Understanding and Minimizing Epidemiologic Bias in Public Health Research

Bernard C.K. Choi, PhD, MSc¹

Anita W.P. Pak, PhD, MA, MEd²

ABSTRACT

Awareness of potential biases is important for both researchers and policy-makers in public health: for researchers when designing and conducting studies, and for policy-makers when reading study reports and making decisions. This paper explains the meaning and importance of epidemiologic bias in public health and discusses how it arises and what can be done to minimize it. Examples of counting participants in a meeting, to which many policy-makers can relate, are used throughout the paper to illustrate bias in general, random error and systematic error, the effect of sample size, the three main categories of bias (selection, information and confounding), stratification and mathematical modeling.

MeSH terms: Bias (epidemiology); epidemiology; selection bias; observer variation; confounding factors (epidemiology)

Bias is explained in great detail in many epidemiology textbooks¹⁻⁹ and research papers¹⁰⁻¹⁴ but remains underappreciated. Often, bias is explained with detailed classification schemes,^{7,13} complicated conceptual frameworks,^{5,9} and exhaustive lists^{10,12} or is accompanied by sophisticated mathematical proof,^{6,8} to the extent that they may not be readily understood by many public health practitioners.

The objective of this paper is to summarize the basic concept of bias in an accessible way – to explain the meaning and importance of bias in the context of public health research relevant to policy-making, and to discuss how it arises and what can be done to avoid it.

SYNOPSIS

Bias is defined in epidemiology as the deviation of results from the truth.¹⁵ For example, if there are 30 people at a meeting, but you keep counting only 20 participants, then there must be something wrong with your counting. Your bias (deviation) is 20 minus 30, which is -10. In public health practice, bias in research can affect the validity of the evidence base used for developing policies.

Random error (or noise) is not considered to be bias. In the above example, if you count 20 people the first time, 40 the second time, 35 the third time, and 25 the fourth time, then it is random error. When you average the four results, you get 30 participants, which is accurate (unbiased). Bias, on the other hand, is systematic error. If you consistently under- or over-estimate the true number (for example, if you keep counting 20, 19, 21, or 20 participants, with an average of 20), then your results are biased. Increasing sample size reduces random error, but has no effect on bias. So, even if you count 100 times, you will still get a biased estimate if there are systematic errors. There are many possible sources of bias, including selection bias, information bias and confounding bias.^{1,4,13}

Selection bias

Selection bias occurs in the study design stage,^{4,13} for example, when subjects who should be included in the study are not. In the example above, consistent undercounting of participants may be due to the fact that 10 people are sitting in a corner of the

La traduction du résumé se trouve à la fin de l'article.

1. Public Health Agency of Canada, Government of Canada, Ottawa, ON; Department of Public Health Sciences, University of Toronto; and Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, ON

2. Office of Institutional Research, University of Ottawa

Correspondence: Bernard Choi, Public Health Agency of Canada, Government of Canada, AL 6701A, 120 Colonnade Road, Ottawa, ON K1A 1B4, E-mail: Bernard_Choi@phac-aspc.gc.ca

room where they cannot be seen. Being hidden in a corner and not seen, those people are erroneously excluded from (not selected for) the study. Selection bias can be minimized by improved study design, such as by sending staff to completely enumerate all participants, including those sitting in hidden corners.

Information bias

Information bias occurs in the data collection stage,^{4,13} for example, when the results are distorted either by measurement errors or by misclassifying the subjects into wrong categories. In our example, if you count the participants in the room by asking them to raise their hands, you may get the wrong count if some of the participants do not understand your question, are not paying attention, or are too lazy to raise their hands. In this case, even though people are included in (selected for) the study, the solicitation and recording of their information can be incomplete or inaccurate. Sometimes there is a social desirability effect¹² – when everyone else in the room raises a hand, one tends to raise a hand as well, in order not to be singled out as the odd person. Information bias can be minimized by improved data collection methods, such as by asking participants to sign attendance as they enter the room.

Confounding bias

Confounding bias occurs in the data analysis stage,^{4,13} when techniques are not used to control for confounding (that is, when a particular factor is associated with both the exposure and outcome of interest).¹⁵ In our example, the counting of participants may be confounded if there are two different meetings being held simultaneously in the same room. One of the meetings represents the exposure of interest, while the other meeting represents the confounding factor. This makes your counting difficult because the other meeting (confounder) is associated both with your meeting (exposure of interest) – being in the same room – as well as the outcome of interest (number of participants).

Stratification is one method for controlling for confounding bias.^{1,16} To stratify, you would simply ask participants attending one meeting to move to one end of the room, and those attending the other meeting to move to the other end. The problem with

TABLE I

Number of Biases by Source (Based on Choi and Pak, 1998¹²)

Source of Bias	No. of Biases
Literature review	4
Study design	
Faulty selection of subjects	
– Defective sampling frame	9
– Non-random sampling process	9
– Non-coverage of target population	5
Non-comparable comparison groups	7
Inappropriate sample size	1
Study execution	3
Data collection	
Defective measuring instruments	14
Wrong data source	4
Error of the observer	5
Error of the subjects	18
Data handling error	5
Analysis	
Confounding	5
Inappropriate analysis strategies	7
Post hoc analysis	3
Interpretation of results	7
Publication	3
Total	109

TABLE II

Number of Questionnaire Biases by Source (Based on Choi and Pak, 2005¹⁰)

Source of Questionnaire Bias	No. of Biases
The way a question is designed	
Problems with wording	7
Missing or inadequate data for intended purpose	4
Faulty scale	4
Leading questions	3
Intrusiveness	2
Inconsistency	4
The way the questionnaire as a whole is designed	
Formatting problem	3
Questionnaire too long	3
Flawed questionnaire structure	1
How the questionnaire is administered	
Interviewer not objective	2
Respondent's subconscious reaction	2
Respondent's conscious reaction	6
Respondent's learning	2
Respondent's inaccurate recall	4
Cultural differences	1
Total	48

stratification, however, is that those data moved to other strata are effectively lost from the analysis. If many potentially confounding factors must be controlled for simultaneously, a stratified analysis will eventually break down.⁹ There will be too many strata and too few data in each stratum.

Mathematical modeling, such as multiple logistic regression, is another method to control for confounding.^{1,17} It makes smoothed estimates in terms of a model via a mathematical formula, without throwing any data away.⁹ Using our example, you would not have to group the participants physically, but just ask them to wear different coloured hats, for example, black to indicate one meeting and white for the other.

DISCUSSION

This paper only describes epidemiologic bias. It does not deal with statistical bias or

with the common use of the word “bias”. There is a difference in the technical definition of bias as used by statisticians and epidemiologists. For a statistician, bias is usually discussed as a property of an estimator (an estimator is any quantity calculated from the sample data which is used to give information about an unknown quantity in the population, for example, the sample mean is an estimator of the population mean). An unbiased estimator is one where the expected value of the estimator is equal to the value one is estimating.^{18,19} An epidemiologist, on the other hand, thinks of bias as systematic error.^{1,2} For a statistician, estimator bias is often reduced by increasing sample size – this is not the case in epidemiologic bias. There is also a difference between the “common” use of the word bias and its technical use by epidemiologists. As pointed out by Last in his dictionary of epidemiology, “The term bias

does not necessarily carry an imputation of prejudice or other subjective factor, such as the experimenter's desire for a particular outcome. This differs from conventional usage, in which bias refers to a partisan point of view".¹⁵

Furthermore, this paper deals with bias in public health research and not all research. The issues discussed in this paper are those relevant to public health research. For example, selection bias is not relevant to bench-top research into the properties of a particular chemical agent. In addition, the quantitative methods discussed, such as stratification and mathematical modeling, do not apply to qualitative research.

If one is intrigued by the subject of bias, an article in the *Encyclopaedia of Biostatistics* provides a catalogue of 109 biases at various stages of research (Table I).¹² It can serve as a useful checklist for determining and minimizing possible sources of bias. Furthermore, if one is involved in a questionnaire survey, there is a catalogue of 48 questionnaire biases (Table II), each complete with a definition, an example of a bad question, and an example on how to improve.¹⁰ It can be used as a checklist to help improve the quality of questionnaire design.

CONCLUSION

It is difficult to provide a simple overview on the topic of epidemiologic bias. There is always a risk of oversimplification and unexplained complexity.

Awareness of potential biases is the responsibility of not only those who design, carry out and analyze research in public health, but also those who read the research reports and make policy and other decisions based on the results. It is therefore important that everyone understand bias and how to minimize it, in order to improve the generation and uptake of scientific knowledge.

REFERENCES

1. Rothman KJ, Greenland S. *Modern Epidemiology*. Philadelphia, PA: Lippincott-Raven Publishers, 1998.
2. MacMahon B, Trichopoulos D. *Epidemiology: Principles and Methods*. Boston, MA: Little, Brown and Company, 1996.
3. Beaglehole R, Bonita R, Kjellström T. *Basic Epidemiology*. Geneva, Switzerland: World Health Organization, 1993.
4. Monson RR. *Occupational Epidemiology*. Boca Raton, FL: CRC Press, Inc., 1990.
5. Breslow NE, Day NE. *Statistical Methods in Cancer Research*, Volume II. Lyon, France: IARC, 1987.
6. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, CA: Lifetime Learning Publications, 1982.
7. Schlesselman JJ. *Case-control Studies: Design, Conduct, Analysis*. New York, NY: Oxford University Press, 1982.
8. Fleiss JL. *Statistical Methods for Rates and Proportions*. New York, NY: John Wiley & Sons, 1981.
9. Breslow NE, Day NE. *Statistical Methods in Cancer Research*, Volume I. Lyon, France: IARC, 1980.
10. Choi BCK, Pak AWP. A catalog of biases in questionnaires. *Prev Chron Dis* 2005;2:1-13 (Available on-line at: http://www.cdc.gov/pcd/issues/2005/jan/04_0050.htm).
11. Delgado-Rodriguez M, Llorca J. Bias. *J Epidemiol Community Health* 2004;58:635-41.
12. Choi BCK, Pak AWP. Bias, overview. In: Armitage P, Colton T (Eds.), *Encyclopaedia of Biostatistics*, Volume 1. Chichester, England: John Wiley & Sons, 1998;331-38.
13. Choi BCK, Noseworthy AL. Classification, direction and prevention of bias in epidemiologic research. *J Occup Med* 1992;34:265-71.
14. Sackett DL. Bias in analytic research. *J Chron Dis* 1979;32:51-63.
15. Last JM. *A Dictionary of Epidemiology*. New York, NY: Oxford University Press, 2001.
16. Choi BCK, de Guia N, Walsh P. Look before your leap: Stratify before you standardize. *Am J Epidemiol* 1999;149:1087-96.
17. Choi BCK, Howe GR. Methodological issues in case-control studies: IV. Validity and efficiency of various analysis strategies for continuous variables using the unconditional logistic regression model. *Int J Epidemiol* 1984;13:526-32.
18. Larson HJ. *Introduction to the Theory of Statistics*. New York, NY: John Wiley & Sons, Inc., 1973.
19. Wonnacott TH, Wonnacott RJ. *Introductory Statistics*. New York, NY: John Wiley & Sons, Inc., 1969.

Received: June 9, 2004

Revisions requested: November 19, 2004 & February 9, 2005

Revised mss: December 3, 2004 & February 16, 2005

Accepted: February 28, 2005

RÉSUMÉ

Il importe que les chercheurs et les décideurs en santé publique soient conscients des risques de biais. Les chercheurs doivent en tenir compte lorsqu'ils conçoivent et effectuent des études, tout comme les décideurs lorsqu'ils lisent des rapports d'étude et prennent des décisions. Le présent article explique la signification et l'importance du biais épidémiologique en santé publique et montre comment il survient et ce qu'on peut faire pour l'atténuer. Des exemples de comptabilisation des participants à une réunion, exemples proches de la réalité de nombreux décideurs, sont utilisés tout au long de l'article pour illustrer le biais en général, l'erreur aléatoire et l'erreur systématique, l'effet de la taille de l'échantillon, les trois principales catégories de biais (sélection, information et facteurs de confusion), la stratification et la modélisation mathématique.