Scand J Work Environ Health 1983;9(5):377-383

<https://doi.org/10.5271/sjweh.2396>

Issue date: Oct 1983

Some remarks on the operation of biases.

by [Nurminen M](#)

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/6673096

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Some remarks on the operation of biases

by Markku Nurminen, DSc¹

NURMINEN M. Some remarks on the operation of biases *Scand j work environ health* 9 (1983) 377-383. Etiologic research is easily invalidated by wrong conclusions based on the use of dubious methods of study design or data analysis. This paper discusses three central types of validity concepts, namely, selection bias, information bias, and confounding bias. Bias is defined in statistical terms as a misrepresentation of an effect measure, such as disease rate ratio. The discussion is presented basically within the framework of fourfold frequency table data to investigate the dependence of the presence or absence of a disease on a dichotomous exposure variate. A brief review of the statistical strategies and techniques available for controlling confounding is given. The issue of properly measuring a true confounder is dealt with in some detail because of its importance for the avoidance of biased estimates. The presentation concentrates on pointing out the direction and magnitude of the distortion caused by the operation of these sources of error.

Key terms: epidemiologic methods, statistics, validity, precision.

This paper reviews some of the ways different biases affect epidemiologic study results. The concepts and principles involved are certainly not new; the motive for presenting them in this context is to remind the reader of their central importance in the process of evaluating comparative, nonrandomized studies.

It has been found useful to distinguish between three basic types of bias, namely, that of selection, information and confounding, in the terminology of Miettinen (8). I shall proceed to discuss selected aspects related to the assessment and possible control of these sources of error with a view towards applying them in nonexperimental (observational) studies.

from the way people are selected from the source population into the study base (or sampled population), which consists of exposed and nonexposed (and other) persons whose actual experience forms the empirical basis for the epidemiologic inquiry. The influence of selection forces can be expressed conceptually, in a reduced situation, through the fourfold table presented in table 1, which relates an exposure with a disease [see, eg, Kleinbaum et al (6)]. The parameters α , β , γ , and δ represent the probabilities of selecting the population from one of the four categories of the source population into the corresponding subdomain in the study base. Thus, for example, α is estimated as $\hat{\alpha} = a^0/a$, the relative frequency with which the

Selection bias

Selection bias is a distortion in the estimate of the effect of a particular exposure on a health parameter of interest. It results

Table 1. Selection probabilities from the source population to the sampled population according to the dichotomous exposure and disease classifications.

Category of disease	Category of determinant	
	Exposed	Referent
Case	α	β
Referent	γ	δ

¹ Department of Epidemiology and Biostatistics, Institute of Occupational Health, Helsinki, Finland.

Reprint requests to: Dr M Nurminen, Institute of Occupational Health, Haartmaninkatu 1, SF-00290 Helsinki 29, Finland.

subjects in the source population, who could and would constitute the exposed-cases portion of the sampled population (a), actually become members of it (a^o). Therefore the observed frequency a^o is a subsample of a .

In occupational health studies selection bias is manifested in that the compared vocational groups – one with the index exposure contrasted with the reference (comparison) experience – tend to enter and leave their particular jobs in dissimilar relative numbers, the dissimilarity being related to the risk of the disease at issue (popularly referred to as the “healthy worker effect”).

Imagine a cohort study on the incidence of Raynaud’s phenomenon among lumberjacks using vibrating chain saws with the interest on the provocative role of cold weather in the appearance of the symptoms. The index group of workers could be selected from a locality with a much lower average temperature than that experienced by the otherwise comparable reference group of forest workers. For example, one could compare lumberjacks in northern Finland to a similar group in the southern part of the country. Were the follow-up to last several years, it could well happen that, owing to the more severe work conditions in the north, those index subjects who were prone to experience finger blanching would seek employment in warmer work environments (indoors or outdoors). Thus, when the incidence of “white fingers” is inquired about, the remaining members of the index cohort would likely be individuals who were more resistant to vasoconstriction in response to cold exposure than those who rejected their jobs, and the actual study samples might represent a biased population contrast. The result would then be an underestimate of the true incidence density ratio.

In a case-referent (case-control) sampling approach the corresponding selective operation, distorts, for example, the referral and admission rates to hospitals by way of the patients being selected on the basis of some exposing condition that, in fact, was connected with the hospitalizing illness (Berkson’s fallacy).

Consider the question of the diagnosis of traumatic vasospastic disease as an occupational disease entitling legal com-

pensation. To differentiate acquired traumatic vasospastic disease from the constitutional Raynaud syndrome, a necessary (but not sufficient) criterion for the diagnosis would be that the individual seeking compensation has used vibrating tools in his/her employment. As a consequence workers suffering from Raynaud’s phenomenon are more readily referred to hospitals if they have been exposed to vibration than they are if they have not been so exposed. Thus the case identification is conditional on exposure to the alleged cause, and the case sample is thereby biased so that it becomes invalid for any quantitative inference.

The described problem is divorced from another biasing mechanism involving other exposing factors with effects interwoven with those of the determinant under study. For example, an attempt to define a “vibration syndrome” is a multifactorial issue, for exposure to vibration in work is often associated with cofactors such as exposure to high levels of noise, heavy physical exertion, etc. The modification and/or confounding of effects due to the simultaneous operation of cofactors may be possible to control in the stage of data analysis (see the following discussion), whereas biased sampling is only avoidable through proper design of the study.

The induced bias (B) in the observed odds ratio (OR^o) can be expressed in relative terms as [see, eg, Kleinbaum et al (6)]:

$$\begin{aligned} B &= (\hat{OR}^o - \hat{OR})/\hat{OR} = \hat{OR}^o/\hat{OR} - 1 \\ &= (a^o/a)(d^o/d)/[(b^o/b)(c^o/c)] \\ &= \hat{\alpha}\hat{\delta}/(\hat{\beta}\hat{\gamma}) - 1. \end{aligned}$$

From this expression it is easily seen that all four selection probabilities do not have to be equal for the bias to be absent. For $B = 0$ to be the case, it is sufficient that either

$$\begin{aligned} \alpha/\beta &= \gamma/\delta \\ \text{or} \\ \alpha/\gamma &= \beta/\delta. \end{aligned}$$

The former equality has the interpretation that the exposure odds for the diseased cases should equal those for the reference (noncase or other) subjects. The latter condition requires that the disease odds

for the people in the index category of (high) exposure should be the same as those for the reference category (low or no exposure).

If one were able to estimate the selection probabilities, an unbiased (corrected) disease or exposure odds ratio could be computed simply by multiplying the biased estimate of OR^o by the factor $\beta\gamma/(\alpha\delta)$. [See Kleibaum et al (6) for a comprehensive treatise of this topic.] Often, unfortunately, one has to be content with estimating the direction of the selection bias only. For example, if $\alpha/\beta > \gamma/\delta$ (or if $\alpha/\gamma > \beta/\delta$), then the odds ratio estimate must exaggerate the ostensible effect, or $B > 0$ (positive bias).

Information bias

While selection bias affects the study base (sampling frame), information bias results in misclassification of the chosen people within the sample. Erroneous information can be acquired when measurement of either the exposure condition or the disease status is inaccurate. Examples of typical sources of such error are untrue professional titles obtainable from census records and wrong diagnostic codes on death certificates. For example, in a register-based study of forest tractor drivers (7) focusing on problems caused by vibration of the whole body, the reference group was chosen from among lumberjacks involved in other activities, and yet over 10 % of the referents reported on a questionnaire that they had been exposed to whole-body vibration during the year preceding the investigation. Thus the information recorded in register files, however "official," may be only partly revealing and should not be relied on as the sole source of data for group classification.

The consequences of information bias can either inflate or deflate the true effect. However, in the special situation of symmetrical information bias between the groups compared, the effects will be diluted, and the parameter values will shift towards their respective null values (eg, $OR = 1$). Symmetrical bias requires that the misclassification of exposure occurs at a nondifferential rate for the disease cases and the noncase referents,

or, alternatively, the misclassification of disease status occurs at the same frequency for both the subjects in the index and those in the reference domain of exposure.

The classification procedure can be completely described [following Kleinbaum et al (6)] with the use of the concepts of sensitivity (Se) and specificity (Sp). If D and $\bar{D}$ denote classified disease and nondisease status, respectively, and E and $\bar{E}$ denote classified exposure and nonexposure status, respectively, then

$$Se(D) = \Pr(D|D) = \text{probability of classifying a truly diseased person as diseased;}$$

similarly,

$$Sp(D) = \Pr(\bar{D}|\bar{D}),$$

$$Se(E) = \Pr(E|E),$$

$$Sp(E) = \Pr(\bar{E}|\bar{E}).$$

To illustrate the effects of symmetrical information bias, the data from a case-referent study conducted by Hardell & Sandström (5) can be used (table 2). An estimate of the apparent exposure odds ratio is $OR^o = (a^o/b^o)/(c^o/d^o) = 4.4$. There exist algebraic formulas to adjust the observed frequencies according to the estimated values of the sensitivity and specificity parameters. The results obtained under various assumptions concerning sensitivity and specificity are given in table 3. It is evident that even the shown slight departures from perfectly valid data result in markedly erroneous estimates.

Again it is stressed that, even if it is not realistic to give exact estimates of the degree of misclassification in a given study, knowledge of the ranges of such bias can greatly facilitate the interpretation of the study results.

As an example, consider a cohort study in which the exposure status is assessed without error. The misdiagnosis bias, expressed in terms of the incidence density ratio or rate ratio parameter ($RR = R/R_r$),

Table 2. Case-referent data used to illustrate the operation of information bias in the observed association between a binary disease outcome (D) and a binary exposure variate (E).

	E	$\bar{E}$
D	19 (= a ^o)	33 (= b ^o)
$\bar{D}$	6 (= c ^o)	46 (= d ^o)

Table 3. Effect of symmetrical information bias — The estimated correct frequencies a, b, c, d; the corrected odds ratio (OR) estimate, $\hat{OR} = ad/(bc)$; and the bias (B) in percent, $B = 100(\hat{OR} - OR)/OR$, for the data of Hardell & Sandström (5). (Se = sensitivity, Sp = specificity, D = binary disease outcome, E = binary exposure variate)

Se (D)	Sp (D)	Se (E)	Sp (E)	a	b	c	d	$\hat{OR}$	B
1.0	1.0	1.0	1.0	19.0	33.0	6.0	46.0	4.4	0
0.9	1.0	1.0	1.0	21.1	36.6	3.8	42.3	6.2	41
1.0	0.9	1.0	1.0	21.1	30.8	6.6	45.3	4.6	5
1.0	1.0	0.9	1.0	15.3	36.6	0.8	51.1	24.0	445
1.0	1.0	1.0	0.9	18.3	27.8	6.6	51.1	5.0	14
0.9	0.9	1.0	1.0	23.4	34.3	4.3	41.9	6.6	50
1.0	1.0	0.9	0.9	15.2	30.9	0.9	56.7	28.2	541
1.0	0.9	1.0	0.9	20.3	25.8	7.4	50.3	5.3	20
0.9	1.0	0.9	1.0	.a	.a	.a	.a	.a	.a
0.8	0.9	1.0	0.7	25.5	16.0	2.2	60.1	43.1	880

^a This combination of values of sensitivity and specificity produced inadmissible (ie, negative or indeterminate) values for the adjusted cell frequencies and thus invalid study results.

and comparing the rate of the exposed (R) and reference (R_r) subcohorts, is defined as

$$B = (RR^o - RR)/RR = RR^o/RR - 1,$$

where $RR^o = R^o/R_r^o$ is estimated in the usual manner by

$$\hat{RR}^o = [a^o/(a^o + c^o)]/[b^o/(b^o + d^o)],$$

and $RR = R/R_r$ is estimated analogously from the corrected figures. The magnitude of the bias satisfies the double inequality

$$0 \leq |B| \leq 1,$$

and the direction of the bias is either positive or negative in conformity with the true rate difference, $RD = R - R_r$. [See appendix 1 for the derivation.]

Confounding bias

Confounding is a ubiquitous concern in epidemiologic research. When the problem of confounding is dealt with, the basic steps of planning and executing a nonexperimental (characteristically nonrandomized) study are (i) to list all the known and plausible confounders in conceptual terms, (ii) to design the study to record (measure) the confounders, (iii) to translate the recordings into statistical variates, (iv) to select those variates that are to be adjusted in either the design or the analysis, and (v) to find some method of removing or reducing the biases that such variates may cause.

The definitional properties of a confounder have been stated (i) as predic-

tive of the health outcome in the base and (ii) as associated with the determinant under study (8). For example it has been hypothesized (10) that the effect of exposure to noise and hand-arm vibration are confounded in the induction of vasoconstriction in both the cochlear and peripheral blood vessels of workers exposed occupationally to vibration because these same workers are also often exposed to excessive noise levels.

A more demanding problem is the verification of the first condition of the preceding definition. The discernment of health predictors (risk indicators) is a matter of substantive judgment requiring the theoretical arguments or the empirical evidence of earlier studies rather than a decision-making process based on statistical criteria.

The dilemma posed by the conceptual identification of every important confounder is that, in addition to those factors brought forward in the study, another factor may have been omitted that, together with the included factors, constitutes a confounder. There is also another problem, that is, even when all the plausible confounders have been found to be (nearly) identically distributed between the compared groups, their joint distribution may still cause disturbance.

Naturally there is a practical limit to the number of variates that can be admitted to a study or at least, simultaneously subjected to statistical adjustment. The problem is compounded by the search for operationally accurate variates or the construction of adequate indices for the con-

founders. Operational imperfections of the empirical measures can result in failure to remove all the confounding. (Appendix 2 gives a theoretical illustration of this point.) This often underrated task is amplified by the fact that controlling extraneous variates actually involves two goals. Primarily, bias should be prevented. Secondly, even if there seems to be no danger of bias, the presence of the disturbing variate may inflate the variance of the estimator of the occurrence or the estimate of the effect parameter between the compared groups to such an extent that the comparison becomes imprecise. Such imprecision should also be avoided.

The selection of the important confounding variates from the list of potential confounders is not part of the nature of classical statistical inference, but, again, it depends mainly on consideration of the subject matter. The notion of confounding inherently belongs to the realm of Bayesian inference in that prior knowledge of various causal links must be on hand before the logical candidates for confounders can be distinguished. Nevertheless some conventional statistical tactics or techniques may prove helpful as guidelines. The confounders are generally arranged according to the following three classes (4): (i) major variates for which some kind of control is considered essential (Their number is preferably kept small in view of the complexities involved in controlling many variates at the same time.); (ii) variates which, ideally, one would like to control, but for which, instead, one must be content with some verification that their effects produce little or no bias of practical importance; (iii) variates whose effects are thought to be minor and which are therefore disregarded. Variates for which no data are available in the study base may have to be included in the last category.

The selection of a sufficient and relevant subset of confounders in a manner similar to that familiarly used in connection with the specification of a model in multiple regression analysis can be extremely misleading. The confusion results from the procedure followed for ordinary stepwise regression techniques that commonly examine the association between extraneous variates and the outcome in terms of partial correlation coefficients (squared), which do not clearly indicate the causal

nature of the associations. An alternative approach would be to search for structure via simultaneous regression equations.

When the confounding potential of a covariable predictive of the health outcome in the selected base is considered, one should ensure that the association between the covariate and the determinant under study is such that the exposure-disease relation has not been appreciably distorted. The assessment of the difference in the covariate distribution between the compared groups in the context of statistical inference tests may not be proper (1), principally because of the consequence of the different rationale in significance test-int - demonstrating that, under the assumption that two samples were drawn from the same (hypothetical) population, an observed difference between the samples is due to variability occurring in the sampling during repeated sampling. Chance fluctuation is no more (or less) harmful than fluctuation of the same magnitude that springs from some other source. In practice, an appreciable difference need not be the same as a statistically significant difference.

There are the well-known strategies for controlling confounders during study design [see, eg, Breslow & Day (2)]. In restriction, matching, and blocking, the samples are selected from the base populations so that the distributions of the confounding variates within the samples are similar in some respects, whereby the outcomes become more or less correlated. Balancing (or frequency matching) does not eliminate confounding, but it does reduce its effect and make it unidirectional within the strata. Methods available in the stage of data analysis include stratification (standardization) and various modeling techniques, such as regression analysis, analysis of covariance, logit analysis, log-linear analysis, and multivariate survival analysis. [For a choice among the procedures, consult chapter 13 in the communication by Anderson et al (1).] A new method of adjustment for a vector of covariates may be applied to observational studies; it includes many of the aforementioned standard methods [see Rosenbaum & Rubin (11)]. This technique is based on the so-called propensity score, that is, the propensity towards a particular exposure given the observed covariates.

Concluding remarks

The intent of this communication has been to serve as a reminder of some of the many ways that biases can enter epidemiologic studies. In addition to the discussed biases inherent in study design there exist several data-analytic biases (pseudo results) which generally derive from (i) specification error in the mathematical form of the model function, (ii) violated assumptions underlying the used statistical methods, (iii) uninhibited use of a multitude of sophisticated methods on a sparse or rather untidy data set, and (iv) misguided interpretation of the computed results. For a comprehensive coverage of the topics touched in this paper the reader is referred to the excellent texts in the following list of references, especially the recent book by Kleinbaum et al (6).

Acknowledgment

I wish to thank Bo Gullberg for his kind cooperation during the drafting of this presentation.

References

1. Anderson S, Auquier A, Haunck W, Oakes D, Vandaele W, Weisberg HI. Statistical methods for comparative studies: Techniques for bias reduction. John Wiley & Sons, Inc, New York, NY 1980.
2. Breslow NE, Day NE. Statistical methods in cancer research. Volume 1 (The analysis of case-control studies). International Agency for Research in Cancer, Lyon 1980. (IARC scientific publication no 32).
3. Chiang CL. Introduction to stochastic processes in biostatistics. John Wiley & Sons, Inc, New York, NY 1968.
4. Cochran WG. The planning of observational studies of human populations. *J r stat soc A* 128 (1965) 234–255.
5. Hardell L, Sandström A. Case-control study: Soft-tissue sarcomas and exposure to phenoxy acetic acids or chlorophenols. *Br j cancer* 39 (1979) 75–81.
6. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: Principles and quantitative methods. Lifetime Learning Publications, Belmont, CA 1982.
7. Korhonen O, Nummi J, Nurminen M, Nygård K, Soininen H, Wiikeri M. Metsätyöntekijä. Osa 3. Metsätraktorin kuljettajien terveysterveystutkimuksia no 127). Institute of Occupational Health, Helsinki 1980. (Työterveyslaitoksen tutkimuksia no 127).
8. Miettinen OS. Principles of epidemiologic research. Harvard University, Boston, MA 1976. (Course text).
9. Miettinen OS. Principles of epidemiologic research. Harvard University, Boston, MA 1982. (Course text).
10. Pyykkö I, Stark J, Färkkilä M, Hoikkala M, Korhonen O, Nurminen M. Hand-arm vibration in the aetiology of hearing loss in lumberjacks. *Br j ind med* 38 (1981) 281–289.
11. Rosenbaum RR, Rubin DD. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1983) 41–45.

Appendix 1

Magnitude of diagnostic bias

Assume that the exposure category is determined without misclassification so that the sensitivity (Se) and specificity (Sp) of exposure (E) assessment is $Se(E) = Sp(E) = 1$. The ostensible disease rate (R^o), subject to diagnostic error, is, in this situation,

$$R^o = Se(D)R + [1 - Sp(D)](1 - R),$$

where R is the correct rate. The ostensible rate ratio (RR^o), with the subindex r representing the reference category, is defined as $RR^o = R^o/R_r^o$, and the relative bias be-

comes

$$\begin{aligned} \frac{RR^o}{RR} - 1 &= \frac{Se(D)R + [1 - Sp(D)](1 - R)}{Se(D)R_r + [1 - Sp(D)](1 - R_r)} / \left(\frac{R}{R_r}\right) - 1 \\ &= \frac{[1 - Sp(D)]/R_r - [1 - Sp(D)]/R}{[1 - Sp(D)]/R_r + Se(D) + Sp(D) - 1}. \end{aligned}$$

Now the validity of a diagnostic test, defined as the sum of the sensitivity and specificity, must exceed unity for it to be of

any value, or $\text{Se}(D) + \text{Sp}(D) \geq 1$. Noting that $0 \leq R_r < \infty$ and $0 \leq \text{Sp}(D) \leq 1$, the inequality $[1 - \text{Sp}(D)]/R_r \geq 0$ follows, and results in the absolute value of the relative bias being between zero and one or

$$0 \leq |B| \leq 1.$$

The bias can be recast as

$$B = RD / \langle R[1 + R_r\{\text{Se}(D) + \text{Sp}(D) - 1\}] / [1 - \text{Sp}(D)] \rangle,$$

where $RD = R - R_r$, so that B is always in the direction of the actual rate difference RD .

Appendix 2

Control of confounding by empirical measure

Suppose each subject has an unobservable correct value C of the confounding factor, with the theoretical mean μ and variance σ^2 , and an empirical measure M of the confounder, which includes measurement error e independent of C , with the mean zero and variance σ_e^2 so that $M = C + e$. If the joint distribution of C and M is bivariate normal and if the coefficient of the correlation between C and M is ρ , it follows that the conditional expectation of C , given a measured value of $M = m$, is

$$E(C|m) = \mu + (m - \mu) \rho \sigma / (\sigma^2 + \sigma_e^2)^{1/2}$$

[see, eg, Chiang (3, p 22)]. Expressing ρ in terms of the covariance between C and M is

$$\rho = \text{cov}(C, M) / [\sigma(\sigma^2 + \sigma_e^2)^{1/2}]$$

where $\text{cov}(C, M) = \text{cov}(C, C + e) = \sigma^2$, since e was taken not to depend on C ,

$$E(C|m) = \mu + (m - \mu) \sigma^2 / (\sigma^2 + \sigma_e^2).$$

Thus, when two populations are compared with the theoretical means μ_1 and μ_2 and the same variance σ^2 , the expected

amount of confounding remaining after the control of M is (9):

$$E(C_1 - C_2|m) = (\mu_1 - \mu_2) \sigma_e^2 / (\sigma^2 + \sigma_e^2).$$

Table A shows the percentage of confounding unremoved as a function of the error distribution relative to that of the correct distribution. Thus, for example, if the distributions of the measurement errors and the true values are equally dispersed, a not-so-uncommon situation in nonexperimental research practice, only half of the confounding will be controlled by adjustment for the imperfect measure.

Table A. Effect of the imprecision of a confounding variate.

Ratio of standard deviations	Percentage of residual confounding
σ_e/σ	$\sigma_e^2/(\sigma^2 + \sigma_e^2) \%$
0.1	1
0.5	20
1.0	50
1.5	69
2.0	80
3.0	90

Received for publication: 1 July 1983