
22. Stokes A, 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. [PubMed: 23448510]
23. 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): 1181–1186. [PubMed: 17898773]
24. White K, Mondesir FL, Bates LM, Glymour MM. Diabetes risk, diagnosis, and control: do psychosocial factors predict hemoglobin A1c defined outcomes or accuracy of self-reports? Ethnicity & disease. 2014; 24(1):19–27. [PubMed: 24620444]
25. Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides in- sights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics. 2012; 44(9): 981. [PubMed: 22885922]
26. Robins, JM. Analytic methods for estimating HIV treatment and cofactor effects. In: Ostrow, DG.; Kessler, R., editors. Methodological Issues of AIDS Mental Health Research. New York: Plenum Publishing; 1993. p. 213-290.
27. Loeys T, Goetghebeur E, Vandebosch A. Causal proportional hazards models and time-constant exposure in randomized clinical trials. Lifetime Data Analysis. 2005; 11:435–449. [PubMed: 16328570]
28. Cuzick J, Sasieni P, Myles J, Tyrer J. 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. 2007; 69:565–588.
29. Nie H, Cheng J, Small DS. Inference for the Effect of Treatment on Survival Probability in Randomized Trials with Noncompliance and Administrative Censoring. Biometrics. 2011; 67:1397–1405. [PubMed: 21385167]
30. 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.
31. Lin DY, Ying Z. Semiparametric analysis of general additive-multiplicative hazard models for counting processes. The Annals of Statistics. 1995:1712–1734.
32. Martinussen T, Scheike TH. A flexible additive multiplicative hazard model. Biometrika. 2002; 89(2):283–298.
33. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. Journal of health economics. 2008; 27(3):531–543. [PubMed: 18192044]
34. Li J, Fine J, Brookhart A. Instrumental Variable Additive Hazards Models. Biometrics. 2014 Available online ahead of print.
35. Martinussen, T.; Scheike, T. Dynamic regression models for survival data. Springer-Verlag: 2006.
36. Vansteelandt S, Martinussen T, Tchetgen Tchetgen E. On adjustment for auxiliary covariates in additive hazard models for the analysis of randomized experiments. Biometrika. 2014; 101:237–244.

Table 1

Observation and IV analysis of HRS data to estimate the effect 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)

* Difference in the hazards of death.

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