International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 16(2): 77–87 (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.205

Estimating causal effects from observational data with a model for multiple bias

MICHAEL HÖFLER,[1] ROSELIND LIEB,[2] HANS-ULRICH WITTCHEN[1]

1 Technische Universität Dresden, Institute of Clinical Psychology and Psychotherapy, Dresden, Germany
2 University of Basel, Institute of Psychology, Basel, Switzerland

Abstract

Conventional analyses of observational data may be biased due to confounding, sampling and measurement, and may yield interval estimates that are much too narrow because they do not take into account uncertainty about unknown bias parameters, such as misclassification probabilities. We used a simple, multiple bias adjustment method to estimate the causal effect of social anxiety disorder (SAD) on subsequent depression. A Monte Carlo sensitivity analysis was applied to data from the Early Developmental Stages of Psychiatry (EDSP) study, and bias due to confounding, sampling and measurement was modelled. With conventional logistic regression analysis, the risk for depression was elevated in the presence of SAD only in the older cohort (age 17–24 years at baseline assessment); odds ratio (OR) = 3.06, 95% confidence interval (CI) 1.64–5.70, adjusted for sex and age. The bias-adjusted estimate was 2.01 with interval limits of 0.61 and 9.71. Thus, given the data and the bias model used, there was considerably more uncertainty about the real effect, but the probability that SAD increases the risk for subsequent depression (OR > 1) was 88.6% anyway. Multiple bias modelling, if properly used, reveals the necessity for a better understanding of bias, suggesting a need to conduct larger and more adequate validation studies on instruments that are used to diagnose mental disorders. Copyright © 2007 John Wiley & Sons, Ltd.

Key words: Causality, observational studies, causal effect, bias, confounding, measurement error, selection bias, mental disorders

Introduction

A causal effect of a factor in an individual can be simply defined as the presence of a difference in a subsequent outcome between the two factor levels. In the present study we attempted to address bias when estimating the effect of social anxiety disorder (SAD) on the probability of subsequently developing depression. Imagine a person with SAD who subsequently develops depression that would not have occurred if the person did not previously have SAD. For this person SAD caused depression. Having not had SAD instead of having had SAD is referred to as a 'counterfactual condition', and the associated outcome (no depression) is be referred to as a 'counterfactual' or 'potential' outcome (see Höfler, 2005a and references therein). When referring to counterfactual causality it is crucial to consider

- whether all individuals with SAD could have had no SAD, and
- whether all individuals who did not have SAD could have had SAD.

The first condition can be assumed because SAD could have been prevented with cognitive behavioural methods. The second condition is uncertain because there could be a genetic vulnerability, that is necessary to develop SAD, which may be lacking in some individuals. Therefore, the analysis reported in the present study refers mainly to a population of persons with SAD. Individuals who do not have SAD are considered to be 'substitutes' for individuals with SAD but not necessarily vice versa. To obtain stable estimates, one must include many individuals under both conditions. Therefore, because an individual can only be observed

Copyright © 2007 John Wiley & Sons, Ltd

under one condition, only population average effects can be estimated. Even if there is no average effect, there might be individual effects that cancel out. This concept of counterfactual causality has become standard in epidemiology (Höfler, 2005a and references therein). It reflects the viewpoint intervention on causality upon which public health decisions depend. For instance: should one intervene in cases of SAD to prevent depression? That is, would intervention in SAD decrease the incidence of subsequent depression? It is important to note that the effects of intervention depend heavily on the mode of intervention in a multivariate framework (Greenland, 2005a). For example, intervention in SAD may be more effective if it occurs at a younger age and even more effective if the incidence of SAD is fully prevented.

Bias

Estimates of causal effects are often biased, for example due to measurement error. For instance, a disorder may not be measured although it is present (Höfler, 2005b). Randomized clinical trials tend to yield lower bias than observational studies but they are often costly and difficult to conduct. A common strategy is to screen for associations between factors and a particular disorder or disease in cross-sectional or case-control studies, which are the easiest and least costly studies, and to further investigate any identified associations with a prospective study. Finally, if the associations are confirmed, a randomized study may be conducted.

Unfortunately, bias and random error can mask effects at any of those stages, so the strategy might fail if bias is not sufficiently considered or understood. An intricate, although invaluable, approach is to simultaneously conduct a randomized trial and an observational study on the same population. This was done in the Women's Health Initiative (Prentice et al., 2005). If bias in the randomized trial is low, it provides a valid basis with which to estimate bias in the observational data. Unbiased estimates of causal effects are only guaranteed if various assumptions are true, including:

- randomized exposure, which guarantees that individuals are exchangeable across conditions despite random error;
- random selection of individuals from the target population, upon which inference is to be made;
- random occurrence of missing information within levels of controlled covariates; and
- absence of any measurement error in exposure, confounder, and outcome variables (for example, Rosenbaum and Rubin, 1983; Holland, 1986; Little and Rubin, 2000; Maclure and Schneeweiß, 2001; Maldonado and Greenland, 2002; Höfler, 2005a). In observational studies, participants are not randomly assigned to exposure levels.

Thus, results may be confounded by factors that affect both exposure and outcome (for example, Rothman and Greenland, 1998). The exposure and the outcome variable may be measured with error, whereas in randomized studies with perfect compliance there is no measurement error for exposure. In experimental and observational studies, individuals are often not randomly selected from the target population and, if they are, there are frequently systematic non-responses or dropouts that may introduce bias (Höfler et al., 2005). Confounding, selection, and measurement often constitute the major sources of bias in observational studies and, in the present study, bias adjustment is restricted to those three types of bias. Some authors have serious concerns about any attempts to estimate causal effects from observational data. However, many of the concerns appear to be of a semantic nature. We do not claim that we are able to remove bias or establish causation by adjusting for supposed bias. However, we can develop multiple bias models that reflect available knowledge about bias and we can make defendable assumptions about bias parameter distributions and use them to compute the distribution of the unknown effect given the data and the bias model used. We can also apply several bias models that reflect a realistic sample of all possible scenarios of bias to determine how sensitive the results of one particular bias model are. These issues were summarized by Greenland (2005b):

> I regard any causal analysis of observational data (or randomized trial with major compliance problems) as just a piece of sensitivity analysis; it is the piece in which results are obtained under the particular assumptions of that analysis. Because we never know that the assumptions are correct (and in fact would wisely doubt them), we had better try more than one analysis.

In the current study, a method that is relatively easy to implement was applied to address bias when estimating the causal effect of SAD on subsequent depression

Copyright © 2007 John Wiley & Sons, Ltd

(major depression or dysthymia), using data from the Early Developmental Stages of Psychiatry (EDSP) study. Our aim was to demonstrate if and how the results of the study changed based on meaningful assumptions made about bias due to confounding, selection, and measurement.

Conventional analyses and methods to address bias

Using conventional statistical methods, data are analysed as if they were generated in a randomized experiment with perfect compliance. That is, subjects are selected randomly from a target population, there is no measurement error, and so forth. Such conventional analyses include well-known methods such as the chi-squared test for independence, analysis of variance and logistic regression. Usually, the only corrections made in epidemiological studies of mental disorders are adjustments for potential confounders like age and sex in regression models and the use of sampling weights to reduce bias due to selection (Höfler et al., 2005). These adjustments, however, do not address confounding and sampling bias due to unconsidered factors or measurement error. In defence of conventional analysis methods, results can be interpreted, although not causally, by expressing them in statements like 'Individuals that meet the criteria for disorder *X* subsequently meet the criteria for disorder *Y* more frequently than would be expected by chance.' It is not clear what 'expected by chance' means if individuals were not randomly selected or randomly assigned to groups (Greenland, 1990). Moreover, researchers are hardly interested in crude associations; they are interested in whether and to what extend *X causes Y* (Soldani et al., 2005). Conventional estimates of causal effects can be false in two ways. First, point estimates may be biased for the reasons mentioned above. Second, interval estimates are often much too narrow because they do not take any uncertainties about unknown bias parameter values into account (for example, misclassification probabilities). The more uncertain the bias parameter values the more uncertain a causal estimate should turn out to be (Greenland, 2003, 2005c). Bias that is not accounted for may yield false positive conclusions or push an estimate toward the null value of no effect. Bias parameters can, at best, be estimated. For instance misclassification probabilities can be estimated from validation studies.

The degree to which bias parameter values are uncertain depends on the depends on the size of the sample from which they are assessed or, if they have to be guessed at, the degree of subjective variance. Importantly, the relative impact of uncertainty on bias parameter values increases as the sample size increases, because random error decreases while uncertainty remains constant (Greenland, 2005c). For instance, suppose that the sample size for the causal analysis increases but the sample size from which misclassification probabilities are estimated remains the same. In such a case, random variation decreases but uncertainty about misclassification probabilities (systematic variation) remains unchanged. In the vast majority of papers, discussions about bias are based on intuition only. If bias is evaluated numerically, however, such discussions frequently turn out to be inappropriate (Greenland, 2004, 2005c).

Methods to correct for bias and take into account uncertainty about bias parameter values include Bayesian bias models and Monte Carlo sensitivity analysis (MCSA). Bayesian methods directly compute the probability distribution of an unknown causal effect given the data and a bias model. Bias models include the types of bias that are supposed to exist, assumptions about how those biases act together, and uncertainties about bias parameter values. These uncertainties are summarized in 'prior distributions'.

Monte Carlo sensitivity analysis (Greenland, 2001, 2004) is easier to implement than Bayesian models, and its results approach those of Bayesian models if the estimator of a causal effect is efficient, the data provide no information about the bias parameters, and the MCSA is modified as described below (Greenland, 2005c). Importantly, these methods account for understood biases only, and results may still be biased by misunderstood or unknown sources of bias (Greenland, 2003, 2005c). Note that conventional analyses, if regarded from the Bayesian point of view, are based on wrong point prior distributions at zero on the parameters that produce bias. That is, they assume that all bias parameter values (for example, misclassification probabilities) equal 0 with 100% certainty (Greenland, 2005c).

Data

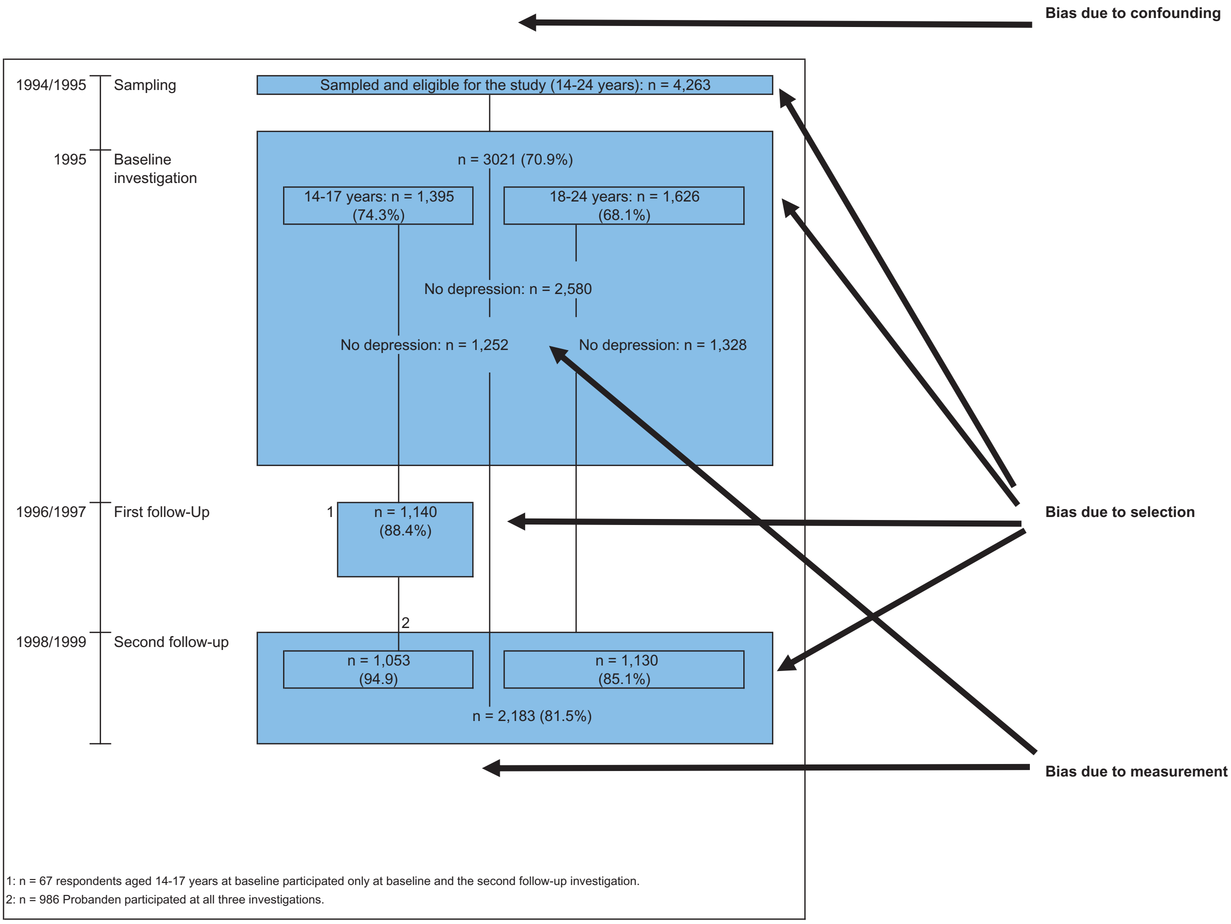
We used data from the EDSP study to estimate the causal effect of SAD on depression. The EDSP is a prospective study of the general population that examined the early developmental stages of mental and substance-use disorders including risk and vulnerability

Copyright © 2007 John Wiley & Sons, Ltd

factors (Wittchen et al., 1998a,b; Lieb et al., 2000). Individuals were randomly sampled from the greater Munich area, Germany. Diagnoses of mental disorders were based on DSM-IV (American Psychiatric Association, 1994) criteria assessed with the Munich version of the Composite International Diagnostic Interview (M-CIDI; Wittchen and Pfister, 1997).

At baseline investigation (T0) study participants were between the ages of 14 and 24 years. The first follow-up assessment (T1) took place 19.7 ± 2.3 (SD) months after T0, and was completed only by those participants who were 14–17 years of age at T0. The second follow-up assessment (T2) which was 41.7 ± 3.0 (SD) months after T0 and was to be completed by all probands. T1 and T2 assessments together encompass the entire T0–T2 follow-up period. Figure 1 summarizes the sampling procedure and non-response and conditional drop-out rates and indicates where bias is likely to occur.

At T0, 441 of the 3021 participants met the criteria for lifetime major depression, dysthymia, hypomanic episodes, or manic episodes and were excluded from analyses. At T2 397 individuals dropped out and were also omitted from analyses. Thus, data from 2183 participants were available for analysis. At T0, SAD also included individuals who did not fulfil the DSM-IV impairment criterion. Depression at follow-up (T1 or T2) was defined as meeting criteria for major depression or dysthymia. Most of the interviewers who performed the T0 SAD assessment were different from those who performed the T1 and T2 depression assessments. Furthermore, the T1 and T2 interviewers were blind to participant T0 SAD status. The M-CIDI ensures high objectivity because the interviewers have no control over the course of an interview. The sequence of questions is fully determined by previous answers, and diagnoses are calculated according to fixed algorithms.



1: n = 67 respondents aged 14-17 years at baseline participated only at baseline and the second follow-up investigation.
2: n = 986 Probanden participated at all three investigations.

Figure 1. EDSP study design and potential biases.

Copyright © 2007 John Wiley & Sons, Ltd
DOI: 10.1002/mpr

Bias model

According to the comprehensive sequential model of bias by Maclure and Schneeweiß (2001), containing 11 sources of bias, bias due to confounding is the first to occur (among the three kinds of bias considered here) followed by bias due to selection and bias due to measurement. This suggests that bias should be accounted for in the inverse order i.e. bias due to measurement, selection and, finally, bias due to confounding (Greenland, 2005c). Basically, we assume that different sources of bias act independently of each other. This means, for instance, that if we know that misclassification probabilities are high, we would not assume particularly high or low probabilities of participation and vice versa. Our aim is to adjust for supposed bias when estimating the odds ratio (OR) of SAD with the subsequent onset of depression. We use the OR because we want to show how the results from a conventional analysis change when bias is taken into consideration. Limitations of the OR are numerical (Kraemer et al., 1999; Kraemer, 2003) and related to causal interpretations (Greenland, 1987, 1993). To address these limitations we present the crude cross tabulations of SAD and depression. Together with the bias-adjusted OR, this allows the reader to calculate any desired index of association from the cross tabulation that one would expect after adjusting for bias. For example, one could replace the frequency of individuals with tabulations of SAD and depression with the frequency corresponding to the adjusted OR while leaving the other cell frequencies fixed, and then multiply each cell size by a constant to match the original sample size. From the posterior OR distribution, based on the data and the bias model, one can compute the probability that SAD increases the risk of depression, which is not possible with conventional frequentist analysis. We attempted to map the observed counts in the 2 × 2 cross-table of reported SAD and depression into the 2 × 2 table that one would expect if our bias model was true, and recalculate the OR. We used a flat prior distribution for the unknown causal effect as Greenland (2005c).

Bias due to measurement

We excluded individuals who had depression at T0 from the analysis, so bias due to measurement might have occurred prior to bias due to selection. This appears to contradict our sequential approach. However, additional analyses demonstrated that sociodemographic variables and diagnoses at T0 did not predict dropout to any considerable extent (Lieb et al., 2000; Höfler et al., 2005). Thus, we assume that bias due to selection had already occurred at T0. Misclassification can be modelled by multiplying a misclassification matrix by the matrix that contains the observed counts (Greenland, 2005c, and references therein). For the observed counts of depression and SAD, a misclassification matrix would aim to map the observed depression and SAD counts onto a matrix of the true depression and SAD counts. The 16 unknown elements of this misclassification matrix can be factored into probabilities for (a) true depression given true SAD, observed depression, and observed SAD; and (b) true SAD given observed depression and observed SAD (see Greenland, 2005c, for the general case). These 16 misclassification probabilities contain 12 degrees of freedom. That is, the remaining four probabilities can be calculated from the joint frequencies of depression and SAD, which can be estimated from the data. The corresponding 12 unknown parameters contain:

- Four positive predictive values (PPVs) for observed depression. That is, the probability that there is true depression given that depression is observed and given the observed and true values of SAD.
- Four negative predictive values (NPVs) for depression. That is, the probability that there is no true depression given that no depression is observed and the observed and true values of SAD.
- Two PPVs for SAD. That is, the probability that there is true SAD given that SAD has been observed and the observed value of depression.
- Two NPVs for SAD. That is, the probability that there is no true SAD given that no SAD has been observed and the observed value of depression.

The OR can be recomputed from the resulting depression-SAD cross tabulation that would be expected given the data and the model of misclassification. We assume that PPVs and NPVs are independent of other variables if we take true and observed SAD (or observed depression, resp.) into account.

Bias due to selection

As mentioned in the introduction, inference was made primarily on the population from which a subset of 112 individuals with SAD at T0 was drawn. This sample is too small for stable assessments and, in prior analyses, there were no considerable differences in dropout rates

Copyright © 2007 John Wiley & Sons, Ltd

DOI: 10.1002/mpr

between individuals with and without SAD at T0 (Lieb et al., 2000; Höfler et al., 2005). Therefore, assessment of bias due to selection is based on the difference in dropout rates between the 2,183 participants in the analysis and those individuals who did not participate at T1 or T2 and had no depression at T0. The OR that is adjusted for bias due to measurement can be further adjusted for bias due to selection by dividing by the following expression (Greenland, 2005c):

$$\frac{\mathrm{R(depression, SAD)\cdot R(nodepression, no\ SAD)}}{\mathrm{R(depression, no\ SAD)\cdot R(nodepression, SAD)}}.$$

Where R(·) is the probability of completing the study given the true status of depression and SAD.

Bias due to confounding

To correct for confounding, we assume there is a latent binary factor, *U*, which affects both SAD and depression. *U* can be illustrated by imagining a hidden variable that represents two latent classes, individuals with low ($U = 0$) and high ($U = 1$) vulnerability for SAD and depression. The OR that is already adjusted for bias due to measurement and selection can be further adjusted for bias due to confounding bias by dividing by the following expression (Greenland, 2005c):

$$\frac{(\mathrm{OR}_{\mathrm{U,SAD}}\cdot \mathrm{OR}_{\mathrm{U}}\cdot\lambda+1)(\lambda+1)}{(\mathrm{OR}_{\mathrm{U,SAD}}\cdot\lambda+1)(\mathrm{OR}_{\mathrm{U,depression}}\cdot\lambda+1)}.$$

Where $\mathrm{OR}_{\mathrm{U,SAD}}$ and $\mathrm{OR}_{\mathrm{U,depression}}$ are the causal ORs between *U* and true SAD and depression. And λ is the odds (probability/(1-probability)) that $U = 1$ if there is no true SAD and no true depression. To compute the OR between true SAD (depression, resp.) and *U*, we only need the probabilities that there is true SAD (depression, resp.) given $U = 0$ and $U = 1$.

Estimation procedure

The MCSA procedure to simulate the distribution of the bias-adjusted OR, $\mathrm{OR}_{\mathrm{bc}}$, is as follows (Greenland, 2005c):

1. Compute the depression – SAD cross tabulations or the expected cross tabulations based on the adjustments already made in conventional analysis.
2. Draw a random number from the prior distribution of each of the 12 + 4 + 5 = 21 bias parameters.
3. Compute the bias-adjusted OR, $\mathrm{OR}_{\mathrm{bc}}$, as described in the sequential procedure above. Calculate the standard error of ln-$\mathrm{OR}_{\mathrm{bc}}$, denoted as $\mathrm{SE}_{\mathrm{bc}}$, under the assumption that the random numbers from Step 2 were the true bias parameter values.
4. Add a normal (0, $\mathrm{SE}_{\mathrm{bc}}$) disturbance to ln-$\mathrm{OR}_{\mathrm{bc}}$, and continue with Step 2 until the $\mathrm{OR}_{\mathrm{bc}}$ distribution remains virtually unchanged. We used 250 000 replications as Greenland (2005c).

Step 4 constitutes the modification of the MCSA procedure that improves the Bayesian interpretability of MCSA methods (Greenland, 2005c).

Prior distributions of bias parameters

We translated the prior distributions of the 21 unknown probabilities into meaningful data equivalents. That is, each prior distribution was expressed as a point estimate and a sample size assumed to be equivalent to the estimate's precision, with a larger sample size indicating greater precision (Greenland, 2006). The uncertainties of the probabilities were modelled with beta distributions. Let *p* be the prior point estimate of an unknown probability and *N* the associated sample size. Our uncertainty about an unknown probability is summarized as a beta(*a*,*b*) distribution with $a = p * (N - 2)$ and $b = (N - 2) * (1 - p)$. Table 1 shows the *p* and *N* values we chose for the 21 probabilities.

Priors for misclassification probabilities are difficult to determine. Self-reports can be faulty for various reasons, including memorization, cognition, social desirability, lying and instrument-related factors, such as wording of questions (Ritter et al., 1998; Kessler et al., 2000; Schwarz and Oyserman, 2001; Hardt and Rutter, 2004). The priors for the unknown misclassification probabilities are based on data from Reed et al. (1997) because this is the only study in which the validity of the CIDI was assessed with DSM-IV criteria. In that study, single and recurrent depressive episodes were assessed rather than major depression. According to cumulative lifetime incidence estimates from the EDSP until T2, the results from individuals with depressive episodes were weighted five times more than those of individuals with dysthymia in the present analysis.

In Reed et al. (1997) PPVs and NPVs were not studied in strata according to comorbid diagnoses so we had to make assumptions about how the probabilities differed according to diagnoses of SAD or depression. For depression, we assumed higher PPVs when the true

Copyright © 2007 John Wiley & Sons, Ltd

DOI: 10.1002/mpr

Table 1. Prior distributions of bias parameters

		Point estimate of probability	
1. Misclassification		%	N
positive predictive values for DEP	P(DEP = 1 ∣ DEP* = 1, SAD = 0, SAD* = 0)	90	12
	P(DEP = 1 ∣ DEP* = 1, SAD = 0, SAD* = 1)	85	12
	P(DEP = 1 ∣ DEP* = 1, SAD = 1, SAD* = 0)	80	12
	P(DEP = 1 ∣ DEP* = 1, SAD = 1, SAD* = 1)	90	12
negative predictive values for DEP	P(DEP = 0 ∣ DEP* = 0, SAD = 0, SAD* = 0)	98	39
	P(DEP = 0 ∣ DEP* = 0, SAD = 0, SAD* = 1)	90	39
	P(DEP = 0 ∣ DEP* = 0, SAD = 1, SAD* = 0)	95	39
	P(DEP = 0 ∣ DEP* = 0, SAD = 1, SAD* = 1)	98	39
positive predictive values for SAD	P(SAD = 1 ∣ SAD* = 1, DEP* = 0)	70	7
	P(SAD = 1 ∣ SAD* = 1, DEP* = 1)	70	7
negative predictive values for SAD	P(SAD = 0 ∣ SAD* = 0, DEP* = 0)	98	27
	P(SAD = 0 ∣ SAD* = 0, DEP* = 1)	98	27
2. Participation probabilities			
	P(participation ∣ SAD = 0, DEP = 0)	60	1265
	P(participation ∣ SAD = 0, DEP = 1)	57	146
	P(participation ∣ SAD = 1, DEP = 0)	60	74
	P(participation ∣ SAD = 1, DEP = 1)	63	27
3. Confounding			
	Prevalence of a latent binary confounder *U*	10	300
	P(SAD = 1 ∣ U = 0)	7	50
	P(SAD = 1 ∣ U = 1)	15	25
	P(DEP = 1 ∣ U = 0)	15	50
	P(DEP = 1 ∣ U = 1)	30	25

and observed SAD status agreed (90% versus 85% and 80%). This assumption was justified by supposing that there were shared factors (for example, personality traits) that decreased both the probability of reporting true SAD and true depression. If SAD was not diagnosed although truly present, we assumed that the NPV was higher (85%) than if SAD was diagnosed and truly present (80%). Similar assumptions were made about NPVs for depression (see Table 1). The PPVs and NPVs for SAD were assumed to be equal in individuals with and without depression because depression was assessed after assessing SAD. The *N*s for the priors were chosen as the number of individuals in the Reed et al. (1997) study from which the PPVs or NPVs, respectively, could be estimated, divided by the number of subgroups required for bias modelling (two subgroups for SAD and four for depression). Differences in participation rates between the four groups defined by the joint SAD and depression status were estimated with the dropout rates from the EDSP at T0 to T2, according to observed SAD and depression, and were weighted as described below. The overall participation rate was estimated as 60%, the EDSP response rate up to T2. The *N*s were chosen as the unweighted *N*s from the EDSP in the four observed cells, and they were divided by four to account for uncertainty about whether dropout rates could be used to estimate rates for completion of T0 to T2. We assumed a 10% prevalence of vulnerability for SAD and depression among individuals with neither SAD nor depression and twofold probabilities for SAD and depression in the presence of vulnerability. The overall rates of SAD and depression that were necessary to compute ORs were estimated with cumulative lifetime incidences in the EDSP up to T2. Our assumptions about the risks for having either of the disorders were supposed to be as precise as a sample size of 25 if vulnerability was present and 50 if vulnerability was not present.

Results

The cross tabulations of SAD at T0 and the incidence of depression at T1 and T2 are presented separately for

Copyright © 2007 John Wiley & Sons, Ltd

DOI: 10.1002/mpr

Table 2A. Unadjusted cross tabulation of the incidence of SAD and depression in the younger cohort (age 14–17 at baseline assessment)

	No depression		Depression	
	N	%	N	%
No SAD	906	89.3	109	10.7
SAD	37	97.4	1	2.6

Table 2B. Unadjusted cross tabulation of the incidence of SAD and depression in the older cohort (age 18–24 at baseline assessment)

	No depression		Depression	
	N	%	N	%
No SAD	968	91.8	87	8.3
SAD	58	77.3	17	22.7

each age cohort (18–24 versus 14–17 years at T0) in Tables 2A and 2B.

The results differ apparently between the two age cohorts. In the older cohort, the risk of depression in the sample was considerably elevated in the presence of SAD (22.7 versus 8.3%). In the younger cohort the risk was considerably smaller (2.6 versus 10.7%). Therefore, we analysed the cohorts separately. The same bias model was applied to both cohorts, although misclassification error might be lower in the younger cohort because two follow-up assessments were conducted for that cohort. On the other hand one could argue that self-reports of younger probands might be less accurate than those of older probands. The conventional OR for SAD at T0 and subsequent depression was adjusted for bias due to confounding according to sex and age by including these variables in the logistic regression equation. We corrected for bias due to selection by weighting the data according to intended differences in sampling probabilities at T0 (younger individuals were over sampled) and differences in participation rates at T0 according to age, sex, and geographical location (Höfler et al., 2005). The Huber-White sandwich estimator was used for robust inference on weighted data (Royall, 1986). The analyses were carried out with

Table 3. Conventional and bias-adjusted results

	Aged 14–17 years at T0			Aged 18–24 years at T0		
	OR	95%	CI	OR	95%	CI
Conventional estimate *(adjusted for observed confounders sex and age and observed selection bias – According to sex, age, and geographical location at T0).*	**0.18**	**0.02**	**1.35**	**3.06**	**1.64**	**5.70**
	Posterior OR distribution*			Posterior OR distribution*		
		Percentiles			Percentiles	
	Median	0.025	0.975	Median	0.025	0.975
Adjusted for bias due to misclassification only	0.86	0.07	3.94	2.56	1.01	9.89
Adjusted for bias due to selection only	0.18	0.02	1.36	2.65	1.37	8.07
Adjusted for confounding only	0.16	0.01	1.38	2.61	1.32	5.26
Adjusted for bias due to misclassification, selection and confounding	**0.72**	**0.07**	**3.94**	**2.01**	**0.61**	**9.71**

* Based on 250 000 replications in the MCSE procedure.

Stata (Stata Corp, 2005) and MCSA was self-implemented as an ado-file. For bias adjustment, the cross tabulation of observed SAD and depression were re-calculated based on:

- the adjusted ORs in the cohorts mentioned above; and
- the observed weighted frequencies of individuals who had neither SAD or depression.

We then proceeded with the estimation procedure outlined above to estimate the OR for each cohort given the data and our bias model. In additional analyses, each of the three types of bias was adjusted for while ignoring the other biases to assess which bias had the greatest impact on the results.

Table 3 shows that in the younger cohort the conventional estimate of for SAD and depression was 0.18 with a 95% confidence interval (CI) of 0.02–1.35. Thus, SAD could be associated with a heavily decreased and a moderately elevated odds of subsequent depression. Bias adjustment pushed the point estimate strongly toward the null value with the OR estimated to be 0.73, but the interval boundaries of 0.07 and 3.89 indicated that, given the data and the bias model, we have little information about the causal effect of SAD on depression here. The change in the point estimate was mainly due to misclassification adjustment. The adjusted results, however, should be interpreted very cautiously because there was only one individual with SAD and depression, and this has probably caused a floor effect in the adjustment.

In the older cohort, the conventional results were compatible with a moderate to strong increase in the odds of depression following SAD (1.64–5.70). After bias adjustment, the point estimate decreased from 3.06 to 2.01. The associated interval estimate of 0.61–9.71 was compatible with moderately decreased to strongly increased odds. The simulated distribution indicated that the probability that SAD increases the risk of subsequent depression (i.e., OR > 1), given the data and the bias model, is estimated at 88.6%.

Discussion

In the present study, we proposed a simple method to account for bias due to confounding, selection, and measurement when estimating the causal effect of SAD on depression. The results showed that both point estimates and interval estimates can change considerably after adjusting for biases with a bias model. The interval estimates in particular were much broader after taking uncertainty about bias parameter values into account.

There is typically insufficient information about biases to estimate causal effects from observational data. Therefore, we made assumptions about biases based on other data sets and thoughtful guesses, not knowing if they truly apply to our data or not. Although uncertainties about bias parameter values are addressed by the variances of their priors, there is unaddressed uncertainty in applying the entire bias model. If more conservative multiple bias models were used, the intervals would have been even broader. The uncertainties in our multiple bias model are:

- Bias might actually operate in a much more complicated way. For instance, misclassification probabilities could be correlated or vary individually according to variables that might also affect participation and produce confounding. As far as we know, such issues have not yet been assessed.
- The priors on bias parameters could be inaccurate. For instance, we took priors on misclassification probabilities from a CIDI clinical validation study (Reed et al., 1997), but the PPVs and NPVs might be different in the general population (Wittchen, 1994). Besides, unlike Reed et al. (1997), our sample contained only adolescents and young adults who did not have depression at T0, and the Reed et al. (1997) study was cross-sectional rather than prospective. Moreover, clinical diagnoses were the standard against which CIDI was evaluated, and it might well be that clinical diagnoses are less accurate than CIDI diagnoses.
- Our priors on confounding parameters are based on subjective, rather than objective evidence.
- We do not know if dropout results can be applied to participation in the entire study.

We found very different results in the two age cohorts, and the differences disappeared only partially after bias adjustment. This difference could be due to methodology (i.e., younger participants were assessed twice during the follow-up period) or due to bias operating differently in both cohorts. One peak period for onset of depression in the younger cohort was yet in the age range 13–18 years, which was not the case in for the older cohort. We also found that, in the younger cohort, approximately half of those who had ever

Copyright © 2007 John Wiley & Sons, Ltd

DOI: 10.1002/mpr

fulfilled the criteria for SAD fulfilled it only after T0. Most of them, however, reported an age of onset that was lower than their age at the T0 assessment. Both points contribute to an explanation of why we did not find an association between SAD and subsequent depression here. We shall examine the age- and age-of-onset-dependent associations between SAD and depression in more detail in another paper.

The general conclusion about bias modelling is that the less information about biases and the higher the uncertainty about applying results from the literature or assumptions on bias parameters, the more uncertainty emerges in the resulting model-based causal estimate. This property validates properly used multiple bias adjustment methods. Unlike in conventional methods, uncertainty about biases in bias correction methods carries over to the interval estimates. In the extreme case, this means that if one knows nothing about biases or has a poor understanding of how data was generated, one will never be able to demonstrate a causal effect, and an accurate interval estimate will always include the null value of no effect.

As long as the bias model and the priors used are not fundamentally wrong, the results from multiple bias adjustment methods can be expected to outperform conventional analyses. This is because the conventional analyses are based on the absurd assumption that there is no bias at all in the data.

Acknowledgements

This work is part of the Early Developmental Stages of Psychopathology (EDSP) Study and is funded by the German Federal Ministry of Education and Research (BMBF) project no. 01EB9405/6, 01EB 9901/6, EB01016200, 01EB0140, and 01EB0440. Part of the field work and analysis was also supported by grants from the Deutsche Forschungsgemeinschaft (DFG) LA1148/1-1, WI2246/1-1, WI709/7-1 and WI 709/8-1. Principal investigators are Dr Hans-Ulrich Wittchen and Dr Roselind Lieb. Core staff members of the EDSP group are Dr Kirsten von Sydow, Dr Gabriele Lachner, Dr Axel Perkonigg, Dr Peter Schuster, Dipl.-Stat., Michael Höfler, Dipl.-Psych., Holger Sonntag, Dipl.-Psych., Tanja Brückl, Dipl.-Psych., Elzbieta Garczynski, Dr Barbara Isensee, Dipl.-Psych., Agnes Nocon, Dr Chris Nelson, Dipl.-Inf., Hildegard Pfister, Dr Victoria Reed, Dipl.-Soz. Barbara Spiegel, Dr Andrea Schreier, Dr Ursula Wunderlich, Dr Petra Zimmermann, Dr Katja Beesdo and Antje Bittner, Dipl.-Psych. Scientific advisors are Dr Jules Angst (Zurich), Dr Jürgen Margraf (Basel), Dr Günther Esser (Potsdam), Dr Kathleen Merikangas (NIMH, Bethesda), Dr Ron Kessler (Harvard, Boston) and Dr Jim van Os (Maastricht).

References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 4 edn. Washington DC: APA, 1994.

Greenland S. Interpretation and choice of effect measures in epidemiological analyses. Am J Epidem 1987; 5: 761–8.

Greenland S. Randomization, statistics, and causal inference. Epidemiology 1990; 1: 421–9.

Greenland S. Basic problems in interaction assessment. Environl Health Perspect Suppl 1993; 101 Suppl 4: 59–66.

Greenland S. Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. Risk Anal 2001; 4: 579–83.

Greenland S. The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukemia. J Am Stat Assoc 2003; 98: 47–54.

Greenland S. Interval estimation by simulation as an alternative to and extension of confidence intervals. Int J Epidemiol 2004; 33:1–9.

Greenland S, Epidemiological measures and policy formulation: lessons from potential outcomes (with discussion). Emerging Themes in Epidemiology 2005a; 2: 1–4.

Greenland S. Discussion on 'Statistical issues arising in the Women's Health Initiative'. Biometrics 2005b; 61: 920–1.

Greenland S. Multiple bias modelling for analysis of observational data. With discussion. J Roy Stat Soc A 2005c; 168: 267–306.

Greenland S. Bayesian perspectives for epidemiological research: I. Foundation and basic methods. Int J Epidemiol 2006; 35: 765–75.

Hardt J, Rutter M. Validity of adult retrospective reports of adverse childhood experiences: review of the evidence. J Child Psychopathol 2004; 2: 260–73.

Höfler M. Causal inference based on counterfactuals. BMC Med Res Methodol 2005a; 5: 28.

Höfler M. The effect of misclassification on the estimation of association: a review. Int J Meth Psychiatr Res 2005b; 14: 92–101.

Höfler M, Pfister H, Lieb R, Wittchen HU. The use of weights to account for non-response and dropout. Soc Psychiatr Psychiatr Epidemiol 2005; 40: 291–9.

Holland PW. Statistics and causal inference. J Am Stat Assoc 1986; 945–60.

Kessler RC, Wittchen HU, Abelson J, Zhao S. Methodological issues in assessing psychiatric disorders with self-reports. In: Stone AA, Turkan JS, Bachrach CA, Jobe JB, Kurtzman HS, Cain VS: The Science of Self-report: Implications for Research and Practice. New Jersey: Lawrence Erlbaum Associates, 2000.

Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of a risk factor for clinicial or policy significance. Psychol Meth 1999; 4: 257–71.

Copyright © 2007 John Wiley & Sons, Ltd

DOI: 10.1002/mpr

Kraemer HC. Reconsidering the odds ratio as a measure of association of 2x2 association in a population. Stat Med 2003; 23: 257–70.

Lieb R, Isensee B, Sydow von K, Wittchen HU (2000) The Early Developmental Stages of the Psychopathology Study (EDSP): A methodological update. Eur Add Res 6: 170–82.

Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes. Annu Rev Publ Health 2000; 21: 121–45.

Maclure M, Schneeweiß S. Causation of bias: the episcope. Epidemiology 2001; 12: 114–22.

Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. Biometrics 2005; 61: 899–941.

Reed V, Gander F, Pfister H, Steiger A, Sonntag H, Trenkwalder C, Hundt W, Wittchen HU. To what degree does the Composite International Diagnostic Interview (CIDI) correctly identify DSM-IV disorders? Testing validity issues in a clinical sample. Int J Meth Psychiatr Res 7: 142–55.

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

Rothman KJ, Greenland S. Modern Epidemiology. 2 edn. Lippincott Williams & Wilkins, Philadelphia, 1998.

Royall RM. Model robust confidence intervals using maximum likelihood estimators. Int Stat Rev 1986; 54: 221–6.

Rutter M, Maughan B, Pickles A, Simonoff E. Retrospective recall recalled. In: Cairns, Bergman LR, Kagan J (eds.): Methods for Studying the Individual. London: Sage, 1998.

Schwarz N, Oyserman D. Asking questions about behavior: cognition, communication, and questionnaire construction. Am J Eval 2001; 22: 127–60.

Soldani F, Ghaemi N, Baldessarini R. Acta Psychiatrica Scandinavica 2005; 112: 1–3.

StataCorp. Stata Statistical Software: Release 9.0. College Station, TX: Stata Corporation, 2005.

Wittchen HU. Reliability and validity studies of the WHO Composite International Diagnostic Interview (CIDI) – a critical review. J Psychiatr Res 1994; 28: 57–84.

Wittchen HU, Perkonigg A, Lachner G, Nelson CB. Early Developmental Stages of Psychopathology Study (EDSP): objectives and design. Eur Addiction Res 1998; 4: 18–27.

Wittchen HU, Pfister H (eds) DIA-X-Interviews. Manual für Screening-Verfahren und Interview; Interviewheft Längsschnittsuntersuchung (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, 1997.

Correspondence: Michael Höfler, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Chemnitzer Str. 46a, 01187 Dresden.
Email: hoefler@psychologie.tu-dresden.de
Telephone (+49) 351-46336921
Fax (+49) 351-46336984

Copyright © 2007 John Wiley & Sons, Ltd