Baseline selection on a collider: A ubiquitous mechanism occurring in both representative and selected cohort studies

Baseline selection on a collider: a ubiquitous mechanism occurring in both representative and selected cohort studies

Lorenzo Richiardi[1,2], Neil Pearce[3], Eva Pagano[2], Daniela Di Cuonzo[1,2], Daniela Zugna[1], Costanza Pizzi[1]

1. Department of Medical Sciences, University of Turin, Italy
2. University Hospital Città della Salute e della Scienza di Torino and CPO-Piemonte, Turin, Italy
3. Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine

Correspondence to: Lorenzo Richiardi, Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin and CPO-Piemonte, Via Santena, 7 – 10126 Torino, Italy. Tel: +390116334673. Fax: +390116334664. Email: lorenzo.richiardi@unito.it

Running head: Selection on a collider in cohort studies.

Word count: abstract, 239; text, 3323; 1 Figure, 7 Tables

Contributors: LR, DZ, NP and CP contributed to the concept and design; EP, DDC and LR contributed to the acquisition and analysis of the data; all authors contributed to the interpretation of the results and to critically revising the manuscript; LR and NP drafted the work; All authors gave final approval and agree to be accountable for all aspects ensuring integrity and accuracy.

Funding: LR, CP and DZ received funding from the European Union's Horizon 2020 Research and Innovation Programme [LifeCycle project, grant agreement number 733206]. NP's involvement in this work was supported by the European Research Council under the European Union's Seventh Framework Programme [FP7/2007-2013 / ERC grant agreement number 668954].

Competing Interests: No, there are no competing interests for any author

Ethics: No research approval was obtained as the included exampes are based on deidentified administrive data

What is already know

- There is debate as to whether cohort studies that are based on a selected source population are more prone to bias than those that are based on a representative source population
- The debate lies mainly on the possibility that in selected cohorts the associations between the exposure of interest and the outcome risk factors are altered by collider bias

What this study adds

- Both representative and selected source populations may be affected by underlying selection processes linked to the exposure of interest and the outcome risk factors; therefore collider bias may occur in representative cohorts as well as in selected cohorts
- To identify which risk factors should be controlled for to obtain an unbiased estimate in prospective cohort studies, regardless whether they are based on a representative or a selected source population, it is essential to consider the mechanisms that led individuals to be members of such a source population

Abstract

There is debate as to whether cohort studies are valid when they are based on a source population that is non-representative of a given general population. This baseline selection may introduce collider bias if the exposure of interest and some other outcome risk factors affect the probability of being in the source population, thus altering the associations between the exposure and those risk factors. We argue that this mechanism is not specific to 'selected cohorts' and also occurs in 'representative cohorts' due to the selection processes that occur in any population. These selection processes are for example linked to the life status, immigration and emigration, which, in turn, may be affected by environmental and social determinants, lifestyles and genetics. We provide real-world examples of this phenomenon using data on the population of the Piedmont Region, Italy. In addition to well-recognised mechanisms, such as shared common causes, the associations between the exposure of interest and the risk factors for the outcome of interest in any source population are potentially shaped by collider bias due to the underlying selection processes. We conclude that, when conducting a cohort study, different source populations, whether 'selected' or 'representative', may lead to different exposure - outcome risk factor associations, and thus different degrees of lack of exchangeability, but that one approach is not inherently more or less biased than the other. The key issue is whether the relevant risk factors can be identified and controlled.

Keywords: collider bias, cohort studies, bias, selection, representativeness

Introduction

The literature on cohort studies distinguishes between 'selected cohorts', based on a selected source population, and 'representative cohorts', based on a given general population,[1-6] tipically defined as a collection of people who share the specific geographical location they inhabit in a specific period of time.[7] Some authors also link representativeness to the concept of a target population to which we wish to draw inference.[8] While the general population is typically defined by geographical boundaries and calendar time, the target population depends on the research question. Some research questions may imply a well-defined target (e.g. 'what is the lung cancer burden due to smoking in Italy in 2018?'), while in other instances the target is less clear (e.g. 'what is the effect of smoking on lung cancer [in humans]?'). In fact, the target population is often not defined, and is usually assumed to be potentially 'all humans' or 'all humans with a specific characteristic'; however, in cohort studies, the concept of representativeness is usually applied to the general population during a particular time period and not to a well-defined target population.

A classic example of a selected cohort would be the British Doctors' Study.[9] The aim was to investigate the health effects of smoking in general and not specifically in British doctors, but this group was chosen for practical and scientific reasons. Another example is the Internet-based NINFEA birth cohort that is restricted to Internet users.[10] Studies such as EPIC,[11] ALSPAC[12] or the Framingham cohort[13] instead are representative cohorts as they recruited a representative sample of a general population (within given age criteria). Regardless of whether they are selected or representative, these studies are intended to yield findings that apply beyond their source populations. We do not intend to address external validity in this paper (our focus is on internal validity), but we would briefly note that the results obtained in a given study population can be: formally generalized, typically using marginal effects, to a specific population from which the study sample originates; formally transported to other general or target populations; or used to make a

scientific generalization, which is described, for example, by Rothman and colleagues as a 'a process of constructing a correct statement about the way nature works'. [2-3,6,14]

The restriction of the source population is an 'intentional' selection, based on criteria established by the researcher.[5] A second form of baseline selection in cohort studies refers to the difference between the source population and the study population, i.e. between those who are eligible to participate and those who are actually recruited in the study. This form of baseline selection is 'unintentional', and depends on characteristics that are not defined by the researcher.[15]

In prospective cohorts, baseline selection is not affected by the outcome under study as, by design, prevalent cases should be excluded from the study sample at baseline.[16] In other words, the members of a cohort should be free from the outcome of interest at the beginning of the follow-up; if this principle is met, baseline selection can be affected by the exposure of interest and its determinants or by outcome risk factors, but cannot be directly affected by the outcome. This assumption would be violated if a cross-sectional analysis is conducted at baseline in a cohort study[17] or if some prevalent cases are undiagnosed. We will focus on cohort studies in which prevalent cases at baseline have been excluded, and the only concern is selection on exposure and/or other risk factors for the outcome. In cohort studies, the study population can also be selected over time because of loss to follow-up. This is a radically different type of selection, as loss to follow-up can be affected directly by the outcome of interest.[18] The paper mainly focuses on intentional baseline selection and we will not further discuss loss to follow-up.

Baseline collider bias in selected cohort studies

Some authors have argued that if the study aims to obtain a valid estimate of the effect of the exposure on the outcome in the study sample, restriction of the source population does not

introduce bias.[6] Other authors argue that, on the top of generalizability issues, cohort studies based on selected source populations are also more prone to bias than representative studies.[1, 17] The debate lies mainly on the possibility that baseline selection may induce collider bias (i.e. "baseline collider bias"): if the exposure of interest and another outcome risk factor (i.e. an outcome determinant – or its proxy - that is not caused by the exposure) are both associated with the probability of being in the selected source population, then membership of the source population becomes a collider that is inherently conditioned on, thus inducing an association between the exposure and the risk factor. This mechanism is shown in Fig. 1 using a directed acyclic graph (DAG) and has been discussed several times.[5, 16-24] In Fig. 1, the association between the exposure E and the outcome risk factor R is induced by baseline collider bias due to restriction to the selected source population S, resulting in lack of exchangeability if the risk factor is not controlled for. As a more general scenario, the exposure and the risk factors may be already associated in the general population, for example if they share common causes.[5] Under those scenarios the use of a selected source population may simply alter those associations, provided that membership in the source population is a collider. Hence the associations between the exposure and the outcome risk factors in a selected source population is due to the potential combination of different mechanisms, including baseline collider bias.

We will argue here that baseline collider bias potentially applies to all cohorts, since all general populations, of which a cohort may be representative, are subject to selection processes and those selection processes may be associated with the exposure of interest and the outcome risk factors. We will thus discuss 'population selection processes' and their impact on lack of exchangeability in the source population.

This paper is organized in three parts: we will first illustrate the concept that baseline collider bias may occur also in representative source populations, we will then provide two real examples of this phenomenon, and finally we will discuss its implications.

Baseline collider bias in representative cohort studies

Let us suppose that we aim to conduct a representative cohort study of residents of a given city in a given time period (the "general population"). Baseline collider bias may occur in this representative source population if selection processes in the underlying general population are linked both to the exposure under study and the risk factors for the outcome of interest. This mechanism can be seen in Fig. 1, if S stands for a representative instead of a selected source population . For example, if E is age and R is sex, both age and sex are likely to affect the probability of being alive and a member of a given general population. Even if age and sex are not expected to be associated (age only depends on year of birth and calendar year, sex is genetically determined) they become associated because of conditioning on a collider (membership of the representative source population). A representative cohort intended to assess the causal effect of sex on a given outcome, say cardiovascular diseases, would thus have to control for age. This is not a novelty (almost all analyses involve adjustment for age) but we note that the reason why we should adjust for age in this example is baseline collider bias, not classical confounding by age.

We can indeed imagine more complicated scenarios involving factors that also share common causes. For example, if E is smoking and R is heavy alcohol drinking, they are associated also because they are affected by socioeconomic position. The association between smoking and alcohol at baseline in a representative cohort would thus depend on the combination of shared causes and baseline collider bias. In addition, individuals leave and join populations and their decisions as to whether to remain in a given city in a given year, or to move to that city, may be affected by a large

number of factors. Therefore, in Fig. 1, E and R could stand for occupational status, smoking, obesity, educational level, mental health, general health status, family composition, genetic factors, air pollution, noise, war, poverty, climate change, etc.

The example of the Piedmont general population

To analyse a real-world example, we extracted from the database of the general practitioners' lists the data of the resident population of the Piedmont region, Italy, on April 30, 2018. For each individual, we obtained information on age in 2018, place of birth (Piedmont, elsewhere in Italy, or abroad) residence, sex, and vital status on August 31, 2018. For simplicity we did not consider migration among municipalities within Piedmont, even if only 25% of the Piedmont population was resident in the same municipality of birth. Data on people who were born in the Piedmont region and had left Piedmont were not available in this database.

We restricted the database to individuals aged less than 90 years to avoid outliers or incorrect registrations. We excluded 52,600 subjects resident in municipalities that after 1991 acquired new areas, were newly established or changed name. We also excluded 6,516 subjects resident in municipalities with less than 150 residents in 2018. As reported in Table 1, out of a population of 4,427,450 individuals, 35.6% had joined the Piedmont population from outside the region. Table 2 reports the data for three selected municipalities in which a hypothetical representative cohort could be conducted: the city of Torino, which is the capital of the region; the municipality of Moncalieri, which is close to Torino and is highly residential; the municipality of Cuneo, that is the main city of a large more rural area. It is likely that the key determinants for the individuals to be resident in their municipality vary over the three municipalities. The direction and strength of collider bias will vary accordingly.

As an example, we show in Table 3 the associations of place of birth with sex and age in the three municipalities. Sex, age and place of birth are potential determinants of being present in the population and they can also interact in this selection process (e.g. migrants from a specific area might be mainly of a specific age). Therefore, associations of place of birth with sex and age in a specific general population may indicate the presence of collider bias in that population.

As reported in Table 3, the direction and strength of the association between sex and place of birth varied over the three municipalities. Similarly, age was strongly associated with being born outside the Piedmont region in Torino and Moncalieri, but less so in Cuneo. To show the potential impact of this source of collider bias, we used a log-binomial model to analyse the association between place of birth and 4-month mortality in the three municipalities (Table 4). In Torino and Moncalieri the risk ratio (RR) of 4-month mortality for being born in other Italian regions than Piedmont compared to being born in Piedmont decreased from an estimate larger than 2.0 to almost 1.0 after adjusting for age, while in Cuneo the age-unadjusted RR was already close to 1.0. Again the potentially novel message of this example is not that the estimates had to be adjusted by age, but that the reason to adjust for age was baseline collider bias.

The example of a representative conception cohort

The population selection process and the consequent collider bias occur because the general population is dynamic, due to immigration, emigration, and death. It could be then argued that a representative conception cohort would be immune to these mechanisms, but conception cohorts are naturally selected by the probability of getting pregnant (i.e. by the choice of the target population). This can be seen in Fig.1 if S stands for being pregnant, which is potentially affected by several factors; the mutual associations among these factors in the population are thus shaped by collider bias.

To analyse a real-world example, we extracted the data of residents of the Piedmont region, Italy, in 2013 who were also listed in the 2011 census, and restricted to women aged 18 to 44 years on the 1st of January 2013 (as a proxy of fertile age). In this population, as a proxy of the period prevalence of pregnacy in 2012-2013, we identified hospital admissions in 2013 for abortion, miscarriage, spontaneous delivery and caesarean section using an algorithm based on combinations of ICD-9-CM codes (main or secondary diagnosis: V27.xx or 650 or 669.7 or 640.x1-676.x1 or 640.x2-676.x2, V30.01, V31.01, V32.01, V33.01, V34.01, V36.01, V37.01, V39.01; procedure: 69.01-69.02, 69.52, 72.x, 73.2x, 73.5x, 73.6x, 73.8, 73.9x, 74.0, 74.1, 74.2, 74.4, 74.99, 75.0, 96.49). It should be noted that this combination of codes is unable to identify all spontaneous abortions that occurred before the identification of the pregnancy. Our approach also missed pregnancies that were initiated in 2012 and led to a premature birth or abortion before the 1st of January 2013. We identified 35,531 women who were pregnant in the period 2012-2013 and concluded the pregnancy in 2013, out of an overall population of 673,821 women. From the records of the 2011 census we obtained pre-pregnancy information on: educational level, occupation, number of family members, and place of birth. These variables were categorized as shown in Table 5. We excluded 14 women because of missing information on educational level and/or the occupational status and 681 women because they had a number of family members above 15, which was arbitrarily chosen to identify unusual family or residential compositions. For these variables we used a log-binomial regression to estimate the prevalence ratio of being pregnant in 2012-2013. Then, to understand the role of collider bias, we investigated the association between age and being born outside Italy in the whole population and in the pregnant population. Finally, retsricted to live births (n= 28,790), we estimated the crude and age-adjusted (using cubic splines and 5 knots) odds ratio of caesarean section (n= 8,932), which, admittedly, is only a proxy of the actual risk ratio of caesarean section as we did not have information on pregnancy duration and could not reconstruct the assumed conception date.

As reported in Table 5, most of the selected variables were associated with the probability of being pregnant. Some of the variables also interacted, as, for example age and educational level, and age and place of birth (p-values for departure from multiplicative interaction: 0.0001; data not reported in the Table).

Table 6 reports the association between age and place of birth in the whole population of women aged 18-44 years and the subgroup of pregnant women of the same age. Age and being born outside Italy were positively associated in the general population, while they were negatively associated among pregnant women. This difference is attributable to the conditioning on being pregnant, i.e. to the choice of the target population. This induced association between place of birth and age translated into a change in the estimate (beyond issues of collapsibility) of the odds ratio of caesarean section for being born outside Italy compared to being born in Italy (Table 7). The crude OR was 0.89 (95% CI: 0.84-0.95), while the age-adjusted OR was 1.01 (95% CI: 0.95-1.07).

Discussion

The associations between the exposure of interest and the outcome risk factors in any source population are potentially shaped by collider bias due to the underlying selection processes in that population. Therefore baseline collider bias may affect both cohorts that are representative of a given general or target population and cohorts that are based on a selected source population.

The potential magnitude and direction of collider bias has been assessed in several studies, based on algebraic calculations[19,20] or simulations[17,23]. The importance of baseline collider bias in shaping the associations between the exposure and the outcome risk factors depends on the strengths of the associations of these variables with the membership in the source population. Sometimes collider bias can induce strong associations,[25] in other instances it may attenuate the associations caused by

the other mechanisms.[24] In our view, the key issue is not whether collider bias occurs or not in a given (representative or selected) source population, but whether we are able or not to obtain exchangeability in that source population.

We have focused on intentional baseline selection, but both selected and representative prospective cohorts can be affected also by unintentional baseline selection due to volunteering or non-response at recruitment. In prospective cohort studies typically the causal structure of unintentional selection resembles that of intentional selection. To depict this graphically it would be enough to consider in Fig. 1 that S stands for 'study population' instead of 'source population'. An exception is when subjects with an undiagnosed outcome of interest cannot be excluded from the source population and early manifestations of that disease directly affect their participation in the study. Under this scenario, the outcome (D) could directly affect the selection (S), thus hampering the possibility to obtain a valid exposure-outcome estimate even if the outcome risk factor R is known and measured.

In our view, the key issue to consider when discussing the impact of baseline selection remains the research question. If a study aims at assessing the effect of an exposure on an outcome in a given population in a given time, then a representative source population is likely to be the best option. The distribution of the mediators, the distribution of the effect modifiers, the confounding structure would reflect those of the general population. In addition, the focus would be on the marginal effects.[26] However, most studies are intended to generalize findings beyond a particular setting and time,[2] because preventive interventions based on those findings are typically carried out several years after the recruitment of a particular cohort. In those situations, we should obtain valid effect estimates, typically conditional,[26] and then understand how they can be transported to other contexts and populations.[3,27]

Both representative and selected source populations may be affected by underlying selection processes linked to the exposure of interest and the outcome risk factors. This leads to collider bias that concurs, together with other mechanisms, including classical confounding, to the definition of which outcome risk factors should be measured and controlled for to aim at exchangeability and an unbiased estimate of the effect of the exposure on the outcome. There is thus no a priori reason to expect that one type of source population will involve more or less bias than the other. Clearly the identification and measurement of potential confounders is a key issue in any observational study including prospective cohorts, and any observational study may be affected by a certain degree of residual confounding. The rules to identify those confounders, for example avoiding to adjust for mediators, are the same in selected and representative cohorts, although a particular selected source population may be chosen because confounding is likely to be small (e.g. in comparisons between different occupational groups), or because information on confounders is available (e.g. one might study participants in a health insurance scheme because confounder information is available, whereas it may not be available for the general population). We argue that it is also relevant to consider the mechanisms that led individuals to be members of a given source population, which involve the selection processes occurring in the underlying population and the potential intentional restriction introduced by the researcher. If the exposure of interest is not (or weakly) linked to those selection processes, baseline collider bias is unlikely. Otherwise, any outcome risk factor, which is not affected by the exposure and could potentially be associated with the selection into the source population, should be considered as an additional variable to adjust for in the study. We conclude that, to take into account baseline collider bias, the a-priori knowledge on the selection mechanisms in the source population should be considered when planning and analysing a prospective cohort.

References

1. Ebrahim S, Davey Smith G. Commentary: Should we always deliberately be non-representative? *Int J Epidemiol* 2013;**42**:1022-6.

2. Keiding, N , Louis, TA. Perils and potentials of self-selected entry to epidemiological studies and surveys. *J. R. Statist. Soc. A* 2016;**179**:319-76.

3. Keiding, N, Louis TA. Web-Based enrollment and other types of self-selection in surveys and studies: consequences for generalizability. *Annu Rev Stat Appl* 2018;**5**:25-47

4. Pizzi C, Pearce N, Richiardi L. Noncollapsibility in studies based on nonrepresentative samples. *Ann Epidemiol* 2015;**25**:955-8.

5. Richiardi L, Pizzi C, Pearce N. Commentary: Representativeness is usually not necessary and often should be avoided. *Int J Epidemiol* 2013;**42**:1018-22.

6. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;**42**:1012-4.

7. Keyes K, Galea S. What is a population? In: Keyes K, Galea S editors. Population health science. New York: Oxford University Press; 2016. p 4

8. Rothman KJ, Greenland S, Lash TL. Generalizability. In: Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. Philadelphia: Lippincott Williams & Wilkins, 2008. p 146-7.

9. Doll R, Hill Ab. The mortality of doctors in relation to their smoking habits; a preliminary report. *Br Med J* 1954;**1**:1451-5.

10. Richiardi L, Baussano I, Vizzini L, *et al*. Feasibility of recruiting a birth cohort through the Internet: the experience of the NINFEA cohort. *Eur J Epidemiol* 2007;**22**:831-7.

11. Riboli E. Nutrition and cancer: background and rationale of the European prospective investigation into cancer and nutrition (EPIC). *Ann Oncol* 1992;**3**:783-91.

12. Golding J, Pembrey M, Jones R; ALSPAC Study Team. ALSPAC--the Avon longitudinal study of parents and children. I. Study methodology. *Paediatr Perinat Epidemiol* 2001;**15**:74-87.

13. Dawber Tr, Meadors Gf, Moore Fe Jr. Epidemiological approaches to heart disease: the Framingham study. *Am J Public Health Nations Health* 1951;**41**:279-81.

14. Lesko CR, Buchanan AL, Westreich D, *et al*. Generalizing Study Results: A Potential Outcomes Perspective. *Epidemiology* 2017;**28**:553-561.

15. Nohr EA, Frydenberg M, Henriksen TB, *et al*. Does low participation in cohort studies induce bias? *Epidemiology* 2006;**17**:413-8.

16. Hatch EE, Hahn KA, Wise LA, *et al*. Evaluation of selection Bias in an internet-based study of pregnancy planners. *Epidemiology* 2016;**27**:98-104.

17. Munafò MR, Tilling K, Taylor AE, *et al*. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol* 2018;**47**:226-35.

18. Taylor AE, Jones HJ, Sallis H, *et al*. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2018. doi: 10.1093/ije/dyy060

19. Cole SR, Platt RW, Schisterman EF, *et al*. Illustrating bias due to conditioning on a collider. *Int J Epidemiol* 2010;**39**:417-20.

20. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003;**14**:300-6.

21. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;**15**:615-25.

22. Jacobsen TN, Nohr EA, Frydenberg M. Selection by socioeconomic factors into the Danish national birth bohort. *Eur J Epidemiol* 2010;**25**:349-55.

23. Pizzi C, De Stavola B, Merletti F, *et al*. Sample selection and validity of exposure-disease association estimates in cohort studies. *J Epidemiol Community Health* 2011;**65**:407-11.

24. Pizzi C, De Stavola BL, Pearce N, *et al*. 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.

25. Preston SH, Stokes A. Obesity paradox: conditioning on disease enhances biases in estimating the mortality risks of obesity. *Epidemiology* 2014;**25**:454-61.

26. Keiding N, Clayton D. Standardizartion and control for confounding in observational studies: a historical perspective. *Statist Sci* 2014;**29**:529-58.

27. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. *Statist Sci* 2014;**29**:579–95.

Fig. 1. Directed acyclic graph of collider bias induced by selection in the source population S. The square around a variable means conditioning on that variable. The dashed line implies induced association. E is an exposure of interest and R is a risk factor for the disease D of interest.

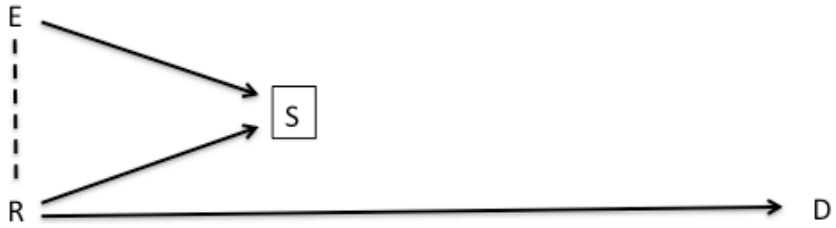

Table 1. Place of birth of the residents of the Piedmont region, Italy, 2018[a].

Place of birth	Number	Proportion
Piedmont region	2,850,610	64.4%
Other Italian regions	972,985	22.0%
Abroad	603,855	13.6%
Total	4,427,450	100.0%

[a] data obtained from the database of the general practitioners' lists

Table 2. Distribution of place of birth by three municipalities of the Piedmont region, Italy, 2018.

Place of birth	**Municipality of residence**[a]		
	Torino (N=942,548)	**Moncalieri (N=59,209)**	**Cuneo (N=55,780)**
	%	**%**	**%**
Piedmont region	52.9%	62.3%	74.1%
Other Italian regions	25.2%	24.3%	12.5%
Abroad	21.9%	13.4%	13.4%

[a]data obtained from the database of the general practitioners' lists

Table 3. Association of place of birth with sex (odds ratio of being a male) and age (odds ratio of older age than the regional median, 47 years) in three selected municipalities of the Piedmont region, Italy, 2018.

Place of birth	**Municipality of residence**		
	Turin	**Moncalieri**	**Cuneo**
	OR (95% CI)	OR (95% CI)	OR (95% CI)
	Sex: male		
Piedmont region	1.00 (ref[a])	1.00 (ref)	1.00 (ref)
Other Italian regions	0.92 (0.91-0.93)	0.92 (0.89-0.96)	1.22 (1.16-1.28)
Abroad	1.01 (1.00-1.02)	0.91 (0.87-0.96)	0.88 (0.84-0.93)
Overall p-value	p<0.001	p<0.001	p<0.001
	Age: 48+ years		
Piedmont region	1.00 (ref[a])	1.00 (ref)	1.00 (ref)
Other Italian regions	6.31 (6.23-6.89)	8.12 (7.73-8.53)	1.95 (1.85-2.06)
Abroad	0.71 (0.71-0.72)	0.80 (0.76-0.84)	0.46 (0.44-0.49)
Overall p-value	p<0.001	p<0.001	p<0.001

[a]OR, odds ratio, CI, confidence interval; Ref, reference

Table 4. Four-month mortality by place of birth, adjusted and unadjusted by age (above and below the median) in three selected municipalities, Piedmont Region, Italy, 2018

Place of birth	**Municipality of residence**					
	Torino (N deaths=2070)		**Moncalieri (N deaths=132)**		**Cuneo (N deaths = 122)**	
	Crude RR[a] (95% CI[a])	Age-adjusted[b] RR (95% CI)	Crude RR (95% CI)	Age-adjusted RR (95% CI)	Crude RR (95% CI)	Age-adjusted RR (95% CI)
Piedmont region	1.00 (ref[a])	1.00 (ref[a])	1.00 (ref)	1.00 (ref[a])	1.00 (ref)	1.00 (ref[a])
Other Italian regions	2.42 (2.11-2.64)	0.96 (0.88-1.05)	2.68 (1.90-3.79)	1.05 (0.75-1.48)	1.07 (0.65-1.77)	0.84 (0.51-1.38)
Abroad	0.23 (0.19-0.29)	0.51 (0.41-0.63)	0.22 (0.07-0.70)	0.56 (0.17-1.78)	0.22 (0.08-0.60)	0.65 (0.24-1.77)

[a] RR, risk ratio estimated using a log-binomial model; CI, confidence intervals; ref, reference
[b] age as a continuous variable, by 10 years increase.

Table 5. Crude prevalence ratios, and corresponding 95% confidence intervals (CI), of being pregnant; women aged 18-44 years, Piedmont region, Italy, 2013.

Variable	**Crude PR (95% CI)**[a]
Age (years)	
18-24	0.34 (0.33-0.36)
25-29	1.00 (ref[a])
30-34	1.26 (1.22-1.29)
35-44	0.48 (0.47-0.49)
Educational level	
< high school	1.00 (ref)
At least high school	1.19 (1.17-1.22)
Occupation	
Employed	1.00 (ref)
Unemployed	0.80 (0.77-0.83)
Other	0.56 (0.54-0.57)
Number of components	
1	0.61 (0.59-0.64)
2	1.00 (ref)
3	0.72 (0.71-0.74)
4	0.33 (0.32-0.34)
5+	0.50 (0.48-0.52)
Place of birth	
Italy	1.00 (ref)
Outside Italy	1.37 (1.34-1.40)

[a]PR, prevalence ratios estimated using a log-binomial model; CI, confidence interval; ref, reference

Table 6. Association between age and being born outside Italy in women aged 18-44 years, Piedmont region, Italy, 2013.

Age (years)	OR (95% CI)[a] of being born outside Italy	
	All women	**Pregnant women**
<25	1.00 (ref[a])	1.00 (ref)
25-29	1.70 (1.67-1.74)	0.72 (0.66-0.78)
30-34	1.89 (1.86-1.94)	0.44 (0.40-0.48)
35-44	1.30 (1.28-1.33)	0.32 (0.30-0.35)

[a] OR, odds ratio; CI, confidence interval; ref, reference

Table 7. Odds ratio of caesarean section vs. natural delivery for pregnant women born outside Italy compared pregnant women born in Italy; 28,790 deliveries of live children in 2013, of which 8,932 were caesarean sections, women aged 18-44, Piedmont region, Italy.

Woman's place of birth	**OR (95% CI)[a] of caesarean section**	
	Crude	**Age-adjusted**
Italy	1.00 (ref[a])	1.00 (ref)
Outside Italy	0.89 (0.84-0.95)	1.01 (0.95-1.07)

[a] OR, odds ratio; CI, confidence interval; ref, reference