Wagner, J., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G. and Ndiaye, S. K. (2012) Use of paradata in a responsive design framework to manage a field data collection. *J. Off. Statist.*, **28**, 477–499.
Weinberg, C. R. and Wilcox, A. J. (2008) Time-to-pregnancy studies. In *Modern Epidemiology*, 3rd edn, pp. 625–628. Philadelphia: Lippincott, Williams and Williams.
Weisberg, H. I. (2015) What next for randomised clinical trials? *Significance*, **12**, no. 1, 22–27.
Weiss, C. O., Segal, J. B. and Varadhan, R. (2012) Assessing the applicability of trial evidence to a target sample in the presence of heterogeneity of treatment effect. *Pharmepidem. Drug Safty*, **21**, suppl. 2, 121–129.
Wilcox, A. J. (2010) *Fertility and Pregnancy: an Epidemiologic Perspective.* New York: Oxford University Press.
Wirth, K. E. and Tchetgen Tchetgen, E. J. (2014) Accounting for selection bias in association studies with complex survey data. *Epidemiology*, **25**, 444–453.
Wise, I. A., Mikkelsen, E. M., Rothman, K. J., Riis, R. H., Sørensen, H. T., Huybrechts, K. F. and Hatch, E. E. A. (2011) A prospective cohort study of menstrual characteristics and time to pregnancy. *Am. J. Epidem.*, **174**, 701–709.
Wise, L. A., Rothman, K. J., Mikkelsen, E. M., Sørensen, H. T., Riis, A. and Hatch, E. E. (2010) An internet-based prospective study of body size and time-to-pregnancy. *Hum. Reprodn*, **25**, 253–264.
Wise, L. A., Rothman, K. J., Mikkelsen, E. M., Sørensen, H. T., Riis, A. H. and Hatch, E. E. (2012) A prospective cohort study of physical activity and time to pregnancy. *Fertil. Steril.*, **97**, 1136–1142.
Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A. and Wang, R. (2011) Comparing the accuracy of RDD telephone surveys and Internet surveys conducted with probability and non-probability samples. *Publ. Opin. Q.*, **75**, 709–747.
Zhang, X., Holt, J. B., Lu, H., Wheaton, A. G., Ford, E. S., Greenlund, K. J. and Croft, J. B. (2014) Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *Am. J. Epidem.*, **179**, 1025–1033.
Zukin, C. (2015) What's the matter with polling? *New York Times*, June 21st, SR1. (Available from `http://nyti.ms/1H00TPy`.)

Discussion on the paper by Keiding and Louis

Miguel A. Hernán (*Harvard University, Boston*)

Keiding and Louis wisely appeal for statisticians and other investigators to join forces. Together they can better address the methodologic problems that are raised by the selection of individuals in surveys, epidemiological studies and other research endeavours involving human subjects. A sensible first step to combine the knowledge accumulated by different disciplines is to develop a common framework for the study of selection biases. Here, besides a vote of thanks to Keiding and Louis, I also propose an outline for that framework.

Fig. 1 shows three levels of selection in human studies:

(a) from humankind to target population,
(b) from target population to target sample and
(c) from target sample to actual sample.

The first two levels are under the investigators' control: investigators decide their targets. The third level is not: whether and when individuals participate in human studies is influenced by their own decisions and by other factors. Considering each selection level separately helps to categorize disagreements between investigators and to determine which disagreements are statistical.

First, the selection of the target population is determined by the (scientific, policy) question at hand. Investigators use their expert knowledge to specify both the parameter of interest and the eligibility criteria that characterize the target population. For example, they may want to describe the mean time to pregnancy among nulliparous women who tried to conceive in Denmark between 2000 and 2010. At this selection level, there may be subject matter disagreements about the relevance of the target population, but there is no selection bias.

Second, the selection of the target sample is guided by the principle that the parameter estimate should be unbiased for the parameter in the target population. No bias is expected under random sampling from the target population, but random sampling is often impractical. Thus investigators use their expert knowledge to define a non-random sampling procedure that, they believe, will result in no bias, i.e. no bias is expected by the investigators under non-random sampling (otherwise they would have chosen a different sampling procedure). In our example, if the target sample is nulliparous women who tried to conceive in Denmark between 2000 and 2010 *and who had Internet access*, the investigators are assuming that the average time to pregnancy is approximately equal among women with and without Internet access. As Keiding and Louis illustrate, disagreements about the sampling procedure lead to claims of selection bias, but again the discussion typically revolves around subject matter, not statistical, issues.

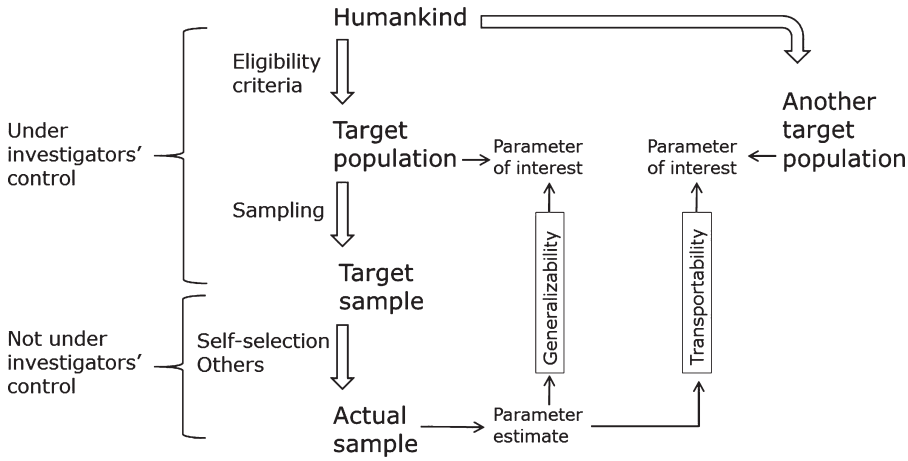

Fig. 1. Framework for selection in human studies

Third, the actual sample may differ from the target sample because of a number of selection processes that were not intended by the investigators. For example, younger women may be more likely to answer a questionnaire. If age affects the probability of conception, the average time to pregnancy in the actual sample differs from the time that would have been observed in the target sample. Because the difference is partly due to systematic differences between older and younger women, we say that there is selection bias. At this level of selection we may find disagreements about

(a) the characterization of the factors that, like age in our example, cause bias and
(b) the optimal procedure to adjust for those factors.

Disagreements of type (a) result from variations in expert knowledge and beliefs among investigators. These are again subject matter, not statistical, disagreements. Disagreements of type (b) can be genuinely labelled as statistical.

The conceptual framework that is depicted in Fig. 1 may facilitate the interdisciplinary conversations about selection bias that Keiding and Louis encourage. Take the concepts of generalizability and transportability, which are sometimes used to denote lack of selection bias. One possible interpretation of generalizability is an unbiased parameter estimate in the actual sample for the parameter in the target population; one possible interpretation of transportability is an unbiased parameter estimate in the actual sample for the parameter in *another* target population.

This discussion has focused on studies with a descriptive aim: estimation of a functional of the distribution of some variable(s) in a target population, e.g. the mean time to pregnancy. In descriptive studies, selection bias is a synonym for lack of external validity. Keiding and Louis consider also studies with a causal aim: estimation of the comparative effect of different courses of action on the distribution of some variable(s) in a target population, e.g. the effect of cigarette smoking on the mean time to pregnancy. In comparative studies, selection bias may mean either lack of external validity or lack of internal validity, and it may arise even when defining the target population (if a collider is incorrectly used as an eligibility criterion).

As Keiding and Louis vehemently argue, it is time to overcome the barriers that have traditionally impeded a cross-disciplinary understanding of selection bias. A framework that explicitly references the level of selection and the aims of the study weakens those barriers. I propose a vote of thanks to Keiding and Louis.

Peter V. Miller (*US Census Bureau, Washington DC*) © US Government

I am grateful for the opportunity to comment on this interesting and important paper. In it, Professor Keiding and Professor Louis provide a thoughtful discussion of the problem of generalization, or external validity, in epidemiology and in survey research. They urge cross-fertilization of methods that are employed in each field to address this common inferential issue. I support their call for interdisciplinary research. At the same time, I want to call attention to issues that need to be addressed in such interdisciplinary collaboration.