Annals of Epidemiology 25 (2015) 955–958

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Brief communication

Noncollapsibility in studies based on nonrepresentative samples

Costanza Pizzi PhD [a,*], Neil Pearce DSc, PhD [b,c], Lorenzo Richiardi MD, PhD [a]

[a] Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, Turin, Italy
[b] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK
[c] Centre for Public Health Research, Massey University, Wellington, New Zealand

ARTICLE INFO

Article history:
Received 19 May 2015
Accepted 14 September 2015
Available online 30 September 2015

Keywords:
Odds ratios
Selection bias
Cohort study

ABSTRACT

Background: It is common to use nonrepresentative samples in observational epidemiologic studies, but there has been debate about whether this introduces bias. In this article, we consider the consequences on noncollapsibility of a sample selection related to a relevant outcome-risk factor.
Methods: We focused on the odds ratio and defined the noncollapsibility effect as the difference between the marginal and the conditional (with respect to the outcome-risk factor) exposure-outcome association. We consider a situation in which the aims of the study require the estimate of a conditional effect.
Results: Using a classical numerical example, which assumes that all variables are binary and that the outcome-risk factor is not an effect modifier, we illustrate that in the selected sample the noncollapsibility effect can either be larger or smaller than in the population-based study, according to whether the selection moves the prevalence of the risk factor closer to or away from 50%. When the outcome-risk factor is also a confounder, the magnitude of the noncollapsibility effect in the selected sample depends on the effects of the selection on both noncollapsibility and confounding.
Conclusions: When a key outcome-risk factor is unmeasured, in presence of noncollapsibility neither a population-based nor a selected study can directly estimate the conditional effect; whether the computable marginal is closer to the conditional in the selected or in the population-based study depends on the underlying population and the selection process.

Introduction

It is common to use nonrepresentative source populations (i.e. those that are not based on the general population of a defined geographical area) in observational epidemiologic research, but there has been considerable debate about whether this introduces bias and to what extent [1–6]. In a recent article on this topic, Rothman et al. [1] emphasized the difference between descriptive studies that describe the specific population in which they are conducted and therefore should rely on representative samples, and studies that aim at “explaining how nature works” and thus focus on scientific inference with no need of representativeness. Ideally, a scientific finding should not be limited to a particular context, but should be applicable (given certain assumptions) to other populations and time periods (see Pearl and Bareinboim [7] for a formal approach on how to transport effects from one population to another). Here, we discuss the consequences of nonrepresentativeness in relation to noncollapsibility, which involve considering the consequences when the selection of the study sample is related to a risk factor for the outcome.

When a binary outcome is not rare and there is a casual effect of an exposure on the outcome, effect measures that are not risk ratios or risk differences, for example, odds ratios (ORs) or rate ratios, are noncollapsible. Formally, a measure of association between an exposure and the outcome is strictly collapsible across a third variable if the marginal effect measure is a weighted average of the stratum-specific (based on the third variable) effect measures [8,9]. On the contrary, in presence of noncollapsibility, the marginal and the conditional effects might differ even when the third variable is neither a confounder nor an effect modifier. It should be emphasized that both the marginal and the conditional effects are interpretable, but only the former is affected by the population-specific distribution of the risk factor. Clearly, the appropriateness of the marginal or the conditional effect depends on the causal structure of the problem investigated and the aim of the study [10]; however in general, if the aim of an epidemiologic study is not descriptive, but is scientific inference, then the conditional effect is more likely to be generalizable and is often the one of main interest. Typically, however, some of the outcome risk factors are unmeasured or unknown, and therefore, only the marginal effect, with respect to

* Corresponding author. Via Santena 7, Turin 10126, Italy. Tel.: +39-011-6334628; fax: +39-011-6334664.
E-mail address: costanza.pizzi@unito.it (C. Pizzi).

http://dx.doi.org/10.1016/j.annepidem.2015.09.007
1047-2797/

the unmeasured/unknown risk factors, can be estimated in the study, even if we were interested in the fully conditional effect (with respect to these risk factors). Under this scenario and assuming no confounding and no effect modification due to these unmeasured/unknown risk factors, when using ORs or rate ratios, the error that we would commit in interpreting the marginal estimate as the conditional one depends on the magnitude of the noncollapsibility effect, that is, the difference between the marginal and the conditional estimate.

In an influential article, Greenland et al. [8] discussed issues of noncollapsibility in epidemiologic studies and described the difference between lack of collapsibility and confounding, providing numerical examples. In our article, we will start from these examples to examine the situation of a nonrepresentative study and to describe the impact of the selection on the noncollapsibility effect, in the specific scenario when the selection is related to an unmeasured and/or unknown outcome risk factor.

Methods

We first consider a scenario involving the effect of an exposure (X) on the outcome (Y) in presence of a risk factor (Z) for Y. This simple scenario is described in Figure 1A, using directed acyclic graphs (DAGs). We focused on the OR, assumed that Z is not an effect modifier on the OR scale, and calculated both the marginal and the conditional (with respect to Z) X-Y associations.

We start with the numerical example presented in Greenland et al. [8] (Table 1), in which X, Z, and Y are all binary variables. From these data, we have generated a corresponding study based on a selected population. We assumed that 60% of subjects with the risk factor ($Z = 1$), and 20% of those without it ($Z = 0$) were included in the restricted cohort ($S = 1$), thus generating a strong positive association between the risk factor and selection into the study (OR = 6.0).

We then changed the scenario, assuming that numbers of the selected sample were the initial population-based numbers, whereas the numbers presented by Greenland et al. [8] were those obtained after the introduction of selection.

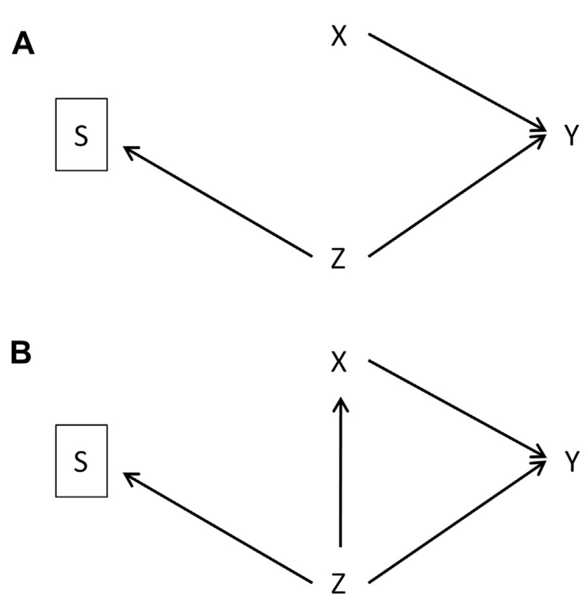

Fig. 1. Diagram of a population-based cohort and of the corresponding selected study. (A) The exposure of interest X affects the outcome Y, which is also caused by the risk factor Z. The probability of being selected as a member of the restricted cohort (S) is affected by the risk factor Z. (B) Z is also associated with the exposure X and therefore acts as a confounder of the X-Y association.

Finally, as in Greenland et al. [8], we considered the scenario in which Z also causes the exposure X and therefore is a confounder for the X-Y association. This scenario is depicted with a DAG in Figure 1B. To generate data for this latter example, we followed the approach used by Greenland et al. [8] and modified the data of Table 1 to induce an association between X and Z. We examined both the scenario with negative confounding, by assuming an OR for the effect of Z on X of 0.5 and the one with positive confounding, by assuming an OR of 2.

Both in the population-based study and in the corresponding selected study (stratum $S = 1$), we calculated the marginal X-Y OR and the two stratum-specific (with respect to Z) X-Y ORs. When investigating the setting of Figure 1B (lack of collapsibility with confounding) to disentangle the confounding bias and the noncollapsibility effect, we calculated the X-Y effect marginalized over Z, using the methods described in the literature [11,12].

Results

The top half of Table 1 (population-based study) shows the same numbers reported by Greenland et al. [8]. The prevalence of each of the three variables X, Z, and Y is 50% with the joint distributions clearly summing to 1 over the Z strata. The marginal and the conditional ORs differ due to lack of collapsibility (marginal OR = 2.25, conditional OR = 2.67). As previously demonstrated, in presence of noncollapsibility, the marginal effect is closer to the null value than the conditional effect (see, e.g., rule 1 in Hauck et al. [13]). The bottom half of Table 1 reports the data that would be obtained after applying the Z-driven selection. In the selected sample ($S = 1$), the prevalence of Z increases to 75%. Noncollapsibility is still present, but its effect is smaller than in the population-based study, as the marginal OR (now equal to 2.33) is closer to the corresponding conditional estimate (OR = 2.67).

If we exchange the population-based sample with the selected sample (i.e., the bottom half of Table 1 now represents the initial population-based sample), then the prevalence of Z is 75%, the stratum specific ORs are equal to 2.67, and the population-based marginal OR is 2.33. The upper part of the table would now represent the selected sample (OR of 0.17 for the effect of Z on S), in which the prevalence of Z would be 50%. The difference between the conditional estimate (2.67) and the marginal estimate (2.25) is now larger in the selected sample ($S = 1$) than in the population-based study. Indeed, when the disease risk factor is binary, a

Table 1
Joint distribution of the exposure (X), risk factor (Z), and outcome (Y) variables. Example of noncollapsibility without confounding of the OR

Study population		$Z = 1$		$Z = 0$		Marginal	
		$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$
Population-based*							
	$Y = 1$	0.2	0.15	0.1	0.05	0.3	0.2
	$Y = 0$	0.05	0.1	0.15	0.2	0.2	0.3
	OR†	2.67		2.67		2.25	
Selected sample‡							
	$Y = 1$	0.3	0.225	0.05	0.025	0.35	0.25
	$Y = 0$	0.075	0.15	0.075	0.1	0.15	0.25
	OR†	2.67		2.67		2.33	

* Data of Table 1 of Greenland et al. [8].
† OR = odds ratios.
‡ 60% of subjects with $Z = 1$ and 20% of subjects with $Z = 0$ have been included in the selected sample.