Paolo Vineis

University of Torino and ISI Foundation, Torino

Series: History of epidemiology

History of bias

Summary

Epidemiologists have always been conscious of the importance of controlling for distortions, although the definition itself of bias has changed over time. Central to this discussions in the past was the relative vulnerability of different study designs to bias and uncontrollable confounding (confounding being clearly distinguishable from bias, as a problem of inter-mixed causal effects due to the non-random distribution of risk factors within the study population). In particular, controversy arose over aspects of case-control study design. Also a formulation of "typologies of bias" during the 1970s helped to define some of the most important sources of distortion in the design, analysis and interpretation of epidemiological studies. The subsequent period – until now – has been characterised by more formal and systematic definitions.

Keywords: Distortion – Study design – Selection – Information – Confounding.

The idea of bias has been associated historically with three main meanings: a) prejudice of the observer (including the influence of a theory upon observation); b) bias as systematic error of an instrument; c) bias as a consequence of an erroneous study design.

Erroneous explanations, based on "fashionable" theories, have been rather common in medicine; e.g., a parasytic or infectious agent was sought for beri-beri and other diseases which turned out to be due to completely different aetiologies. Eijkman, for example, studied 27 jails in Java where prisoners ate refined rice, and 74 where prisoners ate raw rice: the prevalence of beri-beri was 1/39 in the first group and 1/10000 in the second. With a typically epidemiologic design, he showed that the cause of beri-beri had to be sought in dietary habits. However, Eijkman's theory on the dietary origin of beri-beri was accepted only after a long time; many researchers in the second half of the 19th century looked for infectious causes of beri-beri and, in fact, of all diseases. A second example – among many others – of the influence of general theories over disease understanding is represented by New Guinea's kuru, which still in 1958 was considered to be psychogenic.

Theoretical models are very influential in clinical medicine. In the 17th century vomiting and diarrhoea were considered as signs of effectiveness of *digitalis* treatment, while now they are considered signs of intoxication (excessive dose). This different interpretation is due to different underlying paradigms of heart failure: in the 17th century *digitalis* was thought to work because it caused a release of liquids, with a typical inversion of cause and effect.

A short history of the concept of "meter"

The classical distinction between *bias* (as a distortion, or lack of correspondence between observation and putative truth) and *imprecision* (random fluctuation) has been introduced in physics and technology first.

A considerable discussion about measurement took place during the French revolution, in particular about the definition of the "meter". Originally the meter was conceived as an absolute and "objective" measure (40×10^{-6} of the earth's meridian). However, it was soon clear that it was too difficult to measure. The platinum standard that was kept at the National Archives, in fact, was found to be biased (i.e., it did not correspond to the quantity originally meant). In addition, measurements of different "meters" in comparison with this standard were imprecise (subject to small variations, irreproducible). Nevertheless, the standard was kept

Soz.- Präventivmed. 47 (2002) 156–161
0303-8408/02/030156-06 $ 1.50 + 0.20/0

as an "arbitrary" exemplar. Precision was improved by using platinun-iridium and a different shape. Only in 1952 a natural standard unit was introduced, as 1650763.73 times the wavelength of the radiation released by crypton 86. This decision improved precision and allowed an "objective" (natural) standard.

In general *absolute* measures tend to be imprecise but more valid (unbiased). They are more difficult to obtain. *Relative* measures, on the opposite, are frequently precise but may be biased if they are referred to a "natural" standard. The platinum standard was biased but it served as an arbitrary model (the unit of length).

The history of bias in epidemiology

The awareness of bias in medical observation has a long history. For example, William Augustus Guy, a student of Louis and professor of forensic medicine and hygiene, studied the relationship between occupation and health. He found that the ratio between "pulmonary consumption" and other diseases was 1:3.47 in compositors and 1:5.12 in pressmen. The hypothesis that was put forward at the time to explain these observations was that the lack of physical exercise (in compositors) lead to a greater risk of pulmonary consumption. Guy hypothesised that, in fact, the association could be explained by self-selection of workers: the weakest and sick ones were supposed to choose more sedentary works. He interviewed 503 workers and found that only 11 had been influenced by strength or health in their choice of the job. So, selection bias was ruled out (Guy 1843).

Important discussions on bias took place when the concept of study design in epidemiology was refined. In the 1950s, the introduction of the randomised clinical trial into practice and its role as a "gold standard" for medical research lead to the idea that bias could be avoided by randomisation. However, Cornfield already in 1954 stated that the RCT was "one of various inferential instruments" (1954).

In the 1970s, intense discussion started on the merits and limitations of different study designs, in particular the case-control study (which has been considered as a powerful but dangerous tool for a long time). In particular, controversy arose over being the case-control study simply a cohort study seen from the bottom end (the "trohoc" design); over the choice of controls (population vs hospital-based studies); over the criteria for inclusion/exclusion of cases and controls; over criteria for matching, and so on.

In addition to general discussions on the sources of bias, also a *taxonomy* of bias was proposed by Murphy (1976) and Sackett (1979) among others. Murphy (1976) suggested the following examples of bias:

1. Uncontrolled studies
2. Fundamental differences in method of measurement
3. Controlled studies in treatment of "spontaneous" disease (e.g., the growth hormone has profound effects in adolescent boys, while it is not effective in middle-aged dwarfs)
4. Non-simultaneous comparisons
5. Bias of estimation
6. Bias in the assumptions underlying the analysis
7. Bias in hypothesis testing
8. Bias of reporting.

Although "selection" bias is not clearly defined as such by Murphy, he gives an example which can be classified as selection bias: Wood in 1950, in a series of 233 cases of congenital heart disease treated in London found that 5% had a specific disorder, ventricular septal defect. Estimates of 35% and 37% were reported from provincial centres. The explanation given by Wood is that ventricular septal defect is easy to recognise, also at the local level, while difficult diagnoses were referred to London. Therefore, the two series (London vs provincial centres) were not comparable.

The catalogue of 35 types of bias proposed by Sackett (1979) is very detailed, but in fact can be simplified into six categories:

1. Bias in "reading-up" on the field
2. in specifying and selecting the study sample
3. in executing the experimental manoeuvre
4. in measuring exposures and outcomes
5. in analysing the data
6. in interpreting the analysis.

Of the 35 biases, nine are discussed more in detail by Sackett (1979). These deserve some discussion, since they are particularly relevant to the history of epidemiology. The first is the *prevalence-incidence bias*: "a late look at those exposed (or affected) early will miss fatal and other short episodes, plus mild or silent cases and cases in which evidence of exposure disappears with disease onset". *Admission rate bias* refers to the fact that exposed and unexposed cases have different hospital admission rates, so that their relative odds of exposure to the putative cause will be distorted in hospital-based studies. *Unmasking bias* means that "an innocent exposure may become suspect if, rather than causing a disease, it causes a sign or symptom which precipitates a search for the disease". *Non-respondent bias* (i.e.: "non-respondents from a specified sample may exhibit exposures or outcomes which differ from those of respondents") is obviously crucial for all social research. Selection bias seems to be considered by Sackett (1979) as a subgroup of *membership bias*, i.e. "membership in a group may imply a degree of health which differs systematically from that of the general population". *Diagnostic suspicion bias* is described as follows: "a knowl-

Soz.- Präventivmed. 47 (2002) 156–161

Table 1 Effect of nine biases upon observed relative risks in case-control studies (increase or decrease) (from Sackett 1979)

Biases	Observed relative risks increase (+) decrease (–)
Prevalence-incidence bias	+ or –
Admission rate bias	+ or –
Unmasking bias	+
Non-respondent bias	+ or –
Membership bias	+ or –
Diagnostic suspicion bias	+
Exposure suspicion bias	+
Recall bias	+
Family information bias	+

edge of the subject's prior exposure to a putative cause may influence both the intensity and the outcome of the diagnostic process". *Exposure suspicion bias* is defined in this way: "a knowledge of the patient's disease status may influence both the intensity and outcome of a search for exposure". In more modern definitions, exposure suspicion bias and recall bias ("questions about specific exposures may be asked several times of cases but only once of controls") are considered within the same category of information bias. *Family information bias* is a special type of information bias which occurs when families are investigated about both disease status and exposure of their members. Sackett (1979) tries to evaluate the implications of the nine more important biases in terms of distortion of relative risk estimates. The most powerful in distorting measures are diagnostic suspicion bias and different categories of inormation bias (see Tab. 1).

The period of bias "catalogues" has been helpful in defining some of the most important and frequent distortions in the design, analysis, and interpretation of epidemiological studies. The following period is characterised by more formal and systematic definitions.

Modern formal definitions of bias

According to O.S. Miettinen (1985), bias refers to the validity of contrasts we make within epidemiologic studies: "the key to successful design of a non-experimental study in this area, as in general, is the emulation of experimentation". The validity of a randomised clinical trial, considered as the gold standard, rests on three main features:

1. the use of a placebo, i.e. *comparability of effects*
2. the use of randomisation, i.e. *comparability of populations*
3. the use of blinding, i.e. *comparability of information*.

Hence, Miettinen (1985) suggests a classification of bias into comparison, selection and information bias (according to the prevailing type of design error). Such classification is present in other texts of epidemiology, such as Rothman's (1986) and Hennekens and Buring's textbooks (1987). The characteristic of more modern definitions is that they are formal: according to Rothman (1986), selection bias is "a distortion of the effect measured, resulting from procedures used to select subjects that lead to an effect estimate among subjects included in the study different from the estimate obtainable from the entire population theoretically targeted for the study". *Selection bias*, in fact, depends on selection of the exposed/unexposed subjects on the basis (i.e., not independently) of the outcome, or on the selection of the diseased/healthy subjects on the basis of exposure status. Similarly, we have *information bias* when the error of classification on one axis (exposure or outcome) is not independent of the classification on the other axis.

What seems to be common to all types of bias is the fact that the main aspects of study design (selection of subjects, collection of information) are not independent of the *a priori* hypothesis: instead of a factual "truth" we incur in a "logically true" relationship. In statistical terms:

(a) $p(d \mid e \,\&\, d) > p(d \mid \hat{e})$
(b) $p(d \mid e) > p(d \mid \hat{e} \,\&\, d)$

where e = exposure ($\hat{e}$ = lack of)
d = disease ($\hat{d}$ = lack of).

Both (a) and (b) are examples of selection bias: in (a) the exposed group is selected among ill people, like when we start a cohort study by recruiting in the exposed group a "cluster" of exposed cases; in (b) the unexposed group is recruited among healthy people. In both examples an association between exposure and disease is found as a logical, not an empirical truth (i.e., it is necessary, not contingent).

Some historical examples and controversies

Berkson's bias

In 1946, Joseph Berkson published a paper (Berkson 1946) in which he raised doubts about the validity of epidemiologic research within hospital settings. The underlying idea was that the relative frequency of disease in a group of patients who are hospitalised is inherently biased when compared to the population served by the hospital. This phenomenon is attributable to the way in which the probabilities of hospitalisation combine in patients with more than one disease (if you have two diseases, your probability of being hospitalised is greater than the probability associated with either disease separately). Berkson's argument applies in particular to hospital-based case-control studies in which one or more risk factors (especially medicines) are studied in relation to the risk of a specific disease. If, for example, obese people who have hypertension have a higher probability of being hospi-

Soz.- Präventivmed. 47 (2002) 156–161

talised than people with hypertension alone, a spurious association between obesity and anti-hypertensive drugs can be found. People with multiple diseases or conditions become over-represented in the hospital population, and this over-representation affects the distribution of risk factors as well.

Berkson's bias has been considered as an epidemiologic curiosity for a long time, until its reality was empirically demonstrated by Roberts et al. (1978). They re-analysed household surveys designed to capture health utilisation information. Information was gathered for eight clinical conditions and six medications from both hospitalised and non-hospitalised patients. All possible pairs of association were examined in the two groups, and statistically significant differences in relative risks were identified, showing that associations between drug use and specific diseases changed from community-based to hospital-based settings. Examples of associations that might have been distorted by such phenomenon are diabetes mellitus and Bell's palsy, gout and idiopathic heart block, or hypertension and peptic ulcer.

Detection bias: the example of benign breast disease vs breast cancer

When the first studies on the relationship between mammographic patterns and the risk of breast cancer appeared, suggesting that benign breast disease could predispose to cancer, the objection was raised that the observed association could be attributed to "detection bias", i.e., the greater probability that women with benign breast cancer had to undergo detailed examinations, including repeated mammograms, and to have an earlier diagnosis of cancer. This bias was empirically demonstrated by Silber and Horwitz (1986) in a case-control study. They showed that the crude odds ratio for the association between benign breast disease and breast cancer was 2.6 (statistically significant). However, when inequalities in detection of disease were considered by sampling patients according to diagnostic procedures, the association disappeared (OR = 0.9 for mammography patients, 0.8 for biopsy patients).

"Detection bias" is likely to be a common problem in case-control studies in which the risk factor investigated leads to special diagnostic procedures and thus increases the probability that the disease is identified (in contrast to unexposed subjects). Other historical examples are represented by the discussions (1) on the association between the use of reserpine (an anti-hypertensive drug) and the risk of breast cancer and (2) on the risks of women taking sex steroids. A thorough discussion on the methodological aspects of case-control studies on both issues was published in the Journal of Chronic Diseases in 1979 (the same issue in which the classification of biases by Sackett was included), together with a methodological appraisal of detection bias by Feinstein (1979).

Detection bias can be considered within the general category of information bias in that the probability of identifying the diseased people is conditional on the clinical information collected, which is different in the categories of the risk factor.

Healthy worker effect

William Ogle, when studying death rates in different industries (1885), described two difficulties he encountered: the first was "the considerable standard of muscular strength and vigour to be maintained" in order to keep on performing many tasks in the industry. If the individual's health or strength fell below this standard, he was compelled to move to a more suitable activity, or even retire. The second difficulty was that "some occupations may repel, while others attract, the unfit at the age of starting work and, conversely, some occupations may be of necessity recruited from men of supernormal physical condition" (Ogle 1985).

Nearly 100 years after Ogle, Fox and Collier (1976) examined the same phenomenon in terms of standardised mortality ratios with a general population referent. The overall mortality experience of an employed population is known to be more favourable than that of the general population, at least in western countries. The unemployed section of the general population includes people with serious health conditions that hamper their ability to work (this is not necessarily true in the Third World, where manual workers may undergo more superficial pre-hiring visits and may suffer from more severe consequences of workplace exposures). For example, in a study described by Richard Monson, the mortality rate per 1000 per year was 9.1 among white steelworkers, and 15.8 in the general population (non-white: 9.9 and 18.8, respectively) (Monson 1980).

The most widely accepted explanation for the so-called healthy worker effect (HWE), that has been empirically described many times, is selection of the workforce, either as a result of self-selection by the employee or selection by the employer. For this reason the HWE has been considered to be a kind of selection bias by some of the early papers. However, Monson (1980) has claimed that it is not a selection bias, which would occur only if persons with disease were selectively entered into the study cohort. According to Monson, the HWE is a particular type of confounding, since good health status is a determinant of the outcome and is associated with the exposure (employment in industry) (this is the canonical definition of confounding). However, one can argue that selection is based on unknown variables and leads to confounding that cannot be adjusted for. More recently, Arrighi and Hertz-Picciotto (1993) have clarified that the

HWE comprises two complementary processes: (1) an initial selection process whereby healthy people are more likely to seek and gain employment in a specific industry; and (2) a continuing selection process such that those who remain employed will tend to be healthier than those who leave employment. Therefore, modern definitions of the HWE include both of the "difficulties" originally encountered by Ogle.

Bias in clinical research: the Will Rogers phenomenon and "survivor treatment selection bias"

Will Rogers was a humorist-philosopher who described a geographic migration during the American economic depression of the 1930s. He said: "when the Okies (the inhabitants of Oklahoma) left Oklahoma and moved to California, they raised the average intellectual level in both states" (citation from Feinstein et al. 1985). Although the comment is slightly racist, it refers to a general phenomenon: for example, migration of an average soccer player from a very good to a poorly performing team will improve the performances of both. In medicine, the Will Rogers phenomenon refers to better classification of disease stages: if diagnostic sensitivity increases, metastases are recognised earlier, so that the distinction between early and late stages of cancer will improve. Because the prognosis of those who migrated, although worse than that for other members of the good-stage group, is better than that for other members of the bad-stage group, survival rates rise in each group without any change in individual outcomes (Feinstein et al. 1985). The Will Rogers phenomenon has been empirically demonstrated several times.

The *survivor treatment selection bias* is a potentially very serious problem that has been clearly described in AIDS research. In contrast with Berkson's bias, which was described on theoretical grounds decades ago and then empirically demonstrated, this bias has not been identified until recently. The underlying idea is that in an observational study (not in randomised trials), patients who live longer have more opportunities to select treatment, while those who die earlier may be untreated by default. The effect of the bias is to lead erroneously to the conclusion that an ineffective treatment prolongs survival (Glesby & Hoover 1996).

Bias in screening practices

In 1928 two independent investigators, Papanicolaou in the United States and Babes in Rumania, reported that cancer of the cervix could be diagnosed by examining exfoliated cells from the cervical epithelium. Only after the war, however, their technique was systematically introduced in order to detect cervical cancer at early stages and improve survival of the patients. X-ray examination of the breast in asymptomatic women was already advocated in the 1930s by Gershon-Cohen as a means to reduce breast cancer mortality. The first mammography technique was introduced by Egan in the 1960s, and in the same period the first randomised trial (HIP) was started. Among the different methodological issues that were raised concerning mass cancer screening, two peculiar types of bias have been described in early seminal papers by Hutchinson and Shapiro (1968) and Feinleib and Zelen (1969). *Length bias sampling* concerns the fact that individuals who develop a rapidly progressive disease and who are thus more likely to die than the majority of individuals with disease, are unlikely to be found in a population that presents for screening. In other words, the screening programme is likely to select subjects with long-lasting diseases, so that the effectiveness of screening in terms of survival is overstated. Bailar (1978) has indicated how the length bias effect in the US Breast Cancer Detection Demonstration Projects may have loaded the series with individuals who were unlikely to have had much benefit from screening.

A second peculiar kind of bias in screening is the lead time effect. Lead time is defined as the interval between the time of detection by screening and the time at which the disease would have been diagnosed in the absence of screening. It is the time by which screening advances diagnosis of the disease, which does not correspond to the time by which death is postponed.

Screening is affected also by selection bias, since those who participate in screening programmes are not a random sample of the target population, but may be individuals with particularly high or low incidence rates for the disease at issue. For example, participants in screening programmes for cervical cancer are often women with lower risk than the average (high socio-economic status), thus lowering the detection rate of cancer.

Conclusions

Epidemiology is a non-experimental discipline. Although this has been mainly perceived as a weakness, it can also be seen as a strength. In fact, the lack of experiments (except for randomised clinical trials) has lead to a very critical and sophisticated (formal) attitude towards different types and sources of error. The fact that such an attitude was less developed in the social sciences is probably related to the reason that epidemiology has a very practical goal, the identification of single agents for preventive purposes, while social sciences deal with complex realities and do not aim to disentangle simple causal pathways.

Soz.- Präventivmed. 47 (2002) 156–161

Zusamenfassung

Bias

Epidemiologen war schon immer bewusst, wie wichtig es ist, Verzerrungen zu berücksichtigen, auch wenn die Definition von Bias (Verzerrung) sich im Laufe der Zeit änderte. In der Vergangenheit stand die relative Verletzbarkeit verschiedener Studiendesigns durch Bias und unkontrollierbare Störfaktoren (confounding) im Zentrum der Diskussion. Confounding ist dabei klar von Bias zu unterscheiden, als ein Problem von sich vermengenden ursächlichen Effekten, die auf nicht-zufällige Verteilungen von Risikofaktoren innerhalb der Studienpopulation zurückzuführen sind. Die Kontroverse entwickelte sich vor allem um Aspekte der Fall-Kontroll-Studie. Die Formulierung von "Bias-Typologien" während der 1970er Jahre half einige der wichtigsten Quellen für Verzerrung von Design, Analyse und Interpretation epidemiologischer Studien zu definieren. In den darauf folgenden Jahren, d.h. bis heute, ging es dann um mehr formale und systematische Definitionen.

Résumé

Biais

Les épidémiologistes ont toujours été conscients de l'importance de contrôler les distorsions, bien que la définition elle-même du biais ait changé avec le temps. Par le passé, c'était la vulnérabilité aux biais des différents plans d'études qui était au centre de la discussion, ainsi que les effets de confusion non contrôlés (l'effet de confusion en soi est clairement à distinguer du biais, en tant que problème de mélange d'effets causaux dus à la distribution non aléatoire des facteurs de risque dans la population étudiée). En particulier, il y a eu une polémique autour de certains aspects du plan d'étude cas-témoins. La formulation d'une "typologie des biais" au cours des années 1970 a permis de définir certaines des plus importantes sources de distorsion dans le plan, l'analyse et l'interprétation des études épidémiologiques. La période suivante, jusqu'à aujourd'hui, a été caractérisée par des définitions plus formelles et systématiques des biais.

References

Arrighi M, Hertz-Picciotto I (1993). Definitions, sources, magnitude, effect modifiers and strategies of reduction o fthe Healthy Worker Effect. JOM *5*: 890–1.

Bailar JC III (1978). Mammographic screening: a reappraisal of benefits and risks. Clin Obstet Gynecol *21*: 1–14.

Berkson J (1946). Limitations of the application of 4-fold tables to hospital data. Biomet Bull *2*: 47–53.

Cornfield J (1954). Statistical relationhips and proof in medicine. Am Statistician *8*: 19–21.

Feinleib M, Zelen M (1969). Some pitfalls in the evaluation of screening programs. Arch Environ Health *19*: 412–5.

Feinstein AR, Sosin DM, Wells CK (1985). Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. N Engl J Med *312*: 1604–8.

Feinstein AR (1979). Methodologic problems and standards in case-control research. J Chron Dis *32*: 35–41.

Fox AJ, Collier PF (1976). Low mortality rates in industrial cohort studies due to selection for work and survival in the industry. Br J Prev Soc Med *30*: 225–30.

Glesby MJ, Hoover DR (1996). Survivor treatment selection bias in observational studies: examples from the AIDS literature. Ann Intern Med *124*: 999–1005.

Guy WA (1843). Contributions to the knowledge of the influence of employments upon health. J Roy Stat Soc *6*: 197–211.

Hennekens CH, Buring JE (1987). Epidemiology in medicine. Boston, Mass: Little, Brown and Co.

Hutchinson GB, Shapiro S (1968). Lead time gained by diagnostic screening for breast cancer. J Natl Cancer Inst *41*: 665–81.

Miettinen OS (1985). Theoretical epidemiology. New York: J. Wiley.

Monson RR (1980). Occupational epidemiology. Boca Raton: CRC Press.

Murphy EA (1976). The logic of medicine. Baltimore: Johns Hopkins University Press.

Ogle W (1985). Letter to the Registrar-General on the mortality in the registration districts of England and Wales during the ten years 1871–80. Supplement to the 45th Annual Report of the Registrar General of Births, Deaths, and Marriages, in England: xxiii.

Roberts RS, Spitzer WO, Delmore T, Sackett DL (1978). An empirical demonstration of Berkson's bias. J Chron Dis *31*: 119–28.

Rothman KJ (1986). Modern epidemiology. Boston, Mass: Little, Brown and Co.

Sackett DL (1979). Bias in analytic research. J Chron Dis *32*: 51–63.

Silber ALM, Horwitz RL (1986). Detection bias and relation of benign breast disease toi breast cancer. Lancet *i*: 638–40.

Address for correspondence

Prof. Paolo Vineis
University of Torino
via Santena 7
I-10126 Torino

Tel.: +39 011 6706525
Fax: +39 011 6706692

e-mail: paolo.vineis@unito.it

Soz.- Präventivmed. 47 (2002) 156–161