Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2013; all rights reserved.

International Journal of Epidemiology 2013;**42**:1016–1017
doi:10.1093/ije/dyt102

Commentary: Epidemiologists have debated representativeness for more than 40 years—has the time come to move on?

Ellen A Nohr* and Jørn Olsen

Department of Public Health, Section for Epidemiology, Aarhus University, Aarhus, Denmark

*Corresponding author. Aarhus University, Bartholins Allé 2, 8000 Aarhus C, Denmark. E-mail: ean@soci.au.dk

Accepted 25 March 2013

We appreciate the opportunity to take part in the important debate on the use of representative sampling in epidemiology. It helps define epidemiology as a discipline that shares only parts of its history and tradition with biostatistics and sociology.

Rothman *et al.* have convincingly argued that scientific and statistical inference build on two very different sets of logic and need to be clearly distinguished.[1] We concur; representativeness is an overrated principle, although there are legitimate reasons for doing random sampling in epidemiology. In a descriptive study that aims at estimating the occurrence of a disease or a risk factor in a given population, representative sampling is appropriate. The same may be the case if we want to estimate an association for a specific population with a specific mix of component causes according to the component causal model, as it has been presented by Mackie and Rothman.[2,3] In this model the sufficient set of component causes is called a causal field in Mackie's terminology (Rothman used the term a causal pie), and Mackie added the 'INUS' conditions: a component cause is an Insufficient but a Necessary part of a causal field and the causal field is an Unnecessary but Sufficient condition for the event (given there are more than one causal field leading to this event). In this causal model, all component causes will act in a probabilistic manner (unless there is a causal field with only one component cause) and the strength of an association will depend on the frequency of other component causes in the causal fields leading to the disease (a serious setback for meta-analyses). Estimates of association therefore refer to a certain population with a given distribution of component causes and causal fields. Since these causal fields are only partly known, we have to sample at random to get this distribution right for the population we want to study. We also sample at random from the population at risk to get the proper exposure time distribution among controls in a case-control study with incidence density sampling.

Besides these and a few other examples, random sampling and representativeness are concepts that have caused problems, most dramatically and sadly illustrated in the National Children Study—the birth cohort above all birth cohorts (www.nationalchildrensstudy.gov). This extremely costly study from National Institute of Child Health and Development (NICHD) in the USA aims to represent all births in the USA in a certain time period, but needs to include data on causal factors that are collected prior to birth, preferably even before conception. To complicate this sampling strategy even further, the main aim is to study prenatal causes of adult diseases. This means that when these studies can be done, the causal pattern refers to a population of pregnant women that existed more than 20 years ago. Representativeness is time- and place-specific and will therefore always be a historical concept.[4] Representativeness is gone as we speak, as Heraclitus told us more than 2000 years ago: 'You cannot step twice into the same river'.[5]

Even if we could base our study on elegant probability sampling principles, when these principles meet the real world, results are often disappointing. Many compete for getting access to data from the population. The clever ones use information that the people provide themselves when using the internet or their credit cards for shopping etc. The shoe leather epidemiologist seeks permission to ask questions and more and more people say no to this request, which effectively kills the idea of obtaining a representative sample. Neither non-responders nor non-responses are based on random selection processes, and there is a limit to how much can be imputed or based on speculative modelling. Good observations can only partly be replaced by fictitious data, and the old principle of garbage in, garbage out still holds some truth.

Furthermore, the more representative a sample is, the more difficult it may be to get repeated data, and loss to follow-up is often a much more serious concern for our abilities to make counterfactual valid comparisons among the exposed and the unexposed. We should rather argue for a cohort with sufficient exposure contrast and with limited risk for loss to follow-up, although this cohort represents only the members selected for the cohort. Our causal inference addresses general laws in nature, and non-randomly selected cohort members may serve us better than a representative sample. It is only how these laws translate into a disease occurrence that depends on the population studied, and an outdated representative sample may not be an attractive choice. Also, our own experience from the Danish National Birth Cohort indicates that internal comparisons of exposed and unexposed are quite robust to lack of representativeness. Even with a recruitment rate of only 60% of those invited—and with considerable non-participation of about 40–50% in specific follow-ups—we and others so far find minor differences when comparing measures of association from the study population with those from the source population.[6-8]

As Rothman *et al.* rightly note, most scientists are not concerned about representativeness. To that group, one may add epidemiologists doing randomized trials. They often apply strict inclusion criteria in order to maximize compliance to the protocol at the expense of representativeness.

We are, and should be, concerned about whether the exposure is causally related to the disease we study. Our causal criterion is that the exposure is a cause of the disease in this population if it is true for at least one of the diseased that he/she would not have got the disease at this point in time had he/she not been exposed, all other things being equal. Given we have identified such a factor, the next concern is to find out how important it is in a given population which may be different from the population that provided the data. To say something meaningful about this we need to know the expected distribution of the other component causes and causal fields in that population. These causal patterns will change over time. That is why disease epidemics come and go. Just think about infections.

The National Children Study illustrates that this discussion is not only of internal academic interest but has serious consequences, especially for the cost of running such a cohort. Giving advice on design issues is a risky matter that requires much more than theoretical knowledge on sampling theories.

Conflict of interest: None declared.

References

[1] Rothman KJ, Gallacher J, Hatch E. Why representativeness should be avoided. *Int J Epidemiol* 2013;**42:**1012–14.

[2] Mackie JL. *The Cement of The Universe: a Study of Causation*. Oxford: Oxford University Press, 1974.

[3] Rothman KJ. Causes. *Am J Epidemiol* 1995;**141:**90–95.

[4] Olsen J. Random sampling—is it worth it? *Paediatr Perinat Epidemiol* 2013;**27:**27–28.

[5] Harris W. *Heraclitus: The Complete Fragments*. Translation and commentary and the Greek Text; available at http://community.middlebury.edu/~harris/Philosophy/heraclitus.pdf (16 May 2013, date last accessed).

[6] Greene N, Greenland S, Olsen J, Nohr EA. Estimating bias from loss to follow-up in the Danish National Birth Cohort. *Epidemiology* 2011;**22:**815–22.

[7] Howe LD, Tilling K, Galobardes B, Lawlor DA. Loss to follow-up in cohort studies: bias in estimates of socioeconomic inequalities. *Epidemiology* 2013;**24:**1–9.

[8] Nohr EA, Frydenberg M, Henriksen TB, Olsen J. Does low participation in cohort studies induce bias? *Epidemiology* 2006;**17:**413–18.