Population sampling in longitudinal surveys

COMMENT AND DEBATE

Population sampling in longitudinal surveys

Harvey Goldstein University College London and University of Bristol, UK
h.goldstein@bristol.ac.uk
Peter Lynn University of Essex, UK
Graciela Muniz-Terrera University of Edinburgh, UK
Rebecca Hardy University College London, UK
Colm O'Muircheartaigh University of Chicago, US
Chris Skinner London School of Economics, UK
Risto Lehtonen University of Helsinki, Finland

http://dx.doi.org/10.14301/llcs.v6i4.345

When and why do we need population representative samples?

Harvey Goldstein University College London and University of Bristol, UK
h.goldstein@bristol.ac.uk

(Received February 2015 Revised April 2015)

Abstract

The paper questions the need for observational studies to achieve representativeness for real populations, in particular for longitudinal studies. It draws upon recent debates and argues for the need to distinguish scientific inference from population inference.

Keywords

Observational studies, longitudinal, representativeness.

Introduction

In a recent issue of the International Journal of Epidemiology (2013, vol 42, 1012-1028) there was a debate about whether analysts have overrated, in epidemiology and social and medical science more generally, the importance of having representative samples from well-defined 'real' populations. In this paper the arguments are summarised and developed to understand how they might affect, in particular, longitudinal studies.

Setting out the arguments

The lead paper in this collection by Rothman, Gallacher and Hatch (2013a) argues that efforts to obtain samples that are representative of real populations are often misplaced and that scientific research questions in epidemiology (and the human sciences more generally) are usually better tackled by sampling purposively. By this they mean selecting groups for study that are directly relevant for the comparisons or relationships of interest, rather than attempting to estimate such relationships within any specific 'real' population. They claim that the key scientific criterion should be the attempt to replicate (generalise) findings across different populations and groups. Any failure to replicate can then lead to a study of those factors that differ among groups and which might explain

varying relationships. Thus, for example, replication across different ethnic groups, need not involve representative samples from a population containing such groups, but rather ensuring that data are representative of the groups in question and not subject, for example, to selection bias.

They suggest that traditional emphasis on statistical significance and obtaining population-unbiased estimates downplays the importance of the scientific need for generalisation and replication. As an example they talk about sampling equal numbers in age groups rather than attempting to match the distribution to the distribution within a population. This particular argument, however, seems weak since, in fact, like the example of ethnic groups, this can be regarded simply as a stratified population sample which, combined with suitable weights, can also be used to make population inferences. They also appear to be concerned largely with the situation where there are pre-existing hypotheses or comparisons of interest, whereas in reality populations are often representatively sampled in order to allow exploratory analyses that rely on sufficient diversity and heterogeneity within the population.

They also seek to make a clear distinction between descriptive statistics that require representative samples and analytical statistics that attempt to address scientific hypotheses. In fact, this distinction is often far from clear and I shall return to this point later where I also discuss what exactly is meant by a 'population'.

Four sets of authors provide responses to Rothman's paper, three of whom are broadly supportive (Elwood; Nohr & Gleen; Richiardi, Pizzi & Pearce, 2013). I shall deal with these first, then look at the paper (Ebrahim & Davey-Smith, 2013) that takes a somewhat different view and then refer to a rebuttal by Rothman and colleagues (2013b).

Elwood (2013) makes the point that that any real population is a historical entity, and when inferences about it are available it may have changed in important ways. Of course, for enumeration purposes, this may still be the best information available. For scientific purposes, however, the real population serves as an instance of an underlying process that generates a data set at a particular time, and where inference is to all possible instances. This is often referred to as a superpopulation approach and the actual real population is treated as if it were a sample from such a conceptually infinite population. Thus, the actual population serves as a useful data set for exploratory purposes or to test hypotheses within a heterogeneous sample.

All these three respondents point to the importance of taking account of possible confounders and see this as a key concern for scientific purposes. There is some discussion about choosing unrepresentative samples with a high response rate as being preferable to choosing representative samples with a low response rate. The idea of a purposive sample that can achieve a high response rate is an interesting one, but its success depends crucially on knowing the relevant characteristics of the sample. Examples where this might be the case are the use of internet-based surveys and in some cases of clinical trials. In longitudinal studies it is similar to the way in which attrition may be handled. Such studies often settle down to having a fairly stable sample that has a high response rate in repeated waves. Because the initial sample is often fairly representative the characteristics of these initial respondents can be used to 'adjust' subsequent analyses to avoid attrition biases.

The contribution by Ebrahim and Davey-Smith (2013) seeks to disagree with Rothman and colleagues on several points. They discuss the cases where non-representative samples, in particular randomised controlled trials (RCT), give results different to those from representative samples. They suggest that 'volunteer bias' may distort non-representative studies, including RCT's, and that representative sample inferences may be more trustworthy. They point to the example of the United Kingdom biobank which is not only unrepresentative but also has a very low response rate of 6%. They claim that this will not matter in terms of genetic associations since these are unlikely to be associated with selection and not susceptible to influence by confounders such as, for example, social class. Both of these statements, however, seem disputable, especially in terms of gene-environment interactions, and would require strong supporting evidence for general acceptance.

The final rebuttal by Rothman et al. (2013b) reiterates many of the original points. They use the example of the Doll/Hill smoking and lung cancer study to emphasise the importance of representativeness, although this is really an argument about observational studies versus RCTs

and doesn't add anything new. They also discuss the meaning of the statistical term 'bias' and the importance of being clear what this refers to. This is an important issue and I will return to it below.

Defining populations

It is pertinent to ask what is meant by the term 'population' and the associated issue of what is meant by 'bias'. From a statistical viewpoint these are technical terms. Statistical analysis aims to provide estimates for a collection of units (people, institutions etc.) that, at least notionally, can be resampled. Any particular sample is regarded (perhaps conditional on particular variable values, such as belonging to a given age group) as randomly selected in the case of classical inference or as being 'exchangeable' in terms of Bayesian inference. This collection of units is a population. It may be real in the sense that it can repeatedly be sampled or conceptual in the sense that any realised sample is considered to be drawn at random from it (exchangeable with respect to all other possible draws) – a superpopulation. For example, we can define the population of women who smoke in the second trimester of pregnancy as all women who have, or could ever be observed to have, this characteristic. Any scientifically generalisable statement will be one about the distribution of any of their characteristics and relationships. The term 'bias' is defined in terms of the extent to which the estimates obtained from any particular sample differ from the (unknown) distribution in this population. Thus, from a statistical viewpoint, the population does have to be well defined in terms of being able to describe its characteristics, but it does not have to correspond to any actual 'real' population. Unfortunately, there is sometimes confusion between these uses of the term population, but here I use it in the sense of a well-defined collection of units rather than any human population that actually exists or has existed.

In the case of longitudinal data there is a special problem. Suppose we sample randomly from a real population, for example all births in a given country. After the first contact with respondents, the relationship with this real population will change. Thus, some individuals will emigrate and when reporting on relationships across time, in terms of population representativeness we will need to choose whether the relevant population consists of those individuals present in the country at a subsequent occasion, including immigrants, or those who were present at the start and did not emigrate. If it is the latter then we may anticipate that as time goes on the relationships estimated are less and less appropriate for the individuals who currently make up the population (including immigrants). If the former, then we may try to obtain current representativeness, treating unknown early data on immigrants as missing. The problem is that in general such earlier data values may have different distributions from the earlier data values of those present at the start of the study. This issue will be especially important if immigration status is one of the factors under study. In the light of this, thinking about specific comparison groups would seem to be a more useful focus than attempting to decide how to define population representativeness.

The argument about the lack of need for a representative sample has considerable strength. From an analytical (scientific) perspective what is required are statements that are generalisable to specific groups, including of course those people living within a given society or environment at any moment and who happen to constitute a 'real' population, such as is measured by a census. The distinction between scientifically driven data analysis and analysis directed at making estimates for real populations, however, is not always clear. For example, if interest is in prevalence differences between ethnic groups within age categories, there may be scientific interest in whether these are changing over time within the same geographically defined population, and whether any changes can be explained by other factors. In this case successive representative samples would be needed. What would be gained scientifically from such a comparison is information on potentially causal factors that mediate or explain the prevalence differences. The use of 'real' populations for this purpose in effect is to take advantage of 'naturally occurring' changes in such factors that may be happening over time. On the other hand it may be more efficient to choose a heterogeneous sample that allows the same exploration based on having sufficient variation for those factors. Thus, if we were interested in the relationship between pregnancy smoking and neonatal mortality, we would not generally wish to derive estimates for a real population where the structure of that population affected the size of the relationship or the power to detect any effects. Thus, for example,

in a population with high average birth weight this relationship is known to be weak with a very large sample size needed to have reasonable power to detect it (see for example, Goldstein, 1977). Selecting a sample that does not represent a real population but has a high degree of heterogeneity in terms of birth weight, may provide much more power to investigate the hypotheses of interest.

We can illustrate this particular point from an analysis of early studies that looked at the relationship between maternal smoking and neonatal or perinatal mortality. Goldstein (1977) showed that, for different studies representing different populations of pregnant women, the difference in (or ratio of) mortality rates between smokers and non-smokers increased steadily as the average birth weight in the population decreased. Table 1 shows this for six different studies. The simplest explanation for the relationship is that smoking acts on mortality through an average 160g reduction in birth weight. The relationship between mortality and birth weight is nonlinear, with the relationship becoming steeper as birth weight decreases, and this implies that we will observe a greater difference for those populations with more low birth weight babies. In fact, for the two populations with the highest average birth weight, the difference is negligible.

Thus, if we had confined ourselves to the 'marginal' relationship between smoking and mortality, then our inferences would have differed according to the 'real' population studied. From a scientific perspective however, such inferences, especially in terms of a causal relationship, would be inadequate. It illustrates the point that, from a scientific perspective, the real population is of secondary importance: what we need is to understand those factors that could mediate the relationship of interest.

Table 1. Maternal smoking in pregnancy and neonatal/perinatal mortality

Population (1950-1970)	% low birth weight (<2500g)	Mortality ratio: smokers/non-smokers
US private health	3.2	1.03
Sweden	3.5	1.01
US naval wives	4.3	1.32
Ontario	4.5	1.27
UK	5.4	1.28
US general	5.9	1.40

A case where both specific population estimates are required and there is sufficient power to explore scientifically interesting hypotheses, is the British birth cohort known as 'Life Study' (Dezateux et al., 2013). This has a design that studies all 60,000 mothers over a period of time during pregnancy within relatively small but heterogeneous geographic clusters, treated effectively as a random sample from a superpopulation for those geographic strata,

together with a UK random sample over the same time period, of some 20,000 live births, treated as a random sample from the superpopulation defined over the whole country. Both components of the study are followed up during the first year of life (and potentially beyond) with considerable overlap in terms of the information collected. The pregnancy component aims to collect genetic and other biological data not collected in the birth component. The advantage of such a design is that for population estimates using variables collected in the birth component there is additional information available from the numerically larger pregnancy component to improve the accuracy of these, for example using suitable weights that can be computed from nationally available birth data. For many scientific hypotheses the data available from the pregnancy component alone will often suffice, but power can also be increased by using the data from the birth component, within a combined analysis. Furthermore, informative selection, notably as a result of non-response, can be addressed by the existence of comprehensive population birth registry data against which the characteristics of those responding can be checked. This is in effect a special case of purposive sampling.

The ability to exploit such a design requires appropriate software tools that can 'borrow strength' across the two components. Providing such tools for routine data analysis is highly desirable, although it may be practically challenging. The point, however, is that it helps to understand the debate over whether a sample should be purposive or representative since in this case it can efficiently be both.

Conclusions

The idea that population studies, especially longitudinal ones, should strive to be representative of 'real' populations may not always be helpful. While, for certain purposes associated with enumeration and administrative policies, real population representativeness is required, from a scientific perspective this may well be unnecessary. Scientific inferences are concerned with uncovering relationships that can be tested across different contexts and that may eventually attain the status of causal explanations. To ensure validity researchers need to pay attention to selection factors that may lead to biased estimates, where 'bias' is defined in terms of a clearly defined *statistical* (super)population, and much of applied statistical methodology is devoted to this issue. To enhance the effectiveness of any analysis, heterogeneity is generally desirable, and this will often imply purposive sampling that is non-representative of any particular real population. In practice, as is the case with Life Study, an optimum design may well be one that combines such purposive sampling with population representativeness, so serving both enumeration and scientific aims.

Acknowledgements

My thanks for comments are extended to the following: Carol Dezateux, Francesco Sera and Rachel Knowles. The paper is based upon one delivered at the SLLS 2014 conference in Lausanne. This work was supported in part by the Economic and Social Research Council [Grant number ES/L002353/1]. Life Study is supported by the Economic and Social Research Council (ESRC), the Medical Research Council (MRC) and University College London (UCL) and is part of the Birth Cohort Facility Project, which receives funding from the UK Government's Large Facilities Capital Fund.

References

Dezateux, C., Brockelhurst, P., Burgess, S., Burton, P., Carey, A., Colson, D., Dibben, C., Elliot, P., Emond, A., Goldstein, H., Graham, H., Kelly, F., Knowles, R., Leon, D., Lyons, G., Reay, D., Vignoles, A., & Walton, S. (2013). Life Study: a UK-wide birth cohort study of environment, development, health, and wellbeing. *The Lancet, 382,* S31. http://dx.doi.org/10.1016/S0140-6736(13)62456-3

Ebrahim, S. & Davey-Smith, G. (2013). Commentary: should we always be deliberately non-representative? *International Journal of Epidemiology, 42*, 1022-1026. http://dx.doi.org/10.1093/ije/dyt105

Elwood, J.M., (2013). Commentary: on representativeness. *International Journal of Epidemiology, 42,* 1014-1015. http://dx.doi.org/10.1093/ije/dyt101

Goldstein H. (1977). Smoking in Pregnancy: some notes on the Statistical Controversy. *British Journal of Preventive & Social Medicine 31* 13-17. http://dx.doi.org/10.1136/jech.31.1.13

Nohr, E.A., & Gleen, J. (2013). Commentary: Epidemiologists have debated representativeness for more than 40 years – has the time come to move on? *International Journal of Epidemiology, 42,* 1016-1017. http://dx.doi.org/10.1093/ije/dyt102

Richiardi, L., Pizzi, C.F. & Pearce, N. (2013). Commentary: representativeness is usually not necessary and often should be avoided. *International Journal of Epidemiology, 42,* 1018-1022. http://dx.doi.org/10.1093/ije/dyt103

Rothman, K.J., Gallacher, J.E.J., & Hatch, E.E. (2013a). Why representativeness should be avoided. *International Journal of Epidemiology, 42,* 1012-1014. http://dx.doi.org/10.1093/ije/dys223

Rothman, K.J., Gallacher, J.E.J., & Hatch, E.E. (2013b). When it comes to scientific inference, sometimes a cigar is just a cigar. *International Journal of Epidemiology, 42,* 1026-1028. http://dx.doi.org/10.1093/ije/dyt124

Commentary by

Peter Lynn University of Essex
plynn@essex.ac.uk

The need for representative survey samples

Introduction

In any field of scientific endeavour it is healthy to challenge orthodoxy. Standard practice should not be assumed to be best practice without question. Representative sampling is the orthodoxy in many applied fields of survey research and it is pleasing that this special section of *Longitudinal and Life Course Studies* is questioning when and why this should be the case. Let us be clear what this debate is *not* about. It is not about *how* to select a representative sample. There is a long history of debate on that subject, going back at least as far as the foundation of modern survey sampling theory with Kiaer (1897) and Neyman (1934), given prominence following the 1948 United States Presidential Election polling disaster (Mosteller, Hyman, McCarthy, Marks & Truman, 1949), and periodically revisited in various forms ever since. My thoughts on the role of non-probability sampling are recorded in Lynn (2005). That debate is again topical currently, particularly due to the rise of relatively cheap and fast online access panels in the social and political sciences (Bosnjak, Das & Lynn, 2015). However, the topic here is not *how* to select a representative sample but rather *when* and *why* it should be our objective to do so.

What should a sample represent?

Survey samples are rarely if ever of inherent interest. Rather, a sample is used to make broader inferences. Therefore, survey samples should be representative of something broader. But what? Goldstein's article touches upon this question by drawing distinctions between descriptive and analytical statistics and highlighting the role of confounding (or mediating) variables. I would suggest that if the analytical objective is to estimate the association between a particular set of variables, then the sample should be representative of that association. If the objective is to estimate a population distribution of some kind (be that univariate or multivariate) then the sample should be representative of that distribution. And so on. If the sample is not representative of the set of parameters to be estimated, whether those are causal, associative or descriptive, then we risk biased estimation, in the statistical sense outlined by Goldstein. It could therefore be argued that the representativeness objectives for a survey sample should depend on the analytical objectives[1].

To take an extreme example, suppose we want to estimate the association between two variables, when we already know (or assume) this association to be linear and already know (or assume) that there are no (important) confounding variables. If there are truly no confounding variables, the association should hold in any population, so it matters not whether our sample represents any particular population. In fact, we only need two non-identical observations in order to be able to perfectly estimate the bivariate association. This is obviously an unrealistic example for survey research (though it is exactly the type of estimation that takes place in school physics classes, for example), so it should be instructive to consider the ways in which it is unrealistic. First, it is ambitious to suppose that we know in advance the exact form of the association. Sampling just a few observations from each extreme of the distribution should be adequate to estimate a linear association, but if the true association has some curvature, this may be missed unless we have observations from throughout the distribution. Second, a complete absence of confounding variables is unlikely. Thus, to estimate the (conditional) association between our two variables of interest, we need also to identify (and obtain good measurements of) each confounding variable. One could argue, then, that a representative sample is not necessary provided that we can identify in advance all confounding variables of the relationship of interest, and measure them with our survey, and provided we

ensure that the sample broadly covers the distribution of interest. However, this begs the question: which distribution? To be able to truly generalise our findings, we surely mean the distribution of values that could exist in any population to which we wish to claim that our results apply. Thus, we cannot completely get away from the notion of populations.

These criteria for being able to rely on a non-representative sample are quite demanding. It is hard to envisage a realistic social science research example where we can be confident of knowing in advance all possible confounding variables (let alone being able to measure them all well). When the causal mechanism of interest is, say, biological or chemical, one may be able to get closer to meeting these criteria - and that is a possible reason for epidemiologists to have a different take on this debate to social scientists - but the fundamental issues are the same.

Most social surveys – even those tightly focused on a single topic – have multiple analysis objectives. Large numbers of estimates of different kinds are typically required, making it unlikely that all confounding mechanisms are known for all analyses. In this situation, as pointed out by Goldstein, a population representative sample will at least provide a means of identifying the form of unexplained variation, testing in an exploratory way the association of this variation with other variables, and thereby moving towards the advancement of knowledge about hitherto unidentified causal factors. The primary purpose of some surveys – and secondary purpose of many – is to provide a data resource for research by secondary analysts. It is impossible for such research to have been specified prior to the original design of the survey and therefore to have influenced the survey design. In this situation, having a population representative sample can be thought of as a safety mechanism that ensures that the population distribution of the phenomena of interest is covered and also permits estimation of the extent and nature of unexplained variation. Of course, it remains up to the researcher to decide whether the particular population covered is suitably similar to, or representative of, the kind of population to which inferences should be made. I return to this issue below.

Which Population?

The ultimate objective of most survey-based research is to inform policy or practice of some kind. With this in mind, my earlier statement about wanting a sample to be representative of the parameters of interest can be re-cast. The parameters of interest are those in the population(s) that will be affected by policy or practice. Let's refer to this population as the *policy population*[2]. So, broadly, we want our survey sample to be representative of the policy population in terms of the parameters to be estimated. How can we be sure that this is the case? We can't. Not least because the policy population is always, by definition, a future population and we can never perfectly predict the future. But there are two things we *can* do:

a) try to minimise the risk that our parameters of interest differ greatly between the study population and the policy population, by defining the study population appropriately;
b) try to predict or model relevant ways in which the policy population may differ from the study population and incorporate this into our estimation.

Step a) is typically achieved by studying the most recent available equivalent of the relevant future population. Thus, in 2015 we may be able to analyse data from a representative sample of the 2014 population of Great Britain, for example, in order to infer the likely effects of a policy that might be implemented in 2016. Our assumption is that the 2016 population will be broadly similar to the 2014 one in terms of the relevant (causal) parameters. However, we do not expect the population structure to be identical: based on recent trends, we may expect some net ageing and some net immigration, for example, in which case we can implement step b) by projecting our estimated parameters onto the predicted 2016 population structure.

The example of the previous paragraph is an optimistic scenario, where the study population and policy population have a very large overlap, though even in this case the overlap may not be as large as it seems. Policies often remain in place for many years, and can have long-lasting

impacts, so the true policy population perhaps consists of people resident in Britain at any time over the subsequent several years or decades. And often study and policy populations are even further disconnected. For example, if a good survey-based study has been carried out in one country, should researchers and policy-makers in another country assume that the findings will apply to their situation too? This is a common dilemma.

Funders must decide whether it is worth investing considerable resources to replicate a study carried out in a different context. They should be guided by the principles set out above. It is only worth funding the replication study if there is a sufficiently strong probability that the key parameters of interest are substantially different. Interpreting concepts such as "sufficiently strong probability" and "substantially different" will of course be subjective, but can be guided by knowledge of pertinent differences between the two populations and, particularly, by study findings regarding important confounders and unexplained variance.

Relevant policy populations can be very different for different types of research. Medical researchers may often hope that their findings could be generalisable to almost all current and future human populations (barring changes in the underlying etiology), whereas public bodies concerned with administering healthcare, education, housing, social support and so on are generally responsible for populations that are clearly defined by geography, usually at a national, regional, or local level. In the latter case, researchers may use survey samples that are representative of a recent equivalent of the same geographically-defined population or may resort to similarity-of-parameters arguments in using data from a different population (for example, arguing that national findings should apply in each region of the country).

Longitudinal Surveys

The arguments that I have presented so far are rather general and should apply to any sample-based scientific endeavour. However, longitudinal studies in the social sciences have at least three additional distinct characteristics that should influence the answer to the question posed in the title of Goldstein's paper:

a) Longitudinal estimates by definition refer to longitudinal populations;
b) The time interval between data collection and policy impact can be particularly great;
c) During the course of the study, new research agendas can emerge that were not envisaged when the study was initially designed.

I discuss here each of these three points in turn.

Any human population ('real' population, in Goldstein's terms) is dynamic; people will join or leave the population over time. Analysts of cross-sectional surveys tend to ignore this uncomfortable fact and instead claim that their estimates relate to a well-defined population that existed at a moment in time. This may be a reasonable approximation to reality for many purposes, but the longer the period of time over which elements were sampled or data collected, the less accurate the approximation will be.

Longitudinal surveys cannot duck this issue. An estimate of, say, the relationship between a treatment or baseline measurement and an outcome ten years later can only be based on a sample of people who were in the 'real' population at both points in time. People who entered the 'real' population subsequent to the baseline measurement (e.g. through birth, migration or status change) or who left the 'real' population prior to the outcome measurement cannot contribute to the estimate. The study population can therefore be defined as persons who were members of the 'real' population at both time points. Longitudinal parameters are properties of longitudinal populations (Smith, Lynn & Elliot, 2009), whether the population is 'real' or a conceptual superpopulation. The distinction between cross-sectional and longitudinal representativeness is important (Lynn, 2011).

Research based on long-term longitudinal studies is incredibly powerful for understanding dynamics and causality over long periods. The down side of this is that some of the data underpinning the research will be rather old. A study of the influence of infant feeding practices on, say, educational and employment outcomes by age 30 must rely on feeding practice data that is at least 30 years old. The study population and policy population are therefore separated

not by just a couple of years, as in the example of the previous section, but by four decades or more. This makes it harder for the researcher to be confident that key population parameters will remain unchanged: in a rapidly-changing world, not only may feeding practices themselves have changed, but so might the many mediators of their impacts on early-adulthood outcomes.

Research agendas certainly evolve over time, due to new knowledge, new technology, new social problems, and so on. When the sample design for the National Child Development Study (NCDS) was established, in the 1950s, it would have been impossible to envisage the myriad purposes for which researchers would be using the data half a century later. For this reason, the role of population representative sampling in ensuring the sample will contain as much heterogeneity as exists in the population is particularly important. The heterogeneity will be present for any research objective, not just those that were identified when the study was conceptualised.

Conclusion

The omission of the word 'population' from the title of this piece is deliberate: survey samples certainly need to be representative, but not necessarily of a conventionally-defined population. To meet scientific objectives, samples should represent the estimation parameters of interest. How this is best achieved will depend largely on how much is known about these parameters prior to the study. When little is known, and particularly when some research objectives cannot be well specified in advance, population representative sampling provides a mechanism for ensuring representation of extant variance. For multi-purpose surveys, population representative sampling is likely to represent an efficient compromise between the diverse optimal sample distributions for different analytical purposes. The sample should represent a population that is as similar as possible to the future policy population(s) that may be affected by study findings. A good choice may be a recent equivalently-defined population, especially when this maximises overlap between the study population and the policy population.

Longitudinal studies are typically characterised by the features that point towards population representative sampling as an appropriate strategy (limited advance knowledge about estimation parameters, inability to specify all estimation requirements in advance, large time interval between data collection and policy implementation).

References

Bosnjak, M., Das, M. & Lynn, P. (2015). Methods for probability-based online and mixed-mode panels: Selected recent trends and future perspectives. *Social Science Computer Review*. Published online 7 April 2015. http://dx.doi.org/10.1177/0894439315579246

Kiaer, A.N (1897). *The Representative Method of Statistical Surveys*. Translation 1976, Norwegian Central Bureau of Statistics, Oslo.

Kruskal, W. & Mosteller, F. (1979). Representative sampling III: the current statistical literature. *International Statistical Review 47*(3), 245-265. http://dx.doi.org/10.2307/1402647

Lynn, P. (2005). Inferential potential of non-probability samples: discussion. *Bulletin of the International Statistical Institute*, Proceedings of the 55th Session. Sydney: International Statistical Institute.

Lynn, P. (2011). Maintaining cross-sectional representativeness in a longitudinal general population survey. *Understanding Society Working Paper* 2011-04, Colchester: University of Essex. https://www.iser.essex.ac.uk/research/publications/working-papers/understanding-society/2011-04

Mosteller, F., Hyman, H., McCarthy, P.J., Marks, E.S. and Truman, D.B. (1949). *The Pre-election Polls of 1948*. New York: Social Science Research Council.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society 97*, 558-606. http://dx.doi.org/10.2307/2342192

Smith, P., Lynn, P. & Elliot, D. (2009). Sample design for longitudinal surveys. In Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*, 21-33. Chichester: Wiley. http://dx.doi.org/10.1002/9780470743874.ch2

Endnotes

[1] Kruskal and Mosteller (1979) distinguish estimation bias from selection bias. Goldstein notes that unbiased estimators can be constructed from biased samples, provided the biasing selection mechanism is known, as with the case of disproportionate stratified probability sampling. In this brief note I shall fudge this issue: my use of the term population representative sample includes – but is not necessarily limited to – any probability-based sample that covers the whole population.

[2] I deliberately avoid the term *target population*, as this is usually used in a more restrictive sense. However, under an explicit superpopulation model the two concepts converge.

Commentary by

Graciela Muniz-Terrera University of Edinburgh, UK
G.Muniz@ed.ac.uk
Rebecca Hardy University College London, UK

Some thoughts about representativeness

The paper by Goldstein makes an important additional contribution to the ongoing debate about whether and when analytic samples need to be population representative in studies in epidemiology and social and medical research.

The paper outlines the arguments presented by Rothman, Gallacher and Hatch (2013) and the stimulating accompanying commentaries that initiated the recent discussion on the topic. The need to distinguish between a "real" population and a population defined as a statistical concept that refers to any well-defined collection of units, but that may not reflect any actual population is also discussed. Additionally, Goldstein recalls the definition of bias as the difference between estimates obtained from any particular sample and the unknown true parameter of the population under study, emphasising that this population only has to be a statistically defined population and not a "real" population. In this paper, we comment on a number of points which have particular relevance for birth cohort and longitudinal studies.

The discussion of the temporal aspect of the concept of representativeness is, of course, important. Goldstein points out that representativeness is not a static concept that is preserved indefinitely over time, but rather, is a concept affected by the passing of time. Even when all efforts are made to select a representative sample of a given population at the outset of a study, the representativeness of this initial sample is unlikely to be preserved over time as the sample is followed up longitudinally. The real population of which the sample was initially representative will inevitably evolve, while at the same time loss to follow up will alter the characteristics of the study sample. Goldstein cites the example of the 'Life Study', the newest of the British birth cohort studies, where a complex sampling strategy and the use of weighting allows both the estimation of population parameters with adequate accuracy and the investigation of scientific hypotheses in a group with more extensive biological data. Let us now consider the oldest of the British birth cohort studies, the MRC National Survey of Health and Development (NSHD) (Wadsworth, Kuh, Richards & Hardy, 2006). The NSHD followed up a sample of all single births to married women in England, Scotland and Wales which took place in one week in March 1946. This initial sample included all babies born to women with husbands in non-manual and agricultural employment and one in four births to women with husbands in manual employment. This sampling scheme was chosen to keep the national distribution and to achieve a similar proportion of children in each social group (Wadsworth, 1991). Weights have thus been used when calculating prevalence estimates in order to allow for this original sampling. In 2015, the cohort is now aged 69 and the 24th data collection on the whole sample is taking place. Of course, the NSHD sample are no longer representative of the population of individuals aged 69 years old now living in England, Scotland and Wales. Demographic changes have occurred, with both immigration and emigration taking place over the lifetime of the cohort. Hence, any prevalence estimates can only ever be representative of the British-born population of 69 year olds. Furthermore, the diverse origins of immigrants joining the British population will mean that they have been exposed to different early life conditions compared with the British born population. Such differences in early life experience are likely to impact on adult health and mortality patterns and could thus affect estimates of association between early life risk and adult outcomes.

This raises the question of whether national cohort studies should adopt the practice of supplementing the samples to try and maintain study representativeness. Such supplementation was not attempted in the NSHD. In contrast, in the 1958 British Birth Cohort (National Child Development Study) and the 1970 British Birth

Cohort, during childhood, as cohort members could be traced through schools, immigrants born in the reference week were added to the samples. This was no longer possible once cohort members became adults (Power & Elliott, 2006, Elliott & Shepherd, 2006). We appreciate the value of such attempts to retain representativeness, but also see challenges in this practice if the distribution of the subgroups that comprise the original population is also dynamic and vary significantly over time. The innovative design of the 'Life study' (Dezateux et al., 2013) means that the initial sample is both "purposive and representative" and it will be informative to see how appropriate software tools for routine and complex data analysis can be provided. It will also be interesting to see whether representativeness can be maintained as the sample is followed up longitudinally, as loss to follow up and continuous demographic changes to the population occur. Given the richness of the data available in cohort studies and their ability to address unique scientific hypotheses about long term associations, we need to consider whether attempting to retain representatives by sample supplementation or by statistical weighting for investigations of prevalence is the best use of such studies.

In the original exchange between Rothman and others, Elwood (2013) elaborated the concept that any real population is a historical entity and that by the time inferences about the population are available, the initial population may have changed in important ways. We now reflect on how period effects can affect inferences made using historical data. As an example, let us consider the association of smoking and cognitive function in school pupils aged 15. Assume we have data for two samples of children that were representative of the school population aged 15 at the time of data collection, such that one sample comprised of students aged 15 years old in 1982 and the other of students aged 15 in 2013. Smoking prevalence in these two samples born 30 years apart will vary greatly. In 1982, 24 % of pupils aged 15 smoked, a percentage that has been decreasing steadily over time so that by 2013 only 8 % of pupils smoked (www.ash.org.uk) as a consequence of heightened awareness of its negative effects on health and various changes in laws, public health and commercial policies. A lack of power to detect an effect of smoking on cognitive function could therefore result as the prevalence of the risk factor declines. So, even when both samples were chosen to be representative of the population of pupils aged 15, because of a period effect, different conclusions about the association of interest could be drawn. If the researcher is interested in the potential causal association between smoking and cognition, then selecting a population with a higher prevalence of smoking is more important than picking one which is representative. On the other hand a risk factor might become more prevalent over time and thus associations may not be picked up in historical cohorts. For example, the prevalence of childhood obesity was considerably lower in the NSHD compared with cohorts born in the 1990s and later (Johnson, Li, Kuh & Hardy, 2015). It is therefore unclear whether the generally null associations between body mass index (BMI) in early childhood and coronary heart disease (CHD) observed in historical cohorts (Owen et al., 2009) are due to a lack of power. Such historical differences need to be considered and discussed when, for example, synthesizing results in systematic reviews and when implementing evidence based public health policies.

Finally, an interesting argument presented by Goldstein and discussed in the original exchange between Rothman and other commentators is about the value of non-representative samples in the context of replication and generalisation of results across different populations. The importance of a thorough understanding of all the potential sources of heterogeneity across studies, including the representativeness, or not, of samples, and the period effects, as well as differences in data collection methods and analytic methods when evaluating the reproducibility of results is vital. These points are of particular relevance in the heated debate about reproducibility and replicability of results that has entertained the attention of researchers across various scientific areas (Francis, 2012; Ioannidis, Nosek & Iorns, 2012; McNutt, 2014; Mulkay & Gilbert, 1986), particularly when reproducibility is defined as the conceptual replication of experiments as conceived by Drummond (2009). Despite unfortunate publishing practices that discourage publication of reports that aim at testing reproducible research and result in publication biases (Francis, 2012), the concept of

reproducible research has, historically, been at the core of scientific discovery.

From that perspective, the need to generate strong evidence about patterns of associations is at the core of the multi-study work fostered by the Integrative Analysis of longitudinal Studies of Ageing network, a network of longitudinal studies of ageing (www.ialsa.org). Researchers affiliated to the IALSA network independently analyse data from multiple studies employing a coordinated approach that involves the consistent use of the same analytical method (identical analytical model where possible and consistent coding of harmonized variables where possible). This coordinated analytical approach maximises the ability to fairly compare results and enables the examination of consistency of patterns and of associations across samples that may differ in a variety of ways, including differences by geographical location, sample composition and representativeness (Piccinin). The use of the same analytical approach reduces the potential sources of heterogeneity across studies that may emerge from the use of different statistical methodologies to answer similar questions. Consistent results generated from diverse samples are reassuring and provide stronger evidence in support of the hypothesis tested. On the other hand, inconsistent results require a thoughtful evaluation of potential reasons that may explain the divergence of results, including differences that may emerge from features of the data (including representativeness), and sample composition and sampling procedures. For example, in an investigation of the association of the effect of education, age and sex on global cognitive function measured using the Mini Mental State Exam in six international longitudinal studies of ageing, Piccinin and colleagues (2012) found that education was positively associated with performance across all six studies, but was only associated with rate of decline in the cohort containing the oldest participants. In five of the six studies, estimates of rate of decline were also found to be similar, but in the cohort of oldest individuals, individuals were found to decline at a much faster rate than in the other samples. The authors report that an investigation of the sample composition and a better examination of the sampling procedure followed in this outlying study helped them understand that dementia cases had been handled differently in the study compared to the other studies. Indeed, in this study efforts had been made to keep individuals who developed dementia in the study, whereas in all the other studies individuals with dementia were not included in the follow up samples. When individuals with dementia were removed from the sample, the estimated rate of decline aligned to the rate of decline estimated in the other five studies.

The general discussion about representativeness and Goldstein's contribution with particular relevance to longitudinal studies and their historical context is very valuable. This discussion is helpful in raising awareness among researchers to think more about when representativeness is a problem, but also to appreciate when to value a lack of representativeness.

References

Dezateux, C., Brockelhurst, P., Burgess, S., Burton, P., Carey, A., Colson, D., Dibben, C., Elliot, P., Emond, A., Goldstein, H., Graham, H., Kelly, F., Knowles, R., Leon, D., Lyons, G., Reay, D., Vignoles, A., & Walton, S. (2013). Life Study: a UK-wide birth cohort study of environment, development, health, and wellbeing. *The Lancet, 382,* S31. http://dx.doi.org/10.1016/S0140-6736(13)62456-3

Drummond, C. (2009). Replicability is not Reproducibility : Nor is it Good Science. Retrieved from: www.site.uottawa.ca/ICML09WS/papers/w2.pdf

Elliott, J. & Shepherd, P. (2006) Cohort profile: 1970 British Birth Cohort (BCS70). *International Journal of Epidemiology, 35*(4), 836–843. http://dx.doi.org/10.1093/ije/dyl174

Elwood, J.M. (2013). Commentary: On representativeness. *International Journal of Epidemiology, 42*(4), 1014–1015. http://dx.doi.org/10.1093/ije/dyt101

Francis, G. (2012) Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin and Review 19*(6), 975-991. http://dx.doi.org/10.3758/s13423-012-0322-y

Ioannidis, J.P.A., Nosek, B. & Iorns, E. (2012). Reproducibility concerns. *Nature Medicine 18*(12),1736–7. http://dx.doi.org/10.1038/nm.3020

Johnson, W., Li, L., Kuh, D. & Hardy, R. (2015). How Has the Age-Related Process of Overweight or Obesity Development Changed over Time? Co-ordinated Analyses of Individual Participant Data from Five United Kingdom Birth Cohorts. *PLOS Medicine*, *12*(5), e1001828. http://dx.doi.org/10.1371/journal.pmed.1001828

McNutt M. (2014). Reproducibility. *Science*. *343*(6168),229. http://dx.doi.org/10.1126/science.1250475

Mulkay, M.& Gilbert, G.N. (1986) Replication and Mere Replication. *Philosophy of the Social Sciences 16*(1), 21–37. http://dx.doi.org/10.1177/004839318601600102

Owen, C.G., Whincup, P.H., Orfei, L., Chou, Q.A., Rudnicka, A.R., Walthern, A.K. … Cook, D.G. (2009). Is body mass index before middle age related to coronary heart disease risk in later life? Evidence from observational studies. *International Journal of Obesity, 33*(8), 866–877. http://dx.doi.org/10.1038/ijo.2009.102

Piccinin, A.M., Muniz-Terrera, G., Clouston, S., Reynolds, C.A., Thorvaldsson, V. Deary, I.J. …Hofer, S.M. (2012).Coordinated Analysis of Age, Sex, and Education Effects on Change in MMSE Scores. *The Journal of Gerontologly Series B: Psychological Sciences and Social Sciences*, *68*(3), 374-390. http://dx.doi.org/10.1093/geronb/gbs077

Power, C. & Elliott, J. (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology, 35*, 34–41. http://dx.doi.org/10.1093/ije/dyi183

Rothman K.J., Gallacher, J.E.J. & Hatch, E.E. (2013). Why representativeness should be avoided. *International Journal of Epidemiology 42*(4), 1012-1014. http://dx.doi.org/10.1093/ije/dys223

Wadsworth, M. Kuh, D., Richards, M. & Hardy, R. (2006). Cohort profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology 35*(1), 49-54. http://dx.doi.org/10.1093/ije/dyi201

Wadsworth, M.E.J. (1991). *The imprint of time: Childhood, History and Adult Life.* Oxford University Press.

Commentary by

Colm O'Muircheartaigh University of Chicago, US
caomuirc@uchicago.edu

Why we need population representative samples

Goldstein questions the need for observational studies to achieve representativeness for well-defined populations, in particular for longitudinal studies. While he recognises the distinction between the notions of representativeness and proportionality, he fails to acknowledge the importance of distinguishing between samples of convenience and targeted samples from special subpopulations. In this note I emphasise the critical significance of probability sampling, in contrast to purposive sampling, and draw special attention to the artificial distinction between descriptive and analytical statistics. Goldstein (correctly) draws attention to the confusion between disproportional sampling and non-representative sampling but fails to recognise the inferential implications of choosing between probability samples and nonprobability samples. A probability sample is in essence a sample in which every element of the population has a (known) non-zero probability of selection; the definition of the population may be such that it does not correspond to a real population. The structure of a probability sample from a (general) population may exclude some domains from the target population and may be modified by design in order to produce appropriate numbers of cases for particular comparisons of subsamples of that target population. Probability samples have particular strength in making inferences, whether for scientific or for policy purposes.

Defining populations

All inference is, by definition, to a population beyond the sample on which the inference is based. Much of the argument in Goldstein, and in the papers he references, has to do with the definition of this inferential population. I concur that the population must be clearly defined; I accept also that it may not correspond to a "real" population at a point in time. However, unless it can be defined in such a way that a sample may be selected from it, there will be no scientific foundation for inferences to it without untestable assumptions about freedom from bias.

Consider first the case where the purpose is to represent a national population; as an example, consider the selection of a sample for the United States (US) National Children's Study (NCS)(Michael & O'Muircheartaigh, 2008). In designing a nationally representative sample for this study, the purpose is not to address every subpopulation of interest in the US. The purpose is to insure that every element in the population has a non-zero probability of being selected into the sample. This is achieved by identifying a survey population that is defined to be as close to the target population as feasible, such that it reflects both measurable and unmeasurable characteristics of that population

Suppose that we are interested in the relationship between an environmental exposure X and a health outcome Y, which can be modeled (for simplicity) as the linear function $Y=a+bX+e$. If all people in the population have the same b, then the nature of the sample does not matter because as long as X is accurately measured we will have only random measurement error in Y. However, if there are confounding factors Z, which affect Y and are related to X, then our estimate of b may be biased unless the elements of Z are controlled. If Z is known, then model-based estimates of the relationship between X and Y can be obtained that control for Z and yield an unbiased estimate of b, again regardless of the sampling design. However, there may also be moderator variables W, which interact with X in influencing Y. Here, different individuals will have different values of b depending on the elements in W. If W is known, then we can include interactions in the model and the separate estimates of b will also be unbiased.

Unfortunately in practice W and Z are at least to some extent unknown and in the case of longitudinal studies like the NCS are likely to evolve over time. Some elements of Z and W may be known but are unmeasurable and others may simply be unknown at the time. Here, the best that we can do is to provide an average effect b. To do so, however, requires that we create a sample that

fully reflects the population of interest, a probability sample drawn from the population so that our estimate of *b* is an unbiased estimate of the average effect in the population or in a defined subgroup. The probability sample guarantees that we will (in expectation) cover the range of confounding variables proportionately.

It is also possible that the interest is not in the average effect but in the effect on specific subgroups of this general population (as in the birth weight example below). Thus, in the NCS we might wish to focus on particular ethnic groups or on the comparison of these groups. In this case to maximize power for the comparison we would take equal numbers of cases from the groups of interest, rather than numbers proportional to their distribution in the general population. These subsamples would however be chosen to be representative of the groups of interest; their representativeness would be warranted by the fact that they were probability samples from their respective groups. Only the relative sizes of the subsamples would deviate from the parent population, not the intrinsic nature of the sampling process.

Goldstein's example of the relationship between pregnancy smoking and neonatal mortality provides a further illustration of this principle. The six studies he cites (from an analysis by Goldstein (1977)) demonstrate a non-linear relationship between mortality and birth weight, with a negligible effect for the two populations with the highest average birth weight, and an increasingly steep relationship as the population average birth weight decreases.

Goldstein argues that this example demonstrates the secondary importance of the population. To the contrary, the data demonstrate the opposite. Had the range of birth weights across the US been included in a single US study, the analysts might have been more likely to observe the non-linearity in the relationship; this indicates the importance of covering the full range of variation of X, W, and Z in a population rather than accepting the subpopulation that is most convenient. One might indeed argue that there was a failure of both the theoretical basis and the analysis of the studies in not examining the data for possible interactions with birth weig
ht in the model,

At no point in his disquisition does Goldstein suggest that the samples in any of the studies he cites should be "non-representative". The implicit understanding is that the sample in each is in fact representative of the population from which it is drawn. Were it not, neither the partial generalisation within the study would be justified, nor would its incorporation into Goldstein's 1977 meta-analysis.

Two-phase sampling

The case of the British birth cohort known as the 'Life Study' is also subject to an alternative interpretation from that offered in Goldstein. A geographically clustered sample of 60,000 mothers is selected from a set of relatively small but geographically heterogeneous clusters; the 60,000 mothers are assumed to constitute a random sample from a set of geographic strata; there is a parallel (random) UK sample of 20,000 live births. The two samples can be used together to "borrow strength" from each other for different analyses. Comprehensive national (population) data from birth registries can be used to correct for differential nonresponse.

This combining of samples with different characteristics and different intensity of measurement is well recognised as a powerful design. The classic two-phase sampling design (Neyman, 1938) proposes just this combination of general representation and subsample focus; Neyman visualizes both samples as probability samples. Goldstein proposes this as a special case of purposive sampling, though it is not clear what his argument is. Presumably he does not argue that selecting the geographical areas purposively is superior to a design in which the areas were selected on a probability basis from a properly constructed frame of geographical areas. If indeed the selected areas were for some reason the only areas available, then suspicion must attach to them as being unrepresentative even of areas with ostensibly equivalent characteristics.

The extent to which the combined sample can be justifiably used to make inferences to the whole population depends critically on either (i) both samples being probability samples, or (ii) model-based assumptions that allow generalisation from the purposive component to the whole.

Additional benefits of representation through probability sampling

A platform for scientific discovery

Hypotheses about new exposures and gene-by-environment moderation will arise over the next 20 years, and a probability sample provides the best insurance that the study will provide useful numbers of children with variation in those environments and exposures of interest. The probability design also increases the prospects for serendipity by maximizing the spread of W and Z in the sample.

Maximization of scientific acceptability of data and of discoveries across disciplines

While many disciplines do not require probability samples for their inferences, no discipline considers a probability sample to be inferior to an alternative. Thus data based on a probability sample maximize the potential for cross-disciplinary collaboration and publication.

Public and political/policy acceptance

Resource allocation and acceptability of discoveries will be greater if the data are based on a scientifically warranted representative sample of the population.

Full variation in risks and exposures

A probability sample will produce generalisable risk estimates and the capability to estimate policy/intervention benefits from associations discovered and reported from the study.

Conclusion

Investigations of all kinds can make a contribution to science, and samples that are not representative have a place in scientific research, especially at early stages of exploration. I contend however that the superficial message of Goldstein's excellent article is wrong. Ceteris paribus, for both science and policy a probability sample is superior to a non-probability sample, representation trumps convenience, and the best way to obtain representation of the population of interest is through probability methods.

References

Michael, R. & O'Muircheartaigh C. (2008). Design Strategies and Disciplinary Perspectives: the Case of the US National Children's Study. *Journal of the Royal Statistical Society, Series A*, *171*(2) 465-480. http://dx.doi.org/10.1111/j.1467-985X.2007.00526.x

Neyman, J. (1938). Contributions to the theory of sampling human populations, *Journal of the American Statistical Association, 33*, 101-116. http://dx.doi.org/10.1080/01621459.1938.10503378

Commentary by

Chris Skinner London School of Economics, UK
c.j.skinner@lse.ac.uk

Discussion of 'When and why do we need population representative samples?'

There is much wisdom in this paper by Harvey Goldstein which builds on discussion in a set of papers in the *International Journal of Epidemiology* (IJE), and applies the ideas developed to a new British birth cohort study, the Life Study. I shall focus on his main theme, which rejects the need for representative samples, and on his concluding remarks relating to the Life Study. My comments come particularly from a survey statistics perspective.

I was reminded in looking at the papers in IJE of the observation by Kruskal and Mosteller (1979) (and in their three related articles) that the term 'representative sample' has multiple uses and "because of its ambiguities and imprecision", they "recommend great caution" in the use of this term and "usually a more specific expression will add clarity" (p.13). I shall seek to make greater use of the expressions 'population', 'sample' and 'bias' in my discussion.

As I understand Goldstein's main concern about representative sampling, it is that, for scientific purposes, making inference about 'real' populations is of secondary importance. This is a position which I should like to question. The survey statistics literature does make a distinction between descriptive/enumerative and analytic/scientific uses of surveys/studies. Estimation for a single study population is a common primary objective for the former. For the latter, the focus of Goldstein's paper, I think the notion of population will invariably need further refinement, but I think it can still serve a useful purpose to specify collections of units underlying targets for inference. I do not feel the need to downplay the notion of 'real' population.

Perhaps the simplest definition of populations of interest for scientific purposes is where there are two subpopulations to compare. I conceive of these subpopulations as 'real populations' in Goldstein's terminology. Suppose, for example, we wish to undertake a comparison of an outcome Y, according to values of X, given confounding factors Z (say infant mortality by maternal smoking given birth weight in Goldstein's example). For such conditional analysis, it would be natural to define specific subpopulations by X and Z, between which comparisons are to be made. Thus, in the example, one might choose to compare a low birth weight subpopulation and a normal birth weight subpopulation. Such comparisons have many vital roles in scientific research, as Goldstein notes. They may help to elicit and test causal hypotheses, perhaps through control of confounding factors. They may be valuable in assessing the replicability of findings across populations or to learn about interactions.

Given the specification of such subpopulations, it will often make sense to sample these subpopulations with different sampling fractions. For example, as discussed by Goldstein, the power to investigate the analytic objectives may be improved by sampling the low birth weight subpopulation with a higher relative sampling fraction. But I do not see this observation as any reason why the subpopulations (as real populations) are of 'secondary importance'. Their definition seems fundamental. I also do not see any reason why an analysis embracing a comparison of such subpopulations need be weighted to the population of all births (Skinner, 2005, p.84), let alone any need for the analysis to be confined to the 'marginal' relationship between smoking and mortality.

The simple comparison of subpopulations needs extension in various ways. With a continuous variable like birth weight, the definition of subpopulations via cut-points is arbitrary and we may imagine intervals of values of decreasing width and decreasing population counts. In this context, the notion of superpopulation which Goldstein mentions is useful and enables, for example, a regression relationship with continuous covariates to be specified in usual model terms. The longitudinal setting also introduces complexities, as Goldstein notes. A population like a labour force becomes dynamic with people entering and leaving the labour force over time. Even more complexity

arises with, for example, households with the structure of the unit changing over time. In such cases, the term 'population' may seem stretched, but I think it is still reasonable to think in terms of what Goldstein calls a 'well-defined collection of units'. Causal questions cannot be assessed from data on a single case but rather require reference to a set of units. As Holland (1986, p. 947) writes, "the important point is that the statistical solution [to the fundamental problem of causal inference] replaces the impossible-to-observe causal effect of t on a specific unit with the possible-to-estimate average causal effect of t over a population of units". In my view the relevant populations do define 'real' notions of primary not secondary importance, given the need to report scientific findings transparently in terms of the kinds of people or other units to which they apply.

I now turn to the role of sample selection. I have already noted, in agreement with Goldstein's discussion, that it may often be sensible to allocate the sample differentially according to variables of scientific interest (X and Z above) with a view to improving sampling efficiency (i.e. reducing variance). Consider next the question of bias, as arising from differences between the characteristics of sample units and those in the population (as conceived of in the previous three paragraphs). I have in mind bias arising from purposive and other forms of non-probability sampling, for example the volunteer effects described by Ebrahim and Davey-Smith (2013). Such bias is of major concern to survey methodologists today, with the relentless push to adopt non-probability samples, such as in internet panels, for cost and other non-scientific reasons.

In summary, I do think that in the analysis of longitudinal studies it is desirable to specify collections of units as populations, with a clear scientific rationale, and that the potential biasing effects of sample selection are of primary concern.

My final comments will elaborate on these points in the context of the Life Study. Here the basic study populations from which samples are drawn (leaving aside timing aspects) are (a) k populations of pregnant mothers (and partners) associated with k maternity units and (b) the population of all live births in the UK. I am unclear about the value of k (perhaps it remains to be determined) but suppose that it is small (under 10?). Sampling in (a) is by census and in (b) by a standard probability scheme and so, for the purpose of current discussion and leaving aside non-response considerations, I think we can disregard issues of representative sampling **within** these populations.

In the context of the earlier discussion, the key issue relates to the purposive selection of the maternity units. Following Goldstein's discussion, it seems natural to ask what is the scientific rationale for the choice of maternity units? From Goldstein's paper, the rationale seems to be geographic heterogeneity, perhaps associated with differences in distributions of what I have called X and Z variables relevant to the study. This raises the question of how differences in findings between different maternity units are to be interpreted? If, for example mortality ratios vary between units as in table 1 and there is also significant variation between units in a large number of other maternal health and socioeconomic factors, how will the finding be scientifically informative if k is small? Moreover, for some kinds of analyses, interpretation may even be complicated by confounding between the effect of the maternity unit and the nature of the maternal population.

In any case, if the results of analyses of data from a given maternity unit are only to be reported as relating to that population then issues of external generalisability are avoided and I have no concerns about sample selection bias. There do not then seem to be any differences in questions of representativity/generalisability compared to other geographically specific studies, such as the Southampton Women's Survey (Inskip et al., 2006). The fact that scientific studies have some spatial and temporal specificity seems inevitable.

The more difficult questions relate to how the data will be combined across populations. The statistical methodology for standard comparisons would seem straightforward. Thus, in a regression setting, one may construct a categorical covariate representing both the k maternity populations and the general 'birth population', the latter possibly broken down by region or in some other geographical way. I am still unclear how to interpret the coefficients of this covariate and associated interaction terms, but this is just the comparative question I have already asked above.

Much less straightforward seems to me the question of how far it will be possible to increase "the precision of estimates for nationally representative measures" (Dezateux et al., 2013)

using the maternity unit data, that is how to use data from a restricted and purposefully selected set of geographical clusters to make inference about the wider UK population? This 'borrowing of strength' across (a) and (b) is intended to provide, as Goldstein refers to it, an optimum design combining purposive sampling with population representativeness.

A review of non-probability sampling was conducted recently by the American Association of Public Opinion Research, with a summary report and discussion appearing in Baker et al. (2013). The combination of a national probability sample with a small number of geographically clustered 100% samples does not appear to be a standard approach. Baker et al. (2003) do provide some discussion of weighting and note that "the main concern with model-based inferences from non-probability samples is that population estimates are highly dependent on model assumptions" (p.97). A combination of a large non-probability sample (161,000+ web respondents) with a smaller 'nationally representative' quota sample (10,000+ respondents) was used in the Great British Class survey (Savage et al., 2013). Savage et al. (2014) recognised that their design is 'unorthodox', in response to criticisms e.g. by Mills (2014), and emphasised that their work should be seen as part of an 'experiment'. This survey is very different from the Life Study but I mention it just to illustrate that such 'combined' designs seem to me still novel and the extent to which reliable and efficient national estimates can be produced by combining the separate data sources seems to me a topic still in need of further study.

'Borrowing strength' is referred to in the small area estimation literature (e.g. Ghosh & Rao, 1994), but in that context borrowing across geographical units comes from fitting a model across a sufficient number of such units for a reasonable model to be fitted and for valid confidence intervals, taking account of geographic heterogeneity, to be constructed. It is not clear to me that k will be large enough for such an approach to be adopted.

Goldstein suggests a weighted approach will be used. One approach would be to weight inversely by the probability of selection, with weights of one attached to members of the maternity unit sample (since 100% are sampled). However, I would assume this would only increase the effective sample size of the birth sample by a small fraction and that this is not what is conceived. The idea may instead be to construct weighting classes using population registry data (but not geography) and then to make the modelling assumption that observations are exchangeable between the maternity unit and birth populations within weighting classes. Such a modelling assumption will depend upon the relevant analysis and the availability of auxiliary information but, in general, it would seem to me heroic. The assumption should, at least, be testable, although its testing would seem to be similar to testing the hypothesis of no maternity unit effect in the kind of regression analysis I noted above, where weighting variables are included as covariates. In summary, the proposed combined design seems to me to be novel (although perhaps I am unaware of similar designs) and I think there are several methodological questions regarding data combination to explore, even before one gets to the question of software tools referred to by Goldstein.

I am grateful to the Editor for the opportunity to discuss this interesting paper.

References

Baker, R., Brick, M.J., Bates, N., Battaglia, M., Couper, M.P. & Dever, J.A. (2013) Summary report of the AAPOR Task Force on non-probability sampling (with discussion). *Journal of Survey Statistics and Methodology*, *1*, 90-143. http://dx.doi.org/10.1093/jssam/smt008

Dezateux, C., Brockelhurst, P., Burgess, S., Burton, P., Carey, A., Colson, D., Dibben, C., Elliot, P., Emond, A., Goldstein, H., Graham, H., Kelly, F., Knowles, R., Leon, D., Lyons, G., Reay, D., Vognoles, A., Walton, S. (2013). Life Study: a UK-wide birth cohort study of environment, development, health, and wellbeing. (2013). The Lancet, 382, Pp S31. http://dx.doi.org/10.1016/S0140-6736(13)62456-3

Ebrahim, S. & Davey-Smith, G. (2013). Commentary: should we always be deliberately non-representative? *Intl. J.* Epidemiol., 42, 1022-26. http://dx.doi.org/10.1093/ije/dyt105

Ghosh, M. & Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. *Statistical Science, 9,* 55–93. http://dx.doi.org/10.1214/ss/1177010647

Holland, P.W. (1986) Statistics and Causal Inference, *Journal of the American Statistical Association*, *81*, 945-960. http://dx.doi.org/10.1080/01621459.1986.10478354

Inskip, H.M., Godfrey, K.M., Robinson, S.M., Law, C.M, Barker & Cooper, C. (2006) Cohort profile: the Southampton Women's Survey. *International Journal of Epidemiology*, *35*, 42-48. http://dx.doi.org/10.1093/ije/dyi202

Kruskal, W. & Mosteller, F. (1979) Representative sampling, I: Non-scientific literature, *International Statistical Review*, *47*, 13-24. http://dx.doi.org/10.2307/1403202

Mills, C. (2014) The Great British Class Fiasco: A Comment on Savage et al. *Sociology*, *48*, 437-44. http://dx.doi.org/10.1177/0038038513519880

Savage M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J. … & Miles, A. (2013) A new model of social class? Findings from the BBC's Great British Class Survey experiment. *Sociology, 47*, 219–50. http://dx.doi.org/10.1177/0038038513481128

Savage M., Devine, F., Cunningham, N., Friedman, S., Laurison, D., Miles, A. … & Taylor, M. (2014) On Social Class, Anno 2014. *Sociology*, *48* (forthcoming). http://dx.doi.org/10.1177/0038038514536635

Skinner, C.J. (2005) Introduction to Part B. In R.L. Chambers and C.J. Skinner (Eds), *Analysis of Survey Data* (pp. 75-84). Chichester: Wiley.

Commentary by

Risto Lehtonen University of Helsinki risto.lehtonen@helsinki.fi

I want first to congratulate Harvey Goldstein for his inspiring debate paper titled "When and why do we need population representative samples?" Population representativeness versus sample purposefulness has been recently debated in epidemiology and social sciences literature. Rothman, Gallacher and Hatch (2013a) challenge the dominant role of representativeness in epidemiology and social and health sciences by asking why representativeness should be avoided and arguing that "...studies that control skillfully for confounding variables and thereby advance our understanding of causal mechanisms" offer a proper route ahead (1014). According to Rothman, Gallacher and Hatch (2013b) "representativeness, although it may have a place in health surveys, is not a proper goal for scientific studies" (1027). By "scientific studies" he refers to causal studies about how nature operates.

In his debate paper Goldstein addresses several points that remain unclear in Rothman's writing. I agree with many of Goldstein's arguments. Both he and Rothman seem to restrict what they call "representative sampling" essentially to "enumeration" or "population inference" purposes, that is, the sample data set is used to estimate the parameters of a well-defined finite population, for example the prevalence of chronic disease in age-sex-groups in a given real population at a given time point. Later in the paper Goldstein however widens his framework beyond that of Rothman. As an example, he describes the follow-up study design of the British "Life study". For that study he proposes the use of additional register-based population information (sometimes called auxiliary data), supplementing the original study data, for both descriptive (enumeration) purposes and for studying scientifically interesting hypotheses. He considers the combined use of data taken from different sources to represent a special case of purposive sampling. Goldstein thus proposes a kind of hybrid solution: "...an optimum design may well be one that combines such purposive sampling with population representativeness, so serving both enumeration and scientific aims". In my opinion, this is a fruitful view and I will try to elaborate this approach further in my commentary.

What is meant by 'representativeness' and 'purposefulness'?

Population representativeness (representativeness for short) and representative sampling are key concepts in Rothman and Goldstein's papers (29 hits in Rothman and 25 in Goldstein) but the concept itself remains unclear. This is not necessarily a surprise because there is no universally accepted definition of representativeness or representative sampling. In a series of four papers on representative sampling published in the International Statistical Review in 1979 and 1980, William Kruskal and Frederick Mosteller give nine different definitions of representative sampling they have found in scientific literature. All definitions are loose. Freshmen may think population representativeness refers to a miniature population obtained by representative sampling i.e. study subjects are selected from the population with an equal chance of being included. This interpretation is far too simplified because such a design only represents a special case of probability sampling. Even if the term "representativeness" is rarely used in modern survey sampling literature, we might think of population representativeness as a procedure where the study subjects are selected with a specified random mechanism from a well-defined finite population, either with equal or varying probabilities. If drawn with varying probabilities, the structure of the realised sample data set is restored (or forced to be "population representative") by weighting the observations by the inverses of the inclusion probabilities. Obvious benefits of probability sampling are in its flexibility for a controlled selection of the study subjects and in its ability to provide a basis for proper statistical inference under the actual sampling design. For example, oversampling of understudied groups would be covered, as suggested by Rothman. However, the scope of representative sampling in Rothman's paper seems narrower (this also holds for Goldstein's paper). Unequal probability sampling

is not explicitly covered, as can be inferred, for example, from Rothmans's rebuttal (2013b p. 1026). When reading both papers it is hard to disagree with Kruskal and Mosteller who suggest avoiding the use of the concept of representativeness. In epidemiological literature, the term is occasionally used without clarification but it is fair to say that in some cases, a reasonable explanation is given (An example is Rothman, Greenland & Lash, 2008 p. 146).

Purposive sampling is another key concept in Goldstein's paper (Rothman does not use the concept of purposive sampling). This concept is problematic as well. Purposive in what specific sense? In Goldstein's paper, purposive sampling refers to a sample that is "non-representative of any particular real population". Now, it remains unclear whether a probability sample from a real population becomes "purposive" because of serious and informative nonresponse or if, instead of probability sampling, a quota sampling method or a self-selection scheme has been used or, alternatively, if the realised sample data set is being interpreted to be "representative" of a fictitious superpopulation. Later on I will come back to purposive sampling from a survey statistics point of view.

As a curiosity, there is a certain discrepancy between representativeness and purposefulness going back to the infancy of probability sampling. In a seminal paper entitled "Den repræsentative Uildersøgelsesmethode" (The representative method of statistical surveys) published in 1897 by a Norwegian statistician Anders Kiær, the term "representative" appears for the first time in survey sampling literature. His main argument was that it is not necessary to implement a census to obtain useful information on a human population but to carry out a "partial investigation". Fulfilling a well-specified type of representativeness on the population structure, would be enough to make inferences on the whole population. But in fact, the method of Kiær is a kind of combination of representativeness and purposive sampling (see e.g. Langel & Tillé, 2011).

The contradictory nature of the two key concepts, representativeness and purposefulness, has given rise to much debate and misinterpretation for decades (and the contradiction is, implicitly or explicitly, visible in both Rothman and Goldstein's papers). The discriminatory power of the terms is weak as is evident from example in the paper by Goldstein. As the climax of his paper, a certain type of data combination appears effectively to be both purposive and representative, indicating a complete overlap of the concepts. So, we are back in Kiaer!

The hybrid solution revisited

Let me elaborate further Goldstein's hybrid solution by using ideas from modern survey statistics. The key idea is to successfully combine, in one way or another, methods used in the sampling phase for the selection of study subjects and the methods used in the analysis of the study data. In both sampling and analysis phases, auxiliary population data taken from administrative registers or censuses and statistical modelling can play a crucial role. For example, in balanced sampling (Deville & Tillé 2004) the sample is forced to fit with the known population distribution of selected auxiliary variables, in effect representing purposive sampling with properly defined inclusion probabilities. In the analysis phase, the effect of varying inclusion probabilities caused by balancing can be adjusted for by weighting the sample observations with inverse inclusion probabilities, which is a standard survey analysis practice. Alternatively, the effect of balancing can be accounted for by including the balancing variables as potential explanatory variables in the statistical model to be fitted to the study data set, representing a possible model-based way of treating sampling complexities. As an extension for the analysis phase, statistical calibration techniques (e.g. Särndal, 2007) offer methods for the construction of calibrated weights that force the sample distribution of selected auxiliary variables (covariates; e.g. demographic, socioeconomic etc.) to fit with a known population distribution. The weights (possibly combined with the original survey weights) are then supplied to the analysis procedure (as weight variables or covariates). Thompson (2015) addresses complex longitudinal surveys from both a survey analysis and model-based analysis point of view. Gelman (2007) discusses weighting in the context of Bayesian analysis.

In my opinion, the hybrid design of combining the study data, the available auxiliary population data and statistical modelling fulfils many of the properties of an optimal design introduced by Goldstein. There are many favourable properties in

this approach. The combined methodology offers a useful tool for the balancing of the sample distribution of important confounders against the known distributions at the population level, needed in studies based on purposive sampling and in probability samples that suffer from severe and informative nonresponse and selective attrition. Protection against model mis-specification can be attained for superpopulation-based approaches. If the inferential framework is model-based, the auxiliary variables (or the constructed weight variable) - featuring important aspects of the sampling design and nonresponse patterns - might be included as covariates in the statistical model to be fitted in the analysis phase. Effective adjustment for informative nonresponse and attrition can be attained if the auxiliary variables correlate with the response mechanism. Moreover, improved accuracy is possible if the auxiliary variables correlate with the study variables.

Obviously, the hybrid methodology can be very effective in "enumeration studies" where probability sampling (with equal or unequal inclusion probabilities) plays an important role, even if the inferential frameworks may differ. This is because probability sampling offers a firm basis for statistical inference in any empirical science. With certain restrictions, the methods are applicable for non-probability samples as well. In "scientific studies", the approach can be used for example to protect against the possible selection bias of study subjects. Moreover, the methodology toolbox fades out the unnecessary or even harmful confrontation between “scientific studies” and “enumeration studies”, because with appropriate choices the methodology applies to both.

Requirements for data infrastructure

The power of the hybrid machinery described above depends on the data infrastructure accessible to the researcher. Even if there are huge differences in this respect between countries, aggregate-level auxiliary data on demography, health and social affairs are often available in population censuses, official statistics and administrative registers, fulfilling minimum requirements for the methodology. The British "Life study" described by Goldstein offers a good example. Li, Li and Graubard. (2011) illustrate the importance of accounting for the complexities of the study design (stratified multi-stage sampling involving intra-cluster correlation, informative nonresponse accounted for with weighting and calibration to census totals) in order to obtain valid inference in a genetic study. The study shows the potential of the combined methodology in a data infrastructure where aggregate-level census data are available.

In the so-called register countries, notably in the Nordic countries, including Denmark, Finland, Norway and Sweden, unit-level data on various auxiliary variables are available from statistical register and from administrative sources for scientific research in epidemiology and social and health sciences. Examples of data sources are health registers and registers on socio-economic conditions (see e.g. Gissler & Haukka, 2004). In such an infrastructure, the various administrative register files can be linked cross-sectionally at the unit level and also in a panel fashion. The combination of the administrative data sources into integrated statistical registers at the unit level is based on unique identifiers such as personal identification numbers. In many cases, records from the register databases can be linked with the original study data records at the unit level, giving much flexibility in the combined use of the various data sources. Jousilahti, Salomaa, Kuulasmaa, Niemelä & Vartiainen (2005) provides an example of data linkage and the use of combined information in examining drop-out and attrition structures in a health study conducted in a register-based data infrastructure. Fortunately, in many countries such data infrastructures are becoming accessible for scientific research and public statistics purposes.

Conclusion

From the statistical methodology perspective, the dichotomy between “scientific inference” and “population inference” is restrictive and prevents full utilisation of the potential of modern statistical apparatus and today’s emerging data infrastructures. Alongside relaxing this dichotomy, the confrontation of representativeness and purposefulness becomes unnecessary and can be dropped from the researcher’s terminology toolbox. It will also be necessary to introduce up-to-date materials in university courses in epidemiology on such topics as sampling and data integration and statistical record linkage techniques as well as analysis methods for complex study data. I agree with Goldstein’s comment on the importance of

access to suitable statistical software in exploiting a combined study design.

Goldstein seems to neglect somewhat the potential of probability sampling as an important phase of the research process but I think that an obituary for probability sampling is premature.

References

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika 91*, 893–912. http://dx.doi.org/10.1093/biomet/91.4.893

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science 22*, 153–164. http://dx.doi.org/10.1214/088342306000000691

Gissler, M. & Haukka, J. (2004). Finnish health and social welfare registers in epidemiological research. *Norsk Epidemiologi 14*, 113–120.

Jousilahti, P., Salomaa, V., Kuulasmaa, K., Niemelä, M., & Vartiainen, E. (2005). Total and cause specific mortality among participants and non-participants of population based health surveys: a comprehensive follow up of 54 372 Finnish men and women. *Journal of Epidemiology & Community Health 59,* 10–31. http://dx.doi.org/10.1136/jech.2004.024349

Langel, M. & Tillé, Y. (2011). Corrado Gini, a pioneer in balanced sampling and inequality theory. METRON - *International Journal of Statistics LXIX*, 45–65. http://dx.doi.org/10.1007/bf03263549

Li, Y., Li, Z. & Graubard, B.I. (2011). Testing for Hardy Weinberg equilibrium in national household surveys that collect family-based genetic data. *Annals of Human Genetics 75*, 732–741. http://dx.doi.org/10.1111/j.1469-1809.2011.00680.x

Rothman, K.J., Gallacher, J.E.J. & Hatch, E.E. (2013a). Why representativeness should be avoided. *International Journal of Epidemiology 42,* 1012–1014. http://dx.doi.org/10.1093/ije/dys223

Rothman, K.J., Gallacher, J.E.J. & Hatch, E.E. (2013b). When it comes to scientific inference, sometimes a cigar is just a cigar. *International Journal of Epidemiology 42,* 1026–1028. http://dx.doi.org/10.1093/ije/dyt124

Rothman, K.J, Greenland, S. & Lash, T.L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology 33*, 99–119.

Thompson, M.E. (2015). Using Longitudinal Complex Survey Data. *Annual Review of Statistics and Its Application 2*, 305–320. http://dx.doi.org/10.1146/annurev-statistics-010814-020403

Population sampling in longitudinal surveys: a response from Harvey Goldstein

Harvey Goldstein University College London and University of Bristol, UK
h.goldstein@bristol.ac.uk

I am very pleased that my original piece has stimulated an excellent set of thoughtful responses. Reading through these has given me greater insight into the issues and also persuaded me to clarify some of my views, especially on the role of scientific inference. I shall begin by reflecting on the terms that I used since I think that there may be some misunderstanding of my intentions, no doubt through insufficient elaboration originally on my part. I welcome the opportunity to provide elaboration and am grateful to all the contributors for their responses.

I use the term 'real' population, in the same sense as Kish (1965, chapter 1) to mean a finite set of units that, at least in principle, can be enumerated. The intention behind the use of the term 'purposive' sampling is to reflect sampling from a theoretically defined frame of reference but one that does not necessarily correspond to such a population. Thus, the pregnancy component of Life Study is well defined as all the pregnant women attending a set of maternity units. The sample is chosen as those attending over a given time period of four years. Here, the concept of a 'superpopulation' is a key one, namely that any scientific inference that is based upon a sample chosen at a particular time is intended to apply more generally across a time period, a point elaborated by Peter Lynn. We may also wish any inferences we make to apply across space, and both these concerns need to be addressed within a standard scientific framework as I elaborate below.

Such a sample indeed might consist of all the members of a 'real' population, such as the set of children attending primary school in England in year three of their education. Yet scientific inferences about, for example, the relationships between year three childrens' school performance and background factors such as ethnic group, need to postulate a superpopulation model, and we would apply basically the same modelling procedures whether the full population of year three children or a random sample from it, i.e. with known probability of selection, had been chosen. In Life Study, the chosen women would not conventionally be regarded as constituting a probabilistically selected sample from a real geographic population in terms of a fixed time period and by where they live (rather than where they attend), especially as the criteria for being able to attend may change over time, for example in terms of residence or referrals. Nevertheless, for scientific inference purposes, given suitable statistical adjustments, for example to correct for selection biases, we may apply our standard statistical modelling procedures where we attempt to make inferences conditional on individual characteristics such as ethnic origin etc., and we can see that the distinction between a purposive sample and one derived probabilistically from a real population becomes less clear and certainly less important. Let me be clear also that I certainly do not use the term 'purposive', in one of the senses discussed by Risto Lehtonen, namely as a sample that has become biased through selective non-response. As he mentions, an interesting example of purposive sampling is quota sampling where sample members are selected for certain characteristics they happen to possess. Of course, this is not based upon a clear probabilistic mechanism, *but if we are prepared to assume* that the selection process has not differentially sampled individuals who have other characteristics that mediate the relationships of interest, then we will be justified in applying our models to study the relationships of interest. One task for the data analyst is to try to satisfy such an assumption.

The key idea is that it is the underlying social and biological processes that produce an actual set of individuals, that are the real objects of inference, and we are making use of the biological and social realisation of these at a particular historical time to select a sample from which we may make

inferences. Unless we make an assumption of this kind, what we describe is, strictly logically, only of historical interest, although of course, for the purpose of enumeration or, say, resource allocation, this may be appropriate. Furthermore, of course, we also need to assume that the process that generates the actual data is essentially probabilistic in order to make inferences about the parameters in our statistical models, and in addition that we have data that allows us to adjust for factors such as differential non-response that could otherwise lead to biases.

Thus, the Life Study maternity component samples all women over a period attending the maternity units, but it is the superpopulation that 'generates' this group that is of scientific interest and that, conditional on observables, the generation process is assumed random. I accept that there is a vagueness here that contrasts with the strictly defined procedures of the conventional survey framework for selecting probability samples from actual 'real' populations, but it seems to me that we need to accept this in the spirit that whatever inferences we come up with are subject to the strict scientific tests of replication and falsification. These tests are what I was trying to illustrate in discussing the studies of pregnancy smoking and mortality. Thus, I concur with Colm O'Muircheartaigh's remarks about the importance of samples that have a probabilistic basis, since this is fundamental to statistical modelling, but I also contend that such a probabilistic basis is consistent with a superpopulation approach. The points made by Graciela Muniz and Rebecca Hardy about the importance of replication and generalisation are helpful here. I hope that my original intentions may now be clearer, especially in the light of Peter Lynn's useful discussion of real and super - population definitions.

I think my original use of the term 'real population' and a 'purposive' sample may have led to some misunderstandings. Thus Colm O'Muircheartaigh points out that had we taken notice of the study across the whole of the US where there is considerable heterogeneity, rather than the private health one or the Swedish one, we could have observed the positive relationship between percentage low-birth weight and mortality ratio that I presented. Rather than undermining my point, however, that is precisely my contention in that it is the heterogeneity present in the sample rather than the fact that it allows inference to any particular geographically well-defined population, that is of key importance. I particularly welcome Risto Lehtonen's discussion of purposive sampling and how, for example by the use of properly specified inclusion probabilities, weights and covariate adjustment, such samples can be brought within a standard statistical modelling framework.

Turning again to my illustrative example of Life Study, choosing to sample from maternity units was not, as Chris Skinner suggests, based on 'geographic homogeneity', but the practical one that this was the only way to obtain high quality prenatal measurements. He is right that there will no doubt be important differences among maternity units in different parts of the country and one aim of analysis will be to explore and attempt to account for these. This is part of the scientific process of replication. In fact in the case of long term longitudinal studies, apart from the relatively small number of national cohort studies in the UK and elsewhere such as the US, Canada, the Nordic countries, France, Germany, the Netherlands, New Zealand and Australia, most are samples of small geographic regions, institutions or other restricted groups. Indeed, the 1946, 1958 and 1970 British cohort studies sampled all births in just one week, in one sense a real population, but certainly not the target population of interest. The scientific value of such studies lies not primarily in their general representativeness but in their heterogeneity, their ability to explore rich data and ultimately in the possibilities for comparison and replication. In the case of Life Study, access to the national population births register and also to local population data, containing birth and demographic variables, also allows us to post-stratify the sample and to adjust for differential non-response by conditioning on such data. It will also allow us to compute weights so that it can be used together with the parallel national probability sample in Life Study to provide efficient combined analyses, the 'borrowing strength' that Chris Skinner refers to. While he is correct that it adds relatively little *national* information, it will provide the user with a consistent and large combined dataset that contains both sample components. Thus, depending on the purpose of any particular analysis and using appropriate weights, one may certainly treat the overall sample as 'representative' of a real population (over an intended four year period), but

one may also treat it as a realisation of a superpopulation process. I do agree that such designs are non-traditional and would benefit from further study, and Risto Lehtonen's remarks under the 'hybrid model' heading provide a useful elaboration of the basic idea, and his illustration from registers constituting a data 'infrastructure' for removing sample bias in Nordic countries is interesting.

I'm grateful to Graciela Muniz and Rebecca Hardy for usefully illuminating all these issues with their discussion of cohort studies and especially how difficult the concept of representativeness of a real population becomes over time. They also elaborate on the need for replication and reproducibility and how this may be achieved, with some well-chosen examples.

Since preparing my original article, an interesting paper has been read to the Royal Statistical Society by Keiding and Louis (2015) that has a detailed exploration of many of these issues and explicitly comments on the articles by Rothman and colleagues (2013). They argue, I think correctly, that in some respects Rothman and colleagues overstate their case. Keiding and Louis particularly draw attention to the problem of informative differential non-response that can threaten the validity of any inferences, and I fully concur with this as a major issue for all types of study. They also take the view that "The real representativity issue is whether the conditional effects that we wish to transport (to other times and places) are actually transportable". This echoes my remarks about conditioning on known population data to avoid selection bias. I think that the Keiding and Louis paper, however, is less clear about the relationship between scientific inference and inferences to a well-specified population. As I pointed out in the case of the smoking in pregnancy studies, the characteristics of some populations may make them quite unsuitable for purposes of scientific explanation.

Despite remaining differences I am encouraged that there is a general agreement that these issues are useful ones to discuss and I have no doubt that there will be plenty more to say in the future.

Acknowledgements

I would like to thank the editor, John Bynner, for setting up and encouraging this debate and for providing helpful comments on various drafts.

References

Kish, L. (1965). *Survey Sampling*. Wiley: New York

Keiding, N. & Louis, T.A. (2015). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society*, series A. To appear (with discussion) Retrieved from http://www.rss.org.uk/Images/PDF/publications/rss-preprint-keiding-august-2015.pdf

Referencing

The debate should be referenced as:
Goldstein, H., Lynn, P., Muniz-Terrera, G. & Hardy, R., O'Muircheartaigh, C., Skinner, C. & Lehtonen, R. (2015). Population sampling in longitudinal surveys debate. *Longitudinal and Life Course Studies, 6, 447 – 475.* http://dx.doi.org/10.14301/llcs.v6i4.345

Individual contributions may be referenced as:
[Author(s) name(s)]. [Title of contribution], in Goldstein, H., Lynn, P., Muniz-Terrera, G. & Hardy, R., O'Muircheartaigh, C., Skinner, C. & Lehtonen, R. (2015). Population sampling in longitudinal surveys debate. *Longitudinal and Life Course Studies, 6,* 447 – 475. http://dx.doi.org/10.14301/llcs.v6i4.345