Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding

Muhammad Mamdani, Kathy Sykora, Ping Li, Sharon-Lise T Normand, David L Streiner, Peter C Austin, Paula A Rochon, Geoffrey M Anderson

Although confounding is an important problem of cohort studies, its effects can be minimised to enable valid comparison

This is the second of three articles on appraising cohort studies

Institute for Clinical Evaluative Sciences, Toronto, ON Canada
Muhammad Mamdani *senior scientist*
Kathy Sykora *senior biostatistician*
Ping Li *analyst*
Peter C Austin *senior scientist*
Department of Health Care Policy, Harvard Medical School, Boston, USA
Sharon-Lise T Normand *professor of health care policy (biostatistics)*
Department of Psychiatry, University of Toronto, ON, Canada
David L Streiner *professor*
Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care, Toronto, ON, Canada
Paula A Rochon *senior scientist*
Department of Health Policy, Management and Evaluation, Faculty of Medicine, University of Toronto, Toronto, ON Canada
Geoffrey M Anderson *chair in health management strategies*
Correspondence to: G M Anderson, Institute for Clinical Evaluative Sciences, 2075 Bayview Avenue, Toronto, ON M4N 3M5, Canada
geoff.anderson@utoronto.ca

BMJ 2005;330:960–2

In cohort studies, who does or does not receive an intervention is determined by practice patterns, personal choice, or policy decisions. This raises the possibility that the intervention and comparison groups may differ in characteristics that affect the study outcome, a problem called selection bias. If these characteristics have independent effects on the observed outcome in each group, they will create differences in outcomes between the groups apart from those related to the interventions being assessed. This effect is known as confounding.[1] In the first paper in the series we dealt with the design and use of cohort studies and how to identify selection bias.[2] This paper focuses on the definition and assessment of confounders.

What is a confounder?

For a characteristic to be a confounder in a particular study, it must meet two criteria.[1] The first is that it must be related to the outcome in terms of prognosis or susceptibility. For example, in the study of the association between antipsychotic use and hip fracture that we considered in the first paper,[2] age is known to be related to risk of hip fracture and therefore has the potential to be a confounder.

The second criterion that defines a confounder is that the distribution of the characteristic is different in the groups being compared. It can differ in terms of either the mean or the degree of variation or variability in that characteristic. For example, for age to be a confounder in a cohort study, either the average age or the variation in the age in the groups being compared would have to be different. Assessing variation as well as average values is important because groups can have the same average value but very different variation. For example, one group with an average age of 70 could include only people aged 70 and another with the same average age could consist of equal proportions of individuals aged 50 and 90. Nevertheless, even a characteristic that is a strong predictor of outcome will not be a confounder if its distribution is balanced between the comparison groups.

In assessing cohort studies, it is important to identify potential confounders and to examine their distribution in the intervention and comparison groups. Below we describe the three questions that need to be answered.

Has there been a systematic effort to identify and measure potential confounders?

Although currently available evidence helps identify potential confounders, the imperfect state of knowledge means that some characteristics related to the outcome may not have been discovered (unknown confounders). Even if a confounder is known, there may be insufficient data to evaluate it.

In randomised controlled trials, all potential confounders (known or unknown) are expected to be evenly distributed between the groups being compared.[3] Cohort studies, however, have no similar protection against confounding and are especially vulnerable to unknown confounders. This does not mean that all cohort studies are inherently invalid. The unknown potential confounders may not have a large independent effect on the outcome of interest and, therefore, even if unevenly distributed, might not result in much bias. Unknown potential confounders may also be evenly distributed between the groups. Nevertheless, all cohort studies should recognise that unknown confounders could affect the results and, as outlined in the next article in this series,[4] investigators should make an effort to determine how sensitive the results are to unknown confounders.

Although unknown confounders are difficult to deal with in cohort studies, a systematic approach can be used to identify known confounders. This should start with a well designed search of comprehensive databases such as Medline. In the context of the study of the relation between antipsychotic use and the outcome of a hip fracture, a review of the literature suggests that risk factors for hip fracture can be broken down into four categories[5–10]:

- Features of medical history—for example, stroke, osteoporosis
- Exposure to drugs—for example, benzodiazepines, oestrogens
- Demographics—for example, age and sex
- Social and behavioural factors—for example, exercise and diet.

Once the potential confounders have been identified, the next step is to develop ways to measure these in the groups being studied. In many cases, especially when using administrative databases, it may not be possible to measure all known confounders. Even if they are measured, the reliability and validity of the measurement technique may be unclear. In the hip fracture and atypical antipsychotic example (see bmj.com for details of how the cohort was created) we used administrative databases to measure known confounders. These databases are poor sources of information on behavioural and social factors. The failure to include measures of these factors has been identified as a key issue in cohort studies of hip

P+ *Further details on the study cohort and propensity scores are on bmj.com*

fracture,[11] and lack of control for lifestyle factors has been suggested to have a key role in the differences in risk of cardiovascular disease seen in cohort and randomised controlled studies of hormone replacement therapy.[12] Although the administrative databases can provide some information on patient history such as previous falls, they may underestimate their true prevalence. It is important to know which confounders have been measured in the study and how well they have been measured.

Is there information on distribution of potential confounders between groups?

Information on the distribution of potential confounders in the intervention and comparison groups is usually provided in the first table of the paper. Confounding is a problem only if these characteristics are unevenly distributed between the intervention and comparison groups. The table provides information on potential confounders for two comparisons examining the association between atypical antipsychotic use and hip fracture. Tables similar to this should be included in all cohort studies so that the reader can have an overview of the potential for selection bias and confounding.

What methods are used to assess differences in distribution of potential confounders?

Perhaps the most common strategy to identify important imbalances in individual confounders between intervention and comparison groups is to use significance tests such as χ^2 tests (for dichotomous variables) or *t* tests (for continuous variables). A problem with these tests is that the significance levels are sensitive to sample size, and the tests are usually not very meaningful when applied to studies with very large numbers of subjects (as is often the case for cohort studies). Under such circumstances, the differences may be significant but not clinically meaningful. For example, in the comparison restricted to people with dementia in the table, a difference of about three months in mean age between groups is significant ($P<0.001$) but may not be clinically relevant. Alternatively, if the samples are small, differences that are clinically meaningful may not be significant. For these reasons this approach to the assessment of differences is of little value.

An alternative to traditional significance testing is to use standardised differences or effect size to examine between group differences in patient characteristics.

CHRIS STOWERS/PANOS

Cohort characteristics can confound only if they vary between comparison groups

Baseline characteristics of study groups in comparisons of atypical antipsychotic versus no drug in all older people, and atypical versus typical antipsychotic drug in older people with dementia. Values are numbers (percentages) of patients unless stated otherwise

	Comparison 1: All older people				Comparison 2: Older people with dementia			
	Atypical antipsychotic (n=34 960)	No antipsychotic (n=1 251 435)	P value	Standardised difference	Atypical antipsychotic (n=21 427)	Typical antipsychotic (n=33 263)	P value	Standardised difference
Age (years):								
Mean (SD)	80.46 (7.63)	74.50 (6.58)	<0.001	0.90	81.69 (7.11)	81.96 (7.17)	<0.001	0.04
Median (interquartile range)	80 (75-86)	73 (69-79)	<0.001	0.90	82 (77-87)	82 (77-87)	<0.001	0.04
No (%) of women	21 720 (62.1)	714 829 (57.1)	<0.001	0.10	13 406 (62.6)	20 151 (60.6)	<0.001	0.04
Recent drug use								
Oestrogen	1 857 (5.3)	84 364 (6.7)	<0.001	0.06	1 000 (4.7)	983 (3.0)	<0.001	0.09
Bisphosphonates	2 323 (6.6)	48 353 (3.9)	<0.001	0.14	1 417 (6.6)	593 (1.8)	<0.001	0.26
Long acting benzodiazepines	1 177 (3.4)	29 917 (2.4)	<0.001	0.06	532 (2.5)	1 192 (3.6)	<0.001	0.06
Short acting benzodiazpeines	15 722 (45.0)	174 990 (14.0)	<0.001	0.88	9 016 (42.1)	14 267 (42.9)	0.06	0.02
Medical history								
Obesity	1 010 (2.9)	51 306 (4.1)	<0.001	0.06	492 (2.3)	945 (2.8)	<0.001	0.03
Previous falls	3 420 (9.8)	31 712 (2.5)	<0.001	0.45	2 460 (11.5)	3 940 (11.8)	0.196	0.01
Osteoporosis	3 509 (10.0)	84 034 (6.7)	<0.001	0.13	2 119 (9.9)	2 206 (6.6)	<0.001	0.12
Stroke	4 334 (12.4)	44 549 (3.6)	<0.001	0.46	2 779 (13.0)	4 638 (13.9)	0.001	0.03
Parkinsonism	3 613 (10.3)	20 990 (1.7)	<0.001	0.64	2 052 (9.6)	3 154 (9.5)	0.713	0.00
Alcoholism	2 014 (5.8)	18 155 (1.5)	<0.001	0.35	1 355 (6.3)	2 344 (7.0)	0.001	0.03
Hyperthyroidism	148 (0.4)	1 631 (0.1)	<0.001	0.08	83 (0.4)	129 (0.4)	0.993	0.00
Hyperparathyroidism	49 (0.1)	562 (0.04)	<0.001	0.04	31 (0.1)	23 (0.1)	0.006	0.02
Chronic renal failure	2 761 (7.9)	50 478 (4.0)	<0.001	0.19	1 656 (7.7)	2 473 (7.4)	0.204	0.01
Asthma or chronic obstructive pulmonary disease	9 014 (25.8)	240 202 (19.2)	<0.001	0.17	5 155 (24.1)	7 934 (23.9)	0.581	0.00
Rheumatoid arthritis	1 782 (5.1)	57 961 (4.6)	<0.001	0.02	1 014 (4.7)	1 752 (5.3)	0.005	0.02
Visual impairment	978 (2.8)	13 323 (1.1)	<0.001	0.17	623 (2.9)	975 (2.9)	0.873	0.00
Dementia	21 427 (61.3)	58 754 (4.7)	<0.001	2.53				.

Standardised differences reflect the mean difference as a percentage of the standard deviation. To estimate these, differences between groups are divided by the pooled standard deviation of the two groups. This measure of the distribution is not as sensitive to sample size as traditional tests and provides a sense of the relative magnitude of differences. Standardised differences of greater than 0.1 are typically felt to be meaningful.[13]

In the table, traditional significance testing found that all 19 potential confounders were significantly different ($P < 0.001$) in comparison 1, and that 13 of the 19 characteristics had standardised differences greater than 0.1. Of particular note is the large standardised difference for history of dementia. Restriction of the study to people with dementia eliminates the possibility of confounding from this characteristic. For comparison 2, traditional significance tests showed that 8 of the 18 potential confounders were significantly different ($P < 0.001$) but only two had a standardised difference greater than 0.1. The use of the standardised differences technique shows that comparison 1 has substantial selection bias, particularly for dementia, whereas comparison 2 has much less potential for bias.

Both traditional significance testing and standardised differences focus on one potential confounder at a time and do not provide an overall perspective on how the comparison groups differ. For example, two groups could have the same mean age and proportion of women, but one could contain old men and young women and the other old women and young men. An increasingly common approach to the analysis of cohort studies of health care interventions is to use propensity score methods[14 15]—a technique that involves multivariate assessment of confounders (see bmj.com for a brief discussion and an example).

Selection bias in cohort studies can result in confounding. Here we have defined questions that can help identify potential confounders. In the next article we will examine statistical methods that can be used to reduce the effect of confounding and strategies that can be used to determine if the results of a study are plausible.

Key questions

Has there been a systematic effort to identify and measure potential confounders?

Is there information on how the potential confounders are distributed between the comparison groups?

What methods are used to assess differences in the distribution of potential confounders?

We thank Jennifer Gold and Monica Lee for help in preparing the manuscript.

Contributors and sources: The series is based on discussions that took place at regular meetings of the Canadian Institute for Health Research chronic disease new emerging team. MM is a clinician with extensive research experience in cohort studies of prescription drugs who wrote the first draft of this article and is the guarantor. SLTN, DLS, and PCA are statisticians who commented on drafts of this paper. KS and PL programmed and conducted analyses. PAR and GMA conceived the idea for the series and GMA worked on drafts of this article and coordinated the development of the series.

Funding: This work was supported by a CIHR operating grant (CIHR No MOP 53124) and a CIHR chronic disease new emerging team programme (NET-54010).

Competing interests: None declared.

1 Altman D, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663-94.
2 Gurwitz JH, Sykora K, Mamdani M, Streiner DL, Garfinkel S, et al. Reader's guide to critical appraisal of cohort studies: 1. Role and design. *BMJ* 2005;330:895-7.
3 Altman D, Bland MJ. Treatment allocation in controlled trials: why randomise? *BMJ* 1999;318:1209.
4 Normand SLT, Sykora K, Li P, Mamdani M, Rochon PA, Anderson GM. Reader's guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *BMJ* (in press).
5 Lawlor DA, Patel R, Ebrahim S. Association between falls in elderly women and chronic diseases and drug use: cross sectional study. *BMJ* 2003;327:1-6.
6 Feskanich D, Willett W, Colditz G. Walking and leisure-time activity and risk of hip fracture in postmenopausal women. *JAMA* 2002;288:2300-6.
7 Haentjens P, Autier PH, Boonen S. Clinical risk factors for hip fracture in elderly women: a case-control study. *J Orthop Trauma* 2002;16:379-85.
8 Cummings SR, Nevitt MC, Browner WS, Stone K, Fox KM, Ensrud KE, et al. Risk factors for hip fracture in white women. *N Engl J Med* 1995;332:767-73.
9 Masud T, Morris RO. Epidemiology of falls. *Age Ageing* 2001;30(suppl 4):3-7.
10 Ensrud KE, Blackwell T, Mangione CM, Bowman PJ, Bauer DC, Schwartz A, et al. Central nervous system active medications and risk for fractures in older women. *Arch Intern Med* 2003;163:949-57.
11 Schneeweiss S, Wang PS. Association between SSRI use and hip fractures and the effect of residual confounding bias in claims database studies. *J Clin Psychopharmacol* 2004;24:632-8.
12 Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and primary prevention of cardiovascular disease, *Ann Intern Med* 2002;137:273-84.
13 Cohen J. *Statistical power analysis for the behavioural sciences.* Hillsdale, NJ: Academic Press, 1988.
14 Rubin DB, Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;127:757-63.
15 Joffe MM, Rosenbaum PR. Propensity scores. *Am J Epidemiol* 1999;150:327-33.

(Accepted 18 February 2005)