Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2013;**42**:1018–1022
doi:10.1093/ije/dyt103

Commentary: Representativeness is usually not necessary and often should be avoided

Lorenzo Richiardi,[1]* Costanza Pizzi[1,2] and Neil Pearce[3,4]

[1]Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, Turin, Italy, [2]Centre for Statistical Methodology, London School of Hygiene and Tropical Medicine, London, UK, [3]Departments of Medical Statistics and Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK and [4]Centre for Public Health Research, Massey University, Wellington, New Zealand

*Corresponding author. Cancer Epidemiology Unit, Via Santena 7, 10126 Torino, Italy. E-mail: Lorenzo.richiardi@unito.it

Accepted 6 February 2013

We agree with Rothman and colleagues that scientific inference in epidemiology does not require representativeness of the general population or target population in order to be valid. This is an important message and we welcome Rothman and colleagues' paper which has clearly expressed this position.[1]

On the other hand, perhaps Rothman and colleagues go too far in arguing that representativeness should be avoided as a matter of principle, and we consider that there are some situations where representativeness is the most sensible approach. For example, it would be rare for researchers to only study one age group, and to then attempt to extrapolate their findings to other age groups, if sufficient numbers and funding were available to also sample adequate numbers from these other age groups.

In our experience, there are three usual reasons for deliberately opting for non-representativeness in a study design ('intentional' non-representativeness): (i) practical reasons, e.g. it may be most practical to restrict a study to those who have a telephone; (ii) to minimize bias, e.g. by restricting a study to a particular population subgroup (as in the British doctors study[2]) so that there is less likelihood of lifestyle differences between exposed and non-exposed within that group; and (iii) in order to focus on one or more population subgroups, e.g. if we wish to compare exposure-outcome estimates in different ethnic groups.[3] In the first instance, representativeness is not necessary and would usually not improve the feasibility of the study; in the latter two situations it should specifically be avoided.

In addition, non-representativeness may also be 'unintentional', e.g. in longitudinal studies because of low baseline response rates or the recruitment of volunteers rather than a formal sample of a defined population. Such unintentional selection may occur both in studies involving random population samples and in those involving non-representative samples. In this paper we will focus mainly on the issues involved in intentional non-representativeness, but will also consider issues of unintentional non-representativeness. In this latter situation, the potential for bias may be greater. In particular there is potential for large bias if the outcome of interest or its early signs affect the probability of baseline selection. We will argue however that, provided that the outcome does not affect selection, situations of intentional and non-intentional non-representativeness are generally similar in terms of validity. Furthermore, baseline self-selection is likely to create a group of more motivated persons in longitudinal studies, which may result in a better response to follow-up and thus in decreased selection bias. So the possibility of bias from lack of representativeness needs to be balanced against the likelihood of bias from poor response to follow-up in a more representative sample. For example, most researchers, if given the choice, would opt to base a study on 50% of the population and then achieve good follow-up rates, rather than to start with a representative sample and then only achieve 50% follow-up.

We should also note that in some instances the aim of an epidemiological study is primarily descriptive, e.g. to estimate the prevalence of a condition such as asthma in the general population,[4] and in these studies representativeness is necessary to obtain valid estimates. Furthermore, such studies often are not completely descriptive. For example, prognostic research is population- and time-specific, but the identification of a cause of disease progression may add information to the understanding of a biological phenomenon.

We will focus on 'analytical' studies which aim to estimate a particular exposure-disease association, while appropriately controlling for confounding and avoiding other biases. In this situation, we agree that representativeness is not a goal *per se,* but rather needs to be justified in the context of the particular study.

For example, in a clinical trial where we want to understand the efficacy of a treatment for a disease, a random sample is clearly not needed and in many ways it can be inappropriate. Typically we restrict the initial studies to high-risk patients or to patients expected to have a high compliance with the assigned treatment and follow-up.

We have been involved in discussions on representativeness a number of times since 2005, when we started an internet-based birth cohort in Italy (NINFEA cohort, www.progettoninfea.it),[5] followed by a similar study in New Zealand (ELFS cohort, www.elfs.org.nz). Internet-based recruitment has advantages in terms of feasibility, costs and possibilities of reaching traditionally understudied populations. However, this approach is often criticized on the basis of its consequent lack of representativeness of the general population. Internet-based recruitment selects participants who have access to the internet, become aware of the existence of the study and volunteer to participate. Thus, it is based on a restricted source population and the study population is a self-selected sample of the source population (i.e. non-representativeness is both intentional and unintentional).

In this commentary we describe these criticisms and argue, in line with Rothman and colleagues, that restricting a study to a subgroup of the general population does not usually hamper scientific inference, and may often enhance it. We focus on infant cohort studies, but the same arguments on intentional non-representativeness may apply to the corresponding case-control and cross-sectional studies based on the same restricted populations. We focus on the main two arguments which we have received against using non-representative populations in internet-based birth cohorts: (i) lack of heterogeneity; and (ii) the potential for bias. We also consider a third potential criticism relating to selection and a mediating variable.

Criticism 1: Non-representative cohorts lack heterogeneity

One major criticism of the use of non-representative samples is the resulting lack of heterogeneity, with regard to exposures, potential effect modifiers, or both. Although it is true that restriction may decrease the range of exposure levels and the magnitude of the contrasts, we argue that using non-representative samples may often enhance study power to assess main effects and effect modification. To study a rare exposure, for example, either we assemble a very large cohort or we do 'smart selection' of its members. For example, in an internet-based birth cohort study, in which members are characterized by a high socioeconomic status, women having their first pregnancy after their 40s are overrepresented. When high maternal age is the exposure of interest, an internet-based birth cohort becomes more efficient than a birth cohort which is representative of the general population. Similarly, using non-representative samples may enhance our ability to assess heterogeneity with regard to potential effect modifiers, e.g. by ensuring that there are adequate numbers in each of the ethnic groups to be considered if we suspect or are interested in potential modification by ethnicity.

These arguments refer to issues of study efficiency, but lack of heterogeneity among study participants may be an advantage with regard to controlling confounding. Ideally, the best study in terms of scientific validity would be a design involving large heterogeneity in the exposure and complete homogeneity in all other characteristics (provided we did not wish to investigate effect modification and/or the effects of varying population contexts).

Of course it should be acknowledged that lack of heterogeneity is not always an advantage, particularly when there is important effect modification. It can happen that exposure has strong effects in one population subgroup and weaker or non-existent effects in another. If a study is based on the latter subgroup, then the effects of exposure will not be identified. However, once again, to explore such effect modification usually requires non-representative samples, e.g. by studying equal numbers in each age, gender or ethnic group.

Unless we are explicitly interested in, or have a priori reason for, investigating heterogeneity, generalizability is a matter of scientific inference rather than representativeness. There are many situations in which such generalizability is relatively straightforward. Smoking causes lung cancer in every population in which it has been studied, and there was no bias, and considerable practical advantages, to restricting one of the key early studies to British doctors.[6] Similarly, smoking presumably causes lung cancer in those with or without a telephone, those who have registered to vote and those who have not, and in those who use and those who do not use the internet. With rare exceptions, such restrictions may greatly enhance study practicality and thereby response rates and power, and have little or no effect on validity or generalizability.

Criticism 2: If the exposure of interest is associated with the probability of selection, the exposure-outcome associations estimated in a non-representative cohort may be biased

The second major criticism of the use of non-representative samples is the possibility of introducing selection bias. When conducting a cohort study on a selected population, it is likely that there are factors that are associated with selection and are also determinants of the disease of interest. For example, in a cohort study restricted to British doctors, familial history of early mortality from cardiovascular diseases may affect

both the lifetime probability of cardiovascular diseases and the decision to become a doctor. As with other risk factors, the exposure of interest may also be associated with the probability of selection: for example, socioeconomic status may affect both smoking habits and grades at high school (and therefore the probability of being admitted to a medical school). If both the exposure and another risk factor for the disease of interest are associated with the probability of selection, baseline restriction can introduce bias in the exposure-outcome association. This is a type of collider bias that has been discussed extensively in the epidemiological literature, including by us in the context of internet-based cohorts.[7,8] Fortunately, the amount of bias that is expected to be introduced by this phenomenon is small unless all of the associations involved in generating the bias are very strong. Assuming that all relative risks involved are of 2.0, the bias, in logarithmic scale, will be of 0.02 [i.e. a relative risk (RR) for the exposure-outcome association of 1.02, when the true RR is 1.00]; that is, while assuming that all RRs are of 4.0, the bias will be of 0.15 (i.e. a RR of 1.16, when the true RR is 1.00).[7]

However, the exposure of interest is almost always associated with some disease risk factors in the general population, whether or not we study a restricted subpopulation. Indeed each general population, at a given point in time, will have its specific confounding pattern. There is no reason to assume that confounding patterns for, say, the association of smoking with cardiovascular disease in London, UK, in 2012 is the same of that present in Turin, Italy, in 2012: we could, for example, expect that in London smoking is associated with drinking beer whereas in Turin it is associated with drinking red wine. The confounding pattern in the selected cohort may differ from that of the corresponding general population, but we cannot predict whether the amount of confounding will be greater, similar or smaller. The bottom line is that each population, including a selected study population, has its own confounding pattern. Valid scientific inference is achieved if the confounders are controlled for, and there is no reason to believe that control of confounding can be more easily achieved in a population-based cohort than in a restricted cohort. Indeed, we can intentionally restrict the cohort to decrease confounding bias. For example, if we are not able to precisely measure the amount of alcohol consumption in the general population, and we know that alcohol is a relevant confounder of the association of interest, we can restrict the study to non-drinkers and occasional drinkers.

In a recent paper, we compared, for selected exposures and outcomes of interest, the confounding pattern of the NINFEA internet-based cohort with that present in the corresponding general population, showing that the overall confounding was not larger, but was qualitatively different, from that present in the general population.[8]

As mentioned above, it is not impossible to devise situations in which selection bias could occur due to restriction (i.e. non-representativeness), for example when an exposure and an unmeasured risk factor for the disease are independent in the general population but both are associated with the probability of selection. Our argument is not that such bias is impossible, but rather that restricted studies are often likely to be less affected by confounding. Also, any small likelihood of bias from using non-representative samples needs to be balanced against the likelihood of bias if attempts to use random representative samples result in low response rates at follow-up and/or a greater likelihood of information bias. The British doctors study is a relevant example once again, in which the non-representative sample has likely induced better follow-up and greater validity of the smoking information gathered. To insist on doing the study in a random general population sample would have had little or no benefit, and considerable disadvantages in terms of logistics and study validity.

Criticism 3: If an intermediate variable in the causal pathway from the exposure to the outcome is associated with the selection, exposure-outcome associations estimated in a non-representative cohort may be biased

We would therefore argue that the main reasons for opposing the use of non-representative samples—lack of heterogeneity and the potential for introducing selection bias and/or confounding—are rarely valid, and are generally outweighed by the benefits of this approach, although of course this conclusion is highly hypothesis- and study-dependent. In the rest of this paper we will consider an issue which has been less debated, namely the situation in which an intermediate variable (a mediator, that is a variable that is on the pathway from the exposure to the outcome) is associated with the probability of selection.

In most circumstances, baseline selection in cohort studies takes place before the intermediate variable is manifest. For example, in the British doctors study it could be assumed that members of the cohort became doctors before the occurrence of manifest mediators of the effect of the exposure (smoking) on the outcomes of interest. Analogously, in an internet-based birth cohort, having access to the internet likely occurs before pregnancy and, thus, before most of the possible intermediate variables may become manifest. Within this framework, if there is a variable affecting both the intermediate variable and the probability of selection, the use of a non-representative sample could alter the exposure-mediator confounding pattern. This situation is illustrated in Figure 1 using directed acyclic graphs. Figure 1a shows a non-representative cohort in which selection introduces exposure-mediator confounding that was not present in the general

population; Figure 1b shows the case of a representative cohort in which there is already exposure-mediator confounding; in Figure 1c a non-representative cohort study is conducted in the same population as Figure 1b; in Figure 1d the exposure-mediator confounder also affects the probability of selection. An example of the scenarios depicted in Figure 1b and d would be the effect of pre-pregnancy BMI (E) on preterm delivery (O), in which gestational hypertension is a possible mediator (M). Socioeconomic class (C) would be an exposure-mediator confounder, assuming that it affects both pre-pregnancy BMI and gestational hypertension but, in a simplified scenario, it is not a determinant of pre-term delivery otherwise. In a study restricted to internet users, socioeconomic status would also affect selection (S) (as in Figure 1d) and thus restriction would be likely to decrease exposure-mediator confounding due to socioeconomic status.

In summary, some of the scenarios described in Figure 1 increase the overall exposure-mediator confounding, whereas others decrease it. We consider that there is no reason to expect that non-representative cohorts tend to have a larger exposure-mediator confounding than representative cohorts, although we can always plan the selection in order to decrease exposure-mediator confounding. We should acknowledge that a confounder of the exposure-mediator association is often treated as a confounder of the exposure-outcome association, especially when quantifying the role of the mediator is not the focus of the study. In this context, scenarios described in Figure 1 become very similar to those described in the previous section (Criticism 2).

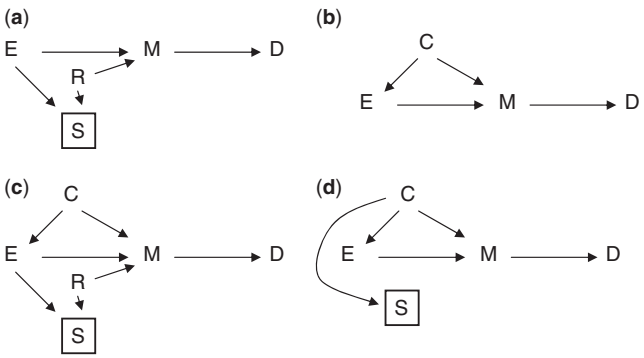

Figure 1 Effect of selection in a cohort study in which a mediator (M) of the effect of the exposure (E) on the outcome (D) becomes manifest after the selection has occurred. Panel (a) shows a non-representative cohort in which the mediator (M) and the selection (S) are affected by a common cause (R), and the exposure (E) is also associated with selection. Panel (b) illustrates a representative cohort in which there is exposure (E)-mediator (M) confounding from (C). Panel (c) illustrates where the scenarios described in a) and b) coexist. Panel (d) shows a similar selection effect to Panel (b), but the confounder (C) also affects selection (S)

It is possible that baseline selection occurs after an intermediate variable becomes manifest. This may typically happen both in representative and in non-representative cohorts when there is unintentional non-representativeness. In a study involving internet-based recruitment, the fact that participants are volunteers who should need to first be aware of the existence of the study may enhance this potential problem. If the intermediate variable has a direct effect on the selection, a number of different scenarios may occur. The simplest scenario is that described in Figure 2, in which there is only a direct effect from the mediator to the selection. According to the causal relationship described in this figure (in which there are no other factors affecting the selection), the effect of the exposure on the outcome of interest would be attenuated. It should be considered, however, that typically the decision to participate in the study depends on a large number of factors and the selection process is poorly predicted by a single intermediate variable. Thus, the situation described in Figure 2 should in most instances introduce a negligible or modest bias in the estimate of the exposure-outcome association. The example of the effect of maternal pre-pregnancy BMI on preterm delivery in which gestational hypertension is an intermediate factor may be used to illustrate also the situation in which selection is directly affected by an intermediate variable. In particular, the decision of pregnant women to participate in the study could depend on whether they have gestational hypertension or not.

The relationship between intermediate variables and selection can become much more complex than described above[9]: for example, the selection could be affected both by the intermediate variable and by the participant's reaction to the intermediate factor. For example, in the hypothetical study of the effect of maternal pre-pregnancy BMI on the risk of pre-term delivery, where gestational hypertension acts as a mediator, we would have to consider that women are usually monitored during the remaining part of pregnancy and may be prescribed blood pressure medications. Participation in the cohort could be affected both by the gestational hypertension and by the consequent activities, e.g. those taking medications being more or less likely to volunteer to participate in the study.

The interplay between intermediate variables and selection, as well as the natural history of disease, will have to be fully explored in a future work.

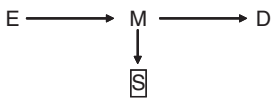

Figure 2 Selection of cohort participants (S) is affected by the mediator (M) of the exposure (E)-outcome (D) association

However, it should be emphasized that, regarding selection, the issue can be solved by taking into account the temporal relationships between the variables under study and, thus, by enrolling the participants before the intermediate variable or its early signs could become manifest. In a birth cohort study involving enrolment during the first trimester of pregnancy, for example, selection cannot be directly affected by intermediate variables acting later in pregnancy or at birth.

Conclusions

In conclusion, we agree with Rothman and colleagues that scientific inference does not require representativeness, and often explicitly requires that study samples should not be representative. Overall, representativeness can be harmful or beneficial depending on the study question and context. There is no reason to embrace representativeness per se, as often restriction can enhance the practicality of a study and/or the validity of the scientific inferences. We acknowledge that further work is needed to fully understand some specific situations, in particular when an intermediate variable directly affects baseline selection. However, leaving aside this specific issue, we consider that the view that studies based on representative samples are clearly better than those based on restricted samples is untenable. Rather, although it is perhaps too strong to argue that representativeness should always be avoided, it is usually not necessary, and often should be avoided.

Conflict of interest: None declared.

References

1. Rothman KJ, Gallacher J, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;**42:**1012–14.
2. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits: a preliminary report. *Br Med J* 1954; **1:**1451–55.
3. Kolonel LN, Henderson BE, Hankin JH *et al*. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol* 2000;**151:**346–57.
4. Asher MI, Keil U, Anderson HR *et al*. International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. *Eur Respir J* 1995;**8:**483–91.
5. Richiardi L, Baussano I, Vizzini L, Douwes J, Pearce N, Merletti F. Feasibility of recruiting a birth cohort through the Internet: the experience of the NINFEA cohort. *Eur J Epidemiol* 2007;**22:**831–37.
6. Doll R, Hill AB. Mortality in relation to smoking: ten years' observations of British doctors. *Br Med J* 1964;**1:** 1399–410.
7. 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* 2012;**65:** 407–11.
8. 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.
9. Howards PP, Schisterman EF, Poole C, Kaufman JS, Weinberg CR. ''Toward a clearer definition of confounding'' revisited with directed acyclic graphs. *Am J Epidemiol* 2012;**176:**506–11.

Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2013;**42**:1022–1026
doi:10.1093/ije/dyt105

Commentary: Should we always deliberately be non-representative?

Shah Ebrahim[1]* and George Davey Smith[2]

[1]Department of Non-communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, London,UK and [2]MRC Centre for Causal Analyses in Translational Epidemiology, University of Bristol, Bristol, UK

*Corresponding author. Department of Non-communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT. E-mail: Shah.ebrahim@lshtm.ac.uk

Accepted 1 May 2013

Rothman and colleagues were invited to submit their piece to our recently established 'Education Corner', but on reading it we felt it merited discussion and debate.[1] Those invited to comment considered that