Addressing unmeasured confounders in cohort studies: InstrumentalVariable method for a time-fixed exposure on an outcome trajectory

Le Bourdonnec Kateline, Cécilia Samieri, Christophe Tzourio, Thibault Mura, Aniket Mishra, David-Alexandre Trégouët, Cécile Proust-Lima

Addressing unmeasured confounders in cohort studies: Instrumental Variable method for a time-fixed exposure on an outcome trajectory

Le Bourdonnec Kateline[1,*], Cécilia Samieri[1], Christophe Tzourio[1], Thibault Mura[2], Aniket Mishra[1], David-Alexandre Trégouët[1], and Cécile Proust-Lima[1]

[1]Univ. Bordeaux, Inserm, BPH, U1219, F-33000 Bordeaux, France
[2]Institute for Neurosciences of Montpellier INM, Univ. Montpellier, INSERM, F-34091 Montpellier, France
[*]Email: kateline.le-bourdonnec@u-bordeaux.fr

Abstract

Instrumental variable methods, which handle unmeasured confounding by targeting the part of the exposure explained by an exogenous variable not subject to confounding, have gained much interest in observational studies. We consider the very frequent setting of estimating the unconfounded effect of an exposure measured at baseline on the subsequent trajectory of an outcome repeatedly measured over time. We didactically explain how to apply the instrumental variable method in such setting by adapting the two-stage classical methodology with (1) the prediction of the exposure according to the instrumental variable, (2) its inclusion into a mixed model to quantify the exposure association with the subsequent outcome trajectory, (3) the computation of the estimated total variance. A simulation study illustrates the consequences of unmeasured confounding in classical analyses and the usefulness of the instrumental variable approach. The methodology is then applied to 6,224 participants of the 3C cohort to estimate the association of type-2 diabetes with subsequent cognitive trajectory, using 42 genetic polymorphisms as instrumental variables. This contribution shows how to handle endogeneity when interested in repeated outcomes, along with a R implementation. However, it should still be used with caution as it relies on Instrumental Variable assumptions hardly testable in practice.

Causality, Instrumental Variable, Repeated data, Cohort study, Mixed model

‘

1 INTRODUCTION

Observational studies are widely used in epidemiology to assess the relation between an exposure X and an outcome Y, with the perspective to identify the causal effect of X on Y. Statistical techniques [1, 2] have been used to derive causal interpretations in the presence of confounding. However, they rely on the assumption that all the sources of confounding have been observed and controlled for. Yet, in many contexts the assumption that all the confounders are observed is unrealistic, and statistical analyses are likely to provide biased estimates of causal associations [3]. For instance, when studying the relation between cardiometablic factors on cognitive aging, so many confounders may intervene [4] that residual unobserved confounding is very likely. The issue of unmeasured confounding relates to the more general problem of endogeneity that occurs when the covariate is partly explained by the system under study. Beyond confounding, endogeneity also encompasses reverse causation which occurs when the outcome or its underlying process may cause a change in the exposure [5].

To handle endogeneity, instrumental variable analysis, first developed in Economics [6], was applied in Public Health from the early 2000s[7]. This method consists in using an exogenous variable, the "Instrumental Variable" (IV), that is not subject to unmeasured confounding and recreates the randomization framework. The principle of the IV methodology can be illustrated in the cross-sectional framework (Figure 1 A). Let us denote Z the IV, X the endogenous exposure variable, Y the outcome, and U the unobserved confounders. To be considered as valid, the IV needs to satisfy 3 assumptions [7]: (1) Z is strongly associated with X; (2) Z is associated with Y only through X; (3) Z is independent of U conditionally on X. Under these assumptions, Z can be used to retrieve the causal association between X and Y. In epidemiology, genetic data have been considered as promising IV since genes are determined from birth, thus not subject to confounding; in this context, IV methodology is called Mendelian Randomization (MR) [8]. Finally, to be interpreted as causal effects, IV analyses require a fourth assumption of homogeneity for the average causal effect or monotonicity for the local average causal effect [9, 10].

The most widely used estimation technique in IV methodology is the two-stage approach, called Two-Stage Least Square (2SLS) method [11]: first the endogenous exposure is regressed on the IV, and second the derived prediction, which is independent of the unmeasured confounders due to the

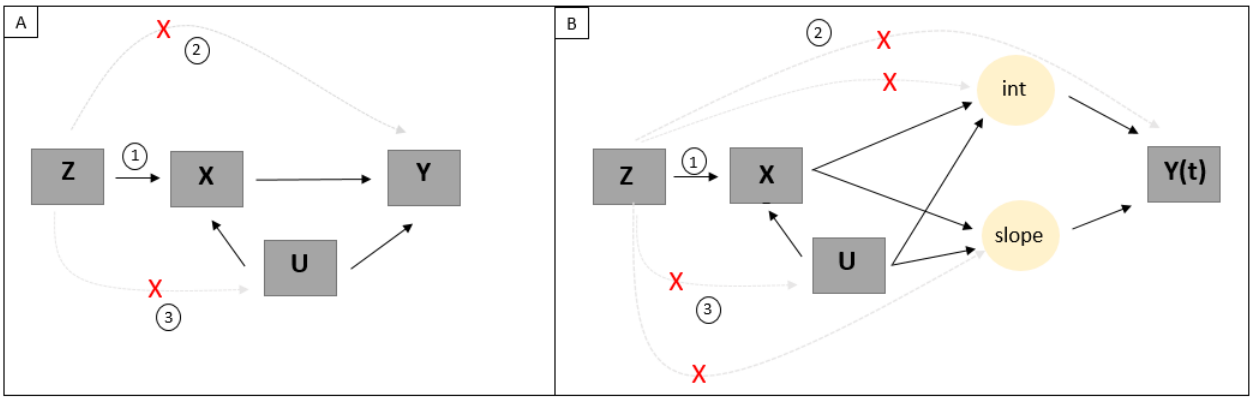

Figure 1: Directed Acyclic Graph for the IV methodology with a cross-sectional outcome Y (panel A) or a longitudinal continuous outcome Y (panel B). X is the exposure, Z the instrumental variable (with 1, 2, 3 the corresponding IV assumptions), and U the (partially) unobserved confounders. Int and slope represent the underlying latent level of Y at baseline, the latent slope of Y over time, respectively.

assumptions of Z, substitutes the exposure in the regression of the outcome to quantify the causal relation between X and Y. First proposed in the cross-sectional framework where X and Y were continuous variables measured at a single time point [11], it was adapted to handle binary exposures and/or binary outcomes [12, 13], and to treat grouped data [14, 15].

Recently, the methodology was extended to handle longitudinal data. Two settings were explored: (i) an exposure repeatedly measured over time and its effect on the concomitant level of a repeatedly measured outcome [16, 17] and (ii) a time-fixed exposure and its effect on the subsequent risk of an event [15, 18, 19]. Yet, another frequent setting encountered in longitudinal studies concerns a time-fixed exposure and its effect on the subsequent trajectory of an outcome repeatedly measured over time.

In the present contribution, we aim to didactically explain how the IV methodology can be used in observational cohort studies to assess the association between an exposure collected at baseline and the trajectory of an outcome repeatedly measured over follow-up in the presence of potential unmeasured confounding. Our solution consists in considering a mixed model for the repeated marker in the second step of the two-stage IV approach. We show how this can solve situations of unmeasured confounding and endogeneity, and we illustrate it in a simulation study considering both a binary and a continuous exposure, and a continuous outcome. We finally apply the methodology to assess the association between type-2 diabetes and cognitive aging in the French cohort "Three city" (3C) [20], by using genetic polymorphisms as the exogenous variable.

2 METHODS

2.1 Framework

Let us consider a classical longitudinal framework (Figure 1B) where X is the time-fixed exposure, $\mathbf{U}$ is a r-vector of confounders and $\mathbf{Z}$ is a p-vector of exogenous (instrumental) variables, all defined and measured at entry in the cohort while the continuous outcome Y is repeatedly measured over time t after baseline. Without loss of generality, we assume $E(\mathbf{U}) = \mathbf{0}$.

To ease the problem description, we first consider the case of a continuous exposure, and we assume Y evolves linearly over time and can be summarized by its latent level at baseline and its latent slope over time, on which the other variables can have an effect. The generalization

to a nonlinear trajectory over time is straightforward by considering a more flexible basis of time functions instead of only intercept and slope.

Let us assume that the true relations schematized in Figure 1B translate for each subject i ($i = 1, ..., N$) of a sample and each occasion j ($j = 1, ...n_i$) in a linear regression for the continuous exposure (1) and a linear mixed model for the outcome (2):

$$X_i = \alpha_0^* + \mathbf{Z_i}^\top \alpha_{\mathbf{Z}}^* + \mathbf{U_i}^\top \alpha_{\mathbf{U}}^* + \epsilon_i^{X*} \tag{1}$$

$$\begin{aligned} Y_{ij} = & \underbrace{\beta_0^* + X_i\beta_e^* + \mathbf{U_i}^\top \beta_{\mathbf{U}}^* + b_{0i}^*}_{Int_i} \\ & + \underbrace{\left(\beta_t^* + X_i\beta_{te}^* + \mathbf{U_i}^\top \beta_{\mathbf{tU}}^* + b_{1i}^*\right)}_{Slope_i} t_{ij} + \epsilon_{ij}^{Y*} \end{aligned} \tag{2}$$

For the sake of readability, conditioning on covariates and random effects, although systematic, is not made explicit in any of the linear regressions throughout the manuscript.

Following classical definitions of the linear mixed model [21, 22], $\mathbf{b_i^*} = (b_{0i}^*, b_{1i}^*)^\top \sim \mathcal{N}(0, \mathbf{B}^*)$ is the vector of individual random effects which accounts for the intra-individual correlation within the repeated Y measures. The measurement error in the exposure regression ϵ_i^{X*} is independent of Z_i and $\mathbf{U_i}$ and the measurement error at time t_{ij} in the outcome regression $\epsilon_{ij}^{Y*} \sim \mathcal{N}(0, \sigma_Y)$ is independent of all the other measurement errors at different times $\epsilon_{ij'}^{Y*}$ with $j' \neq j$, and of X_i, $\mathbf{U_i}$ and $\mathbf{b_i^*}$. The random effects $\mathbf{b_i^*}$ are also independent of X_i and $\mathbf{U_i}$. In Equations (1) and (2), superscript * refers to the parameters and latent variables under the true model.

The parameters of interest are β_e^* and β_{te}^* corresponding to the effect of X on the level of Y at inclusion and the effect of X on the subsequent change of Y over time, respectively. Since all confounders are included through U in model (2), we can interpret these parameters in a causal way. The fundamental problem is that this model and these parameters can not be directly estimated when some of the confounders $\mathbf{U}$ are not observed. Let's split $\mathbf{U} = (\mathbf{U^o}, \mathbf{U^m})$ with $\mathbf{U^o}$ the observed confounders and $\mathbf{U^m}$ the unobserved confounders.

2.2 Naive approach neglecting unobserved confounding

In the presence of unobserved confounding, a naive solution consists in estimating the association between X and the trajectory of Y by considering the model which includes $\mathbf{U^o}$ but omits $\mathbf{U^m}$:

$$\begin{aligned} Y_{ij} = & \beta_0^N + \beta_e^N X_i + b_{0i}^N + \mathbf{U_i^{o\top}} \beta_{\mathbf{Uo}}^{\mathbf{N}} \\ & + (\beta_t^N + \beta_{te}^N X_i + \mathbf{U_i^{o\top}} \beta_{\mathbf{tUo}}^{\mathbf{N}} + b_{1i}^N) t_{ij} + \epsilon_{ij}^{NY} \end{aligned} \tag{3}$$

The estimation of this model relies on the same distributions and independence assumptions as defined for model (2). Yet, those are not satisfied anymore in the presence of unobserved confounding: the neglected confounders $\mathbf{U^m}$ are absorbed by the individual random-effects: $b_{0i}^N = b_{0i}^* + \mathbf{U_i^{m\top}} \beta_{\mathbf{Um}}^*$ and $b_{1i}^N = b_{1i}^* + \mathbf{U_i^{m\top}} \beta_{\mathbf{tUm}}^*$, so that $\mathbf{b_i^N} = (b_{0i}^N, b_{1i}^N)^\top$ is not independent of X_i anymore, and is not homoscedastic anymore. Of note, $\mathbf{U_i^m}$ induces a correlation between b_{0i}^N and b_{1i}^N even when b_{0i}^* and b_{1i}^* were initially independent.

When $\mathbf{U^m}$ is not a confounder, $(\hat{\beta}_e^N, \hat{\beta}_{te}^N)$ is an unbiased estimate of $(\beta_e^*, \beta_t e^*)$ from Equation (2), and under the assumption that $E(\mathbf{U^m}) = 0$, $E(Y_{ij}|X_i, \mathbf{Z_i}, \mathbf{U_i}, t_{ij}) = E(Y_{ij}|X_i, \mathbf{Z_i}, \mathbf{U_i^o}, t_{ij})$. However, when $\mathbf{U^m}$ includes confounders, $E(Y_{ij}|X_i, \mathbf{Z_i}, \mathbf{U_i}, t_{ij}) \neq E(Y_{ij}|X_i, \mathbf{Z_i}, \mathbf{U_i^o}, t_{ij})$ since $E(b_{0i}^N|X_i, \mathbf{Z_i}, \mathbf{U_i^o}, t_{ij}) \neq 0$ and $E(b_{1i}^N|X_i, \mathbf{Z_i}, \mathbf{U_i^o}, t_{ij}) \neq 0$, and $(\hat{\beta}_e^N, \hat{\beta}_{te}^N)$ is not an unbiased estimator of $(\beta_e^*, \beta_{te}^*)$ anymore.

2.3 Instrumental variable approach

The two-stage IV methodology aims at correcting the bias due to residual unmeasured confounding. We show here how it can be adapted to the longitudinal framework described above by replacing the second-stage least square regression by a second-stage linear mixed model.

For clarity, we distinguish below the case of a continuous endogenous exposure from the case of a binary endogenous exposure. The method relies on the independence between the regressors ($\mathbf{Z}$, $\mathbf{U^o}$) and the unobserved variables $\mathbf{U^m}$. As this assumption may likely be violated between $\mathbf{U^m}$ and $\mathbf{U^o}$, we consider below the total vector $\mathbf{U} = (\mathbf{U^m}, \mathbf{U^o})$ as being unobserved to ensure independence.

X continuous

With a continuous endogenous exposure, the two-stage methodology is defined as follows:

$$X_i = \alpha_0 + \mathbf{Z_i}^\top \alpha_{\mathbf{Z}} + e_i^X \tag{4}$$

$$\begin{aligned} Y_{ij} = \beta_0 &+ E(X_i|\mathbf{Z_i})\beta_e + b_{0i} \\ &+ (\beta_t + E(X_i|\mathbf{Z_i})\beta_{te} + b_{1i})t_{ij} + \epsilon_{ij}^Y \end{aligned} \tag{5}$$

This model relies on the same distributions and independence assumptions as model (2).

From the IV conditional independence assumption (3), the conditional expectation $E(X_i|\mathbf{Z_i}) = \tilde{X}_i = \alpha_0^* + \mathbf{Z_i}^\top \alpha_{\mathbf{Z}}^*$ and the residual $X_i - E(X_i|\mathbf{Z_i}) = \mathbf{U_i}^\top \alpha_{\mathbf{U}}^* + \epsilon_i^{X*}$.

When rewritting Equation (2) according to $E(X_i|\mathbf{Z_i})$, one obtains:

$$\begin{aligned} Y_{ij} &= \beta_0^* + X_i\beta_e^* + \mathbf{U_i}^\top \beta_{\mathbf{U}}^* + b_{0i}^* + \\ &\quad \left(\beta_t^* + X_i\beta_{te}^* + \mathbf{U_i}^\top \beta_{\mathbf{tU}}^* + b_{1i}^*\right) t_{ij} + \epsilon_{ij}^{Y*} \\ &= \beta_0^* + E(X_i|\mathbf{Z_i})\beta_e^* + (X_i - E(X_i|\mathbf{Z_i}))\beta_e^* + \mathbf{U_i}^\top \beta_{\mathbf{U}}^* + b_{0i}^* + \\ &\quad \left(\beta_t^* + E(X_i|\mathbf{Z_i})\beta_{te}^* + (X_i - E(X_i|\mathbf{Z_i}))\beta_{te}^* + \mathbf{U_i}^\top \beta_{\mathbf{tU}}^* + b_{1i}^*\right) t_{ij} + \epsilon_{ij}^{Y*} \end{aligned} \tag{6}$$

And using that $X_i - E(X_i|\mathbf{Z_i}) = \mathbf{U_i}^\top \alpha_{\mathbf{U}}^* + \epsilon_i^{X*}$ from model (1),

$$\begin{aligned} Y_{ij} &= \beta_0^* + E(X_i|\mathbf{Z_i})\beta_e^* + (\mathbf{U_i}^\top \alpha_{\mathbf{U}}^* + \epsilon_i^{X*})\beta_e^* + \mathbf{U_i}^\top \beta_{\mathbf{U}}^* + b_{0i}^* + \\ &\quad \left(\beta_t^* + E(X_i|\mathbf{Z_i})\beta_{te}^* + (\mathbf{U_i}^\top \alpha_{\mathbf{U}}^* + \epsilon_i^{X*})\beta_{te}^* + \mathbf{U_i}^\top \beta_{\mathbf{tU}}^* + b_{1i}^*\right) t_{ij} + \epsilon_{ij}^{Y*}. \end{aligned} \tag{7}$$

which reduces to:

$$Y_{ij} = \beta_0^* + E(X_i|\mathbf{Z_i})\beta_e^* + b_{0i} + (\beta_t^* + E(X_i|\mathbf{Z_i})\beta_{te}^* + b_{1i})\, t_{ij} + \epsilon_{ij}^{Y*} \tag{8}$$

with $b_{0i} = \mathbf{U_i}^\top(\alpha_{\mathbf{U}}^*\beta_{\mathbf{e}}^* + \beta_{\mathbf{U}}^*) + \epsilon_i^{X*}\beta_e^* + b_{0i}^*$ and $b_{1i} = \mathbf{U_i}^\top(\alpha_{\mathbf{U}}^*\beta_{\mathbf{te}}^* + \beta_{\mathbf{tU}}^*) + \epsilon_i^{X*}\beta_e^* + b_{1i}^*$. By definition, $E(X_i|\mathbf{Z_i})$ and $\mathbf{U_i}$ are independent so $\mathbf{b_i} = (b_{0i}, b_{1i})^\top$ is independent of the covariates in the model, as required in a linear mixed model. The model defined in Equation (5) is thus equivalent to the target model in Equation (2), except that the variance of the random-effects is not homoscedastic anymore.

Maximum likelihood estimates of the fixed effects in a mixed model being unbiased even when the covariance structure is misspecified (following the same principle as with generalized estimating equations [23]), $\hat{\beta}_e$ and $\hat{\beta}_{te}$ are unbiased estimators of β_e^* and β_{te}^*; they may be used to quantify the causal relation between X and Y. However, their variance needs to be corrected for the heteroscedasticity and the use of an IV. By applying the same principle of robust variances [24, 25] as in IV methods for cross-sectional studies (e.g. in ivtools R package [26]), we define the following sandwich estimator:

$$V_{\text{2-S}}\left(\hat{\boldsymbol{\beta}}\right) = \left(\sum_{i=1}^{N} \hat{\mathbf{W}}_\mathbf{i}^\mathbf{T} \hat{\mathbf{V}}_\mathbf{i}^{-\mathbf{1}} \hat{\mathbf{W}}_\mathbf{i}\right)^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{W}}_\mathbf{i}^\mathbf{T} \hat{\mathbf{V}}_\mathbf{i}^{-\mathbf{1}} \mathbf{V}_\mathbf{i} \hat{\mathbf{V}}_\mathbf{i}^{-\mathbf{1}} \hat{\mathbf{W}}_\mathbf{i}\right) \left(\sum_{i=1}^{N} \hat{\mathbf{W}}_\mathbf{i}^\mathbf{T} \hat{\mathbf{V}}_\mathbf{i}^{-\mathbf{1}} \hat{\mathbf{W}}_\mathbf{i}\right)^{-1} \tag{9}$$

where $\hat{\mathbf{W}}_\mathbf{i}$ is the matrix of variables associated with the vector of fixed effects $\boldsymbol{\beta}$ (in our example in equation (5), $\hat{\mathbf{W}}_\mathbf{i}$ is a $n_i \times 4$- matrix with intercept, time, $E(X_i|\mathbf{Z_i})$ and its interaction with time, and $\beta = (\beta_0, \beta_t, \beta_e, \beta_{te})^\top$), $\hat{\mathbf{V}}_\mathbf{i} = M_i \hat{\mathbf{B}} M_i^T + \hat{\sigma}_y^2 I_{n_i}$ with M_i the matrix of variables related to the random effects (in our example a $n_i \times 2$ with intercept and time), I_{n_i} is the identity matrix and, $\hat{\beta}$, $\hat{\mathbf{B}}$, $\hat{\sigma}$ are the estimates obtained in the second-stage model (5). Finally $\mathbf{V_i}$ is the empirical covariance matrix of Y, that is $\mathbf{V_i} = \text{Cov}(\mathbf{Y_i} - \mathbf{W_i}^\top \hat{\beta}, \mathbf{Y_i} - \mathbf{W_i}^\top \hat{\beta})$ where $\mathbf{W_i}$ is the $n_i \times 4$ matrix with intercept, time, X_i and its interaction with time.

The robust variance $V_{\text{2-S}}\left(\hat{\beta}\right)$ quantifies the second stage variability in the estimates but it neglects the first-stage uncertainty. To compute the total variance that accounts for the variability in the two stages, we use a parametric bootstrap [27]: instead of running the 2^{nd}-stage analysis once from the maximum likelihood estimates $\hat{\alpha}$, the 2^{nd}-stage is replicated M times from 1^{st}-stage parameters α_m $(m = 1, .., M)$ randomly drawn from their asymptotic normal distribution with mean $\hat{\alpha}$ and variance $\widehat{V(\hat{\alpha})}$. The total variance estimate of $\hat{\boldsymbol{\beta}}$ can then be derived with the Rubin's rule [28] from the M 2^{nd}-stage estimates $\hat{\boldsymbol{\beta}}_m$ as:

$$V_{\text{tot}}(\hat{\boldsymbol{\beta}}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{V_{\text{2-S}}(\hat{\boldsymbol{\beta}}_m)} + \frac{(M+1)(M-1)}{M} \sum_{m=1}^{M} \left(\hat{\boldsymbol{\beta}}_m - \overline{\hat{\boldsymbol{\beta}}_m}\right) \left(\hat{\boldsymbol{\beta}}_m - \overline{\hat{\boldsymbol{\beta}}_m}\right)^\top$$

X binary

The absence of bias demonstrated for the continuous exposure comes from the use of additive models in both stages. Although not frequent, a linear model could also be considered for a binary exposure. Called linear probability model [13], it translates into the exact same inference technique as described for the continuous exposure with $E(X_i|Z_i)$ derived from a linear model for X and included into the second-stage linear mixed model, and the same variance estimator.

Alternatively, the more classical logistic model can also be considered:

$$\text{logit}(E(X_i|\mathbf{Z_i})) = \alpha_0 + \mathbf{Z_i}^\top \alpha_\mathbf{Z} \tag{10}$$

with the derived $E(X_i|Z_i)$ included in the second-stage linear mixed model in (5), and the same total variance estimator used. However, due to the nonlinear nature of the logistic regression, $E(X_i|Z_i, U_i)$ does no longer equal $E(X_i|Z_i)$, and the convergence of the estimates of β_e and β_{te} to β_e^* and β_{te}^* in (2) is not ensured anymore. To further account for the residual effect of the unmeasured confounders, some authors recommended to replace the substitution of X by $E(X_i|Z_i)$ by the combination of X and the residual $X - E(X_i|Z_i)$ in the second-stage. We call these three options linear/substitution, logistic/substitution and logistic/residual-inclusion, respectively.

2.4 Software

The IV estimation technique for a binary or continuous time-fixed exposure and a continuous repeatedly measured outcome is implemented in the R package **IVmm** available at *url of the package - blinded version*. It relies on the hlme function of lcmm R package for the linear mixed model estimation [29].

3 SIMULATION STUDY

We ran a simulation study to illustrate the behaviour of the naive approach and of the IV methods in the presence of unmeasured confounding.

3.1 Simulation design

The simulation setting followed the DAG of Figure 1 B. The procedure of data generation including parameters values considered are fully summarized in supplementary Table S1. For each individual i in a sample of size N, we first generated an exogenous instrumental variable Z_i and an unobserved confounder U_i according to standard Gaussian distributions, and random visit times $t_{ij} = j + u_{ij}$ around theoretical annual visits j (with $j = 1, .., 6$) with u_{ij} a visit-and-subject-specific random Gaussian departure ($\mathcal{N}(0, 0.05)$). We then generated the endogenous continuous exposure X_i according to model (4) (for a binary, a logistic version of (4) was considered) and the repeated measures of the outcome Y_i according to model (2).

We considered scenarios with different sample sizes (N=2000, 6000 or 20,000) and different strengths of association between the IV and the exposure α_z resulting in different strengths of the instrumental variable. As common in the IV literature, the strength of association between the IV and the exposure was quantified with the F-statistic (ratio of the explained variance and the residual variance) [30] and the Nagelkerke R^2 for a continuous and binary exposure, respectively. For each scenario, 500 datasets were simulated.

3.2 Simulation results

Table 1: Simulation results for continuous exposure (over 500 replicates) for the association between the exposure and the trajectory of Y (summarized by the effect on the baseline level and the slope over time) according to the sample size, and strength of the instrumental variable (α_Z).

		$\alpha_Z = 0.5$					$\alpha_Z = 1$				
			baseline level		slope over time			baseline level		slope over time	
N	Methods	Strength*	RB	CR	RB	CR	Strength*	RB	CR	RB	CR
2000	Naive	-	44.3	0.0	44.3	0.0	-	33.3	0.0	33.2	0.0
	IV	251	-0.1	93.6	0.3	95.6	1003	0.1	96.8	0.1	95.6
6000	Naive	-	44.5	0.0	44.5	0.0	-	33.4	0.0	33.3	0.0
	IV	757	0.9	95.4	0.4	95.0	3003	-0.1	96.8	-0.1	96.2
20000	Naive	-	44.4	0.0	44.5	0.0	-	33.3	0.0	33.3	0.0
	IV	2503	0.08	96.2	-0.0	94.6	10009	-0.0	95.2	0.0	93.4

* Strength of association is assessed with the F-statistic for continuous X

Abbreviations: N=sample size, RB=Relative bias (defined as the average percentage of difference between the estimate and the true parameter value), CR=Coverage rate of the 95% confidence interval

The results of the naive and the IV approaches are reported in Tables 1 and 2; they are also displayed in Figure 2 for the slope with time (and in Supplementary Figure S1 for the initial level).

As expected, whatever the sample size and the strength of the IV association with the exposure, the naive method showed very large bias and null coverage rate for the association between the exposure and the change over time in all cases. In contrast, the 2-stage IV methods retrieved the true causal association without any bias for the continuous exposure, and for the binary exposure when using the linear/substitution and logistic/substitution methods, even for the scenarios with a weak instrument. In contrast, the logistic/residual methodology for a binary exposure showed

Table 2: Simulation results for binary exposure with naive method, linear/substitution and logistic/substitution IV methods (over 500 replicates) for the association between the exposure and the trajectory of Y (summarized by the effect on the baseline level and the slope over time) according to the type of exposure, the sample size, and strength of the instrumental variable (α_Z).

		$\alpha_Z = 2$					$\alpha_Z = 3$					$\alpha_Z = 4$				
			baseline level		slope over time			baseline level		slope over time			baseline level		slope over time	
N	Methods	Str*	RB	CR	RB	CR	Str*	RB	CR	RB	CR	Str*	RB	CR	RB	CR
2000	Naive	-	135.9	0.0	135.5	0.0	-	106.9	0.0	106.7	0.0	-	67.6	0.0	67.7	0.0
	Log/Res	14.3	100.3	0.0	100.2	0.0	35.0	82.7	0.0	82.5	0.0	58.6	67.9	0.0	67.7	0.0
	Log/Sub	14.3	-1.6	94.6	-2.0	95.2	35.0	-0.8	94.8	-1.4	95.4	58.6	-0.4	94.6	-1.0	95
	Lin/Sub	10.3	-1.0	95.4	-1.4	95.4	25.1	-0.1	96.0	-0.1	93.8	41.6	0.0	94.0	0.2	94.0
		(229)					(676)					(1406)				
6000	Naive	-	135.9	0.0	135.5	0.0	-	106.7	0.0	106.3	0.0	-	68.0	0.0	67.8	0.0
	Log/Res	14.3	100.4	0.0	100.2	0.0	35.0	82.4	0.0	81.8	0.0	58.6	-21.6	0.0	16.2	0.0
	Log/Sub	14.3	-1.3	94.6	-1.2	93.8	35.4	-1.0	94.6	-0.9	94.0	58.6	-0.7	94.0	-0.7	94.4
	Lin/Sub	10.3	-1.0	94.8	-0.1	95.4	25.1	-0.6	96.8	-0.4	96.4	41.6	-0.1	93.0	0.2	96.0
		(692)					(2025)					(4218)				
20000	Naive	-	135.7	0.	135.7	0.0	-	106.7	0.0	106.8	0.0	-	67.9	0.0	67.9	0.0
	Log/Res	14.3	100.4	0.0	100.4	0.0	35.0	82.2	0.0	82.3	0.0	58.6	67.4	0.0	67.4	0.0
	Log/Sub	14.3	-0.3	93.8	0.0	95.6	35.4	-0.6	93.8	-0.3	95.6	58.6	-0.5	94.0	-0.4	95.4
	Lin/Sub	10.3	-0.6	94.0	-0.2	95.0	25.1	0.2	93.8	0.2	94.6	41.6	-0.2	94.6	-0.1	94.6
		(2301)					(6763)					(14037)				

* Strength of association is assessed with the R^2 expressed in % (and F-statistic) for the linear regression, and the R^2 of Nagelkerke for the logistic regression also expressed in %.

Abbreviations: N=sample size, RB=Relative bias expressed in % (defined as the average percentage of difference between the estimate and the true parameter value), CR=Coverage rate expressed in % of the 95% confidence interval, Str = Strength, Log/Sub = Logistic/Substitution Method, Lin/Sub = Linear/Substitution Method

large bias and null coverage rate. In the following, we thus did not investige this method further. The simulation study also validated the proposed estimate of variance with reported coverage rate of the 95% confidence interval very close the nominal value in both the continuous and binary case. However, although correct, the 2-stage IV method showed substantial variability in the estimates when the IV was weaker.

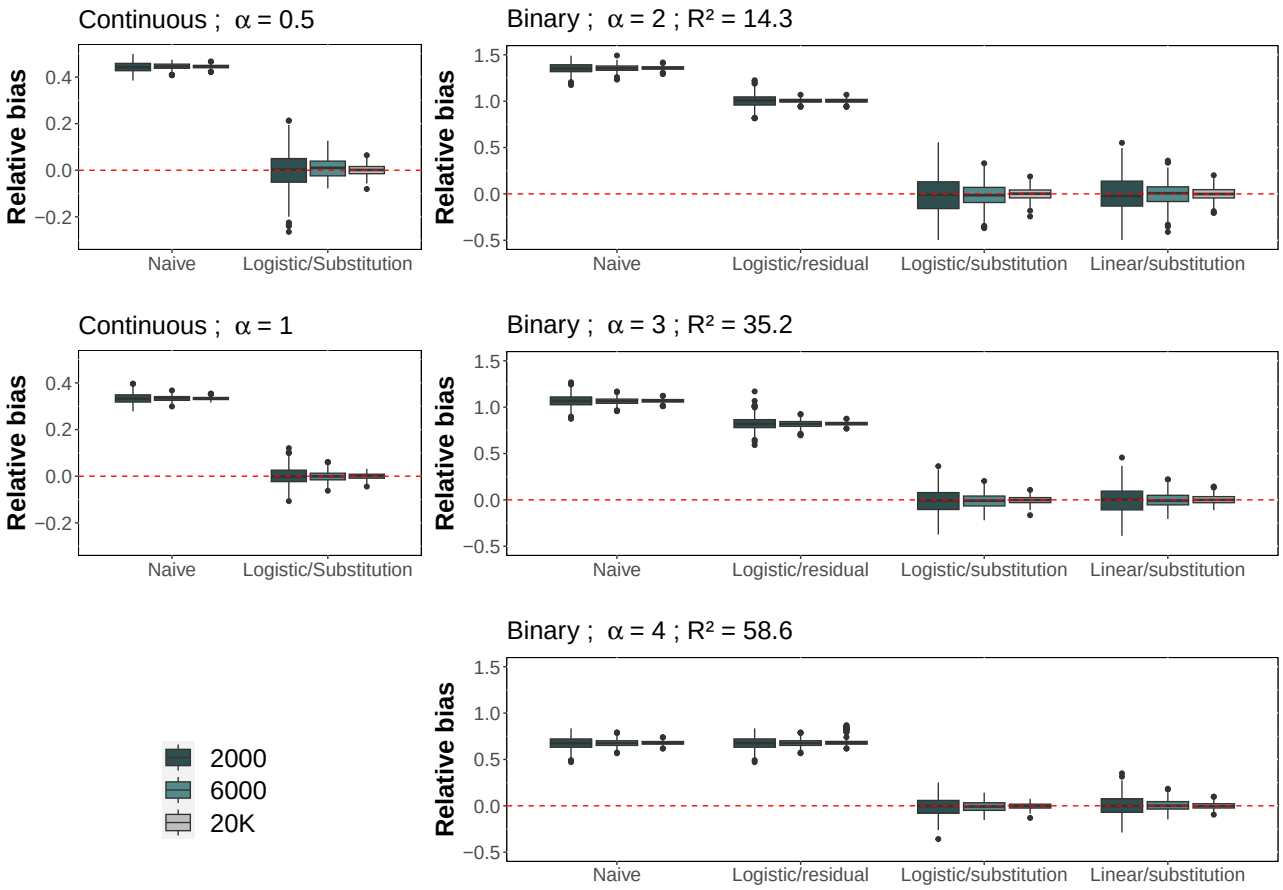

Figure 2: Association estimates (over 500 replicates) of the continuous exposure or the binary exposure with the change of the outcome over time using the naive or the IV approaches (logistic/residual, linear/substitution and logistic/substitution in the binary case) for different sample sizes (N) and different intensities of association (through the regression coefficient α). In the binary case only, the Nagelkerke R^2 is also reported to further illustrate the strength of the IV in comparison with the application setting.

4 APPLICATION

We aimed to assess the relation between type-2 diabetes measured at baseline and subsequent cognitive trajectory in the elderly population. Indeed, biological mechanisms suggest an implication of type-2 diabetes on cognitive aging [31] but unmeasured confounders can interfere with this process. To handle this, we used a genetic instrumental variable defined by the 42 single nucleotide polymorphisms (SNP) (listed in supplementary materials) that were previously identified in genome-wide association studies of type-2 diabetes [32, 18].

4.1 The Three-City study

The 3C study is a population-based prospective cohort which aimed at assessing the relation between vascular diseases and dementia in the elderly [20]. Participants, aged 65 years and older, were randomly selected in 1999 from the electoral lists of three French cities. In total, 9,294 participants underwent an in-depth examination of their health and risk factors at baseline, and were then followed every 2-3 years for up to 20 years with an extensive interview and a neuropsychological battery. Among them, 6,948 participants have been typed on genome-wide genotyping arrays and

further imputed from Haplotype Reference Consortium panel [33]. Genotype data were retained in the study are those with an imputation quality greater than 0.70. Type-2 diabetes was determined from blood glucose level (fasting glucose level$\geq$ 7.0 mmol/L) or the use of antidiabetic treatment at baseline. We studied the cognitive trajectory through the Isaacs Set test (IST), which measures verbal fluency and has been shown to differentiate early in the pathological process towards dementia [34]. The score is the total number of words given in four semantic categories in 15 seconds.

The final sample size included 6,224 participants whose type-2 diabetes was ascertained at baseline, who were genotyped, and had at least one IST measure during the follow-up. Participants were 74 years old at baseline on average, 61 % were women and 38% had an educational level higher than secondary school (Table 3). Among them, 598 (9.6 %) were ascertained with diabetes at baseline; those with diabetes were more often male, more likely to have a low educational level. Participants were followed up for 8 years on average with a mean of 4 repeated measures of IST.

Table 3: Characteristics of the 6224 participants of 3C sample according to their type-2 diabetes and overall

Characteristics	Diabetics (N=598)		No diabetics (N=5626)		Overall (N = 6224)	
	Number (%)	Mean (SD)	Number (%)	Mean (SD)	Number (%)	Mean (SD)
Sex						
female	285 (47.7)		3498 (62.2)		3783 (60.8)	
male	313 (52.3)		2128 (37.8)		2441 (39.2)	
Education level						
no education	78 (13.0)		458 (8.1)		536 (8.6)	
primary school	112 (18.7)		924 (16.4)		1036 (16.7)	
secondary school	218 (36.5)		2086 (37.1)		2304 (37.0)	
high school	99 (16.6)		1138 (20.2)		1237 (19.9)	
university	91 (15.2)		1020 (18.1)		1111 (17.9)	
Age at entry		74.44 (5.4)		74.29 (5.5)		74.31 (5.5)
IST score at baseline		30.48 (6.8)		32.24 (7.0)		32.08 (7.0)
Number of IST measures/subject		4.06 (1.8)		4.47 (1.9)		4.42 (1.9)
Years of follow-up		7.08 (4.6)		8.12 (4.8)		8.02 (4.7)

Abbreviations: N=sample size, IST=Isaacs Set Test, SD=standard deviation

4.2 The IV analysis

We primarily used the logistic/substitution method. The R^2 of 4.8% showed a weak association between type 2 diabetes and genetic polymorphisms. The linear mixed model for the IST trajectory included a basis of four natural cubic splines on the time from baseline to account for the nonlinear trajectories over time. Diabetic status (in the naive model) or its expectation based on the 42 polymorphisms (in the IV model) was included in interaction with each spline function. For the naive model, we considered both no adjustment or adjustment on measured potential confounders (educational level, age at baseline). Parameter estimates are given in Supplementary Table S2. Predicted trajectories of IST according to diabetic status are displayed in Figure 3(a) (corresponding differences over time between groups in Figure 3(b)).

The naive method, whether it was adjusted or not for potential confounders, highlighted a difference at inclusion according to the type-2 diabetes but no differential change over time. At any time, the mean IST score was lower for participants with type-2 diabetes than for those without type-2 diabetes (mean difference in the adjusted model of -1.20 [-1.77;-0.64], -1.36 [-1.94;-0.79], -1.31 [-1.84;-0.78] points at 0, 5 and 10 years). In contrast, the logistic/substitution IV method did

not show evidence of substantial difference in cognitive trajectory according to the type-2 diabetes although the point estimates suggested a higher level at baseline for participants with type-2 diabetes (mean difference of 1.26 [-2.66;5.18] points at baseline) and a steeper cognitive decline in the first years for participants with type-2 diabetes (mean difference of -1.20 [-5.50;3.10], -0.48 [-5.51;4.55] points at 5 and 10 years, respectively). Results were similar when using the linear/substitution IV model (see Supplementary Figure S3).

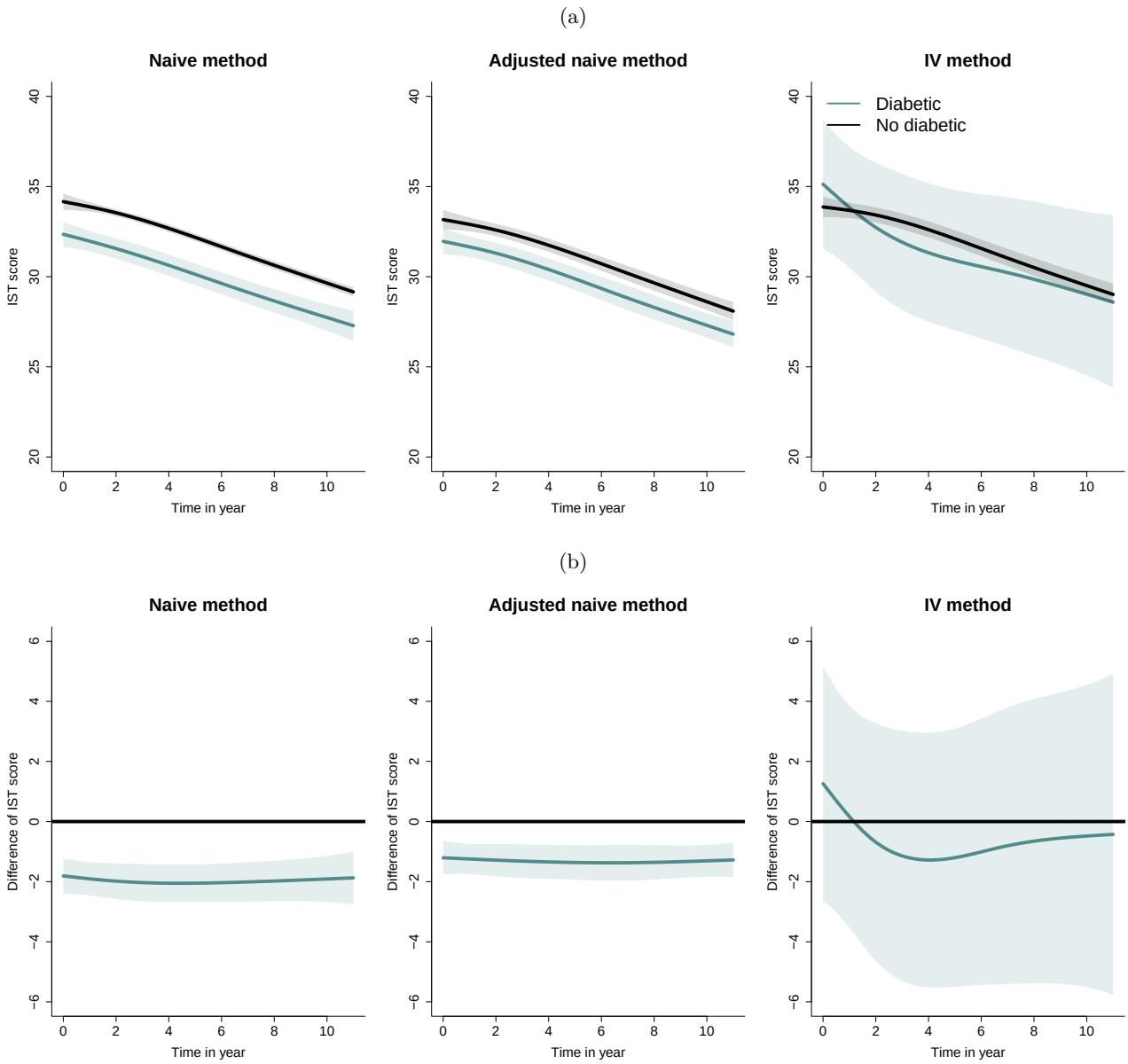

Figure 3: (a). Predicted trajectories of IST score according to type-2 diabetes at baseline and associated 95% confidence interval. (b). Estimated difference in IST score over time for diabetic compared to non-diabetic using the naive method (not adjusted or adjusted on gender, educational level and age) and the logistic/substitution instrumental variable method

5 DISCUSSION

The IV method has gained interest in observational studies to address unmeasured confounding. Yet, although the framework is very common in observational longitudinal studies, an IV solution for the assessment of an exposure collected at baseline on the subsequent trajectory of a repeated outcome had not been previously described in the medical statistics literature. We showed in this work how the two-stage approach frequently used in IV methodology for cross-sectional or survival outcomes [11, 18] could be adapted to study the association between a time-fixed exposure and the subsequent trajectory of an outcome using the mixed model theory. Previous contributions dealing with repeated data over time had systematically focused on time-dependent exposures (rather than time-fixed) and associations with either the level of a time-fixed outcome [35] or the level of a repeated outcome at a given time using distributed lag models [16, 17]. To our knowledge, the use of a mixed model with an instrumental variable approach in epidemiology was limited to the analysis of a complex clinical trial to treat non-compliance over time [36], the issue of measurement error of time-dependent exposures with regression calibration [37] and the issue of between/within unmeasured confounding in cross-sectional grouped data [15].

The conducted simulation study emphasized the highly biased estimations obtained when ignoring unmeasured confounding. They also showed the correct inference our IV solution could provide for assessing the causal association between a time-fixed continuous or binary exposure and a continuous longitudinal outcome in the presence of endogeneity. However, we noticed a very high variance for moderate sample sizes (a few thousand subjects) when the IV was weakly associated with the exposure. For simplicity of result reporting, we focused in the methodology and in the simulations on scenarios with a linear trajectory for the outcome. However, the methodology applies equivalently to any scenario with a nonlinear trajectory provided the mixed model remains linear in the fixed and random effects, and random effects are included for each time function. This is what was done in the application considering natural splines to approximate the nonlinear cognitive trajectory.

The IV methodology highly relies on additive model properties to eliminate the association with the unmeasured confounders. The use of nonlinear models may prevent from a total elimination of this association and induce biased estimates. When considering a binary exposure, we explored linear and nonlinear regressions. Our simulations showed that the causal association could be correctly retrieved when using the linear probability model for the binary exposure but also when using the nonlinear logistic model combined with a substitution method in the second-stage. In the application, both methods also gave the same results. In contrast, the logistic regression combined with the residual inclusion in the second stage [12] showed large bias in our simulation setting with a linear mixed model in the second stage and was not further investigated. Regarding the outcome, we restricted our framework to continuous longitudinal outcomes with linear mixed models and leave extensions to other types of outcomes to future research.

Our motivating application aimed at evaluating the causal association between type-2 diabetes and cognitive decline by using 42 genetic polymorphisms associated with type-2 diabetes as IV. While the classical (naive) regression ignoring unmeasured confounders highlighted a lower cognitive level for type-2 diabetics at all times, the IV methodology which handles unobserved confounding suggested a different and time-varying association. However, the analysis by IV does not allow to reach a conclusion as the confidence intervals were excessively large because of the limited sample size for an IV application with a binary exposure (N=6224), and the weakness of the association between genetic polymorphisms and type-2 diabetes ($R^2 = 4.8\%$). These results were similar when considering logistic and linear models in first step.

MR studies had already been conducted to assess the causal association between type-2 diabetes and cerebral aging. Cross-sectional studies had focused on cognitive level [38] and dementia risk

[39, 40], and one longitudinal survival study had investigated the association with dementia risk [18]. None had identified a causal association between genetically-predicted type-2 diabetes and cerebral aging. Our work goes one step further by considering the association with prospective cognitive decline. Although in accordance with the literature, the highly variable results call for a replication in a much larger sample to overcome a potential lack of power. Additional simulations based on a similar instrument as in our application (Supplementary Figure S2) showed the substantial the gain in accuracy when considering for instance 20,000 subjects rather than 6000 subjects.

The method we proposed relies on assumptions coming from both the IV theory and the mixed model theory. First, the method is based on the fundamental assumptions that define valid instruments: (1) Z is strongly associated with X; (2) Z is associated with Y only through X; (3) Z is independent of U conditionally on X (Figure 1). In our application as in many MR analyses, the genetic IV explains only a small part of the exposure (assumption (1)) leading to a weak instrument, high variances and need for very large sample sizes. The simulation study did not reveal any issue of bias or coverage rate with weak instruments. However, it showed a huge variability that can make the IV method inconclusive, except when carried out on very large samples (20,000 subjects for instance in our case). To better address assumption (1) and not rely on a pre-determined set of IVs, Fan and Zhong [41] proposed an adaptive lasso technique that simultaneously selects the IV variables from a high-dimensional set of candidates. Developed for cross-sectional data, an extension to longitudinal outcome data using our mixed modeling strategy could be possible.

As fixed at birth, the genetic IV can not be affected by the confounders (Assumption 3). However, to guarantee assumptions (2) and (3), we further need to assume that the SNPs associated with type-2 diabetes are not associated with other diseases (pleiotropy). Moreover, the use of genetic variants as an IV for a later in life study relies on the implicit assumption that the genetic variants are not associated with the probability to be alive at the timing of eligibility definition, exposure and outcome collection [42, 43]. Our application was performed under the assumption that genetic polymorphisms and type-2 diabetes were not associated with mortality prior to cohort entry. Finally, causal interpretation of the IV analysis requires a fourth assumption, either the homogeneity for the average causal effect or monotonicity for the local average causal effect [9, 10].

Note that with binary exposures, the interpretation of IV analyses may not be straightforward, especially when the binary exposure reflects an underlying continuous process that should be considered instead [44]. This is however unlikely the case with diabetes. In particular its definition differs from blood glucose since a diabetic person under treatment may be controlled for hyperglycemia.

Our methodology also relies on classical assumptions of longitudinal analyses. We considered the linear mixed model theory rather than marginal models as they better handle selection over time for etiological studies [45]. Our methodology is robust to missing data under the missing at random mechanism (i.e., missingness can be fully determined by the observations) [46] for both the intermittent missing outcome and study dropout. In case of informative dropout linked to the outcome process, the methodology can be easily extended by jointly modeling the risk of dropout according to the trajectory of the outcome [47]. In the application, we performed such a sensitivity analysis where death and dropout from the study were modelled along with the cognitive decline (Supplementary Table S3); it showed concordant results.

To conclude, we provided a full methodology and associated software solution to apply the IV technique to the frequent framework of an exposure measured at baseline and the subsequent trajectory of a continuous marker. It must be used with caution due to the strong and hardly controllable assumptions IV methods must satisfy. However, as illustrated with the causal association between type-2 diabetes and cognitive decline, it constitutes a useful statistical tool to take into account unobserved confounders in prospective cohort studies.

Acknowledgments

Computer time was provided by the computing facilities MCIA (Mesocentre de Calcul Intensif Aquitain) at the University of Bordeaux and the University of Pau and Pays de l'Adour.

Funding

This work was funded by the French National Research Agency (Project DyMES - ANR-18-CE36-0004-01) and was carried out in the framework of the INSERM GOLD Cross-Cutting program. D.-A.T. is supported, in part, by the EPIDEMIOM-VT Senior Chair from the University of Bordeaux initiative of excellence Initiative d'Excellence.

Competing interests

The authors declare that they have no conficts of interest. This paper have not been published previously in whole or in part.

Data availability

Scripts for replicating the application and the simulation runs are provided in supplementary material. This replication material does not include the application data. Anonymized data can only be shared by reasonable motivated request to the 3C scientific committee (not to the authors).

References

1. Ertefaie A, Small DS, Flory JH, Hennessy S. A tutorial on the use of instrumental variables in pharmacoepidemiology. *Pharmacoepidemiology and Drug Safety.* 2017;26(4):357–367.

2. Hernan MA, Robins JM. *Causal Inference : What if.* Boca Raton: Chapmann & Hall/CRC 2020.

3. Fewell Z, Davey Smith G, Sterne JAC. The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology.* 2007;166(6):646–655.

4. Rawlings AM, Sharrett AR, Schneider AL, *et al.* Diabetes in midlife and cognitive change over 20 years: the Atherosclerosis Risk in Communities Neurocognitive Study. *Annals of internal medicine.* 2014;161(11):785–793.

5. Wagner M, Dartigues JF, Samieri C, Proust-Lima C. Modeling Risk-Factor Trajectories When Measurement Tools Change Sequentially During Follow-up in Cohort Studies: Application to Dietary Habits in Prodromal Dementia. *American Journal of Epidemiology.* 2018;187(4):845–854.

6. Wright PG. *The Tariff on Animal and Vegetable Oils.* Macmillan 1928.

7. Greenland S. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology.* 2000;29(4):722–729.

8. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ (Clinical research ed.).* 2018;362:k601.

9. Swanson SA, Hernán MA. The challenging interpretation of instrumental variable estimates under monotonicity. *International Journal of Epidemiology.* 2018;47(4):1289–1297.

10. Hernán MA, Robins JM. Instruments for Causal Inference: An Epidemiologist's Dream?. *Epidemiology.* 2006;17(4):360–372.

11. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research.* 2017;26(5):2333–2355.

12. 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.

13. Li C, Poskitt DS, Windmeijer F, Zhao X. Binary outcomes, OLS, 2SLS and IV probit. *Econometric Reviews.* 2022;41(8):859–876.

14. Li Y, Lee Y, Port FK, Robinson BM. The impact of unmeasured within- and between-cluster confounding on the bias of effect estimatorsof a continuous exposure.

15. Li J, Fine J, Brookhart A. Instrumental variable additive hazards models. *Biometrics.* 2015;71(1):122–130.

16. O'Malley AJ. Instrumental variable specifications and assumptions for longitudinal analysis of mental health cost offsets. *Health Serv Outcomes Res Methodol.* 2012;12(4):254–272.

17. Hogan JW, Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Stat Methods Med Res.* 2004;13(1):17–48.

18. Tchetgen Tchetgen EJ, Walter S, Vansteelandt S, Martinussen T, Glymour M. Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.).* 2015;26(3):402–410.

19. Martínez-Camblor P, MacKenzie TA, Staiger DO, Goodney PP, James O'Malley A. An instrumental variable procedure for estimating Cox models with non-proportional hazards in the presence of unmeasured confounding. *Journal of the Royal Statistical Society: Series C (Applied Statistics).* 2019;68(4):985–1005.

20. Alperovitch A. Vascular Factors and Risk of Dementia: Design of the Three-City Study and Baseline Characteristics of the Study Population. *Neuroepidemiology.* 2003;22:316–325.

21. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics.* 1982;38(4):963–974.

22. Commenges D, Jacqmin-Gadda H. *Modèles biostatistiques pour l'épidémiologie.* De Boeck Superieur 2015. Google-Books-ID: twEtDwAAQBAJ.

23. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986;73(1):13–22.

24. White H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica.* 1980;48(4):817–838.

25. Royall RM. The Prediction Approach to Robust Variance Estimation in Two-Stage Cluster Sampling. *Journal of the American Statistical Association.* 1986;81(393):119–123.

26. Sjolander A, Martinussen T. Instrumental Variable Estimation with the R Package ivtools. *Epidemiologic Methods.* 2019;8(1):20180024.

27. Efron B, Tibshirani R. *An introduction to the bootstrap.* No. 57 in Monographs on statistics and applied probabilityNew York: Chapman & Hall 1993.

28. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* John Wiley & Sons 2019. Google-Books-ID: BemMDwAAQBAJ.

29. Proust-Lima C, Philipps V, Liquet B. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software.* 2017;78:1–56.

30. Andrews I, Stock JH, Sun L. Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annu. Rev. Econ..* 2019;11(1):727–753. Publisher: Annual Reviews.

31. Frison *Diabète et risque de démence.* Thèse de doctorat, Spécialité Santé Publique, option épidémiologieUniversité de Bordeaux 2019.

32. 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):981–990.

33. Lambert JC, Heath S, Even G, *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature Genetics.* 2009;41(10):1094–1099.

34. Amieva H, Mokri H, Le Goff M, *et al.* Compensatory mechanisms in higher-educated subjects with Alzheimer's disease: a study of 20 years of cognitive decline. *Brain : a journal of neurology.* 2014;137(Pt 4).

35. Sánchez BN, Kim S, Sammel MD. Estimators for longitudinal latent exposure models: examining measurement model assumptions. *Statistics in medicine.* 2017;36(13):2048–2066.

36. Bond SJ, White IR, Sarah Walker A. Instrumental variables and interactions in the causal analysis of a complex clinical trial. *Statistics in Medicine.* 2007;26(7):1473–1496.

37. Strand M, Sillau S, Grunwald GK, Rabinovitch N. Regression calibration for models with two predictor variables measured with error and their interaction, using instrumental variables and longitudinal data. *Stat Med.* 2014;33(3):470–487.

38. Ware EB, Morataya C, Fu M, Bakulski KM. Type 2 Diabetes and Cognitive Status in the Health and Retirement Study: A Mendelian Randomization Approach. *Frontiers in Genetics.* 2021;12:634767.

39. Østergaard SD, Mukherjee S, Sharp SJ, *et al.* Associations between Potentially Modifiable Risk Factors and Alzheimer Disease: A Mendelian Randomization Study. *PLoS Medicine.* 2015;12(6).

40. Walter S, Marden JR, Kubzansky LD, *et al.* Diabetic Phenotypes and Late-Life Dementia Risk: A Mechanism-specific Mendelian Randomization Study. *Alzheimer Disease & Associated Disorders.* 2016;30(1).

41. Fan Q, Zhong W. Nonparametric Additive Instrumental Variable Estimator: A Group Shrinkage Estimation Perspective. *Journal of Business & Economic Statistics.* 2018;36(3):388–399.

42. Swanson SA. A Practical Guide to Selection Bias in Instrumental Variable Analyses. *Epidemiology.* 2019;30(3):345–349.

43. Vansteelandt S, Dukes O, Martinussen T. Survivor bias in Mendelian randomization analysis. *Biostatistics (Oxford, England).* 2018;19(4):426–443.

44. Burgess S, Labrecque JA. Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. 2018. arXiv:1804.05545 [stat].

45. Rouanet A, Avila-Rieger J, Dugravot A, *et al.* How Selection Over Time Contributes to the Inconsistency of the Association Between Sex/Gender and Cognitive Decline Across Cognitive Aging Cohorts. *American Journal of Epidemiology.* 2022;191(3):441–452.

46. Little RJA, Rubin DB. *Statistical Analysis With Missing Data.* Wiley 1987.

47. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data.* Boca Raton: Routledge1er édition ed. 2012.