Noncollapsibility in studies based on nonrepresentative samples

Non-collapsibility in studies based on non-representative samples

Costanza Pizzi[1], Neil Pearce[2,3], Lorenzo Richiardi[1]

[1] Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, Via Santena 7, 10126, Turin, Italy
[2] Departments of Medical Statistics and Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK
[3] Centre for Public Health Research, Massey University, Wellington, New Zealand

Corresponding author:

Costanza Pizzi
Address: Via Santena 7, 10126, Turin, Italy
Email: costanza.pizzi@unito.it
Tel: +39 011 6334628
Fax: +39 011 6334664

ABSTRACT

The potential consequences of using non-representative study samples in observational epidemiological research on validity have been discussed, but some specific issues remain to be explored. In this paper we discuss the consequences of selecting the study sample on a relevant outcome risk factor in relation to non-collapsibility.

In presence of non-collapsibility due to an outcome risk factor, the conditional estimates are the same in the selected and the corresponding population-based study, while the marginal estimates differ. To explore this phenomenon, we focused on the odds ratio estimate and defined the non-collapsibility effect as the difference between the marginal and the conditional (with respect to the outcome risk factor) exposure-outcome association.

Starting from a classical numerical examples used in the literature on non-collapsibility, we illustrate that in the selected stratum the collapsibility effect differs from that found before the selection. It can either be decreased or increased, according to whether the selection moves away from or closer to 50% the prevalence of the outcome risk factor among the selected subjects. When the outcome risk factor is also a confounder, the difference between the conditional and the marginal effect in the selected sample depends on the combination of the effects that the selection has on non-collapsibility and control of confounding.

Keywords: Representativeness, collapsibility, selected sample, odds ratio, marginal effect, conditional effect

INTRODUCTION

Attention is been given to the use of non-representative source populations (i.e. those that are not based on the general population of a defined geographical area) in observational epidemiological research. In a series of debate papers on this topic, Rothman et al [1] emphasized the difference between studies with descriptive purposes, which 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 directly applicable to other populations and time periods.

The potential consequences of using non-representative samples on validity have been examined [2-6], especially in the framework of cohort study designs, but some specific issues remain to be explored. Here we discuss the consequences of non-representativeness in relation to non-collapsibility, which implies discussing the consequences of selecting the study sample on 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 Hazard Ratios (rate ratios), face the mathematical problem of being non-collapsible. Briefly, this means that, if strata based on another outcome risk factor are created, the marginal effect of the exposure may differ from the weighted average of the stratum-specific effects (conditional effect) even when this other risk factor is neither a confounder nor an effect modifier [6]. 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. Typically, however, some of the outcome risk factors are unmeasured or unknown, and therefore only the marginal effect, with respect to the unmeasured/unknown risk factors, can be estimated in the study, even if we would be interested in the fully conditional (with respect to these risk factors) effect. Under this scenario, and assuming no confounding due to these unmeasured/unknown risk factors, when using ORs or HRs, the error that we would commit in interpreting the marginal estimate as the conditional one depends on the magnitude of the non-collapsibility effect, i.e. the difference between the marginal and the conditional estimate.

In an influential paper, Greenland et al discussed issues of non-collapsibility in epidemiological studies, and described the difference between lack of collapsibility and confounding, providing numerical examples [6]. In our paper, we will start from these examples to examine the situation of a non-representative study, and to describe the impact of the selection on the non-collapsibility effect, in the specific scenario when the selection depends on an unmeasured/unknown outcome risk factor.

METHODS

We initially considered a scenario aiming to assess the effect of an exposure (E) 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 started form the numerical examples presented in Greenland et al [6] (Table 1), in which X, Z and Y are all binary variables. From these data we 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 shifted the scenario, assuming that numbers of the selected sample were the initial population-based numbers, while the numbers presented by Greenland et al. [6] were those obtained after the introduction of selection.
Finally, as in Greenland et al [6], we considered the scenario in which Z causes also 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 [6] 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) in order to disentangle the confounding bias and the non-collapsibility effect we calculated the X-Y effect marginalized over Z, using the formula described in Pang et al [7].

RESULTS

The top half of Table 1 (population-based study) reports the same numbers used by Greenland et al [8]. The prevalence of each of the three variables X, Z and Y is 50% and the marginal and the conditional ORs differ due to lack of collapsibility (marginal OR=2.25, conditional OR=2.67). As previously demonstrated, in the presence of non-collapsibility, the marginal effect is closer to the null value than the conditional effect (see, for example, Rule 1 in Hauck et al [8]). 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%. Non-collapsibility 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 population based-sample to start with), we deal with a scenario in which 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 now would 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 prevalence of 50% maximizes the non-collapsibility effect [7]. Hence selection increases non-collapsibility among the selected subjects if it brings the prevalence of Z closer to 50% and decreases it if it moves the prevalence of Z away from 50%.

The numbers reported in Table 2 refer to the scenario of Figure 1B and have been created by modifying the data of Table 1 to induce negative confounding. Due to the joint impact of negative confounding and non-collapsibility, the marginal effect of X on Y (OR=1.71) is now even further away from the conditional one (OR=2.67). The total difference between the conditional and the marginal crude effect can be decomposed in two parts [7]: i) confounding bias, i.e. the difference between the crude marginal (OR=1.71) and the unconfounded marginal effect (OR=2.25 Table 2), and ii) the non-collapsibility effect, i.e. the distance between the unconfounded marginal and the conditional effect. In the corresponding selected sample, the crude marginal OR is 1.96, while the unconfounded marginal OR is 2.37 (bottom part of Table 2), thus showing a reduction of both the confounding bias and the non-collapsibility effect. Note that the prevalence of Z goes from 57% in the population-based study to 80% in the selected sample, thus explaining the decreased non-collapsibility effect. The decrease in confounding bias is due to partial control of the confounder Z through conditioning on S. This always holds for binary variables, provided that Z does not qualitatively interact with the exposure X [9]. More stringent assumptions are needed for polytomous risk factors [10].
When Z is a positive confounder (data not shown in Tables), due to the opposite directions of the confounding bias and the non-collapsibility effect, in the population-based study the crude marginal (OR=3.00) is larger than the conditional estimate (OR=2.67). When computing the same estimates in the selected sample, the crude marginal OR is 2.97, so that the distance between the crude marginal and the conditional effects (2.97 vs 2.67) is only slightly smaller than the same difference obtained in the population-based study (3.00 vs 2.67). This happens because the confounding bias and the collapsibility effects cancel out instead of summing up. In this scenario, in the selected sample confounding decreases due to partial conditioning and the collapsibility effect also slightly decreases because the prevalence of Z is closer to 50% in the population study (40%) than in the selected sample (67%).

DISCUSSION

In this paper we have described the consequences on non-collapsibility of restricting the study to a sample selected with respect to an outcome risk factor. In presence of non-collapsibility, the conditional estimates (if all relevant risk factors are taken into account) are the same in the selected and the population-based study, while the marginal estimates differ. The difference is appreciable when selection is strongly affected by the risk factor and the collapsibility effect is not negligible.

We argue that, in presence of non-collapsibility, the causal parameter of main interest in non-descriptive epidemiological studies is the conditional estimate, as this is less time and population specific, and could thus be more easily generalized. If a strong outcome risk factor is unknown/unmeasured or, anyhow, not controlled for in the analysis, the matter is the distance between the marginal and the conditional estimate. As we have illustrated, when selection depends on the risk factor, among the selected subjects this distance can either be smaller or larger than in the corresponding population-based study. For example, if smoking were the (unmeasured) risk factor introducing non-collapsibility and the population prevalence of smoking were, say, 30%, a cohort study in which smokers are less likely to participate would be less affected by non-collapsibility than the equivalent population-based study.

Often, in a specific population the risk factor that is introducing non-collapsibility problems is also a confounder (but not an effect modifier). Again, in this scenario, the best approach is to control for the risk factor, but, if this is not possible, a study selected on the risk factor is likely to be less affected by confounding bias due to partial control of the confounder and therefore, at least when the risk factor is binary [9-10], is expected to produce on average marginal estimate closer to the true conditional effect than the corresponding unselected study. As we have illustrated, the overall gain in validity depends on the combination of the effects that the selection has on non-collapsibility and control of confounding.

ACKNOWLEDGMENTS

We wish to thank Bianca De Stavola and Rhian Daniel (LSHTM), and the ICE (Inferenza Causale in Epidemiologia) working group for many relevant insights.

This work was supported by Compagnia SanPaolo/FIRMS. The Centre for Public Health Research is supported by a Programme Grant from the Health Research Council of New Zealand.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ETHICAL STANDARDS

The manuscript does not contain clinical studies or patient data.

BIBLIOGRAPHY

1. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. Int J Epidemiol. 2013;42:1012-4.
2. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology. 2003;14:300-6.
3. Pizzi C, De Stavola B, Pearce N, Lazzarato F, Ghiotti P, Merletti F, Richardi L. Selection bias and patterns of confounding in cohort studies: the case of the NINFEA web-based birth cohort. J Epidemiol Community Health. 2012;66:976-81.
4. Richiardi L, Pizzi C, Pearce N. Commentary: Representativeness is usually not necessary and often should be avoided. Int J Epidemiol. 2013; 42:1018-22.
5. Bareinboim E, Tian J, Pearl J. Recovering from Selection Bias in Causal and Statistical Inference. In Proceedings of The Twenty-Eighth Conference on Artificial Intelligence (CE Brodley and P. Stone, eds.). AAAI Press, Menlo Park, CA. 2014.
6. Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. Statistical Science. 1999;14:29-46.
7. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Stat Methods Med Res. 2013. [Epub ahead of print]
8. Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol. 1991;44:77-81.
9. Ogburn EL, VanderWeele TJ. On the Nondifferential Misclassification of a Binary Confounder. Epidemiology. 2012; 23(3):433–439.
10. Brenner H. Bias due to non-differential misclassification of polytomous confounders. J Clin Epidemiol. 1993; 46(1):57-63.

Tables

Table 1. Joint distribution of the exposure (X), risk factor (Z) and outcome (Y) variables. Example of non-collapsibility without confounding of the Odds ratios (OR).

		Z=1		*Z=0*		**Marginal**	
		X=1	*X=0*	*X=1*	*X=0*	*X=1*	*X=0*
Population-based study [a]							
	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 [b]	**2.67**		**2.67**		**2.25**	
Selected sample [c]							
	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 [b]	**2.67**		**2.67**		**2.33**	

[a] *Data of Table 1 of Greenland et al [7]*
[b] *OR = Odds ratios*
[c] *60% of subjects with Z=1 and 20% of subjects with Z=0 have been included in the selected sample*

Table 2. Joint distribution of the exposure (X), risk factor (Z) and outcome (Y) variables. Example of non-collapsibility with negative confounding of the Odds ratios (OR).

		Stratum-specific				**Marginal**		
		Z=1		*Z=0*		Crude		Unconfounded [a]
		X=1	*X=0*	*X=1*	*X=0*	*X=1*	*X=0*	
Population-based study [b]								
	Y=1	0.2286	0.1714	0.1143	0.0286	0.3429	0.2	
	Y=0	0.0571	0.1143	0.1714	0.1143	0.2285	0.2286	
OR [c]		**2.667**		**2.667**		**1.71**		**2.25**
Selected sample [d]								
	Y=1	0.32	0.24	0.0533	0.0133	0.3733	0.2533	
	Y=0	0.08	0.16	0.08	0.0533	0.16	0.2133	
OR [c]		**2.667**		**2.667**		**1.96**		**2.37**

[a] *Marginal (over the confounder Z) effect analytically calculated using the formula as described in Pang et al [9]*
[b] *Data derived from Population-based study of Table 1 allowing for an OR for the effect of Z on X of 0.5*
[c] *OR = Odds ratios*
[d] *60% of subjects with Z=1 and 20% of subjects with Z=0 have been included in the selected sample*

Figure 1. Diagram of a population-based cohort and of the corresponding selected study. A) In the population 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) In the population Z is also associated with the exposure X, and therefore acts as a confounder of the X-Y association.