**Selection on a collider: a ubiquitous mechanism that shapes the confounding structure in both representative and selected cohort studies**

Lorenzo Richiardi1,2, Neil Pearce3, Eva Pagano2, Daniela Di Cuonzo1,2, Daniela Zugna1, Costanza Pizzi1

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

**Abstract**

There is debate on whether the validity of cohort studies is threatened when they are based on a selected source population that is non-representative of a given general population. Baseline selection may introduce collider bias if the exposure of interest and other outcome risk factors affect the probability of being in the source population. This produces an altered confounding pattern in the cohort at baseline. We argue that this collider bias is not specific to ‘selected cohorts’, and also occurs in ‘representative cohorts’ due to the inherent selection processes that occur in any population. These selection processes are linked to the life status, immigration and emigration, which, in turn, may be affected by environmental and social determinants as well as by lifestyles and genetics. We provide real-world examples of this phenomenon using the data on the population of the Piedmont Region, Italy. The confounding structure for the exposure-outcome association of interest in any source population will be always potentially shaped by three mechanisms: the collider bias due to the underlying selection processes, the shared common causes between the exposure and the risk factors and the effects of the risk factors on the exposure. We conclude that, when conducting a cohort study, different source populations, whether ‘selected’ or ‘representative’ (of a geographically-defined general population), may lead to different confounding structures, but that one approach is not inherently more or less biased than the other. The key issue is whether those confounders can be identified and controlled for, irrespectively of the nature of the source population.

**Word count:** abstract, 240; text, 3913; 3 Figures, 5 Tables; 2 Supplementary Figures; 1 Supplementary Table

**Introduction**

The literature on cohort studies, including recent debate papers, distinguishes between ‘selected cohorts’ (based on a selected source population), and ‘representative cohorts’ (based on a given general population) [1, 2, 3, 4, 5, 6]; Although the term is ill-defined, we typically think of a ‘general population’ as a collection of people who share the specific geographical location they inhabit in a specific period of time [7]. For example a cohort study recruiting a random sample of people living in Turin, Italy, in 2018 would be typically considered as a representative cohort.

Some authors also refer 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 concept of 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?’), 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 assumed to be potentially ‘everyone living on the planet’, with some specification of the risk period (e.g. a study may be conducted in 2018, but its results may be used also in the future). In this instance, any general population can be regarded as a subset of the intended target population, although the latter may include specific restriction criteria (e.g. ‘patients with a specific disease’). The concept of representativeness is usually applied to the general population and not to the target population, as very few studies involve a random/stratified sample of everyone living on the planet, or a random/stratified sample of some other well-defined target population. We will therefore restrict ourselves to discussing representativeness in terms of a specific general 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]. Again, Internet users were chosen for practical and scientific reasons, but the findings are intended to be applicable more widely. Finally, studies such as EPIC [11], ALSPAC [12] or the Framingham cohort [13] are typically considered as representative cohorts as they recruited a sample of a specified general population (within given age criteria). All of these studies (whether selected or representative cohorts) are intended to yield findings that apply beyond their source populations.

As previously described [5], the restriction of the source population could be labelled as ‘intentional’ selection, as the selection criteria are established by the researcher. 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 we would like to study and those who are actually recruited in the study. This other form of baseline selection, which is due to baseline non-response or volunteering, is ‘unintentional’ and depends on characteristics that are not a-priori decided by the researcher [14]. We assume that, in prospective cohorts, baseline unintentional selection does not depend on the outcome under study as, by design, prevalent cases should be excluded already from the source population. This assumption would be violated if a cross-sectional analysis is conducted at baseline in a cohort study [15] or if some prevalent cases are undiagnosed. We will focus on cohort studies (selected or representative) 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 one of the most relevant determinants of study validity, as loss to follow-up can be affected directly by the outcome of interest [16]. In other words, since this form of selection occurs after recruitment into the cohort, the new events occurring with time can affect the selection process. 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*

With regards to baseline selection, some authors argue that if the study aims to understand whether an exposure affects the outcome or, in other words, if the aim is to try to explain ’how nature works’, there is no need that the source population is representative [6]. Other authors argue that cohort studies based on selected source populations are more prone to bias than representative studies [1, 15]. The debate lies mainly on the possibility that baseline selection may induce collider bias (i.e. “baseline collider bias”). In particular, if the exposure of interest and another outcome risk factor are both associated with 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 Figure 1 using a directed acyclic graph (DAG) and has been discussed several times in the literature [5, 15, 17, 18, 19, 20, 21, 22, 23]. In the example shown in Figure 1, the risk factor is not a potential confounder of the exposure-outcome association in the general population, while it becomes, because of baseline collider bias, a confounder in the selected source population. We have discussed before the more general scenario in which the exposure and the risk factors may be already associated in the general population [5]. Under those scenarios the use of a selected source population simply alters the confounding pattern – both the general population and the selected source population have confounding, and there is no a priori reason to believe than one is ‘better’ than the other. Hence the confounding structure in a given selected source population is due to the potential combination of three general mechanisms: (i) the risk factors for the outcome of interest may affect the exposure of interest; (ii) the outcome risk factors may share common causes with the exposure of interest; (iii) the outcome risk factors and the exposure of interest may be associated with the probability of being in the source population.

The first two criteria apply to all cohorts, whether selective or representative. We will argue here that the third criterion also 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. 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 in a given city in a given time period (e.g. Turin in 2018). The source population for this cohort would typically include residents of that city in that year with no further restriction criteria apart from the possible definition of relevant age groups of interest.

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 is shown in Figure 2, which resembles closely Figure 1 with the exception that the source population is now representative. For example, if E is age, R is sex and the study is carried out in a representative source population, both age and sex are likely to affect the probability of being alive and a member of the source population in a given calendar year. Even if age and sex are not theoretically expected to be associated (age only depends on year of birth and calendar year, sex is genetically determined) they become associated because of the collider bias. If a representative cohort study aims at assessing the causal effect of sex on a given outcome, say cardiovascular diseases, it should control for age, as age becomes a potential confounder due to the baseline collider bias.

It is relatively easy to 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 both affected by socioeconomic position. The association between smoking and alcohol at baseline in a given representative cohort would thus depend on the combination of shared causes and baseline collider bias. In addition, life status is not the only condition for being a member of a given general population in a given time. Individuals leave and join populations all the times, and their decisions on 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 Figure 2, E and R could stand for occupational status, smoking, obesity, educational level, mental health, general health status, family composition, air pollution, noise, war, poverty, climate change, etc. Also genetic factors may be determinants of the membership in a representative source population if they affect survival, behaviours or manifest phenotypes (e.g. height and weight).

**The example of the Piedmont general population**

To analyse a real-world example, we extracted the data of the whole resident population of the Piedmont region, Italy, in 2018. The Piedmont region has more than 4,000,000 inhabitants and includes more than 1,000 municipalities. For each individual, we obtained information on age in 2018, the ISTAT code of the municipality of birth, the ISTAT code of the municipality of residence, and sex. Information on place of birth and place of residence was combined to obtain a 4-level variable: born in the same municipality of residence, born in the Piedmont region but in a different municipality than residence, born in another Italian region than Piedmont, and born abroad. We also created a 3-level variable in which we distinguished whether residents of the Piedmont region were born in Piedmont, in other Italian regions or abroad. This categorization is more robust than the 4-level variable as it is not affected by subjects moving between bordering municipalities or by small administrative changes in the borders of the municipalities over time. Data on people who were born in the Piedmont region and had left Piedmont were not available in this database.

Out of 4,573,258 residents, we excluded 54,149 subjects resident in municipalities that after 1991 acquired new areas, were newly established or changed name. We also excluded 6,765 subjects resident in municipalities with less than 150 residents in 2018.

Table 1 reports the joint distribution of place of birth and place of residence among the Piedmont residents. Out of a population of 4,512,362 individuals, 75.5% were born in a different municipality than that of residence; 35.7% had joined the Piedmont population from outside the region. The corresponding distributions of the 4-level and the 3-level variables for the 15 largest municipalities of the Piedmont region (with at least 35,000 residents) are reported in the supplementary Figures S1 and S2.

Table 2 reports the data for four selected municipalities in which a hypothetical representative cohort could be conducted: the city of Turin, 957,475 inhabitants, which is the capital of the region; the municipality of Moncalieri, 60,001 inhabitants, which is close to Turin and is highly residential; the municipality of Settimo Torinese, 51,202 inhabitants, that was characterized by a high level of industrialization; and the municipality of Cuneo, 56,674 inhabitants, that is the main city of a large rural area.

It is likely that the key determinants for the individuals to be resident in their municipality vary over the four municipalities. The direction and strength of collider bias will vary accordingly. As an example, we show in Table 3 the association between sex and place of birth in the four municipalities. The sex ratio is fairly constant among populations, so theoretically sex is not expected to be associated with place of birth. Sex and place of birth can be both 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 males or mainly females), but they do not cause each other and they do not have common causes. Therefore, an association between sex and place of birth in a specific general population indicates the presence of collider bias in that population.

As reported in Table 3, sex was associated with place of birth in all four populations, and the direction and strength of the association differed over municipalities. The heterogeneity was even larger when we used the 4-level variable that further distinguishes between same and different municipalities than residence within the Piedmont region (Table 1S). A hypothetical representative cohort aiming at assessing the effect of sex on, for example, cardiovascular diseases, would have to adjust for place of birth as this variable may become a confounder due to collider bias; in this scenario, lack of adjustment for place of birth would induce bias in different directions in the four general populations.

**The example of a representative conception cohort**

The selection process and the consequent collider bias that we have just described occur because the general population is dynamic: individuals may join or leave a specific population for different reasons, including immigration, emigration, and death.

It could be argued that a representative birth cohort would be immune to these mechanisms, as emigration and immigration do not involve individuals at birth. However they are still affected by selection processes due to abortions and miscarriages. Hence individuals would have to be selected at conception. In this section we show that collider bias may occur even in a representative conception cohort.

The selection mechanism in a conception cohort is linked to the probability of getting pregnant. As shown in Figure 3, becoming pregnant is a collider, which is potentially affected by several factors; the mutual associations among these factors are thus altered by collider bias. In this example, the restriction is related to the choice of a specific target population, which includes only those who are pregnant (i.e. in a conception cohort there is, by definition, no interest in those who do not get pregnant). In this respect, this example differs from the example of the previous section, in which the inherent conditioning on a collider was due to selection processes related to the choice of a particular general population.

To analyse a real-world example, we extracted the data of residents of the Piedmont region, Italy, in 2013, restricted to women aged 18 to 44 years on the 1st of January 2013 (as a proxy of fertile age). In this population we identified hospital admission 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, which, theoretically, would have to be part of the conception cohort. Our procedure identified 35,531 women who were pregnant in the period 2012-2013 and concluded the pregnancy in 2013, out of an overall population of 673,821women. We then linked these data with records of the 2011 census from which we obtained pre-pregnancy information on: educational level (classified as less than high school, at least high school), occupation (employed, unemployed, other), number of family members, and place of residence (city of Turin, other municipalities). Age was categorized into four categories: less than 25, 25 to 29, 30 to 34, 35 to 44 years. 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. We used logistic regression to analyse whether these variables were associated with the probability of being pregnant in 2012-2013. Then, to understand the role of collider bias, we investigated the association between educational level and age in the whole population and in the pregnant population. This second analysis was restricted to women aged at least 25 years (N=546,869, 32,099 pregnancies), so that all women had had the opportunity to reach the higher educational level (i.e. at least high school).

As reported in Table 4, 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 (p-value for departure from multiplicative interaction: 0.0001; data not reported in the Table).

Table 5 reports the association between age and educational level in the whole population of women aged 25-44 years and the subgroup of pregnant women of the same age. Age and education were negatively associated in the general population, while they were positively associated among pregnant women. In a representative conception cohort carried out in the Piedmont population, if maternal age were the exposure of interest maternal educational level would act as a potential confounder in the opposite direction than that expected on the basis of the confounding structure of the general population. This difference is attributable to collider bias due to the conditioning on being pregnant.

**Discussion**

Even when 100% of the source population is included in the study, baseline collider bias may affect the confounding structure of a cohort study irrespectively of the nature of the source population (whether representative or selective). The confounding structure for the exposure-outcome association of interest in any source population will be always potentially shaped by three mechanisms: the collider bias due to the underlying selection processes, the shared common causes between the exposure and the risk factors, and the effects of the risk factors on the exposure. These three potential biases apply to both selected and representative cohorts.

The potential magnitude and direction of collider bias has been assessed in several studies, based on algebraic calculations [17, 18] or simulations [15, 22]. The importance of baseline collider bias in shaping the confounding structure depends on the strengths of the associations of the risk factors and the exposure of interest with the membership in the source population and their interactions. For example, baseline collider bias can have extreme effects on the confounding structure in clinical cohorts [24]. However, in other situations, a representative cohort may have strong confounding whereas a selective cohort may have less - the use of selective cohorts does not necessarily introduce confounding, but rather changes the confounding structure [23]. 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 control for confounding in that source population.

In this paper, we have focused on intentional baseline selection and its effects on the confounding structure. As mentioned in the introduction, both representative and selected cohorts can be affected also by unintentional baseline selection due to baseline volunteering or non-response. Theoretically, in cohort studies, the source population at baseline should not include subjects who already have the outcome of interest and the latter cannot affect baseline participation. Thus, in cohort studies typically the causal structure of unintentional selection closely resembles that of intentional selection, i.e. if both the exposure and the risk factors are associated with participation, the resulting collider bias shapes the confounding structure. To depict graphically this source of collider bias (i.e. non-response), for example, it would be enough to replace in Figure 1 ‘source population’ with ‘study 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, although it is possible to apply an initial period of follow-up to identify cases that were undiagnosed at baseline. In addition, sometimes researchers carry out cross-sectional analyses nested at baseline in cohort studies. In those analyses, subjects with the outcome of interest already at baseline are not excluded from the source population, and it is thus possible that the outcome directly affects baseline participation. For this reason, cross-sectional analyses of cohort studies at baseline should be interpreted with caution: cross-sectional studies require a high participation proportion at recruitment to be valid, while this requirement is not needed in cohort studies, which often have low baseline response proportions. The key issue in cohort studies is loss to follow-up which, unlike selection at baseline (provided that pre-existing disease at baseline is excluded), can be strongly related to both exposure and outcome.

The issue of selection into the source and/or study population in cohort studies may become critical when the selection process depends on a manifest intermediate between the exposure and the outcome [5]. For example, in a conception cohort that aims at studying maternal socioeconomic position in association with birth outcomes, maternal obesity would be a potential intermediate that may affect the probability of being resident of a given general population, and (more likely and more strongly) the probability of participating in the study. Over-representation of children of non-obese women in the study population would bias the estimate of the total effect of maternal socioeconomic position on birth weight. However, again this might happen regardless of whether the source-population is selected or representative. The issue of possible baseline selection on the mediators does not apply in case of inception cohorts, i.e. when the recruitment and follow-up start before or at the same time of the beginning of the exposure. An example would be maternal use of medications in a conception cohort.

The key element 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, provided that unintentional selection is low and a high completeness of follow-up is possible. The distribution of the mediators, the distribution of the effect modifiers, the confounding structure would reflect those of the target population. 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, the key issue is to obtain valid estimates and then understand how they can be transported to other contexts and populations [25].

In conclusion, different source populations, whether representative of a geographically-defined general population or selected, may be affected by collider bias and may have different confounding structures. There is no a priori reason to expect that one type of source population will involve more or less confounding than the other. In fact, 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). All cohorts, whether selective or representative, have potential confounding; the key issue is thus whether confounders can be identified and controlled for, irrespectively of the nature of the source population.

**Acknowledgements**

Neil Pearce’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].

**References**

1. Ebrahim S, Davey Smith G. Commentary: Should we always deliberately be non-representative? Int J Epidemiol. 2013;42(4):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(2):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(12):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(4):1018-22.
6. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. Int J Epidemiol. 2013;42(4):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(4877):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(12):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(10):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(1):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(3):279-81.
14. Nohr EA, Frydenberg M, Henriksen TB, Olsen J. Does low participation in cohort studies induce bias? Epidemiology. 2006;17(4):413-8.
15. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. Int J Epidemiol. 2018;47(1):226-35.
16. 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
17. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, Poole C. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39(2):417-20.
18. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology. 2003;14(3):300-6.
19. Hatch EE, Hahn KA, Wise LA, et al. Evaluation of selection Bias in an internet-based study of pregnancy planners. Epidemiology. 2016;27(1):98-104.
20. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25.
21. Jacobsen TN, Nohr EA, Frydenberg M. Selection by socioeconomic factors into the Danish national birth bohort. Eur J Epidemiol. 2010;25(5):349-55.
22. 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(5):407-11.
23. 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(11):976-81.
24. Preston SH, Stokes A. Obesity paradox: conditioning on disease enhances biases in estimating the mortality risks of obesity. Epidemiology. 2014 May;25(3):454-61.
25. Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. Statist Sci. 2014;29: 579–95.

**Figure 1**. Directed acyclic graph of collider bias induced by a restricted source population. 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.



**Figure 2.** Directed acyclic graph of collider bias in a cohort based on a representative source population. 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.

****

**Figure 3.** Directed acyclic graph of collider bias in a *conception* cohort based on a representative source population. The square around a variable means conditioning on that variable. The dashed line implies induced association. E is a *maternal* exposure of interest and R is *maternal* risk factor for the *child* outcome D of interest.

****

**Table 1.** Combination of place of birth and place of residence in residents of the Piedmont region, 2018.

|  |  |  |
| --- | --- | --- |
| **Place of birth** | **N** | **%** |
| Piedmont region | 2,903,092 | 64.4% |
| (same municipality of residence) | (1,103,293) | (24.5%) |
| (different municipality than residence) | (1,799,799) | (39.9%) |
| Other Italian regions | 1,002,554 | 22.2% |
| Abroad | 606,698 | 13.5% |
| Total | 4,512,344 | 100.0% |

**Table 2.** Distribution of the combination of place of birth and place of residence in three municipalities of the Piedmont region, 2018.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Municipality of residence** | | | |
| **Place of birth** | Turin  (N=957,475) | Moncalieri  (N=60,001) | Settimo Torinese  (51,202) | Cuneo  (N=56,674) |
|  | % | % | % | % |
| Piedmont Region | 52.8% | 62.1% | 58.2% | 74.2% |
| (same municipality of residence) | (43.0%) | (19.3%) | (4.4%) | (52.4%) |
| (different municipality, Piedmont region) | (9.8%) | (42.8%) | (53.7%) | (21.8%) |
| Other Italian regions | 25.5% | 24.6% | 33.5% | 12.6% |
| Abroad | 21.7% | 13.3% | 8.4% | 13.3% |

**Table 3.** Odds ratio of being a male for place of birth in four selected municipalities of the Piedmont region, 2018.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Municipality of residence** | | | |
| **Place of birth** | Turin | Moncalieri | Settimo  Torinese | Cuneo |
|  | OR (95% CI) | OR (95% CI) | OR (95% CI) | OR (95% CI) |
| Piedmont region | 1.00 (ref) | 1.00 (ref) | 1.00 (ref) | 1.00 (ref) |
| Other Italian regions | 0.91  (0.90-0.92) | 0.91  (0.88-0.95) | 0.90  (0.87-0.93) | 1.22  (1.16-1.28) |
| Abroad | 1.02  (1.01-1.03) | 0.92  (0.87-0.96) | 0.86  (0.80-0.91) | 0.90  (0.85-0.94) |
| Overall p-vale | p<0.001 | p<0.001 | p<0.001 | p<0.001 |

aOR, odds ratio, CI, confidence interval

**Table 4**. Crude and mutually adjusted odds ratios (OR), and corresponding 95% confidence intervals (CI), for being pregnant; women aged 18-44 years, Piedmont region, Italy, 2013.

|  |  |
| --- | --- |
| **Variable** | **Crude OR (95% CI)** |
| Age (years) |  |
| 18-24 | 0.33 (0.31-0.34 |
| 25-29 | 1.00 (ref **a**) |
| 30-34 | 1.28 (1.25-1.32) |
| 35-44 | 0.46 (0.45-0.57) |
|  |  |
| Educational level |  |
| < high school | 1.00 (ref) |
| At least high school | 1.20 (1.18-1.23) |
|  |  |
| Occupation |  |
| Employed | 1.00 (ref) |
| Unemployed | 0.79 (0.76-0.82) |
| Other | 0.54 (0.53-0.57) |
|  |  |
| Number of components |  |
| 1 | 1.00 (ref) |
| 2 | 0.59 (0-57-0.62) |
| 3 | 0.71 (0.69-0.72) |
| 4 | 0.31 (0.30-0.32) |
| 5+ | 0.48 (0.46-0.50) |
|  |  |
| Residence |  |
| City of Turin | 1.00 (ref) |
| Other cities in Piedmont | 0.99 (0.96-1.01) |

aRef, reference

**Table 5**. Odds ratio (OR), and corresponding 95% confidence interval (CI), of higher educational level (at least high school vs. less than high school) for age categories; women aged 25-44, Piedmont region, Italy, 2013.

|  |  |  |
| --- | --- | --- |
| Age (years) | OR (95% CI)a of high educational level | |
|  | All women | Pregnant women |
| 25-29 | 1.00 (ref) | 1.00 (ref) |
| 30-34 | 0.82 (0.81-0.84) | 1.55 (1.46-1.64) |
| 35-44 | 0.54 (0.53-0.55) | 1.40 (1.32-1.49) |

a OR, odds ratio; CI, confidence interval

**SUPPLEMENTARY FUGURES AND TABLES**

**Figure S1.** Distribution of place of birth among residents of the 15 largest municipalities in the Piedmont region in 2018. Municipalities are labelled through their ISTAT code

**Figure 2s.** Distribution of place of birth among residents of the 15 largest municipalities in the Piedmont region in 2018. Municipalities are labelled through their ISTAT code

**Table 1S.** Odds ratio of being a male for place of birth in four selected municipalities of the Piedmont region, 2018.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Municipality of residence** | | | |
| **Place of birth** | Turin | Moncalieri | Settimo  Torinese | Cuneo |
|  | OR (95% CI) | OR (95% CI) | OR (95% CI) | OR (95% CI) |
| Piedmont region – same municipality of residence | 1.00 (ref) | 1.00 (ref) | 1.00 (ref) | 1.00 (ref) |
| Piedmont region – other municipality than residence | 0.79  (0.77-0.80) | 0.97  (0.93-1.01) | 1.16  (1.07-1.27) | 0.73  (0.70-0.76) |
| Other Italian regions | 0.87  (0.86-0.88) | 0.90  (0.85-0.94) | 1.03  (0.94-1.13) | 1.11  (1.05-1.17) |
| Abroad | 0.98  (0.97-0.99) | 0.90  (0.85-0.95) | 0.98  (0.89-1.09) | 0.82  (0.78-0.86) |

aOR, odds ratio, CI, confidence interval