Calculating control variables with age at onset data to adjust for conditions prior to exposure

Michael Höfler, Tanja Brueck, Roselind Lieb, Hans-Ulrich Wittchen

Abstract
Background: When assessing the association between a factor X and a subsequent outcome Y in observational studies, the question that arises is what are the variables to adjust for to reduce bias due to confounding for causal inference on the effect of X on Y. Disregarding such factors is often a source of overestimation because these variables may affect both X and Y. On the other hand, adjustment for such variables can also be a source of underestimation because such variables may be the causal consequence of X and part of the mechanism that leads from X to Y.
Methods: In this paper, we present a simple method to compute control variables in the presence of age at onset data on both X and a set of other variables. Using these age at onset data, control variables are computed that adjust only for conditions that occur prior to X. This strategy can be used in prospective as well as in survival analysis. Our method is motivated by an argument based on the counterfactual model of a causal effect.
Results: The procedure is exemplified by examining of the relation between panic attack and the subsequent incidence of MDD.
Conclusions: The results reveal that the adjustment for all other variables, irrespective of their temporal relation to X, can yield a false negative result (despite unconsidered confounders and other sources of bias).

Key words: confounding, causality, causal inference, age at onset, logistic regression, survival analysis, mental disorders, epidemiology

Introduction

A frequent aim in clinical and epidemiological studies is to examine the association between two variables, say X (e.g., panic attack) and Y [e.g., major depression (MDD)]. Researchers thereby often aim at separating "real" from "spurious" relations between these phenomena. This distinction is closely related to the question of what other factors should be matched or controlled for in statistical models. The meaning of "real" or "spurious", however, often remains obscure unless it arises from the investigator's substantive aims. As we will see, for causal inference, X has to precede Y, and other factors should only be considered if they precede X. The temporal order, for instance between exposure X and outcome Y, can be established (a) in an experiment where X is a manipulated condition, (b) in a prospective cohort study where X is measured at an assessment prior to the assessment of Y, or (c) by collecting age at onset data on both X and Y in a cross-sectional study. Clearly, (a) is preferable to (b), and (b) is preferable to (c) with regard to minimizing measurement errors. In this paper, we propose a simple method to calculate control variables to adjust for factors occurring before X. For this purpose, age at onset data on X and other variables are required. These other variables contain events that precede X and events that follow X. Age at onset data allows the establishment of the temporal relation. In the next section, the counterfactual model of a causal effect is reviewed, which provides the basis for our adjustment method. After discussing some aspects of bias due to confounding, the proposed method is introduced. The last section illustrates our procedure through an example on the relation of panic attack with subsequent onset of MDD.

The counterfactual model of a causal effect

We assume that X is a binary factor, with the values X=0 and X=1 representing two different treatment or exposure levels to be compared (e.g., X=1, one or more panic attacks vs X=0, no panic attack). Let i denote the index for an individual. The standard model for a causal effect in an individual i at fixed time of assignment to X=1 vs X=0 is based on the difference in an outcome Yi under Xi=1 (Yi1) and Xi=0 (Yi0):
Yi1 – Yi0.
For a binary outcome, this model is equivalent to the occurrence of the event Yi=1 under one but not under the other condition. In the example above, an individual would develop subsequent MDD if she or he had a panic attack up to a fixed time, but did not develop MDD in the absence of panic attack by that time. Panic attack would have caused MDD in that individual. It has been argued that the counterfactual model of a causal effect captures most aspects of causal questions in health studies [7, 11].
Clearly, for a fixed individual at a fixed time, the outcome is observable only for one of both exposure levels. A person either has a panic attack up to a certain point in time or not. Thus, the outcome under the unobserved or counterfactual condition (counterfactual or potential outcome) has to be estimated. Note that there are usually no objective criteria to determine individual causal effects in psychiatric studies. Hence, the best thing to do is often to estimate population average effects in the target population while aggregating several individuals. Only under several assumptions, including lack of confounding (see below), absence of measurement, and selection bias, the average causal effect of X on Y can be estimated without bias [12, 19].

Bias due to confounding

In observational studies, the assignment to X=0 vs X=1 (e.g., no panic attack vs panic attack) is not randomized, and the difference in Y (e.g., subsequent onset of MDD) between X=1 vs X=0 is typically biased as an estimate of the causal effect of X on Y. One reason is that a vector of other variables Z (here, e.g., prior disorders, childhood conditions, genetic factors) may affect both X and Y (bias due to confounding; e.g., [19]). In practice, however, usually not all components of Z are observed or even known. This is also the case for the relation between panic attack and subsequent MDD. Moreover, it is often impossible to decide whether adjusting for a certain observed component Zk reduces the bias in the estimation of the causal effect or not. This is, for instance because one does not know how the remaining components of Z distribute across X=0 and X=1 before and after adjustment for Zk (e.g., [15]). However, there exist criteria for sets of confounders that are sufficient to adjust for [6, 15].
We will focus only on the observed components Z1,..., Zm of Z assuming that controlling for these variables reduces the bias due to confounding. Note that for normally distributed outcomes, adjustment that does not alter the bias due to confounding is nevertheless recommended because of variance reduction [19]. For non-normally distributed outcomes such as binary events in logistic regression or survival analysis, however, this adjustment rather leads to unnecessarily broad confidence intervals (CIs) [19].
In practice, one observes a set of variables other than the exposure X and the outcome Y, denoted with W1,..., Wm. These might be related to both X and Y. For simplicity, the following discussion assumes that there is only one such variable, denoted with Wk. There are two commonly used strategies to address this problem. The first strategy is to ignore Wk. Not considering potential bias due to confounding caused by other confounders and other sources of bias, this strategy tends to overestimate the causal effect if Wk is related either positively or

negatively with both X and Y. The second strategy is to adjust for Wk irrespective of its temporal relationship with X. This may underestimate the causal effect because Wk might be the causal consequence of X and may itself affect Y. Here, controlling for Wk would partition away a part of the mechanism that leads from X to Y (again disregarding other potential confounders and other sources of bias). For a more sophisticated presentation of these issues, see the literature on graphical models (e.g., [2, 23]) and on causal graphs (e.g., [6, 14, 15]).

Calculation of control variables

Often, the temporal relation among different variables is unknown or it varies across individuals. For mental disorders and syndromes, however, information about their temporal ordering is often provided through age at onset data. With instruments like the Composite International Diagnostic Interview [24], the age at onset is assessed retrospectively for all disorders for which the diagnostic criteria are met.

Prospective analyses

We propose a simple method to adjust only for events that occur before the exposure X because events that happen between exposure and outcome might be part of the causal mechanism leading from X to Y. Suppose that we want to carry out a prospective analysis of the following kind: Assume that the exposure X represents a condition (e.g., panic attack) assessed at the baseline investigation of a prospective study. X is the lifetime status of that condition; X=1 indicates that an individual has met the criteria for the condition until the age at assessment; X=0 means that diagnostic criteria were not met by the age at assessment, and thus, no age at onset is available. For individuals with X=1, the age at onset of X, denoted with AOX, is observed in terms of completed age years (e.g., age at onset of panic attack). The outcome Y is defined as first incidence of another condition (e.g., MDD) during the follow-up period of a study. Thus, all individuals meeting the criteria for Y already at baseline assessment are omitted from the analysis. Finally, other conditions W1,..., Wm are measured at baseline assessment, each of which, Wk (e.g., specific phobia), having an age at onset variable AOk (among those with Wk=1).
The aim is now to calculate a control variable Zk for each pair of variables Wk and AOk. With such a control, we aim to adjust individuals with X=0 to those with X=1 with respect to the occurrence of disorder k before the age at onset of disorder X (AOX) (e.g., persons with and without panic attack with respect to prior occurring specific phobia). Herewith, the counterfactual behavior of individuals with X=0 is estimated with respect to the outcome Y (e.g., incidence of major depression) if they had the same rate of disorder k before the age at which X "typically" begins in individuals with X=1. Thus, among individuals with X=0, the control variable is coded as "being present" (Zk=1) only if it occurs prior to the typical age at onset of X. Note that coding all individuals with X=0 and Wk=1 as Zk=1 would tend to underestimate the causal effect of X on Y because a longer risk period for developing Wk=1 (namely, the entire lifetime period until baseline assessment) would be assigned to individuals with X=0. For individuals with X=1, Zk is coded as 1 if the individual age of onset of Wk precedes the one of X. Adjustment will then be done by entering the control variables Z1, Z2,..., Zm into the regression equation (e.g., logistic regression, see below) as main effect terms.
For the typical age at onset, we choose the median (medX) instead of the mean because the median is more robust against outliers or unrealistic values, which are sometimes observed in age at onset data (e.g., age at onset <5 years what, however, cannot be definitely ruled out to

be true for some disorders). Such unusual values can have a strong impact on the mean particularly in small samples, a drawback not shared by the median.

Coding of the control variables

We propose to code Zk as follows:
Among individuals with X=1:
Zk=1 if Wk=1 (disorder k is present) and AOk≤ AOX (age at onset precedes or equals individual onset of X),
Zk=0 if Wk=0 (disorder k is not present) or AOk> AOX (age at onset later than the individual age at onset of X).
Among individuals with X=0:
Zk=1 if Wk=1 (disorder k is present) and AOk≤ medX (age at onset precedes or equals the median age at onset of X computed among all cases with disorder X),
Zk=0 if Wk=0 (disorder k is not present) or AOk> medX (age at onset later than the median age at onset of X).
A problem with this method is how to handle "tied" observations with AOk=AOX (or AOk=medX, respectively), that is, individuals where the disorder Wk occurred at the same age than disorder X (or at the same age than medX, respectively). In general, it would be more appropriate to randomize the coding here (code Zk=0 or Zk=1 with equal probabilities). However, this method is more difficult to implement, and correct statistical inference would require carrying out multiple randomizations (i.e., repeat the random coding for tied observations). Each randomization would yield a new dataset, and statistical inference would be based on a kind of "average" and a joint variance estimate across these multiple datasets as done in inference based on multiple missing data imputation [21]. The method used here to code Zk as 1 if AOk=AOX (or AOk=medX, respectively) might tend to yield conservative results for the effect of X on Y (despite other biases) because Wk might have occurred after X (or medX, respectively) in some of these individuals, thus possibly being the consequence, not the cause of X.
For studies with a broad age range of respondents at the baseline assessment, we propose to apply our coding strategy stratified by birth year groups. This accounts for variation in age at onset according to the year of birth. This variation might be caused by "agecensoring" (for young individuals, a high age at onset cannot be observed, see below) or by cohort effects.

Survival analysis

In some applications, it will be preferable to use survival analysis, for example, Cox regression (e.g., Thernau and Grambsch 2000), instead of carrying out a prospective analysis. In survival analysis, the data are organized as follows: For each individual, different rows in the dataset correspond to different years of observation. The number of rows is given by (a) the age at onset of Y if the event coded with Y=1 has ever happened to the individual or (b) the age at assessment if that event has never happened. For instance, suppose an individual 20 years of age at the assessment has developed MDD at the age of 15. This person would have 15 lines in the dataset indicating the age periods of 0–1, 1–2,..., and 14–15 years. The remaining age period after onset of depression (16–20 years) in that individual is not informative for the question which factors affect the onset of Y. The rows in the dataset also contain the values of the covariates, which may depend on age (see below).
Survival analysis has the advantage that it accounts for age censoring in the outcome Y; that is, the individuals differ in age at assessment. For example, if Y is the onset of MDD, some individuals at lower age might turn to MDD cases after the time of assessment (and up to the

maximum age observed in the study). Age censoring is accounted for by assigning a higher weight to higher age periods in older individuals—age periods that are unobserved for younger individuals. The variable X, which is to be evaluated against the subsequent onset of Y, might also depend on age. Imagine that an individual experiences the first panic attack at the age of 10. Before that age, the person is a non-case of lifetime panic attack, and from that age, the person is a case. As X has to precede Y, the definition of X(age) has to be shifted backward by 1 year. Hence, X(age) should be coded as follows:

X(age)=1 if the condition was ever present by age-1, X(age)=0 otherwise. The situation is analogous to the one in prospective analyses described above at fixed age: The covariate is the presence of disorder X before the specific age, and the outcome is the incidence of Y in the subsequent year of age. One can compute overall or age-specific differences in onset rates of Y between X(age)=1 and X (age)=0, for example, with hazard ratios from Cox regression (having an analogous interpretation as risk ratios). Control variables Wk(age) can be calculated exactly in the same way as above, but separately for each year of age. Note that more sophisticated methods also take feedback mechanisms into account: The outcome Yat age t might not only be affected by factors occuring before t (confounding), but it may also impact later time-dependent causes of the outcome Y after t [3, 16, 17]. Our approach is restricted to the onset of prior conditions and disregards such feedback mechanisms. However, it is easy to conduct with standard software (as compared to g-estimation which requires special software) and improves the predominant strategies to adjust for all or none other disorders.

The differences between survival and prospective analysis are rather interpretational ones. First, in survival analysis, the temporal relation between X and Y (not only the one between X andW1,...,Wm) also has to be assessed with age at onset data and cannot be established by the presence of two assessments. Establishing the temporal order by age at onset data is likely to be more prone to errors. On the other hand, the prospective analysis is restricted to the individuals who never had the disorder Y at baseline. Depending on the age at assessment, many individuals with an early onset might be lost for the analysis. Likewise, the results would be restricted to those with a late onset, those who might constitute a special group of individuals.

The effect of using different adjustment strategies: an example from the EDSP Study

Data for the following analysis were taken from the Early Developmental Stages of Psychopathology (EDSP) Study [9, 27]. The EDSP Study is a prospective, general population study on mental and substance use disorders among adolescents and young adults in Greater Munich, Germany. We assessed Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnoses (DSM-IV; American Psychiatric Association [1]) using the Munich version of the CIDI (M-CIDI, DIA-X; [25]). The probands were between 17 and 28 years old at the last assessment (T2).

We analyzed the prospective association between panic attack at the baseline investigation (T0) and the onset of MDD during the follow-up period (T1/T2). The average follow-up time was 41.7 months (range 34.1–49.6 months, SD=3.0). Among the older cohort (18–24 years old at T0), the follow-up period was assessed at T2 only, and among the younger cohort, T1 and T2 together covered the entire follow-up period. A total of 2,548 participants completed T0 and T2, of whom N=326 met the criteria for MDD or a hypomanic or manic episode at baseline, leaving N=2,222 individuals for the analysis. As the study was designed with special interest in early stages of mental and substance use disorders, probands 14–15 years old at T0 were sampled at twice the probability of individuals 16–21 years old, and 22- to 24-year-olds were sampled at half this probability. To account for these intended differences in sampling

probabilities as well as non-response at T0 according to age, sex, and location, data were weighted to match the distribution of these variables in the registry database of the source population. This was done with an iterative procedure resulting in (weighted) distributions of age, sex, and location that approximately equal those in the population [7, 9, 27].

For the calculation of control variables, we used the following seven diagnoses, including substance use disorders and the most frequent anxiety disorders: alcohol dependence, alcohol abuse, nicotine dependence, abuse or dependence of any illicit substance, social phobia, any specific phobia, and phobia NOS. The control variables were calculated stratified by the age groups 14–15, 16–17, 18–19, 20–21, and 22–24 years at T0 and were entered as main effect terms into the model equation. Odds ratios (ORs) were computed from logistic regressions, and robust confidence intervals that accounted for the weighting scheme are based on the Huber–White sandwich estimator [20]. Analyses were carried out with the LOGISTIC procedure of Stata 8.0 [22].

Table 1 is the cross-table between panic attack at baseline and the onset of an MDD at follow-up. Only a weighted 8.4% of the individuals without panic attack at baseline met the criteria for an incident MDD at follow-up, compared to 20.7% of those with at least one panic attack at baseline.

As shown in Table 2, the OR adjusted for sex and age was 2.67 (95% CI=1.34–5.32). After further adjusting for the seven disorders at baseline while disregarding their temporal relation with panic attack, the OR decreased considerably to 1.80 (95% CI=0.89–3.67) and was no longer statistically significant.

When adjusting for these disorders only if they occurred prior to panic attack, using control variables as defined above, an OR emerged that lay between the above two ORs (2.11, 1.02–4.37), again being statistically significant. However, results should not be reduced to a dichotomy, and requiring p<0.05 is nothing but a social convention. Adjusting only for precedent disorders yields a higher OR with a higher left boundary of the CI, thus, more evidence for a causal effect (Höfler and Seaman, unpublished paper) than if we adjusted for all disorders irrespective of their age at onset. Note that this moderate association might easily be explained by unconsidered confounders, other sources of bias (e.g., measurement and selection), and uncertainty on these biases because the left boundary of the confidence interval is close to 1 [5, 18, Höfler and Seaman, unpublished paper].

Comment

In this article, we proposed a method of adjusting for factors prior to the exposure factor under investigation (X) using age at onset data. The performance of this method depends strongly on psychometric properties of age at onset data. Wittchen et al. [26] examined validity and reliability of age at onset data from the CIDI. The study used different raters at both occasions and a duration of 1 to 3 days between the assessments. Age at onset questions yielded at least fair agreement for most questions among DSM-III cases (intraclass correlations of 0.69 or higher; besides questions on "nervousness", "mania 2", "first drunk" and "depression 2"). In the same paper, validity was assessed among former psychiatric inpatients over a 7-year follow-up period and was compared to the age at the index admission. Again, here among ICD cases, intraclass correlations were at least satisfactory for most questions (besides among "neurotic" individuals). However, it should be mentioned that the agreement rates (equal age at onset at both occasions) for both reliability and validity were sometimes lower than 50%. Studies on age at onset of substance use have shown that individuals tended to shift the reported age at onset toward the time of assessment ("forward telescoping"; Johnson et al. [4, 8, 13]. Hence, there seems to be substantial uncertainty in the exact reporting of age at onset, especially when the onset occurred a long time ago. This is

primarily of a concern when the temporal relation between two disorders with similar ages at onset is to be established.

Hence, the possibility to use age at onset data for covariate adjustment as proposed in this paper does not replace the necessity of studies that establish the temporal relation of X to other variables through the study design because "you can't fix by analysis what is screwed up in design" [10]. However, when such studies are not feasible, our method provides a sound procedure to account for conditions prior to the exposure X.

Table 1 Association between panic attack at baseline assessment (T0) and incidence of major depression (MDD) at follow-up (T1, T2), N=2,222

Onset of MDD at follow-up	Panic attack baseline					
	No		Yes		Total	
	N	%w[a]	N	%w	N	%w
No	1,969	91.6	52	79.3	2,021	91.2
Yes	187	8.4	14	20.7	201	8.8

[a] %w indicates weighted percentage

Table 2 Results of different models for the association between a panic attack and subsequent major depression

	OR	95% CI	p Value
Association adjusted for sex and age[a]	2.67	1.34–5.32	0.005
Association adjusted for sex, age, and ten disorders[a] at baseline	1.80	0.89–3.67	0.102
Association adjusted for sex, age, and seven substance use and anxiety disorders[b] at baseline prior to panic attack	2.11	1.02–4.37	0.043

[a] OR from logistic regression
[b] These include alcohol dependence, alcohol abuse, nicotine dependence, abuse or dependence of any illicit substance, social phobia, and any specific phobia and phobia NOS

References

1. American Psychiatric Association (1994) Diagnostic and statistical manual of mental disorders, 4th edn. Washington DC

2. Cox DR, Wermuth N (1996) Multivariate dependencies. Models, analyses and interpretation. Chapman and Hall, London

3. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL (2002) Analysis of longitudinal data, 2nd edition, chap. 12. Oxford University Press, Oxford

4. Golub A, Johnson BD, Labouvie E (2000) On correcting biases in self-reports of age at first substance use with repeated cross-section analysis. J Quant Criminol 16:45–68

5. Greenland S (2005) Multiple-bias modelling for analysis of observational data. J Roy Stat Soc, Series A 168:267–291

6. Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiological research. Epidemiology 10:37–48

7. Höfler M, Pfister H, Lieb R, Wittchen HU (2005) The use of weights to account for non-response and drop-out. Soc Psychiatry Psychiatr Epidemiol 40:291–299

8. Johnson RA, Gerstein DR, Rasinski KA (1998) Adjusting survey estimates for response bias. Public Opin Q 62:354–377

9. Lieb R, Isensee B, Sydow von K, Wittchen HU (2000) The early developmental stages of the psychopathology study (EDSP): a methodological update. Eur Addict Res 6:170–182

10. Light RJ, Singer JD, Willett JB (1990) By design: planning research on higher education. Harvard University Press, Cambridge, MA

11. Little RJA, Rubin DB (2000) Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annu Rev Public Health 21:121–145

12. Maclure M, Schneeweiß S (2001) Causation of bias: the episcope. Epidemiology 12:114–122

13. Parra GP, O'Neill SE, Sher KJ (2003) Reliability of self-reported age of substance involvement onset. Psychol Addict Behav 3:211–218

14. Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–710

15. Pearl J (2000) Causality-models, reasoning and inference. Cambridge University Press, Cambridge

16. Robins JM, Blevins D, Ritter G, Wulfsohn M (1992) G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology 13:319–336

17. Robins JM, Hernan MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11:550–560

18. RosenbaumPR (2002) Observational studies, 2nd edn. Springer, Berlin Heidelberg New York

19. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal inference. Biometrika 70:41–55

20. Royall RM (1986) Model robust confidence intervals using maximum likelihood estimators. Int Stat Rev 54:221–226

21. Rubin DB, Schenker N (1991) Multiple imputation in health care databases: an overview and some applications. Stat Med 10:585–598

22. StataCorp. (2003) Stata statistical software: release 8.0. Stata Corporation, College Station, TX

23. Trichtler D (1999) Reasoning about data with directed graphs. Stat Med 18:2067–2076

25. WHO (1997) Composite international diagnostic interview (CIDI, version 2.1). World Health Organisation, Geneva

24. Wittchen HU, Pfister H (eds) (1997) DIA-X-Interviews. Manual für Screening-Verfahren und Interview; Interviewheft Längsschnittuntersuchung (DIA-X Lifetime); Ergänzungsheft (DIA-X Lifetime); Interviewheft Querschnittsuntersuchung (DIA-X 12 Monate); Ergänzungsheft (DIA-X 12 Monate); PC-Programm zur Durchführung der Interviews (Längs- und Querschnittsuntersuchung); Auswertungsprogramm, Frankfurt, Swets & Zeitlinger

26. Wittchen HU, Burke JD, Semler G, Pfister H, Cranach Von M, Zaudig M (1989) Recall and dating of psychiatric symptoms. Arch Gen Psychiatry 46:437–443

27. Wittchen HU, Perkonigg A, Lachner G, Nelson CB (1998) Early developmental stages of psychopathology study (EDSP): objectives and design. Eur Addict Res 4:18–27