ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2016)
179, *Part* 2, *pp.* 319–376

Perils and potentials of self-selected entry to epidemiological studies and surveys

Niels Keiding

University of Copenhagen, Denmark

and Thomas A. Louis

Johns Hopkins Bloomberg School of Public Health, Baltimore USA

[Read before The Royal Statistical Society *at the 2015 Joint Statistical Meetings of the American Statistical Association in Seattle on Wednesday, August 12th, 2015, the President*, Professor P. J. Diggle, *in the Chair*]

Summary. Low front-end cost and rapid accrual make Web-based surveys and enrolment in studies attractive, but participants are often self-selected with little reference to a well-defined study base. Of course, high quality studies must be internally valid (validity of inferences for the sample at hand), but Web-based enrolment reactivates discussion of external validity (generalization of within-study inferences to a target population or context) in epidemiology and clinical trials. Survey research relies on a representative sample produced by a sampling frame, prespecified sampling process and weighting that maps results to an intended population. In contrast, recent analytical epidemiology has shifted the focus away from survey-type representativity to internal validity in the sample. Against this background, it is a good time for statisticians to take stock of our role and position regarding surveys, observational research in epidemiology and clinical studies. The central issue is whether conditional effects in the sample (the study population) may be transported to desired target populations. Success depends on compatibility of causal structures in study and target populations, and will require subject matter considerations in each concrete case. Statisticians, epidemiologists and survey researchers should work together to increase understanding of these challenges and to develop improved tools to handle them.

Keywords: External validity; Internal validity, Non-probability samples; Representativity; Transportability; Unmeasured confounders; Web-based enrolment

1. Introduction

Participation and response rates in follow-up studies and surveys are decreasing; compliance can be low; costs are increasing. Low front-end cost and relatively rapid accrual make Web-based, self-selected Web enrolment into epidemiological studies and into surveys very attractive and its use rapidly increases. However, self-selection dramatically departs from the traditional, so-called *gold standard* approaches of targeted enrolment to scientific studies and sampling-frame-based surveys. Traditionalists argue that we must adhere to the values of planned accrual and follow-up for all studies and identification of a sampling frame for surveys and possibly also for epidemiological and other such studies. Others propose that we should stop worrying about it and open up accrual, using modern approaches (covariate adjustments, find instrumental variables, 'big data',...) to make the necessary adjustments.

Address for correspondence: Niels Keiding, Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, POB 2099, Copenhagen K DK-1014, Denmark.
E-mail: nike@sund.ku.dk

 0964–1998/16/179319

It is useful to begin by outlining some differences in scientific terminology and practice between classical statistical analysis, survey methodology and epidemiology. The concepts of *internal* and *external* validity are key (see Shadish *et al.* (2002) for formal definitions). Internal validity refers to validity of inferences for a given parameter or estimand (such as a sample mean) for the sample at hand. External validity refers to the degree to which within-study inferences generalize or can be generalized to a target population or context.

In survey research the basic task is to learn about a population by designed sampling so that properties learnt by statistical analysis of measurements in the sample may be generalized to the population. An important focus here is the representativity properties of the sample and the consequent methodology for the generalization, e.g. weighting methods. Classical statistical analysis such as developed by R. A. Fisher takes a similar approach. Here the population is represented by the statistical model and the sample by the observations, and the statistical inference specifies the properties of the generalization from sample to population.

The next obvious question (with a long historical tradition; see Keiding (1987) and Keiding and Clayton (2014)) is whether the findings in the study population are also valid in other populations. In important recent development Pearl and Bareinboim (2014) provided exact criteria for *transportability* based on causal graphs. A general empirical observation is that marginal effects are rarely transportable whereas conditional effects more often can be expected to be transportable, provided that all relevant confounders have been accommodated.

The field of epidemiology does contain descriptive studies much like surveys, a central example being prevalence studies, where it is desired to learn about the distribution of individuals with some disease in a given population on the basis of some sample from that population. However, the major task in analytical epidemiology is to assess the possible effect of some exposure (e.g. air pollution) on some outcome (e.g. lung cancer), and here the general analytical strategy is slightly but importantly different from the standard in survey analysis and statistics. The first priority is to obtain validity of the inferences in the study group, i.e. internal validity. The statistical analysis takes place in this study group, which thus plays the role of the sample in survey analysis and general statistical inference. Threats to internal validity include selection bias (generated from biased (exposure- and/or outcome-dependent) selection of subjects into the study group), not always with a clear specification of the origin from which this selection takes place. One type of selection bias is self-selection (which is our focus), which is generated if 'reasons for self-referral may be associated with the outcome under study' (Rothman *et al.* (2008), page 134).

Assessment of external validity, i.e. generalization to the population from which the study subjects originated or to other populations, will in principle proceed via formulation of abstract laws of nature similar to physical laws, whereas sampling properties (as in survey analysis) or statistical properties are considered less relevant (Rothman *et al.* (2008), pages 146–147). As elaborated below, this view was forcefully formulated by Miettinen (1985), and we shall also mention a recent discussion in the *International Journal of Epidemiology* introduced by Rothman *et al.* (2013a), which almost unanimously claimed that 'Representativity should be avoided'.

More generally, experimental studies in human populations have seldom directly addressed identification of a reference population other than what is implied by inclusion and exclusion criteria. The primary focus is on internal validity conferred by randomization (and careful conduct), without formal identification of a sampling frame or collecting baseline information that could be used to develop a weighted analysis that would allow 'exporting' the internal associations (e.g. treatment effects) to an identified population. Implicit in this approach, and as discussed below made explicit by Miettinen (1985), is the assumption that interactions with demographic attributes (e.g. gender, race or age) and other study features (measurement

methods, follow-up protocols,. . .) are sufficiently small relative to the main effects that they can be ignored. Experiments in other domains, e.g. in agriculture, pay more attention to an external reference, possibly because it is more broadly accepted that outcomes can substantially depend on plant species, soil type, fertilization, temperature and many other factors.

In the survey context, government and other high quality surveys have depended on a well-documented sampling frame, a carefully designed sampling process and weighting to ensure that results are relevant to the reference population. Though developing a sampling frame and purposefully sampling from it, or at minimum identifying a target population and developing explicit entry and exclusion rules, may be the gold standards, in practice these are difficult to accomplish. The accrued sample is never fully representative; item non-response and dropouts in longitudinal surveys make most surveys 'tarnished gold'. Partial fixes are available via covariate adjustment, reweighting and the like but, as we discuss in Section 5.3, these are very unlikely to be completely successful, moving virtually all studies to the middle ground between the gold standard and a completely haphazard enterprise. Consequently, rather than discount all self-enrolment or other departures from the ideal epidemiological study or sample survey, careful evaluations are needed to see whether, and if so when, these studies maintain sufficient quality as measured against the real world performance of studies that attempt the gold standard.

The focus in many studies is on within-study comparisons, e.g. estimation of an exposure–response relationship or the association of an attitude with personal characteristics. However, if the relationship depends on individual attributes their role must be properly modelled, or weights used in a weighted analysis must accurately account for sample inclusion propensities (the unit-specific probabilities of being in the study) to make sample selection non-informative and thereby align the estimated relationship with the population value. Even with a sampling frame and valid weights for making inferences to the frame, validity of inferences to other populations is at risk because the conditional effects might not generalize. Unmeasured confounders that are associated with sample inclusion propensities cannot be incorporated in the weights and, even if all confounders are measured, it is challenging to develop appropriate weights. We consider these issues in more detail in Sections 7.2 and 7.4.2.

We compare and contrast the epidemiological and survey cultures, considering only well-intended investigators and investigations. In this context, both cultures require well-defined goals; design, conduct and analysis to meet those goals, with the principal distinction being the epidemiological focus on internal and the survey focus on external validity. Against this background we discuss recruitment or accession methods and their association with study quality including bias and precision. We focus on studies of human populations, but many of the issues are relevant to studies in agriculture, ecology and other fields. We review recent research, identify issues and research needs, discuss examples of epidemiological studies and surveys, report on some methodological innovations and speculate on the future.

Section 2 gives an overview of the epidemiological context, Section 3 considers transportability and Section 4 the survey context, Section 5 addresses self-selection and Web-based studies, Section 6 briefly mentions the possibility of selection effects induced by requiring informed consent, Section 7 considers overarching inferential goals, Section 8 discusses convergence of the epidemiological and survey cultures, and Section 9 provides a summary of issues and a call to action. Some details on the survey method are in Appendix A.

2. The epidemiological context

Participation rates in epidemiologic studies have been declining for the past 30 years, and the

decline is accelerating (see the literature review by Galea and Tracy (2007)). This situation has stimulated two questions.

(a) How much does this matter for study validity?
(b) As the Internet approaches universal coverage, are competitive Web-based study designs emerging?

Key issues include external validity, representation and transportability, as the following case-studies help to clarify.

2.1. Case-study: the Danish Web-based pregnancy planning study—'SnartGravid'

The primary purpose of time-to-pregnancy (TTP) surveys is to estimate fecundability (defined as the probability that a couple with unprotected intercourse will conceive in a given menstrual cycle), in an attempt to approach biological fecundity (the ability to obtain a pregnancy) in humans (Wilcox, 2010; Weinberg and Wilcox, 2008). The ideal prospective TTP survey would recruit couples at the time (initiation) that they decide to try to become pregnant and follow them prospectively until pregnancy happens, the couple gives up, or the study ends. Such studies are rare and very costly, and have low participation rates and usually rather uncertain representativity status, if it is at all possible to identify a study base (Buck Louis *et al.*, 2011). There are other designs for TTP studies, but they give less direct results; see Keiding *et al.* (2012).

On this background it made much sense to attempt a new way of creating a prospective sample of pregnancy seekers, using the Internet not only for follow-up, but also for recruiting. This was initiated in Denmark in 2007 by a collaborative group of researchers from Boston University, USA, and Aarhus University, Denmark (Mikkelsen *et al.*, 2009). Recruitment was via on-line advertisements, primarily on non-commercial health sites and social networks, supplemented by press releases, blogs, posters and word of mouth. By June 1st, 2014, more than 8500 women had been recruited (E. M. Mikkelsen, personal communication). Women were recruited shortly after initiation and followed until whatever comes first of pregnancy, giving up trying or 12 cycles after initiation. Follow-up rates were satisfactory, with more than 85% responding to each questionnaire and more than 80% of the cohort still included in the follow-up after 1 year (Huybrechts *et al.*, 2010). Relevant exposures which could all be measured before the end of the attempt of conception included, among other, body size (Wise *et al.*, 2010), menstrual characteristics (Wise *et al.*, 2011), consumption of caffeine, soda etc. (Hatch *et al.*, 2012), physical activity (Wise *et al.*, 2012), age and volitional factors (Rothman *et al.*, 2013) and oral contraceptives (Mikkelsen *et al.*, 2013). Using appropriate delayed entry survival analysis, this study should deliver directly interpretable estimates of the (prospective) distribution of TTP for given premeasured exposures, all within the sample of study participants. (Participants were censored at the start of fertility treatment; whether that may be considered independent censoring is never discussed, but that is not a central issue here).

Huybrechts *et al.* (2010) gave a detailed discussion of the representativity issue in self-selected Internet-based studies like this. They acknowledged that

> 'Internet-based recruitment of volunteers has raised concerns among critics because the demographics (e.g., age, socio-economic status) of those with ready internet access differ from those without it. Furthermore, among those with internet access, those who choose to volunteer for studies may differ considerably in lifestyle and health from those who decline.'

But they went on to state that

> 'Volunteering to be studied via the Internet does not, however, introduce concerns about validity beyond those already present in other studies using volunteers. Differences between study participants and non-

> participants do not affect the validity of internal comparisons within a cohort study of volunteers, which is the main concern. Given internal validity, the only problems with studying Internet users would occur if the biologic relations that we are studying differed between Internet users and non-users, a possibility that seems unlikely. The primary concern should therefore be to select study groups for homogeneity with respect to important confounders, for highly cooperative behavior, and for availability of accurate information, rather than attempt to be representative of a natural population.
>
> 'Scientific generalization of valid estimates of effect (i.e., external validity) does not require representativeness of the study population in a survey-sampling sense either. Despite differences between volunteers and non-participants, volunteer cohorts are often as satisfactory for scientific generalization as demographically representative cohorts, because of the nature of the questions that epidemiologists study. The relevant issue is whether the factors that distinguish studied groups from other groups somehow modify the effect in question.'

We note that this text quite precisely illustrates several points about the epidemiological approach to inference as made in Section 1: the 'study population' is what statisticians would call the sample, and the results from analysis of this study population should be directly used to create the abstract laws to be used for generalization, without requiring that it represents all Danish women starting pregnancy attempts. The only exception is if the process (called sampling by statisticians) leading to the study population contains effect modifiers: in other words, if the creation of the study population has generated selection bias. Here, the statement that it is unlikely that Internet users would not have similar biologic relations to those of non-users is imprecise: the question is rather whether volunteers (who here must be Internet users but are not necessarily a representative subset of these) have different such relations from those of non-volunteers.

Rothman *et al*. (2013) studied the age-related decline in fecundability. Here, the search for possible selection bias in the volunteer study group included the permitted delayed entry of up to 3 months (possibly excluding fast conceivers) and underrepresentation of women in the older age groups:

> 'Other factors, such as reproductive history, could have affected participation and be related to fecundability. If these factors were also associated with age, they could have distorted the estimates of fecundability ratios by age.'

This point is a good example of the issue of whether *conditional* effects are transportable, which is a principal focus of our presentation.

However, Rothman *et al*. (2013) noted a stronger decline in fecundability with age for nulliparous (never given birth) women without mentioning the important possible survivor selection generated by the earlier pregnancies of the highly fecund pointed out, for example, by Howe *et al.* (1985). And, even though the title of the paper emphasizes 'volitional determinants', there is no discussion on using self-selected participants to study such 'subjective' determinants. See Section 7.3 regarding a much more focused interest by Rothman *et al*. (2013b) in controlling for differential health awareness by using a sampling frame.

As a postscript, we have been informed through personal communication with E. M. Mikkelsen that this group of researchers has recently initiated a general check of representativity of main parameters against register data for all births in Denmark.

2.2. Case-study: smoking and time to pregnancy—a classic

It is interesting to contrast the current formulation of analytical epidemiology as exemplified above to the preliminary communication in the *Journal of the American Medical Association* by Baird and Wilcox (1985) about a pregnancy-based, self-selected study of the possible importance of smoking for TTP. Informed in pregnancy classes, through posters etc., pregnant women were

encouraged to volunteer for a 15-min telephone interview if they had stopped using birth control to become pregnant and had taken no more than 2 years to conceive. After exclusions 678 women were left for analysis. The study indicated a clear effect of cigarette smoking, delaying TTP.

Baird and Wilcox (1985) did not take the self-selection issue lightly:

> '. . . volunteers were generally affluent and educated. These characteristics of the study design and study population raise questions about the generalizability of the findings. Of primary concern is any source of bias that might result in finding an association in our study population even if no true association exists in the general population.'

They went on to perform a revealing sensitivity analysis, long before such analyses were common, on the question of differential occurrence of accidental pregnancies among smokers and non-smokers. As they explained, this might artificially generate an apparent delaying effect of smoking through differential survivor selection. The sensitivity analysis made such an artefact unlikely. We note that the *SnartGravid* researchers in their recent study on smoking and TTP (Radin *et al.*, 2014) incorporated the sensitivity analysis by Baird and Wilcox (1995) in the discussion.

2.3. Case-study: Web capture of acute respiratory and gastrointestinal infections

An important motivation behind the *SnartGravid* study was that, since it is distinctively difficult to obtain reliable epidemiological information on TTP, the innovative Web-based design should be tested. The pilot study by Mall *et al.* (2014) provided another example where a Web-based design may outperform conventional data acquisition. They used the Internet to collect information on symptoms, number and intensities of the usually relatively harmless episodes of acute respiratory and gastrointestinal infections. This was done within the framework of a large prospective cohort (the German National Cohort) and participants received weekly e-mails asking about new episodes and their symptoms. Mall *et al.* (2014) noted that

> 'Participants of the Web-based study were slightly younger and better educated than non-participants, so selection bias is possible and must be kept in mind when discussing generalizability of the results.'

2.4. External validity

Most of the literature on external validity takes a concrete empirical view. Thus, Galea and Tracy (2007) in their literature review concluded that non-participation bias does not alone indicate a high level of bias in the estimates of effects of exposures, since

> 'It is the difference between participants and non-participants that determines the amount of bias present. Reassuringly, most studies have found little evidence for substantial bias as a result of non-participation.'

This point turns out to be central in what follows: there will be bias only if the participation rate and effect interact. Whereas two recent empirical studies of non-participation bias (heavily quoted in the *SnartGravid* project) by Nilsen *et al.* (2009) and Nohr *et al.* (2006) found that non-participation at the study outset had little influence on effect estimates, two other recent empirical studies of non-participation bias were less optimistic: Nummela *et al.* (2011) showed that a health study in Finland among aging people had differential response rates as well as differential health outcomes according to socio-economic factors, precisely generating bias, and Langhammer *et al.* (2012) documented in a Norwegian health study that participation was associated with survival and depended on socio-economic status, although the precise effects of these associations on final effect estimates would depend on further disease-specific studies. It is

noteworthy that all these validation studies originate from the Nordic countries, where individual linkage of information in population registers allows unusually detailed empirical evidence.

We note that, in addition to the above more analytic studies, Galea and Tracy (2007) added as their second concern the problems for epidemiological studies that use population-based sampling and with attempts to obtain estimates from population-representative samples that are generalizable to a clearly defined reference population. As mentioned in Section 1, such *prevalence* studies are important parts of epidemiology and should not be ignored; the issues that are connected to them are, however, quite similar to many surveys and will be described further below.

3. Transportability

Most empirical science is about generalizing findings beyond the particular setting. In our view a useful framework for discussing generalization (external validity) in epidemiology is that the purpose of epidemiological studies is to obtain information on exposure–outcome relationships in one (standard) population with the aim of transporting them to other (target) populations. Classical standardization of mortality rates focused on avoiding confounding from different age structures in standard and target populations; see Keiding (1987) and Keiding and Clayton (2014) for surveys of this development. Considering a single-sex stratum, by reweighting the age×exposure-specific mortality rates in the target population with the age distribution in the standard population (direct standardization), a direct comparison could be made between the total (i.e. average) exposure-specific mortality in standard and target populations. Validity depended, of course, on an implicit assumption that age × exposure-specific mortality rates were transportable between these populations, although this assumption has traditionally been discussed surprisingly little (see footnote 12 in Pearl and Bareinboim (2014)). If the relationship between age × exposure and mortality is modified by a third factor (e.g. a mortal disease) differently in the two populations, transportability breaks down unless there is also control for this third factor.

Pearl and Bareinboim (2014) initiated a systematic attempt at developing a theory of transportability within Pearl's framework for causal inference based on the counterfactual approach and using directed acyclic graphs as an important tool. The aim was to formalize the kind of scientific knowledge about the causal structures in the standard and target populations that is required for assessing whether transportability is possible and, if so, concrete mathematical formulae for how the knowledge that is obtained in the standard population may be transported to the target or that transporting is not possible. Indirect adjustment is the most basic case, and other computations are far less intuitive. In fact, generally a recursive algorithm is needed to develop a valid mapping. In Pearl and Bareinboim (2014) the starting knowledge was assumed to have been obtained from a randomized study, but the authors also have work under way where an observational study forms the basis of the knowledge in the standard population.

Finally, we note that sometimes a comparison that is not transportable in one scale is transportable in another. For example, the relative risk may be transportable, when the difference in risk is not. However, if the difference in risk is the relevant parameter, then transportability of the relative risk, although of scientific interest, is essentially irrelevant for policy, and extrapolation must be based on the difference in risk, ideally using Pearl and Bareinboim's (2014) technology. A statistical model for the log(relative risk) can reduce the need for interaction terms and produce a parsimonious structure, but it must then be converted to a model for the difference in risk.

3.1. The Miettinen declarative position

As we have indicated, an alternative view on external validity was formulated forcefully by

Miettinen (1985), page 47:

> 'In science the generalization from the actual study experience is not made to a population of which the study experience is a sample in a technical sense of probability sampling. In science the generalization is from the actual study experience to the abstract, with no referent in place or time.'

This has been followed up by the paraphrases of this statement in successive editions of *Modern Epidemiology*: see Rothman (1986), page 95, Rothman and Greenland (1998), pages 133–134, and Rothman *et al.* (2008), pages 146–147. This position stipulates that epidemiology is a science that is much elevated above statistics and more specifically survey design and analysis. Rigid support of this position implies that measurement systems are stable and accurate, that responses or outcomes are recorded accurately and reliably, and that the measurement error process is constant across clinical, demographic, chronological and technological contexts. In the *SnartGravid* case-study that was mentioned in Section 2.1, this declarative position comes close to questioning the relevance of empirical studies of external validity. However, a closer look shows that the authors do address representativity of the study sample, although phrased in terms of absence of selection bias in the internal analysis of the study group.

Importantly, these and other researchers appear to equate 'representative' with 'self-weighting' (i.e. summaries computed by using equal weights for each unit produce unbiased estimates of population values). The more general and more generally accepted definition requires only that appropriate weights are available to produce valid population estimates. Very few surveys could be conducted if they needed to be self-weighting, and most well-designed and well-executed surveys have a sampling frame and sampling plan that support weighting results so that they apply to a reference population. Availability of the relevant weights qualifies the sample as representative. Such representation is a worthy goal for surveys and epidemiological, clinical and other studies that intend to produce findings that are generalizable (henceforth we use 'epidemiological' as a shorthand for all explanatory studies).

We shall return to a general discussion of Miettinen's influential statement, but we emphasize here that Miettinen apparently equated targets of most epidemiological research with physical laws, for which the generalizability issue is not a question of representativity of the sender population (e.g. from which relationships are to be exported) and receiver population (e.g. into which relationships are to be imported). In our view, however, many epidemiological efforts have rather more modest and practical concrete targets, which are expressed not as new general physical laws, but as properties of some but not all human populations. There is therefore a genuine generalizability problem in most epidemiological studies.

4. The survey context

In this section we address standard survey methods, the need to accommodate departures from the ideal and the consequences of departures. In Appendix A we work through a basic example of estimating a population mean and identify links to epidemiological analyses.

4.1. The classical sample survey

Simplifying to a considerable degree, the classical sample survey identifies a sampling frame based on attributes of the reference population, develops a sampling plan that efficiently and effectively achieves study goals and in the analysis uses inverse propensity weights derived from the sampling plan to make inference to the reference population. The weights and propensities need to be modified to reflect refusals and item non-response but, if the adjustments are valid, the analysis will deliver unbiased or at least consistent estimates. Furthermore, in addition to

inference to the current survey frame or population, survey goals can include transportation to other frames or populations. So, even when the ideal can be achieved for a specific reference population, confounding and effect modification must be taken into account for inferences to populations other than the initial referent.

Conducting a gold standard survey with a well-developed sampling frame and sampling plan, a close to 100% response rate, a nearly zero attrition rate and no missing data is a worthy, but unattainable, goal. At the other extreme, a Web-based survey with self-enrolment may give the appearance of a high participation rate, but without an identified reference population there is no way to estimate the rate or coverage of the survey. The true participation rate and population coverage are unavailable and, unless additional information is available, findings pertain only to the actual participants, which is somewhat analogous to internal comparisons in a clinical or epidemiological study. Some might argue that a sampling-frame-based survey with a low response rate is no better than a self-selected sample, in that the respondents in both contexts are self-selected. That is of course true, but the sampling frame provides information on the relationship of respondents to non-respondents and generalization from the sample is possible.

4.2. Survey analysis

Traditional survey analysis is design based, with the population values considered fixed constants and the sample inclusion indicators being the random variables (see for example Särndal *et al.* (1992) and Särndal (2007)). Expectations, standard errors and other features are computed in this framework, with the sampling design providing the inferential basis. The sampling plan is centrally important and, if sample inclusion is informative (associated with attributes of interest), then failure to accommodate it will produce fundamental bias. The advantage of design-based inference is that it is model free, producing valid inferences if the (at least pairwise) sample inclusion probabilities are known. The sample does not need to be 'self-weighted', which is the restrictive definition of representativeness that was communicated by Rothman *et al*. (2013a). It distracts from the central point that, with a known sampling plan, appropriate weights can be constructed and valid inferences are available (see the classic Horvitz and Thompson (1952)).

5. Self-selected and Web-based studies

Faced with the declining rates of participation in traditional epidemiological studies and in surveys, and with the developments in popular use and technical possibilities of the Internet, it has become increasingly attractive to try to recruit and accommodate willing and careful respondents by meeting them right there, where they already spend much good time and energy. Such studies will often be wholly or at least mostly self-selected, and our focus here will be on the changed emphasis on the validity issues that this entails. For example, should the Miettinen (1985) declaration (see Section 3.1) motivate that researchers completely ignore the composition of their study sample? Which tools are available or should be developed to aid researchers wishing to develop self-selected Web-based explanatory studies and surveys further?

5.1. Non-probability sampling

Departures from the ideal survey can be either intended or inadvertent. Intended departures include maximizing internal precision in a mechanistic study, random-digit dialling, quota sampling and self-selected Internet accrual. Inadvertent includes a poorly constructed sampling frame, non-participation and item non-response. As reported by Battaglia (2008), there

is a wide range of sampling plans that depart at least to a degree from the gold standard, each with its potential benefits and drawbacks. Threats and benefits depend on the type of study or survey and availability of a well-documented reference population frame with sufficient covariates to develop (approximate) sampling weights or to conduct covariate adjustments. We focus on Internet surveys, because they are the survey equivalent of Web-based enrolment in epidemiological studies such as Mikkelsen *et al.* (2009). Keeter (2014) crystallized the issue:

> 'The debate over probability vs. nonprobability samples is about representation'.

Self-selection into any study, irrespective of survey or explanatory goals, threatens validity. If there is no information on the propensity for enrolment, safe inferences must be limited to those in the sample, with broadening to a reference population based on pure speculation. In all contexts, extreme care is needed to transport conditional effects, trends and other relationships for which transportability is more delicate, requiring careful design, conduct, analysis and reporting. Pearl and Bareinboim (2014) provided elementary examples showing how transportability of conditional effects depends inherently on the comparability of the compositions of the 'sender' and 'receiver' populations. For example, the ability to make valid inferences for an identifiable population is compromised and in extreme cases validity rests on the fragile assumption of relatively small effect modification by attributes of the population of interest.

5.2. *Internet surveys*

The following discussion focuses on Internet surveys but applies as well to epidemiological studies. Advantages of Internet surveys include lower cost and the ability to reach some hard-to-reach populations (and failing to reach others!). Disadvantages and challenges include difficulty in obtaining probability samples and potentially poor coverage. Regarding this, Leenheer and Scherpenzeel (2013) reported, for the Dutch Longitudinal Internet Studies for the Social Sciences, a random sample of households with Internet provided to those who do not have it, that 'older households, non-western immigrants, and one-person households are less likely to have Internet access'. However, as Internet access increases, this may become less of an issue. In any case, McCutcheon *et al.* (2014) noted that

> 'Internet surveys have emerged rapidly over the past decade or so.... *Inside Research* estimates that in 2012, the online survey business had grown from nothing a decade and a half earlier to more than $1.8 billion.... This represents 43% of all surveys in the U.S. Almost all (85%) of that growth came at the expense of traditional methods.'

Rao *et al.* (2010) and McCutcheon *et al.* (2014) discussed recruitment and other aspects of administering Internet surveys. Self-selection via the Internet can be completely unstructured (take all comers) or can filter for demographic or other attributes, but absent a reference population sampling frame even the latter approach will not provide the information that is needed to evaluate representativeness or to adjust results. Participants in a controlled trial are to some degree self-selected in that they need to agree to participate, but this is very different from taking all comers. Of course, even with a sampling frame, if the sample is far from representative, weighting adjustments will unduly inflate variance (see Gelman (2007)) and generally are not fully effective (see Chang and Krosnick (2009) and Yeager *et al.* (2011)) in part because the decision to respond may depend on unmeasured confounders. However, as we shall see in Section 5.3 probability sampling confers some protection. We propose that this protection is conferred at least in part by sampling frames constructed from demographic attributes that are the principal correlates with responses, with attributes that are not used to produce the frame having a relatively small, residual association with response. This view is supported to a degree by

indications that paradata (context information collected during a survey) although associated with response propensity are only weakly associated with responses (see Wagner *et al.* (2012)).

5.3. Recruitment effects

In the survey context there is considerable research evaluating the effects of recruitment on study outcomes. The American Association of Public Opinion Research (Baker *et al.*, 2013) discussed the issues and recent studies compared random-digit dialling and Internet surveys (Chang and Krosnick, 2009; Yeager *et al.*, 2011). Yeager *et al.* (2011) nicely laid out the issues:

> 'The probability sample surveys were consistently more accurate than the non-probability sample surveys, even after post-stratification with demographics. The non-probability sample survey measurements were much more variable in their accuracy, both across measures within a single survey and across surveys with a single measure. Post-stratification improved the overall accuracy of some of the nonprobability sample surveys but decreased the overall accuracy of others.'

> 'The present investigation suggests that the foundations of statistical sampling theory are sustained by actual data in practice. Probability samples, even ones without especially high response rates, yielded quite accurate results. In contrast, non-probability samples were not as accurate and were sometimes strikingly inaccurate, regardless of their completion rates. . . .'

> 'This is not to say that non-probability samples have no value. . . . The continued use of non-probability samples seems quite reasonable if one's goal is not to document the strength of an association in a population but rather to reject the null hypothesis that two variables are completely unrelated to each other throughout the population. . . .'

These and other researchers (including very pointedly Zukin (2015)) urged caution and noted the protection that is provided by probability-based sampling. We extend this need for caution to epidemiological studies, because self-selection has a straightforward effect on the validity of cross-sectional analyses. However, as Pearl and Bareinboim (2014) and Ebrahim and Smith (2013) showed, the effect on longitudinal trends and relationships between responses can also be substantial, and it is important to entertain designs that target the middle ground. Studies can benefit from use of the Internet and social media to attract potential participants, but that is just the first step.

5.4. Are longitudinal analyses protected?

When selection effects bias prevalences and other cross-sectional population attributes, it may still be that in follow-up studies changes over time are less vulnerable to selection effects. Strictly, this protection requires that level and change are only weakly related, possibly after adjusting for baseline attributes. However, there are many examples of strong association, e.g. the 'horse-racing' effect wherein the pace of change for an individual at the front of the pack is greater than the typical change (Enright *et al.* (2002) noted this in a lung function study). More generally, if a longitudinal relationship depends on individual attributes that either are not used in the assessment or are inadequately modelled, the estimated slope will not align with the population value, sample selection will be 'informative' and even within-study assessments can be compromised (Ebrahim and Smith, 2013). Of course, as important, dropout effects are *prima facie* associated with change. We return in Section 7.5 to the interplay between representativity, cross-sectional classification and longitudinal effects revealed in the detailed reanalyses of the Women's Health Initiative's (WHI's) results on side effects of postmenopausal hormone therapy.

5.5. Relationship to missing data

Many enrolment issues are cognate or identical to those for missing data. If sample inclusion

does not depend on attributes that associate with the attributes of interest, then the sampling process is ignorable relative to those attributes. A sampling plan that depends on measured attributes that associate with target outcomes is analogous to missingness at random, and the sampling process can be made ignorable by weighting by correct propensities, or use of a model that correctly relates these attributes or the related sampling propensities to the target outcome. However, computing these propensities depends on development of a sampling frame and an explicit sampling plan, neither of which is available for self-enrolment studies. More generally, even if propensities are available for identified attributes, if sample inclusion also depends on unmeasured attributes or in the extreme case on the target attribute, the structure is analogous to missingness not at random. In this case, validity is by no means assured.

6. Selection effects induced by informed consent

Informed consent can exert strong selection effects on study or survey participation, but the final verdict has yet to be delivered about their magnitude and type. Put simply, the consent process may be a filter that lets through a population that is different from that of the desired referent. Tu *et al.* (2004) reported strong effects:

> 'Obtaining written informed consent for participation in a stroke registry led to important selection biases, such that registry patients were not representative of the typical patient with stroke at each center. These findings highlight the need for legislation on privacy and policies permitting waivers of informed consent for minimal-risk observational research. Variation in consent process and content affects participation and therefore representation.'

However, Rothstein and Shoben (2013), their discussants and an editorial communicated a variety of views and examples that range from weak to strong selection effects associated with consent. Some commentators have faith in statistical adjustment 'cures', whereas others do not; some propose ways to reduce the bias; some argue that consent is an unnecessary filter for some types of research.

A consent process is by no means the only factor that is associated with refusal to participate. We do not extensively explore this issue but stress the importance of identifying a target population, collecting information on its attributes that potentially associate with the decision to participate and with outcome(s), and thereby be able to conduct weighting adjustments and sensitivity analyses. However, ethical requirements in obtaining informed consent may make it difficult to obtain sufficient information about non-participants to identify the sampling frame.

7. Overarching inferential goals and approaches

The foregoing discussion leads to a discussion of population-based inferential goals and methods. Epidemiological studies traditionally focus on internal validity; surveys are primarily focused on validity for a more general reference population of which the sample is only a subset. Internal and external validity are to a degree in conflict, because within-sample precision is enhanced by studying relatively homogeneous participants or animals under well-controlled conditions, whereas external validity generally requires a study sample that is more representative of the external world. At least two factors have reduced the contrast. As mentioned, pure form surveys are difficult or impossible to conduct, and policy goals encourage being able to make at least reasonable inferences for an identified population. Methodological advances in statistics (propensity weighting, double-robustness,...) and in computer science (record matching) coupled with the availability of 'big data' empower addressing population goals.

7.1. Randomization's roles

A principal role of randomization in clinical and field studies is to eliminate or to reduce confounding substantially, especially with respect to unmeasured attributes. Similarly, survey data collected by using a predetermined sampling plan confer this benefit in that weights are available to eliminate confounding and lack of representation via either a design-based or model-based analysis. Royall (1976) provided a discussion of the fundamental issues including the role of randomization as a basis for inference, as a way of balancing on unmeasured potential confounders and of protecting from unconscious bias, and as a way to assure fairness. He discussed the problems with using randomization as the basis for inference and proposed model-based alternatives, with superpopulation models as an attractive unification. His paper is altogether a rewarding read.

Epidemiological and clinical studies that purport to make generalizable conclusions need to operate at least to a degree as a survey. For example, Mumford *et al.* (2015) addressed the question how long does it take to become pregnant in the real world and not in an artificially constructed environment (an effectiveness rather than an efficacy question)? A survey with a related goal would have to generate a sample that can be mapped back to a population that represents such a world (see Keiding and Slama (2015)). With the complexities of the real world, it would be daunting at best to obtain an effective sample without reliance on random sampling (not necessarily simple random sampling) to deal with unmeasured (actually, unmeasurable) confounders.

7.2. Definition and history of representative sampling

The concept of representativeness is central to our discussion, and we refer to the treatment by Kruskal and Mosteller (1979a, b, c, 1980). In this series of four detailed papers they surveyed the use of the term 'representative sampling' in the non-scientific literature, scientific literature (excluding statistics), current statistical literature and the history of the concept in statistics, 1895–1939. 'Representative' had been taken to mean many things, more or less connected to technical meanings of the word, but always with a positive connotation. Kruskal and Mosteller (1979c) enumerated and explained nine different meanings in the statistical literature:

(a) general acclaim for data (the term representative essentially used in a positive rhetorical fashion);
(b) absence of selective forces (in the sampling process);
(c) the sample as a miniature of the population;
(d) representative as typical;
(e) coverage of the population's heterogeneity;
(f) representative sampling as a vague term that is to be made precise;
(g) representative sampling as a specific sampling method;
(h) representative sampling as permitting good estimation;
(i) representative sampling as sufficiently good for a particular purpose.

The final paper (Kruskal and Mosteller, 1980) outlined the history of representative sampling in statistics 1895–1939. The Norwegian official statistician Anders Nicolai Kiær created considerable controversy in official statistical circles, as particularly expressed in discussions in the International Statistical Institute starting with Kiær (1896), by pioneering the study of a sample rather than recording the full population. During the first decades of the 20th century sampling was gradually accepted in official statistics: not only simple random sampling, but also cluster sampling and in particular stratified random sampling. Stratified random sampling was contrasted with 'purposive selection' in the landmark Royal Statistical Society discussion paper

by Neyman (1934) with which Kruskal and Mosteller ended their historical survey. Neyman quoted from Jensen (1926):

> 'In the selection of that part of the material which is to be the object of direct investigation, one or the other of the following two principles can be adopted: in certain instances it will be possible to make use of a combination of both principles. The one principle [simple random sampling] is characterized by the fact that the units which are to be included in the sample are selected at random. This method is only applicable where the circumstances make it possible to give every single unit an equal chance of inclusion in the sample. The other principle [purposive sampling] consists in the samples being made up by purposive selection of groups of units which it is presumed will give the sample the same characteristics as the whole.'

The later method of 'quota sampling'—aiming at obeying equal marginal proportions in sample and population—is a version of purposive sampling. The use of quota sampling is regarded as a main component in the famous failure of the opinion polls ahead of the US Presidential election in 1948 (Mosteller, 2010).

7.3. Should representativeness be avoided?

Contrary to the many positive meanings of representativity collected by Kruskal and Mosteller, the Miettinen (1985) declaration quoted in Section 3.1 has generated a strong scepticism about representativity among some epidemiologists. Rothman *et al*. (2013a) opened a 'point–counterpoint' debate in the *International Journal of Epidemiology* with a contribution with the title of the above subheading. Rothman and his colleagues (and most of the other contributors to this discussion) apparently equated 'representative sampling' to simple random sampling, as seen, for example, in their several recommendations of a much better idea—what we would call stratified random sampling, for which there is still a clear, and important, sampling frame. The explanation

> 'Thus, if you have a sample that is representative of the sex distribution in the source population, the results do not necessarily apply either to males or to females, but only to a hypothetical person of average sex . . .'

reveals a clear misunderstanding of generalizability of findings from surveys. Furthermore, Rothman *et al*. (2013a) concluded, that

> 'As initial steps, surveys may help to seed hypotheses and give a push toward scientific understanding, but the main road to general statements on nature is through studies that control skillfully for confounding variables and thereby advance our understanding of causal mechanisms. Representative sampling does not take us down that road.'

This statement does not make sense in the ordinary meaning of representative sampling as a survey with a known sampling frame and transparent sampling structure and it forgets to explain how to handle unmeasured confounders that hamper transportability—the place where randomization can often play a pivotal role.

Commentaries by Elwood (2013) and Nohr and Olsen (2013) were generally supportive of the views of Rothman *et al*. (2013a). A somewhat more reflected contribution was by Richiardi *et al.* (2013) where a key statement seems to be that

> 'Valid scientific inference is achieved if the confounders are controlled for, and there is no reason to believe that control of confounding can be more easily achieved in a population-based cohort than in a restricted cohort.'

Again, what do we do about the unobserved confounders and their role in generalizing the findings?

The editors of the *International Journal of Epidemiology* in Ebrahim and Smith (2013) seemed to be overwhelmed by the unanimity of their discussants and apparently tried to cool the iconoclasm a little:

> 'We are concerned that this notion will become accepted wisdom in epidemiology without its implications having been thought through, and feel that representativeness should neither be avoided nor uncritically embraced, but adopted (or not) according to the particular questions that are being addressed'.

The editors went on to specify their objections under the following five headings, all illustrated with concrete epidemiological examples: 'Some uses of epidemiology require representative samples'; 'Non-representative study groups may produce biased associations'; 'Scientific generalization: animals and randomized controlled trials'; 'The road to non-representative studies'; 'Epidemiology in the big data world'. They ended with the following somewhat tame statement:

> 'We feel that representativeness should neither be avoided nor uncritically universally adopted, but its value evaluated in each particular setting.'

In their rebuttal, Rothman *et al*. (2013b) among other things responded to a hypothetical example by Ebrahim and Smith (2013) about the necessity of controlling for differential health awareness between self-selected participants and non-participants. Rothman *et al*. (2013b) relied on a sampling frame to defend their approach:

> 'The bias could be controlled, with or without representative sampling, by measuring and controlling for health awareness, using information about health-seeking behavior such as medical screening visits, influenza vaccinations and other indicators of the selection factor underlying their concern'.

In an unrelated paper, motivated by the difficulties in obtaining generalizable evidence from hidden and hard-to-reach populations, Wirth and Tchetgen Tchetgen (2014) provided a clear counterpoint to the majority view in the above discussion by nicely summarizing the need to attend to survey goals. In the opening sentence of their discussion they stated that

> 'It has been argued that, despite the unequal selection induced by the design of complex surveys, analyses that treat the sampled data as the population of interest remain valid. Using a DAG [directed acyclic graph] framework, we show that this will depend on knowledge about the relationships among determinants of selection, exposure, and outcome. If the determinants of selection are associated with exposure and outcome, failure to account for the sampling design may result in biased effect estimates. This includes settings where determinants of selection are the exposure or outcome under study.'

7.4. The role of 'big data'

There is the potential for big data to evaluate or calibrate survey findings, to help to broaden an inferential frame by providing weights that transport within-study findings, to supplement or complement information gathered by traditional surveys and to help to validate cohort studies. The following examples are included to encourage increased use, with the potential increasing as the breadth, depth and accessibility of big data also increase.

Japec *et al.* (2015) provided a comprehensive survey of the promise and cautions that are associated with use of big data in the survey context, with most issues applying more generally. Japec *et al.* (2015) noted that

> 'The term Big Data is used for a variety of data as explained in the report, many of them characterized not just by their large volume, but also by their variety and velocity, the organic way in which they are created, and the new types of processes needed to analyze them and make inference from them'.

Their report gives examples of how data from the 'PriceStats index' tracks well with the official consumer price index, of information provided by monitor-collected, time-of-day vehicle

passings that can be used for assessing infrastructure needs, and a host of others. They summarized potential benefits.

(a) 'The benefits of using Big Data to improve public sector services have been recognized but the costs and risks of realizing these benefits are non-trivial.'
(b) 'Big Data offers entirely new ways to measure behaviors, in near real-time. Though behavioral measures are often lean in variables.'
(c) 'Big Data offers the possibility to study tails of distributions.'

Ansolabehere and Hersh (2012) reported on very sophisticated and careful analyses of the discrepancies between actual and survey-reported voting behaviour in the USA, showing that

> '... the rate at which people report voting in surveys greatly exceeds the rate at which they actually vote. For example, 78% of respondents to the 2008 National Election Study (NES) reported voting in the presidential election, compared with the estimated 57% who actually voted.'

The 57% came from voting records (a form of 'big data'). Explanations for the discrepancy include misreporting (possibly due to the social desirability of reporting to have voted), sample selection and poor record keeping. On the basis of a deep dive into causes, they reported that

> 'We validate not just voting reports (which were the focus of the NES validation), but also whether respondents are registered or not, the party with which respondents are registered, respondents' races, and the method by which they voted. Validation of these additional pieces of information provides important clues about the nature of validation and misreporting in surveys. Several key findings emerge from this endeavor. First, we find that standard predictors of participation, like demographics and measures of partisanship and political engagement, explain a third to a half as much about voting participation as one would find from analyzing behavior reported by survey respondents.'

Note that the magnitude of associations between personal attributes and voting participation computed by using the survey data do not transport to those computed by using administrative records. This lack of transportability identified via administrative records is probably quite general and shows the value of using 'big data' to conduct research on surveys (as distinct from survey research).

7.4.1. Case-study: big data to validate a clinical trial in Denmark

We now move to more fully developed examples, starting with the Danish Breast Cancer Cooperative Group which was started in 1978 with the dual aims of improving therapy of primary breast cancer in Denmark and facilitating scientific studies of breast cancer treatment (Blichert-Toft *et al.*, 2008a). In Denmark, breast cancer is overwhelmingly treated at the public hospitals, which are free of charge. The programme registers almost all primary breast cancer cases in Denmark, with about 80 000 cases registered by the 30-year anniversary in 2008. For each case extensive details on the tumour and the treatment are stored. Several waves of randomized trials of surgical techniques and of adjuvant therapy have been conducted within this framework, all in principle with the complete Danish population of women (usually stratified by age and/or menopausal status) as sampling frame. One such trial (DBCG-82TM) ran from 1982 to 1989 and regarded breast conserving surgery against total mastectomy (Blichert-Toft *et al.*, 2008b). On the basis of the trial results the Group decided in 1989 to recommend breast conserving treatment as a standard treatment option for suited breast cancer patients in Denmark. The question was, as always, how this general recommendation would work in the real world beyond the trial setting.

The national character of the Danish Breast Cancer Cooperative Group allowed a population-based study (Ewertz *et al.*, 2008), since almost all cases of primary breast cancer in Denmark were

registered in the Group's database and follow-up to death of all patients was possible through the Danish personal registration system. The results were encouraging; women younger than 75 years and operated on during the first 10 years after the recommendation (1989–1998) were followed up for 15 years. The results on survival, locoregional recurrences, distant metastases and benefit from adjuvant radiotherapy closely matched those of the clinical trial.

7.4.2. Representativity of cohort studies in the Nordic countries

The detailed population registries in the Nordic countries facilitate studies of representativity of key demographic variables for cohort studies. We quoted in Section 2.4 two recent validation studies, Nummela *et al.* (2011) from Finland and Langhammer *et al.* (2012) from Norway, casting some doubts on the representativity of their cohort studies as well as two other validation studies, Nohr *et al.* (2006) from Denmark and Nilsen *et al.* (2009) from Norway, which were more optimistic.

Andersen *et al.* (1998) compared mortality among participants in three cohorts recruited in the Copenhagen area to relevant background mortality to elucidate the problem that

> 'Often, the calculated relative risk of being exposed may be correct even in highly selected populations, but there is a risk of bias if other causes for the disease under study or confounders not taken into account in the analysis are differently distributed among the participating subjects and in the population that is target for generalization (see Rothman, 1976). Many factors associated with disease and death differ between participants and non-participants either because they are implicit in the selection criteria or because of the self-selection.'

(Note the focus on unmeasured confounders.) The analysis showed survivor selection in all cohorts (recruited participants being healthier at baseline than non-recruited individuals), which persisted beyond 10 years of observation for most combinations of age and sex.

7.5. Case-study: representativity issues in the Women's Health Initiative

We discuss representativity issues in the reanalyses of the WHI studies of possible side effects of postmenopausal hormone replacement therapy (HRT). By the early 1990s several observational studies had suggested that HRT reduces the risk of coronary heart disease (CHD) by about 40–50%, with similar effects for oestrogen-alone and oestrogen-plus-progestin treatment. However, there was also substantial observational evidence of increased breast cancer risk, particularly for the combination treatment.

On this basis two randomized trials (one for each type of treatment) as well as an observational cohort study on HRT were included in the WHI (see Prentice *et al.* (2005b) for an introduction from a statistical viewpoint). More than 10000 women were randomized to the oestrogen-only trial whereas more than 16000 women were randomized to the combination treatment trial. In 2002, the data and safety committee judged the health risks to exceed benefits in the second of these trials, which was stopped early after an average of about 5.5 years. For details see Rossouw *et al.* (2002), which is a landmark paper with 8311 citations by June 14th, 2015, in the *Web of Science*. The first trial was also stopped early, in 2004. These results had a profound effect on the use of HRT, which decreased dramatically worldwide.

At first sight the results from the randomized trials seemed to be at substantial odds with the earlier as well as the concurrent observational evidence. However, through hard work, skill and patience, the combined efforts of WHI researchers and colleagues outside the project have resulted in almost complete transparency: it does seem feasible to interpret all the evidence as being consistent (see Vandenbroucke (2009) for an easy-going general introduction).

We indicate some of the major findings of these extended post-publication activities. Of par-

ticular interest for our main focus of representativity is the debate about the deviation of the results on risk of CHD between the clinical trial of combination therapy and the previous evidence based on observational studies, as well as the comparison with the parallel observational cohort. As stated by Lawlor *et al.* (2004) with reference to Rossouw *et al.* (2002),

> 'Women in the WHI trial were older than the typical age at which women take HRT and were more obese than the women who have been included in the observational studies',

which is a genuine representativity issue that played an important part in the following discussions.

The CHD risk was reanalysed carefully by Prentice *et al.* (2005a, b), who found that the traditional analysis technique using the Cox proportional hazards model was insufficient, since the hazard ratio between the treatment and control groups was strongly dependent on the current duration of treatment. Introduction of time-stratified hazard ratios cleared up this issue with the conclusion that there was no significant difference between effect estimates in the randomized trial and the observational studies, when due consideration was taken of the widely different distributions of time since initiation of oestrogen-plus-progestin treatment. In other words, the apparently different results for the clinical trial and the observational study could be explained when statistical analysis accommodated the different sampling frames that the study samples represented, exemplifying very convincingly that it may be dangerous to ignore representativity.

A further, highly innovative reanalysis by Hernán *et al.* (2008) (who had no part in the original WHI analysis) attempted to 'emulate' an intention-to-treat analysis (known from randomized trials) of results from the observational Nurses' Health Study. We refrain from reporting details here, and we conclude only that this effort also reconciled results which had so far been considered widely different. A companion analysis by Toh *et al.* (2010a, b) took in a sense the opposite approach by developing adherence-adjusted analyses (which are standard in observational studies) of results from the randomized WHI trial, again with the conclusion that the randomized trials and the observational studies yield compatible results.

Recent authoritative clinical overviews include the detailed report by Manson *et al.* (2013) on the clinical trials and a concise, broader survey by Rossouw *et al.* (2013). These build in essential ways on the careful statistical and epidemiological work that was outlined above. As might be expected, it would be wrong to summarize the complex conclusions from these studies very briefly, and we must refer to the original references for the detailed substantive results. Here we quote the conclusion of the abstract of Rossouw *et al.* (2013) summarizing the clinical recommendations:

> 'Based on Women's Health Initiative data, the use of menopausal HT for fewer than 5 years is a reasonable option for the relief of moderate to severe vasomotor symptoms. The risks seen with estrogen plus progestin therapy suggest careful periodic reassessment of the ongoing therapy needs for women taking estrogen plus progestin therapy. The more favorable profile of estrogen therapy allows for individualized management with respect to duration of use when symptoms persist. For both estrogen therapy and estrogen plus progestin therapy, the baseline risk profile of the individual woman needs to be taken into account. Menopausal HT is not suitable for long-term prevention of CHD given risks of stroke, venous thromboembolism, and breast cancer (for estrogen plus progestin therapy) found in both clinical trials and in observational studies.'

8. Convergence of goals and methods

The goals of epidemiological studies and sample surveys are compatible. In each domain, high quality studies need to be internally valid, with sufficient precision to address the primary

objectives successfully. As we discuss in Section 1, traditional surveys *prime facie* address external validity by taking a probability sample with known selection probabilities so that weighting can transport internal prevalences and associations to a well-defined reference population. Traditionally, epidemiological studies have not explicitly given external validity a high priority, the premise being that the internal world is a good surrogate for the external. However, there is considerable convergence; use of Internet-based surveys and Internet-based enrolment poses similar challenges for both domains due to selection effects and the inability to develop sampling weights for explicit extrapolation. Furthermore, the prevalence of epidemiological studies that identify external validity as at least a secondary goal increases, with the attendant need to approximate a reference population and sampling weights. Innovations in data collection and analysis have the potential to broaden inferences, and doing so should be a top consideration in design, conduct, analysis and reporting of surveys and epidemiological and clinical studies. There are encouraging trends in this direction, but there is a long way to go. We elaborate below.

8.1. Survey goals and methods in epidemiological and experimental studies

Historically, experimental studies in humans have focused on internal validity and only infrequently given a high priority to identifying a reference population, but there is always at least an implicit need to broaden inferences beyond the study population. Without such broadening, it would not be worth conducting the study. The implicit hope is that, though levels (e.g. of blood pressure) may not be generalizable, internal comparisons (e.g. cross-sectional randomization group comparisons, longitudinal, within-individual changes and regression slopes) to a good approximation generalize. However, this transportability is by no means guaranteed, and there is increasing attention to the reference population in both design and analysis, especially for studies that address both scientific and policy goals (we quoted a concrete example in Section 7.4.1 about the Danish Breast Cancer Cooperative Group).

Although randomization in some form is very beneficial, it is by no means a panacea. Trial participants are commonly very different from the external patient pool, in part because of self-selection, but also because of the location of the study centre, and availability of resources that are needed to participate (transportation, child care, ...). Weisberg (2015) encouraged direct attention to generalization in design, conduct and analysis:

> 'The primary emphasis in most RCTs is on internal validity (establishing causality). Toward this end, it may be necessary to impose restrictive entry criteria in recruiting patients. Much less attention is paid in both analysis and reporting to the implications for patients who may differ in varying ways and degrees from the specific homogeneous population studied. However, such considerations of external validity are vital for the practising physician.'

When informing policy, inference to identified reference populations is key, and both methods development and application to achieve this goal burgeon (Frangakis, 2009; Greenhouse *et al.*, 2008; Pressler and Kaizar, 2013; Schumacher *et al.*, 2014; Stevens *et al.*, 2007; Stuart *et al.*, 2011; Stuart, 2014; Turner *et al.*, 2009; Weiss *et al.*, 2012). Stuart *et al.* (2011) discussed weighting and other methods, including computing a propensity score difference between those in and those not in the trial, and illustrated with application to a behaviour modification study. Of course, these analyses require that data are available to estimate propensities via a sampling frame. Stuart (2014) provided an excellent review of issues and provided examples from studies of suicide and human immunodeficiency virus and acquired immune deficiency syndrome. She noted that

> 'there are three main design strategies used to increase the generalizability of randomized trial results: (1) actual random sampling from the target population of interest, (2) practical clinical trials or pragmatic trials (e.g, Chalkidou *et al.*, 2012), which aim to enroll a more representative sample from the start, and

(3) doubly randomized preference trials (Marcus, 1997; Marcus *et al.*, 2012), which allow researchers to estimate the effect of randomization itself.'

Greenhouse *et al.* (2008) provided another example using a suicide study, and Pressler and Kaizar (2013) outlined methods including a discussion of generalizability bias, and illustrated with a comparison of two obstetric procedures. Sutcliffe *et al.* (2012) confirmed that some epidemiologists do attend to survey goals. The broader community should take note and take action.

Experimental and epidemiological studies can benefit from use of the Internet and social media to attract potential participants. For example, the study of Schisterman *et al.* (2014) used Facebook® as a recruiting mode (E. F. Schisterman, personal communication), but enrolment still depended on qualifying for the study. This use of the Internet causes little concern over that associated with traditional recruitment methods and may be the most effective way to accrue to well-designed studies.

Many studies occupy the middle ground between focus only on inference to the studied population and inference that depends on accommodating the sampling plan. For example, Zhang *et al.* (2014) used multilevel regression and post-stratification using data from the 'Behavioral risk factor surveillance system' to assess prevalence of chronic obstructive pulmonary disease in small areas. Many other studies, especially those using information from surveys, are careful to incorporate weights, at least for targeting the defined reference population. Hartman *et al.* (2015) put extrapolation in a causal analysis framework when combining information from experimental and observational studies to estimate population treatment effects. This work, although independent of Pearl and Bareinboim (2014), has a similar spirit. The increasing incidence of these types of study synthesis is a pleasing and beneficial trend.

8.2. Epidemiological or experimental goals and methods in surveys

Research in the survey domain is comprised of two distinct, but related, activities; survey research (asking questions of respondents) and research on surveys (evaluating innovations in survey conduct and analysis). Both activities are adopting, in some cases readopting, the goals and methods from the epidemiological and experimental domains, but technology transfer is greater for research on surveys, because its goals are essentially identical to those for epidemiological and experimental studies. For example, research on surveys entails observational and experimental studies, with internal validity of at least coequal status to external. The full armamentarium of observational and experimental designs and analyses is gaining traction (see, for example, Biemer and Peytchev (2013) and Luiten and Schouten (2013)).

8.2.1. Current trends in survey analysis

Traditionally, surveys were analysed by using the design-based paradigm, but the trend is towards an eclectic combination of design- and model-based approaches. In all situations, modelling is needed to accommodate non-response, dropouts and other forms of missing data. For example, non-response requires a combination of imputation and adjustment of weights, both relying on models dealing with missing data (see Rao *et al.* (2008)). Even with complete data a design-based estimate for a small demographic or geographic domain can have an unacceptably high variance, and modelling is needed to stabilize estimates (see Bell *et al.* (2013) and Opsomer *et al.* (2008)). Stabilization is driven by legal requirements and the need to make inferences for small geographical or sociodemographic domains. And, calibrated, design consistent, Bayesian methods that respect the sampling process are (slowly) gaining favour (see Little (2004, 2012) for examples).

Use of modelling has moved the survey culture considerably towards the epidemiological, and we encourage movement of the latter towards the former. The format and goals of model-assisted, design-based inference are fundamental to strategic approaches for dealing with confounding in epidemiological studies. For example, doubly robust approaches (Kang and Schafer, 2007; Lunceford and Davidian, 2004) are the progeny of model-assisted inference. So, it is a little odd that some epidemiologists dismiss the need to attend to the sampling plan in advance of or following data collection. However, approaches to causal analysis including standard covariate adjustment, principal strata (see Cuzick *et al.* (1997) for an early example), potential outcomes and g-estimation depend on selection propensities or stratification and so are examples of model-assisted, design-based inference. They provide an effective bridge between the survey and epidemiological communities.

8.3. Some optimism and some caution

There is a growing literature on participation factors and consequences; the conclusions are neither uniformly positive nor negative. Some studies judge that the trend towards self-enrolment and Internet-based studies can be valid; others identify cautions. For example, in a recent issue of the *Statistical Journal of the International Association of Official Statistics* (the Association is part of the International Statistical Institute), former President of the International Statistical Institute D. Trewin reported on a session 'What are the quality impacts of conducting high profile official statistical collections on a voluntary basis?' at the World Statistics Congress organized by the International Statistical Institute in Hong Kong in 2013 (see Trewin (2014)). A main issue is the increased risk of survey non-response, where Trewin pointed out that non-response

> '... is one of many sources of error in a survey.... The non-response rate is not necessarily a good proxy for non-response bias. The bias depends on the extent to which the characteristics of respondents differ from those of non-respondents.... The lesson is that rather than focusing just on response rates, there is a need to focus on representativeness.'

Trewin acknowledged the existence of methods (such as post-stratification) for correcting for non-response at the analysis stage, but commented that

> 'In my view, adjusting for non-response at the estimation stage is the non-preferred option. The emphasis should be on the design stage. This includes consideration of which auxiliary variables should be used in stratification.'

The presentations by other participants in the International Statistical Institute session (see Bethlehem and Bakker (2014), Hamel and Laniel (2014), Kott (2014) and Lynch (2014)) provided some support for and some disagreement with Trewin.

In their assessment of participation in a longitudinal, nutrition cohort study, Méjean *et al.* (2014) documented the motives of self-selected participants:

> 'The use of the Internet, the willingness to help advance public health research, and the study being publicly funded were key motives for participating in the Web-based NutriNet-Santé cohort'.

And

> 'These motives differed by sociodemographic profile and obesity, yet were not associated with lifestyle or health status.'

We encourage such documentation in other contexts.

Keeter (2014) discussed the trend towards low response rates and declining participation in telephone election polls, which is the bad news. However, participation rates via mobile phones appears to be stable, which is the good news. Galea and Tracy (2007) documented

declining participation rates, with reasons including the proliferation of studies, a decrease in volunteerism in the USA and the need for personal salience. This decline is counterbalanced to a degree by the finding that the decline in participation does not seem to be strongly associated with end points. All in all, the final verdict has not yet been delivered about the consequences of Web-based and other self-enrolment methods.

9. Discussion

We summarize our thesis with three points.

(a) Justifying epidemiological generalization is difficult concrete work (the WHI being a star example) and declarations that encourage bypassing this are unhelpful.
(b) The real representativity issue is whether the conditional effects that we wish to transport are actually transportable.
(c) Increased cross-fertilization between the epidemiological and survey domains will benefit science and policy.

Valid transportability is challenging. It depends on collecting and accurately measuring the principal attributes that associate with both enrolment and outcomes, and then using these appropriately in the analysis with some combination of using propensity scores, covariate adjustment and instrumental variables. Getting these right is demanding even in well-designed studies let alone those based on self-enrolment via the Internet or other routes. These issues must be addressed in epidemiological studies as well as in surveys.

Epidemiologists are leading developers and users of propensity models, doubly robust approaches and other model-assisted analyses, many of which have their roots in survey sampling. Although these are used primarily to adjust within-study comparisons, we have been surprised about the energetic resistance on the part of influential epidemiologists of the importance of designing a study to increase the validity of these and related approaches in making inferences to a reference population. Similarly, though the survey community does pay careful attention to representation within a usually narrow frame, it needs to adopt and develop methods that support transportability.

All communities need to consider the perils and potentials of self-selection. Those conducting studies and surveys on which policy or other important decisions are based must maintain quality and trust. Doing so requires anchoring to protocol-driven enrolment or probability-based sampling, departing from these only when absolutely necessary and when quality can be maintained.

Harris *et al.* (2015a), debriefing on lessons learned in on-line recruiting, and the associated commentary (Allsworth, 2015) and response (Harris *et al.*, 2015b) provided a reprise on issues including whether representation is necessary, the too narrow definition of it by many epidemiologists and the similarity of issues in the epidemiological and survey worlds. In sum, epidemiological studies can benefit from incorporating survey goals; surveys can benefit from epidemiological analyses. Both can benefit from clearer identification of goals, a broadening from beyond the sample under study or the preidentified reference population. We do not expect or require that the goals and approaches of the epidemiological and survey communities will completely converge, but we encourage them to adopt a common set of principles that structure and empower convergence.

Acknowledgements

We thank the editor and reviewers for their comments and suggestions. TAL prepared this paper

while serving as Associate Director for Research and Methodology at the US Census Bureau under an Interagency Personnel Agreement with Johns Hopkins University.

Appendix A: Estimating a population mean

We consider the basic example of estimating a population mean, specifically the average length of stay (LOS) for hospitals in a specific domain. The issues generalize to estimation of a regression slope, a longitudinal change or other parameters of interest. As an additional simplification, we assume that the target population consists of five hospitals, that a random sample of medical records for each hospital is obtained (in a more complex design the five hospitals would be a sample from a larger universe) and that the within-hospital LOS variance σ^2 is a known constant.

Table 1 displays the observed and the population information and Table 2 three estimates each addressing a different inferential goal:

(a) to produce the minimum variance estimate of the population LOS,
(b) to give each hospital equal weight or
(c) to produce the minimum variance unbiased estimate MVUE.

The minimum variance estimate is directly available via inverse variance weighting, producing the first row in Table 2. The second row of Table 2 is a straightforward consequence of equal weighting, with these weights possibly reflecting a policy goal.

Computing the unbiased estimate requires knowing the (relative) size of each hospital. These are the 'Population information' in Table 1 and, if known, the population weights are available as are the patient-specific relative propensities f_j/p_j. Using population weights on the hospital means (equivalently, weighting individual patients by reciprocal propensity) produces the third row of Table 2. Because the relative

Table 1. Constructed LOS data from five hospitals

Hospital	*Observed information*				*Population information*		*Patient relative propensity* f_j/p_j
	Number sampled n_j	*% of sample* $100f_j$	*Mean LOS* Y_j	*Variance* σ_j^2	*Hospital size*	*% of total population* $100p_j$	
1	30	20	25	$\sigma^2/30$	100	10	2.00
2	60	40	35	$\sigma^2/60$	150	15	2.67
3	15	10	15	$\sigma^2/15$	200	20	0.50
4	30	20	40	$\sigma^2/30$	250	25	0.80
5	15	10	10	$\sigma^2/15$	300	30	0.33
Total	150	100			1000	100	

Table 2. Weights, weighted averages and relative variances†

Estimator	*Hospital-specific weights* $\mathbf{w}$	$\hat{\mu}(\mathbf{w})$	*Variance ratio 100(variance/minimum variance)*
Minimum variance	0.20 0.40 0.10 0.20 0.10	29.5	100
Equally weighted	0.20 0.20 0.20 0.20 0.20	25.0	130
Unbiased	0.10 0.15 0.20 0.25 0.30	23.8	172

†Reciprocal variance weights produce the minimum variance estimate; population weights (the p_j) produce the unbiased estimate; equal weights may address the policy goal of giving each hospital equal weight. The first two rows are available from the sample information; the third row requires population information.

propensities are far from 1.0, using these weights induces a variance increase of 72% over the minimum variance estimate, which in many contexts is a high price to pay for unbiasedness. Rather than pay this price, targeting a low mean-square error $\mathrm{MSE} = \mathrm{variance} + \mathrm{bias}^2$ estimate is attractive. If the p_j are available, relatively low MSE can be achieved either by using a compromise between the minimum variance and the MVUE-weights, or by stabilizing the hospital-specific estimates and applying the population weights (see Gelman (2007) and Pfeffermann (1993) for discussion of this and related issues).

Even if the hospital sizes, and therefore the p_j, were not available when the sample was taken, administrative data (a form of 'big data') might be available to provide good estimates of them, allowing computation of MVUE or a compromise estimate. However, as Chang and Krosnick (2009) and Yeager *et al.* (2011) reported, in many contexts use of such information to improve estimates can be effective but is not competitive with a carefully conducted, probability-based survey.

Finally, we note that the sample in Table 1 is not 'representative' in the narrow sense that was used by Rothman *et al.* (2013a) because it is not self-weighting, i.e. MVUE uses weights that are different from those producing the minimum variance estimate. However, if the p_j are known, the sample is representative in the commonly accepted sense.

References

Allsworth, J. E. (2015) Recruiting for epidemiologic studies using social media. *Am. J. Epidem.*, **181**, 747–749.

Andersen, L., Vestbo, J., Juel, K., Bjerg, A., Keiding, N., Jensen, G., Hein, H. and Sørensen, T. (1998) A comparison of mortality rates in three prospective studies from Copenhagen with mortality rates in the central part of the city, and the entire country. *Eur. J. Epidem.*, **14**, 579–585.

Ansolabehere, S. and Hersh, E. (2012) Validation: what Big Data reveal about survey misreporting and the real electorate. *Polit. Anal.*, **20**, 437–459.

Baird, D. D. and Wilcox, A. J. (1985) Cigarette smoking associated with delayed conception. *J. Am. Med. Ass.*, **253**, 2979–2983.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. and Tourangeau, R. (2013) Summary report of the AAPOR task force on non-probability sampling. *J. Surv. Statist. Methodol.*, **1**, 90–143.

Battaglia, M. P. (2008) Nonprobability sampling. In *Encyclopedia of Survey Research Methods*, pp. 523–526. New York: Sage.

Bell, W. R., Datta, G. S. and Ghosh, M. (2013) Benchmarking small area estimates. *Biometrika*, **100**, 189–202.

Bethlehem, J. and Bakker, B. (2014) The impact of non-response on survey quality. *Statist. J. Int. Ass. Off. Statist.*, **30**, 243–248.

Biemer, P. P. and Peytchev, A. (2013) Using geocoded census data for nonresponse bias correction: an assessment. *J. Surv. Statist. Methodol.*, **1**, 24–44.

Blichert-Toft, M., Christiansen, P. and Mouridsen, H. T. (2008a) Danish Breast Cancer Cooperative Group—DBCG: history, organization, and status of scientific achievements at 30-year anniversary. *Acta Oncol.*, **47**, 497–505.

Blichert-Toft, M., Nielsen, M., Düring, M., Møller, S., Rank, F., Overgaard, M. and Mouridsen, H. T. (2008b) Long-term results of breast conserving surgery vs. mastectomy for early stage invasive breast cancer: 20-year follow-up of the Danish randomized DBCG-82TM protocol. *Acta Oncol.*, **47**, 672–681.

Buck Louis, G. M., Schisterman, E. F., Sweeney, A. M., Wilcosky, T. C., Gore-Langton, R. E., Lynch, C. D., Barr, D. B., Schrader, S. M., Kim, S., Chen, Z. and Sundaram, R. (2011) Designing prospective cohort studies for assessing reproductive and developmental toxicity during sensitive windows of human reproduction and development—the LIFE Study. *Paed. Perntl Epidem.*, **25**, 413–424.

Chalkidou, K., Tunis, S., Whicher, D., Fowler, R. and Zwarenstein, M. (2012) The role for pragmatic randomized controlled trials (prcts) in comparative effectiveness research. *Clin. Trials*, **9**, 436–446.

Chang, L. and Krosnick, J. A. (2009) National surveys via RDD telephone interviewing vs. the Internet: comparing sample representativeness and response quality. *Publ. Opin. Q.*, **73**, 641–678.

Cuzick, J., Edwards, R. and Segnan, N. (1997) Adjusting for non-compliance and contamination in randomized clinical trials. *Statist. Med.*, **16**, 1017–1029.

Ebrahim, S. and Smith, G. D. (2013) Should we always deliberately be non-representative? *Int. J. Epidem.*, **42**, 1022–1026.

Elwood, J. M. (2013) On representativeness. *Int. J. Epidem.*, **42**, 1014–1015.

Enright, R. L., Connett, J. E. and Bailey, W. C. (2002) The fev1/fev6 predicts lung function decline in adult smokers. *Respir. Med.*, **96**, 444–449.

Ewertz, M., Kempel, M. M., Düring, M., Jensen, M.-B., Andersson, M., Christiansen, P., Kroman, N., Rasmussen, B. B. and Overgaard, M. (2008) Breast conserving treatment in Denmark, 1989–1998: a nationwide population-based study of the Danish Breast Cancer Co-operative Group. *Acta Oncol.*, **47**, 682–690.

Frangakis, C. (2009) The calibration of treatment effects from clinical trials to target populations. *Clin. Trials*, **6**, 136–140.
Galea, S. and Tracy, M. (2007) Participation rates in epidemiologic studies. *Ann. Epidem.*, **17**, 643–653.
Gelman, A. (2007) Struggles with survey weighting and regression modeling (with discussion). *Statist. Sci.*, **22**, 153–188.
Greenhouse, J. B., Kaizar, E. E., Kelleher, K., Seltman, H. and Gardner, W. (2008) Generalizing from clinical trial data: a case study: the risk of suicidality among pediatric antidepressant users. *Statist. Med.*, **27**, 1801–1813.
Hamel, M. and Laniel, N. (2014) Producing official statistics via voluntary surveys—the national household survey in Canada. *Statist. J. Int. Ass. Off. Statist.*, **30**, 237–242.
Harris, M. L., Luxton, D., Wigginton, B. and Lucke, J. C. (2015a) Recruiting online: lessons from a longitudinal survey of contraception and pregnancy intentions of young Australian women. *Am. J. Epidem.*, **181**, 737–746.
Harris, M. L., Luxton, D., Wigginton, B. and Lucke, J. C. (2015b) Harris *et al*. respond to "social media recruitment". *Am. J. Epidem.*, **181**, 750–751.
Hartman, E., Grieve, R., Ramsahai, R. and Sekhon, J. S. (2015) From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *J. R. Statist. Soc.* A, **178**, 757–778.
Hatch, E. E., Wise, L. A., Mikkelsen, E. M., Christensen, T., Riis, A. H., Sørensen, H. T. and Rothman, K. J. (2012) Caffeinated beverage and soda consumption and time to pregnancy. *Epidemiology*, **23**, 393–401.
Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E. and Robins, J. M. (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, **19**, 766–779.
Horvitz, D. and Thompson, D. (1952) A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Ass.*, **47**, 663–685.
Howe, G., Westhoff, C., Vessey, M. and Yeats, D. (1985) Effects of age, cigarette smoking, and other factors on fertility—finding in a large prospective study. *Br. Med. J.*, **290**, 1697–1700.
Huybrechts, K. F., Mikkelsen, E. M., Christensen, T., Riis, A. H., Hatch, E. E., Wise, L. A., Sørensen, H. T. and Rothman, K. J. (2010) A successful implementation of e-epidemiology: the Danish pregnancy planning study 'Snart-Gravid'. *Eur. J. Epidem.*, **25**, 297–304.
Japec, L., Kreuter, F., Bert, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. and Usher, A. (2015) *AAPOR Report on Big Data*. Deerfield: American Association for Public Opinion Research.
Jensen, A. (1926) Report on the representative method in statistics. *Bull. Int. Statist. Inst.*, **22**, 359–380.
Kang, J. D. Y. and Schafer, J. (2007) Demystifying double-robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, **22**, 523–539.
Keeter, S. (2014) Change is afoot in the world of election polling. *Amstat News*, no. 448, Oct., 3–4.
Keiding, N. (1987) The method of expected number of deaths. *Int. Statist. Rev.*, **55**, 1–20.
Keiding, N. and Clayton, D. (2014) Standardization and control for confounding in observational studies: a historical perspective. *Statist. Sci.*, **29**, 529–558.
Keiding, N., Hansen, O. H., Sørensen, D. N. and Slama, R. (2012) The current duration approach to estimating time to pregnancy (with discussion). *Scand. J. Statist.*, **39**, 185–213.
Keiding, N. and Slama, R. (2015) Time-to-pregnancy in the real world. *Epidemiology*, **26**, 119–121.
Kiær, A. N. (1896) Observations et expériences concernant des dénombrements représentatifs. *Bull. Int. Statist. Inst.*, **9**, 176–183.
Kott, P. (2014) On voluntary and volunteer government surveys in the United States. *Statist. J. Int. Ass. Off. Statist.*, **30**, 249–253.
Kruskal, W. and Mosteller, F. (1979a) Representative sampling, i: Non-scientific literature. *Int. Statist. Rev.*, **47**, 13–24.
Kruskal, W. and Mosteller, F. (1979b) Representative sampling, ii: Scientific literature, excluding statistics. *Int. Statist. Rev.*, **47**, 111–127.
Kruskal, W. and Mosteller, F. (1979c) Representative sampling, iii: The current statistical literature. *Int. Statist. Rev.*, **47**, 245–265.
Kruskal, W. and Mosteller, F. (1980) Representative sampling, iv: The history of the concept in statistics, 1895-1939. *Int. Statist. Rev.*, **48**, 169–195.
Langhammer, A., Krokstad, S., Romundstad, P., Heggland, J. and Holmen, J. (2012) The HUNT study: participation is associated with survival and depends on socioeconomic status, diseases and symptoms. *BMC Med. Res. Methodol.*, **12**, article 143.
Lawlor, D. A., Davey Smith, G. and Ebrahim, S. (2004) The hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? *Int. J. Epidem.*, **33**, 464–467.
Leenheer, J. and Scherpenzeel, A. C. (2013) Does it pay off to include non-internet households in an internet panel? *Int. J. Internet Sci.*, **8**, 17–29.
Little, R. J. (2004) To model or not to model?: competing modes of inference for finite population sampling. *J. Am. Statist. Ass.*, **99**, 546–556.
Little, R. J. (2012) Calibrated Bayes, an alternative inferential paradigm for official statistics (with discussion). *J. Off. Statist.*, **28**, 309–372.

Luiten, A. and Schouten, B. (2013) Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction. *J. R. Statist. Soc.* A, **176**, 169–189.
Lunceford, J. K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statist. Med.*, **23**, 2937–2960.
Lynch, J. (2014) The evolving role of self-report surveys on criminal vicitmization in a system of statistics on crime and the administration of justice. *Statist. J. Int. Ass. Off. Statist.*, **30**, 165–169.
Mall, S., Akmatov, M. K., Schultze, A., Ahrens, W., Obi, N., Pessler, F. and Krause, G. (2014) Web-based questionnaires to capture acute infections in long-term cohorts: findings of a feasibility study. *Bundesgesundheitsblatt*, **57**, 1308–1314.
Manson, J. E., Chlebowski, R. T., Stefanick, M. L., Aragaki, A. K., Rossouw, J. E., Prentice, R. L., Anderson, G., Howard, B. V., Thomson, C. A., LaCroix, A. Z., Wactawski-Wende, J., Jackson, R. D., Limacher, M., Margolis, K. L., Wassertheil-Smoller, S., Beresford, S. A., Cauley, J. A., Eaton, C. B., Gass, M., Hsia, J., Johnson, K. C., Kooperberg, C., Kuller, L. H., Lewis, C. E., Liu, S., Martin, L. W., Ockene, J. K., O'Sullivan, M. J., Powell, L. H., Simon, M. S., Van Horn, L., Vitolins, M. Z. and Wallace, R. B. (2013) Menopausal hormone therapy and health outcomes during the intervention and extended poststopping phases of the Women's Health Initiative randomized trials. *J. Am. Med. Ass.*, **310**, 1353–1368.
Marcus, S. M. (1997) Assessing non-constant bias with parallel randomized and nonrandomized clinical trials. *J. Clin. Epidem.*, **50**, 823–828.
Marcus, S. M., Stuart, E. A., Wang, P., Shadish, W. R. and Steiner, P. M. (2012) Estimating the causal effect of randomization versus treatment preference in a doubly-randomized preference trial. *Psychol. Meth.*, **17**, 244–254.
McCutcheon, A. L., Rao, K. and Kaminska, O. (2014) The untold story of multi-mode (online and mail) consumer panels: from optimal recruitment to retention and attrition. In *Online Panel Surveys: an Interdisciplinary Approach*. Hoboken: Wiley.
Méjean, C., Szabo de Edelenyi, F., Touvier, M., Kesse-Guyot, E., Julia, C., Andreeva, V. A. and Hercberg, S. (2014) Motives for participating in a Web-based nutrition cohort according to sociodemographic, lifestyle, and health characteristics: the Nutrinet-Santé cohort study. *J. Med. Internet Res.*, **16**, article e189.
Miettinen, O. S. (1985) *Theoretical Epidemiology.* New York: Wiley.
Mikkelsen, E. M., Hatch, E. E., Wise, L. A., Rothman, K. J., Riis, A. and Sørensen, H. T. (2009) Cohort profile: the Danish web-based pregnancy planning study 'Snart-Gravid'. *Int. J. Epidem.*, **38**, 938–943.
Mikkelsen, E. M., Riis, A. H., Wise, L. A., Hatch, E. E., Rothman, K. J. and Sørensen, H. T. (2013) Pre-gravid oral contraceptive use and time to pregnancy: a Danish prospective cohort study. *Hum. Reprodn*, **28**, 1398–1405.
Mosteller, F. (2010) Why did Dewey beat Truman in the pre-election polls of 1948? In *The Pleasures of Statistics: the Autobiography of Frederick Mosteller*, ch. 1. New York: Springer.
Mumford, S. L., Schisterman, E. F., Cole, S. R., Westreich, D. and Platt, R. W. (2015) Time at risk and intention to treat analyses: parallels and implications for inference. *Epidemiology*, **26**, 112–118.
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). *J. R. Statist. Soc.*, **97**, 558–625.
Nilsen, R. M., Vollset, S. E., Gjessing, H. K., Skjaerven, R., Melve, K. K., Schreuder, P., Alsaker, E. R., Haug, K., Daltveit, A. K. and Magnus, P. (2009) Self-selection and bias in a large prospective pregnancy cohort in norway. *Paediatr. Perinatl Epidem.*, **23**, 597–608.
Nohr, E. A., Frydenberg, M., Henriksen, T. B. and Olsen, J. (2006) Does low participation in cohort studies induce bias? *Epidemiology*, **17**, 413–418.
Nohr, E. A. and Olsen, J. (2013) Epidemiologists have debated representativeness for more than 40 year—has the time come to move on? *Int. J. Epidem.*, **42**, 1016–1017.
Nummela, O., Sulander, T., Helakorpi, S., Haapola, I., Uutela, A., Heinonen, H., Valve, R. and Fogelholm, M. (2011) Register-based data indicated nonparticipation bias in a health study among aging people. *J. Clin. Epidem.*, **64**, 1418–1425.
Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008) Non-parametric small area estimation using penalized spline regression. *J. R. Statist. Soc.* B, **70**, 265–286.
Pearl, J. and Bareinboim, E. (2014) External validity: from do-calculus to transportability across populations. *Statist. Sci.*, **29**, 579–595.
Pfeffermann, D. (1993) The role of sampling weights when modeling survey data. *Int. Statist. Rev.*, **61**, 317–337.
Prentice, R. L., Langer, R., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G., Barad, D., Curb, J. D., Kotchen, J., Kuller, L., Limacher, M., Wactawski-Wende, J. and Women's Health Initiative Investigators (2005a) Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *Am. J. Epidem.*, **162**, 404–414.
Prentice, R. L., Pettinger, M. and Anderson, G. L. (2005b) Statistical issues arising in the Women's Health Initiative (with discussion). *Biometrics*, **61**, 899–941.
Pressler, T. R. and Kaizar, E. E. (2013) The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Statist. Med.*, **32**, 3552–3568.
Radin, R. G., Hatch, E. E., Rothman, K. J., Mikkelsen, E. M., Sørensen, H. T., Riis, A. H. and Wise, L. A. (2014) Active and passive smoking and fecundability in Danish pregnancy planners. *Fertil. Steril.*, **102**, 183–191.

Rao, R. S., Glickman, M. E. and Glynn, R. J. (2008) Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statist. Med.*, **27**, 2196–2213.
Rao, K., Kaminska, O. and McCutcheon, A. L. (2010) Recruiting probability samples for a multi-mode research panel with internet and mail components. *Publ. Opin. Q.*, **74**, 68–84.
Richiardi, L., Pizzi, C. and Pearce, N. (2013) Representativeness is usually not necessary and often should be avoided. *Int. J. Epidem.*, **42**, 1018–1022.
Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., Jackson, R. D., Beresford, S. A. A., Howard, B. V., Johnson, K. C., Kotchen, J. M., Ockene, J. and Writing Group for the Women's Health Initiative Investigators (2002) Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *J. Am. Med. Ass.*, **288**, 321–333.
Rossouw, J. E., Manson, J. E., Kaunitz, A. M. and Anderson, G. L. (2013) Lessons learned from the Women's Health Initiative trials of menopausal hormone therapy. *Obstet. Gyn.*, **121**, 172–176.
Rothman, K. J. (1976) Causes. *Am. J. Epidem.*, **104**, 587–592.
Rothman, K. J. (1986) *Modern Epidemiology*. Boston: Little, Brown.
Rothman, K. J., Gallacher, E. E. and Hatch, E. E. (2013a) Why representativeness should be avoided. *Int. J. Epidem.*, **42**, 1012–1014.
Rothman, K. J., Gallacher, E. E. and Hatch, E. E. (2013b) Rebuttal: When it comes to scientific inference, sometimes a cigar is just a cigar. *Int. J. Epidem.*, **42**, 1026–1028.
Rothman, K. J. and Greenland, S. (1998) *Modern Epidemiology*, 2nd edn. Philadelphia: Lippincott Williams and Wilkins.
Rothman, K. J., Greenland, S. and Lash, T. L. (2008) *Modern Epidemiology*, 3rd edn. Philadelphia: Wolters Kluwer.
Rothman, K. J., Wise, L. A., Sørensen, H. T., Riis, A. H., Mikkelsen, E. M. and Hatch, E. E. (2013) Volitional determinants and age-related decline in fecundability: a general population prospective cohort study in Denmark. *Fertil. Steril.*, **99**, 1958–1964.
Rothstein, M. A. and Shoben, A. B. (2013) Does consent bias research? *Am. J. Bioeth.*, **13**, 27–37.
Royall, R. (1976) Current advances in sampling theory: implications for human observational studies. *Am. J. Epidem.*, **104**, 463–474.
Särndal, C. E. (2007) The calibration approach in survey theory and practice. *Surv. Methodol.*, **33**, 99–119.
Särndal, C. E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer.
Schisterman, E. F., Silver, R. M., Lesher, L. L., Faraggi, D., Wactawski-Wende, J., Townsend, J. M., Lynch, A. M., Perkins, N. J., Mumford, S. L. and Galai, N. (2014) Preconception low-dose aspirin and pregnancy outcomes: results from the EAGeR randomised trial. *Lancet*, **384**, 29–36.
Schumacher, M., Rücker, G. and Schwarzer, G. (2014) Meta-analysis and the Surgeon General's report on smoking and health. *New Engl. J. Med.*, **370**, 186–188.
Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, 2nd edn. Boston: Houghton Mifflin.
Stevens, J., Kelleher, K., Greenhouse, J., Chen, G., Xiang, H., Kaizar, E., Jensen, P. S. and Arnold, L. E. (2007) Empirical evaluation of the generalizability of the sample from the multimodal treatment study for ADHD. *Adm. Poly Mentl Hlth Serv. Res.*, **34**, 221–232.
Stuart, E. A. (2014) Generalizability of clinical trials results. In *Methods in Comparative Effectiveness Research.* New York: Taylor and Francis.
Stuart, E. A., Cole, S. R., Bradshaw, C. P. and Leaf, P. J. (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Statist. Soc.* A, **174**, 369–386.
Sutcliffe, C. G., Kobayashi, T., Hamapumbu, H., Shields, T., Mharakurwa, S., Thuma, P. E., Louis, T. A., Glass, G. and Moss, W. (2012) Reduced risk of malaria parasitemia following household screening and treatment: a cross-sectional and longitudinal cohort study. *PLOS One*, **7**, article e31396.
Toh, S., Hernández-Díaz, S., Logan, R., Robins, J. M. and Hernán, M. A. (2010a) Estimating absolute risks in the presence of nonadherence: an application to a follow-up study with baseline randomization. *Epidemiology*, **21**, 528–539.
Toh, S., Hernández-Díaz, S., Logan, R., Rossouw, J. E. and Hernán, M. A. (2010b) Coronary heart disease in postmenopausal recipients of estrogen plus progestin therapy: does the increased risk ever disappear?: a randomized trial. *Ann. Intern. Med.*, **152**, 211–217.
Trewin, D. (2014) What are the quality impacts of conducting high profile official statistical collections on a voluntary basis? *Statist. J. Int. Ass. Off. Statist.*, **30**, 231–235.
Tu, J. V., Willison, D. J., Silver, F. L., Fang, J., Richards, J. A., Laupacis, A., Kapral, M. K. and Investigators in the Registry of the Canadian Stroke Network (2004) Impracticability of informed consent in the registry of the Canadian stroke network. *New Engl. J. Med.*, **350**, 1414–1421.
Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S. and Thompson, S. G. (2009) Bias modeling in evidence synthesis. *J. R. Statist. Soc.* A, **172**, 21–47.
Vandenbroucke, J. P. (2009) The HRT controversy: observational studies and RCTS fall in line. *Lancet*, **373**, 1233–1235.

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.

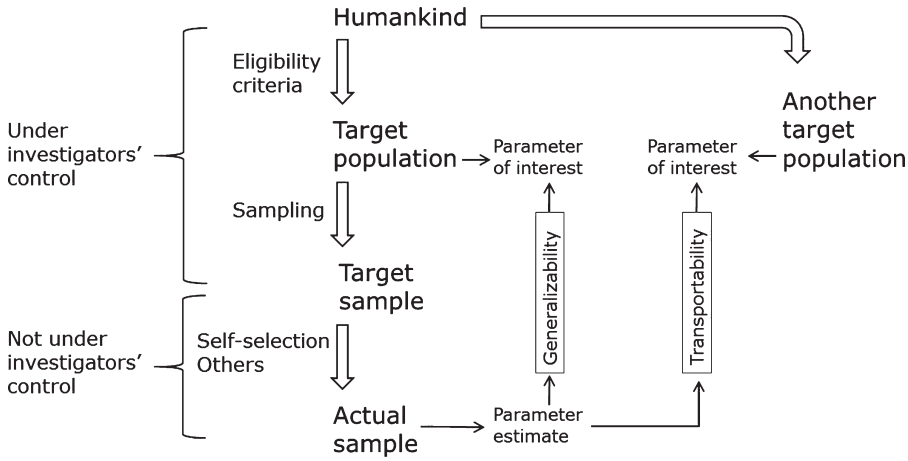

Fig. 1. Framework for selection in human studies

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*)

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.

Keiding and Louis focus most of their attention on the implications of self-selection for the generalizability of research findings. This is a major concern in survey research now, as volunteer respondent panels proliferate, in response to the growing expense and declining rates of participation in probability sample surveys. Are findings from volunteer and probability sample surveys comparable? Can we substitute one for the other? The authors review research that has tried to answer such questions, but they do not mention that there are important confounders that must be considered when comparing these two types of studies. Selection is not the only issue. Volunteer-based surveys often involve self-administered Web questionnaires, whereas probability sample surveys often involve questionnaires that are administered by interviewers, face to face or by telephone. Such mode differences can make it difficult to distinguish the effect of volunteerism from the effects of different measurement conditions. We need to pay more attention to the internal validity of volunteer–probability sample comparisons.

The authors observe rightly that there is a conflict between internal and external validity, between the aim of measuring causal relationships in a study and the aim of generalizing the study's findings to a target population or context. The conflict lies in the need to control, as much as is possible, confounding factors that introduce error into causal inferences and the opposing need to ensure that measures that are taken to limit confounding factors do not restrict broader application of the findings. Special study conditions, which are desirable for isolating causal relationships, pose threats to the ability to generalize the findings.

Keiding and Louis argue that experience in epidemiology with randomized trials could aid the progress of survey methodology. There is a long history of experiments in surveys, in which the conflict between internal and external validity has been played out. Surveys are 'noisy' environments for experiments. The complex interplay between the various components of survey design (sampling, mode of survey contact, questionnaire design and respondent recruitment) makes it difficult to control all confounding factors that threaten internal validity (as in the case of volunteer–probability sample comparisons, noted above). Simplifying the survey design—e.g. using one mode of contact, sampling from a homogeneous population or shortening the questonnaire—to isolate experimental effects better impairs external validity. Kalton and Schuman (1982), in a paper that was read to the Royal Statistical Society, observed that discerning the effect of a single question on responses (*internal* validity), in the context of a larger questionnaire, is a slow, daunting and laborious task. They also cited with approval Cannell's series of experiments on survey question format and interviewing techniques. But they argued that the findings from these studies lacked sufficient *external* validity for them to be widely applied in surveys, since Cannell's experiments were conducted on homogeneous samples, with newly trained interviewers, and in unusual field conditions (see Cannell *et al*. (1981)). One hopes that lessons from clinical trials can enable survey practitioners to find a more acceptable balance of internal and external validity in survey experiments.

On a more mundane practical level, experiments and experimental evidence are not automatically welcome in on-going survey operations. Experiments take resources from survey operations, which may already be short on funding and personnel. Experiments need special conditions to improve internal validity that may be deemed unworkable within the capabilities of organizational routines and systems that are designed for the *status quo*. Since the workings of on-going surveys will be disturbed if experiments are conducted within them, and since, as Dillman (1996) noted, the operational culture in some government survey organizations is antagonistic to methodological research, conducting randomized trials of new methods in on-going surveys can be difficult to accomplish. Strong leadership is needed to overcome organizational obstacles to experimentation.

Finally, in the important work of assessing the perils and potentials of self-selection that Keiding and Louis have recommended, we need more research on 'truth' benchmarks, that can help to adjudicate the quality of data collections under study. Without benchmarks, we can only say that studies based on different respondent recruitment methods produce similar or different results. The authors reference some work using population registers in Norway and Denmark, but clearly more research of this kind is needed.

It is an honour for me to second the vote of thanks.

The vote of thanks was passed by acclamation.

Guanglei Hong (*University of Chicago*)

I thank Professor Keiding and Professor Louis for addressing a major topic that has implications beyond epidemiological research. The rapid increase of Web-based surveys requires a deliberate response from the scientific community. The paper argues for a legitimate place for such studies, despite their potential pitfalls, 'when absolutely necessary and when quality can be maintained'. So what type of quality control might apply when we see Web-based studies in grant applications and journal submissions? What standards and procedures might be recommended for such studies?

The primary issue under discussion is the potential lack of transportability of statistical results. These include, for example, univariate distributions of exposures or outcomes, multivariate associations, causal linkages and longitudinal trends. Two empirical checks for transportability may be conducted ideally before rather than after one launches a Web-based study on its full scale.

Empirical check A
Is there any previous evidence from survey research indicating important heterogeneity in the patterns and trends of interest across theoretically defined subpopulations? A tentative null finding of moderation in previous research, however, cannot rule out possible heterogeneity across subpopulations yet to be identified.

Empirical check B
Is there a pilot study—with a sampling frame and with sufficient power—that compares volunteers for the proposed Web-based study and non-volunteers? Does such a study reveal important differences between the volunteers and non-volunteers in the patterns and trends of interest? Because volunteer status is the only moderator of concern to transportability, passing *empirical check B* makes a much more convincing case than passing *empirical check A*.

Arguably, a Web-based study receives some justification if previous research has shown a high level of homogeneity across many subpopulations and, more importantly, if a pilot study has established that the patterns and trends of interest are independent of volunteer status.

When major threats to transportability are detected, the information that is collected from these empirical checks may nonetheless enable *post hoc* adjustment of results from a Web-based study. In particular, in empirical check B, one may estimate the propensity score for volunteering and then employ propensity-score-based weighting to transform the composition of volunteers. This strategy requires the assumption that, conditioning on the observed covariates, volunteer status becomes independent of the patterns and trends of interest. It also requires the positivity assumption—i.e. there are no *never-takers* of the Web-based study in the target population.

Good news: these assumptions can actually be tested with the pilot data!

Lance A. Waller (*Emory University, Atlanta*)
I thank the authors for a thoughtful statistical view of a complicated and important issue. I particularly appreciate the examples relating to the Women's Health Initiative—an excellent example and case-study for similar efforts in progress or in planning.

The points regarding generalizability that were discussed by the authors extend beyond recruitment of epidemiologic cohorts and provide linkage across many areas of application. I mention two specific extensions: understanding and modelling the observation process and linking multiple sources (e.g. surveys) of overlapping information.

The modelling (or quantification) of the observation process is an essential concept, especially in an age of ubiquitous observational data. Statistical adjustments such as inverse probability weighting, EM algorithms and propensity scores all build on 'models' of the observation process by quantifying and adjusting for the probability of observation (inclusion in the data set). Too often our literature focuses on technical details rather than developing a general conceptual basis for such approaches. Briefly, do we understand how our data arise and can we incorporate our understanding into better inferential methods? Wikle (2003) summarized a helpful hierarchical data model–process model framework from the climatology and ecology literatures, providing a link between a deterministic *process model* (e.g. a global climate model) with a probabilistic *data model* describing distributions of observed quantities, typically including a specific model of the *probability of observation* for each data element.

A second extension relates to linking multiple surveys (or, more generally, multiple sources of data) collected by different agencies for different purposes. Recent examples include estimation of mortality in the Syrian conflict (Price *et al*., 2015) and the number of modern slaves in the UK (Bales *et al*., 2015). Concepts such as *multiple-systems estimation* (Bales *et al*., 2015) provide analytic frameworks for modelling the observation process with a specific goal of accurate pooling of disparate sources of information.

I encourage continued work to develop similar concepts across our field.

J. Michael Brick (*Westat, Rockville*)
I congratulate Professor Keiding and Professor Louis for their important contribution on a very timely topic. Their case-studies are especially helpful in understanding how self-selection may have variable effects on study findings. The linkages between observational studies and sample surveys that they discuss so well

are extremely useful, yet surprisingly unexplored. I fully agree with their call for wider use of the tools that are available in both types of studies as nicely exemplified in Breslow *et al*. (2009). A key feature that both surveys and epidemiological studies stress is the value of design, and finding some way to bring this concept to self-selected samples would be a major advance.

The differences in the focus of epidemiological studies and surveys that Keiding and Louis note is long standing (Kish, 1959), although at least some of this may be due to terminology (Groves (1989), chapter 1). In surveys representation and measurement are both essential for valid inference, but this may be obscured because survey sampling texts concentrate so heavily on representation. Without appropriate attention to measurement, survey inferences to a population are useless. In recent years survey researchers have made progress in integrating both representation and measurement ideas into total survey error. In epidemiological studies, I can only hope that some of the apparent disdain for the role of external validity is also partially terminological. Henrich *et al*. (2010) gave vivid examples supporting Keiding and Louis's contention that sociological, psychological and medical 'laws' are rarely universal.

I would also like to comment on an important difference in the effect of self-selection for epidemiological and sample surveys that was not discussed by Keiding and Louis. In epidemiological studies, the difference between the treated and untreated ($\Delta = E(Y_t) - E(Y_u)$) is usually the parameter of interest; in sample surveys, estimates of totals and means are typically the primary focus. If self-selection has common effects on both components of Δ as might be hoped, then differencing may reduce the self-selection bias in the key estimate. With Internet sample surveys, a common practice is to use a reference probability sample to weight the self-selected Internet sample but the reference sample is not subject to the Internet self-selection mechanism. Survey estimates of totals from self-selected samples thus do not enjoy any bias mitigating effects of differencing. As a result, surveys and observational studies of prevalence are likely to be even more subject to the effects of self-selection than epidemiological studies of treatment effects.

Miron L. Straf (*Virginia Tech, Arlington*)

Not only can epidemiological studies and surveys each benefit from the other but also scientific inference can benefit from both. Keiding and Louis address the former but in doing so inform the latter. The issues that they raise are so fundamental and important to medicine, public health and public policy that our journal should invite and publish continued discussion.

I suggest here topics for further exploration and discussion. One is non-probability sampling. In market research, low response rates are tolerated because the results are *actionable*. To what extent are statistical surveys with greater validity superior in this context and at what cost? Other topics are false discoveries, replicability and the decline of effects over time, as documented by Ioannidis (2005a, b).

Another topic is application to public policy. As noted by Haskins and Margolis (2014), evaluations of social programmes with randomized controlled trials often show at best only modest effects. What does that imply about combining experiments and observational studies for public policy?

Another topic is the referent or relevant population to which we seek to generalize. Different purposes have different referent populations, may require different analytical approaches (Leek and Peng, 2015) and may focus on different measures of an effect size. Spillover effects (Angelucci and Di Maro, 2015) and networks (Aral, 2015) provide particular challenges for identifying the referent population.

Finally, I suggest exploration of a systems approach. Epidemiological studies and their counterparts for public policy often focus attention on one dependent variable, such as a single effect size. But studies often take place in a complex, often dynamic, system, in which all variables are interdependent.

Today, epidemiology, sample surveys and policy analysis are at a watershed. New data and analytical approaches are coming to the fore, in particular in bioinformatics and policy informatics, as well as methods to analyse complex dynamic systems and data from networks. With this increased ability, we can design treatments, programmes and practices particularized for individuals that are more likely to succeed by taking into account myriad individual, social and contextual factors.

With their consummate review, Keiding and Louis have shown why we must take new approaches and have laid the foundation for doing so.

John L. Eltinge (*US Bureau of Labor Statistics, Washington DC*)

Keiding and Louis have presented a fascinating overview and synthesis of internal and external validity, and related questions of transportability. To complement their development and conclusions, one could consider three additional topics. First, as an extension of issues identified in Sections 3.1 and 4.1, it would be of interest to integrate the authors' ideas with the literature on 'total survey error' and related work with survey data quality, e.g. Andersen *et al*. (1979), Brackstone (1999), Groves and Lyberg (2010), Kenett and Shmueli (2014), Biemer *et al*. (2014) and references cited therein.

Second, one might interpret the discussion of internal and external validity as part of a broader debate regarding optimal allocation of resources within a given area of scientific research, and related communication with stakeholders. The competing approaches that were reviewed by the authors lead to distinct profiles of information quality, cost and risk. For example, a traditional sample survey approach will often require allocation of substantial resources to frame development and fieldwork, but it may reduce risks that are associated with selection bias. Elaborating on a point raised by the authors, degradation of some components of the traditional survey environment (e.g. declining response rates) and development of alternative approaches to unit recruitment and data collection (e.g. through Internet panels) may lead to substantial changes in the prospective balance of quality, cost and risk provided by competing approaches. Thus, it would be of interest to explore in additional detail the practical ways in which changes in data quality, and in related risk profiles, may affect the value that a given scientific study provides to its primary stakeholders. For example, if a certain model coefficient is mistakenly treated as 'transportable' between an observed sample and a patient population, to what extent would a given magnitude of bias lead to degradation of clinical practice?

Third, the empirical research under consideration is intended to inform policy that affects a wide range of stakeholders (e.g. health practitioners, patients and funding sources), and the statistical information that is obtained through that research shares many of the characteristics of a 'public good'. Consequently, in exploration of the linkage between quality of data and stakeholder value, it may be of interest to build on previous literature on cost–benefit analyses for public goods including the distinctions between 'use value' and 'option value' that are attributed to such goods, e.g. Samuelson (1954), Arrow and Fisher (1974), Kahneman and Knetsch (1992), Hamilton *et al.* (2003) and Hess and Ostrom (2006).

(The views expressed in this discussion are those of the author and do not necessarily reflect the policies of the US Bureau of Labor Statistics.)

Xiao-Li Meng (*Harvard University, Cambridge*)

Imagine that you are given two data sets by a data scientist DS. DS swears that the one with $n_r = 100$ is a simple random sample (SRS), which he essentially exhausted his funding to collect. He wants to use its average $\bar{x}_r$ to estimate the population mean $\bar{X}_N$, where N is the population size, but he is concerned that n_r is too small. He therefore decided to purchase the best possible Web-based data set that he could afford. Let us imagine that our civilization has advanced to such a degree that all such Web-based data sets come with a Royal Statistical Society certified *data defect index* ρ, which is the correlation between the included value x and the inclusion propensity or probability $p(x)$.

DS was happy that he bought one with $n_s = 50\,000\,000$ and $\rho_s = 0.05$ (subscript 's' for 'self-selected'). But his happiness was short lived: the average of the 50 million data points, $\bar{x}_s$, is outside the 99% confidence interval obtained from his SRS. 'Which one should I trust more?' This is where you come in. What is your intuition?

50 millions are huge compared with 100, which surely should compensate, in mean-squared error (MSE), for the bias caused by the tiny $\rho_s = 0.05$, right? Seriously wrong, because the sample size comparison is not even relevant. For unbiased estimators such as an SRS average, the MSE is the same as variance and hence it is controlled by the *absolute sample size* n_r. But, for biased estimators resulting from a large self-selected sample, the MSE is dominated (and bounded below) by the (squared) bias term, which is controlled by *the relative sample size* $f_s = n_s/N$. As shown in Meng (2014), to *guarantee* $\mathrm{MSE}(\bar{x}_s) \leqslant \mathrm{MSE}(\bar{x}_r)$, we must require (ignoring the finite population correction $1 - f_r$)

$$f_s \geqslant \frac{n_r \rho_s^2}{n_r \rho_s^2 + 1},$$

or equivalently

$$n_r \leqslant \frac{f_s}{1 - f_s} \frac{1}{\rho_s^2} \equiv \frac{n_s}{N - n_s} \rho_s^{-2}, \tag{1}$$

which means $f_s \geqslant 20\%$ for DS's problem. Therefore, DS's question is unanswerable without knowing N. For example, if his intended population is the USA, then $N \approx 320\,000\,000$, and hence he will need $n_s \approx 64\,000\,000$ to place more trust in $\bar{x}_s$. In other words, his 50 million self-selected observations are equivalent to no more than 75 SRS data points (from expression (1)). If $\rho_s = 0.1$, he will need $f_s = 50\%$ or $n_s \approx 160\,000\,000$ to dominate $n_r = 100$.

In reality DS's SRS will also have a non-zero ρ_r (but hopefully $\rho_r^2 < \rho_s^2$) for all the reasons discussed in

this extremely timely paper, and hence the comparison is more nuanced and less dramatic (Meng, 2015). But the general message is the same: when dealing with self-reported data sets, do not be fooled by their apparent large sizes or by common wisdom from studying probabilistic samples.

Roderick J. Little (*University of Michigan, Ann Arbor*)
I appreciate the authors' thoughtful, nuanced paper. The role of probability sampling was widely argued in early debates over the design of a massive longitudinal epidemiologic study: the US National Children's Study. The study eventually collapsed, perhaps under the weight of excessive ambition and expense. I was a member of the US Federal Advisory Committee for the study in its early days, and I quoted Sir Maurice Kendall as arguing powerfully for probability sampling as the 'scientific' design, in the context of the World Fertility Survey in the 1970s. The Federal Advisory Committee, consisting largely of prominent epidemiologists, voted decisively in favour of probability sampling. For arguments for probability sampling in the National Children's Survey context, see Michael and O'Muircheartaigh (2008), Ellenberg (2010) and Little (2010).

Survey samplers distinguish between descriptive estimands—finite population quantities—and analytic estimands—parameters of a superpopulation model. Some believe that probability sampling is important for the former but not the latter. I disagree. Consider the finite population quantity that is obtained by fitting the model to the whole population. If the parameter estimate is not close to that quantity, then what is the analytic estimate estimating? Measures of association may be less subject to selection bias than means and totals but, when there is significant effect modification with observed or unobserved population characteristics, bias is clearly possible. Perhaps we should focus on 'unobserved effect modifiers' as well as the more standard 'unobserved confounders'.

The formal model-based framework for assessing potential bias, in an age where probability sampling is increasingly unattainable, can be found in Rubin's (1974,1978) work on ignorable selection and treatment allocation. Also, in our 'big data' future, probability surveys must be designed to include variables that link to available administrative data sources.

Robert F. Bordley (*University of Michigan, Ann Arbor*)
Keiding and Louis lead an excellent discussion on the degree to which on-line survey responses are reliable. I would like to comment on a certain kind of on-line survey which was designed—under the leadership of former American Statistical Association President and former Census Director Vince Barabba—while both of us were in the marketing and strategy group at General Motors. The firm has found this particular on-line survey to be very effective.

The survey can be accessed by using the link `myproductadvisor.com` and has been in use for more than 10 years. The survey is packaged as a tool designed to help individuals to identify the vehicles that are most consistent with their needs. As with many on-line surveys, any individual interested in obtaining vehicle advice can take the survey.

In the survey, individuals are asked to provide information about their preferences for different attributes of an automobile. Questions are organized into nine categories: brand, price, appearance, comfort, etc. Within each category, they may be asked potentially many questions. In some cases, they are asked about the kinds of vehicle brands that they will consider. In other cases, they are asked to use a slider bar to indicate the relative importance of various attributes. Individuals choose which categories of questions that they wish to answer. They can skip any question (or category of question) that they are uninterested in answering.

Once the user has provided this information, the tool searches through an on-line up-to-date database of vehicles and provides the individual with a list of the 10 vehicles which are more consistent with those preferences. The usefulness of the recommendations increases with the number of questions that individuals answer and the reliability of the information they provide. If individuals find that the recommendations do not seem appropriate, they are free to change their answers to previous questions and to rerun the tool. Since individuals seek to obtain reliable product recommendations, they have strong incentives to provide good information.

The results of the tool are consistent with the findings of more traditional clinics where individuals are shown different vehicles and interviewed about their preferences. The reliability of the information clearly reflects the value of designing a survey as a by-product of an on-line instrument designed to provide a useful service to the respondent.

The following contributions were received in writing after the meeting.

Garnet Anderson (*Fred Hutchinson Cancer Research Center, Seattle*)
I thank Professor Keiding and Professor Louis for their thoughtful review of the issues that attend with self-selection in surveys and epidemiological studies. I share their scepticism regarding survey results where inadequate attention has been paid to the design, sampling frame and response rates. The availability of powerful and inexpensive tools to collect survey data has outpaced the development of appropriate designs to use them and discounted the need for representativeness.

Epidemiologic studies seem different. Here internal validity seems key. The potential for strong effect modification is perhaps the most compelling reason for considering population-based sampling for epidemiologic studies but this is relatively rare.

The authors cite the Women's Health Initiative hormone trial and the novel efforts by Dr Prentice and colleagues (Prentice *et al*., 2005, 2006, 2008a, b, 2009) to explain the discrepancies between the randomized trial and the parallel observational study as an example. A couple of additional points should be noted. This effort was facilitated by the parallel design and implementation of these two study components—a feature that could be employed more often to expand and improve inference. The observational study participants were also volunteers, subject to the same self-selection issues as trial participants, so generalizability to the larger population of postmenopausal women is not technically established through this work. The primary means of assuring representativeness was through making eligibility as broad as possible. The statistical alignment of the trial and observational study results relied heavily on detailed modelling of timing of exposure, and these effects varied by disease. Traditional confounders or effect modifiers such as age were less influential. No population-based sampling plan at the time would have anticipated these aspects in the trial design (which was age stratified to improve power). To recruit for this trial on a sampling basis would have been impractical.

In the USA where health systems often preclude access to large populations without additional consent, to require epidemiologic studies routinely to start from population-based sampling may negatively impact feasibility and cost without obvious benefits. Nations with better population-based health data systems have the advantage.

Statistical leadership regarding use of these new data collection tools is sorely needed. I hope that the paper by Keiding and Louis serves as the stimulus for developing new designs such as embedded trials and multistage studies to take better advantage of these resources and to facilitate assessment of broader inference.

Bernard Baffour (*University of Queensland, Brisbane*)
A point that has not been made clear, although alluded to by Keiding and Louis, is the issue of representativeness (representativity) and the role of sampling frames in fully covering the population. Surveys have had to evolve because, increasingly, significant sections of the population are not covered. One example of this has been the move to dual frame surveys that incorporate both a list of land-line numbers and mobile numbers as a way of adjusting for the (coverage–non-coverage) bias that is introduced through failing to include the mobile-phone-only population in 'traditional' land-line-based surveys (e.g. Blumberg and Luke (2015) in the USA, Barr *et al*. (2014) in Australia and Arcos *et al*. (2014) in Spain). However, a related issue is that of response (or non-response) bias where there is a differential profile of respondents and non-respondents to the survey, and furthermore who have differential views on the particular topic of interest (e.g. health outcomes and behaviours (Gundersen *et al*., 2014; Livingston *et al*., 2013) or public opinion (Pew Research Center, 2012)). Since technically we have little or no information about non-respondents what tends to happen is to post-stratify to known population information and effectively to adjust and readjust the sample distribution to be close to the population distribution. This population information is sourced from external sources such as census or administrative data (hence external validity). Both types of bias influence representativeness. And, although traditional surveys have made significant inroads in trying to compensate for these biases, it is not entirely clear how these newer survey approaches with self-selection and non-probability sampling can adequately (and transparently) adjust for these and other forms of bias. There is certainly value in learning from the experiences of these surveys (e.g. high response, low cost) and also the rigour of the traditional surveys to develop newer, innovative and efficient surveys in the future.

Jelke Bethlehem (*Leiden University*)
Traditionally, data for general population surveys data are collected by means of face-to-face (computer-assisted personal interviewing) or telephone (computer-assisted telephone interviewing) surveys. Experience has shown that they provide the best quality of data and the highest response rates. Times they are

changing, however. These interviewer-assisted surveys are expensive whereas budgets in official statistics are increasingly limited, they are time consuming and response rates drop. Not surprisingly, researchers look for other modes of data collection, like on-line surveys.

On-line surveys have rapidly become popular as they offer an easy, cheap and fast way of collecting large amounts of data. On-line general population surveys are only possible if Intemet coverage is sufficiently high in the country. This not a problem in the Netherlands and Scandinavia, where the coverage is over 95%. In other countries undercoverage may lead to severely biased estimates.

Another problem for on-line surveys is drawing a probability sample. According to the fundamental principles of survey sampling, a random sample should be selected. This requires a sampling frame of e-mail addresses. This is usually not available for general population surveys. One way to solve this is by sending a letter by ordinary mail with an invitation to visit a specific Web site for filling in the questionnaire. This cumbersome procedure undoes some of the advantages of on-line surveys. To avoid these problems, many on-line surveys rely on self-selection of respondents. People are recruited through commercials, banners and pop-up windows. Usually, respondents are people who like to do surveys or are interested in the topic of the survey. This leads to surveys that are not representative of the target population, and thus to biased estimates; see for example Bethlehem (2010).

Lack of representativity has at least three causes:

(a) people outside the target population can complete the survey questionnaire,
(b) respondents can fill in the questionnaire multiple times and
(c) groups of people may attempt to manipulate survey results.

Several cases of survey manipulation have been reported in the Netherlands; see for example Bethlehem (2015).

Attempts can be made to correct the lack of representativity by applying adjustment weighting. Weighting requires auxiliary variables that are measured in the survey and for which the population distribution is available. Weighting is only effective if two conditions are satisfied:

(a) the auxiliary variables must have a strong correlation with the survey variables and
(b) the auxiliary variables must have a strong correlation with participation behaviour.

See also Bethlehem and Callegaro (2014). Usually, these conditions are not satisfied. Van den Brakel *et al.* (2015) explored the accuracy of 18 self-selection on-line panels. They showed that probability samples perform much better than self-selection samples. They also showed that weighting does not help to solve problems.

It will be clear that, for obtaining accurate estimates of population characteristics, self-selection surveys cannot be used.

Saskia le Cessie (*Leiden University Medical Centre*)

My compliments go to Keiding and Louis for this important paper; self-selection is ubiquitous. It is not only present in studies based on volunteers, but also in all studies where participants' consent is requested, or when specific actions of participants are requested (like filling in a questionnaire), i.e. in all traditional epidemiological studies and clinical trials.

Large amounts of time and statistical efforts are spent to obtain internally valid results. In contrast external validity and generalizability are typically only discussed briefly in the discussion section of a paper. Causal modelling is to a certain degree able to focus on the type of population for which a treatment estimate is valid, by distinguishing between average treatment effects and treatment effects in subgroups like the treated. Still, generalizing treatment effects to external populations remains challenging. Keiding and Louis suggest the use of instrumental variables (IVs) in this process. Whether the use of IVs leads to useful estimates is, however, questionable. In an IV analysis, very strict assumptions such as no treatment heterogeneity are needed to obtain population effects, even if the study sample is randomly selected and perfectly representative of the population of interest. Otherwise IV estimators are local treatment effects in an (unidentifiable) subpopulation or even a weighted average of treatment effects in multiple subgroups of patients (Swanson *et al.*, 2015; Boef *et al.*, 2016).

Keiding and Louis encourage epidemiologists, statisticians and survey researchers to develop new methodology to deal with self-selection bias. Variation in the self-selection processes may be of use here, analogously to using different control groups in a case–control study, each control group with its own (self-) selection mechanism (Pomp *et al.*, 2010). An example of a study with different self-selection processes is the Netherlands 'Epidemiology of obesity' study (De Mutsert *et al.*, 2013), which is a large cohort

study in the Netherlands aiming to investigate pathways to obesity-related diseases. Three recruitment approaches were used: first general practitioners invited their patients to participate; because of the small rate of accrual, additional participants were recruited through advertisements in local newspapers and through posters; finally all inhabitants of three municipalities were invited to participate by using civil registries. This was deemed not to be a major problem, given that the focus was on outcomes after follow-up. This assumption would be confirmed if the differently self-selected subgroups yield similar results; otherwise the variation in the three self-selected subgroups could be used to derive more accurate population estimates. It is a challenge to derive the optimal methodology for doing this.

Anna L. Choi (*Harvard School of Public Health, Boston*) **and Tze L. Lai** (*Stanford University*)

Professor Keiding and Professor Louis have presented a comprehensive and thought-provoking review of the challenges and opportunities in Web-based epidemiological studies. They point out in Section 3.1 the tension between traditionalists upholding the gold standard of sampling-frame-based surveys, and epidemiologists who consider sampling or statistical variations to be of questionable relevance to generalization 'from the actual study experience to the abstract with no referent in place or time'. To the former camp, self-selection in Web-based studies creates insurmountable difficulties for the transportability of the survey results to the population desired. To the latter camp, the Web-based approach opens up accrual much more easily, and there are also methods (such as instrumental variables, inverse probability weighting and multiple imputations) to self-weight the sample values, but Section 3.1 emphasizes the need for

> 'a sampling frame and sampling plan that support weighting results so that they apply to a reference population'

qualifying the sample as representative. Despite the conceptual gap between the two disciplines, Sections 8 and 9 conclude that innovations in data collection and analysis in the present 'big data' era have the potential to bring epidemiology and statistics together to develop improved tools to handle the challenges. We agree and want to add that the burgeoning field of population health sciences is highly interdisciplinary and encompasses not only epidemiology and statistics, but also engineering, information technology and biomedical sciences. Indeed, population health sciences involve Web-based observational data, point-of-care comparative effectiveness trials, adaptive randomization, mobile health and personalized care; see Lai and Lavori (2011), Lai (2014) and Shih *et al.* (2015) for some of these developments.

The 'external validity' of an epidemiological study is often provided out of sample by subsequent studies that reinforce the study's major findings; see Grandjean *et al.* (1997) and Choi *et al.* (2015) for exemplary epidemiological studies that eventually led to a United Nations agreement to control global mercury pollution and advise susceptible populations on fish consumption (United Nations Environment Programme, 2009; US Environmental Protection Agency, 2004). Whereas Section 2.1 criticizes the *Snart-Gravid* study for its self-selection bias, our major issue with its analysis by Mikkelsen *et al.* (2013) is the use of confidence intervals, which are based on the sample being drawn from a target population, for fecundability ratios. Similar studies in other regions and over different time periods can provide more convincing out-of-sample measures of the reproducibility of the study's findings than questionable confidence intervals.

Peng Ding (*University of California at Berkeley*)

I congratulate the authors on their scholarly review of external validity, representation and transportability, and I shall comment on some recent advances related to these topics.

The causal inference literature focused mainly on generalizing experimental results based on pretreatment covariates (Stuart *et al.*, 2011; Hartman *et al.*, 2015) under some *ignorability* assumptions. Intuitively, these *ignorability* assumptions require that the observed pretreatment covariates control 'confounding' between the sampling process and treatment effect heterogeneity. Recently, Ding *et al.* (2016) have proposed randomization-based tests for treatment effect heterogeneity by using potential outcomes. Although rejecting the null hypothesis of no treatment effect heterogeneity beyond observed covariates does not falsify the *ignorability* assumptions (Stuart *et al.*, 2011; Hartman *et al.*, 2015), it does warn us about the existence of latent treatment effect heterogeneity that could potentially jeopardize the external validity of the experimental results. Because the key *ignorability* assumptions for external validity cannot be verified by the observed data, sensitivity analysis by allowing for violations from these assumptions seems an attractive analytic tool. However, the current literature does not provide systematic ways to conduct sensitivity analysis for external validity. Furthermore, Ding *et al.* (2016) reprove Reuter's theorem mentioned in Cox

(1984) and highlight the importance of the scale of the outcome in defining treatment effect heterogeneity and transportability, as discussed by Keiding and Louis.

There is also a growing literature on generalizing experimental results based on post-treatment variables (Pearl and Bareinboim, 2014; Pearl, 2015; Jiang *et al*., 2015), including 'surrogate' as a special case (Prentice, l989; Chen *et al*., 2007; Ju and Geng, 2010). We can evaluate the causal effect of a treatment on a surrogate from a randomized trial and gather information about the causal or associative relationship between the surrogate and outcome from other studies. The final goal is to predict the causal effect of the treatment on the outcome. The randomized trial and other studies may not represent the same population, and combining these two sources of information requires some untestable assumptions about transportability. Even if we agree with Keiding and Louis's statement that 'conditional effects more often can be expected to be transportable', we still face some practical questions: what to condition on?; what remains the same across populations? Jiang *et al*. (2016) suggest conditioning on the joint potential values of the surrogate, i.e. the *principal stratification* (Frangakis and Rubin, 2002) and assume that the values or signs of the causal effects within principal strata remain the same across populations. Jiang *et al*. (2016) give sufficient conditions that ensure correct prediction of the value or the sign of the causal effect on the outcome by using the causal effect on the surrogate. In practice, we may have other types of post-treatment variables that could help to explain the causal mechanisms and aid transportability across populations (VanderWeele, 2015).

David Draper (*University of California at Santa Cruz*)

I commend Keiding and Louis for focusing on such a crucial—and yet insufficiently discussed—topic, and I offer two comments.

(a) Table 3, from Draper (1995), reinforces the authors' position on the vital but awkwardly worded concept of *representativity*, by cross-tabulating *strong ignorability of treatment assignment* (Rosenbaum and Rubin, 1983) in experiments against *exchangeability of sampled and unsampled units in the target population* (e.g. Greenland and Draper (2014)) in sample surveys. When both of these assumptions are unjustifiable, it is still possible to do a weak form of *calibration inference* (see Draper (l995)) on the observational units in a given study, but outward generalization is not supported. If one of the two assumptions is justifiable, *specific causal inference* to the units in the experiment or *sampling inference* to the unsampled units in the population are valid, but it is only with both assumptions justified that general causal inference to the entire population is supported. The Web-based surveys and observational, studies of which the authors speak unfortunately typically fall into the upper-left-hand cell of Table 3, and yet it is all too frequent to see people making (much) stronger inferential claims than such data sets actually justify.

(b) Given the crucial nature, for valid inference, of the ignorability and exchangeability assumptions in Table 3, I conclude with an immodest proposal. *Every* paper, in any discipline, that employs statistical inference to draw real world conclusions should be required to address explicitly the following question, perhaps as part of a structured abstract:

> What is the broadest scope of valid generalizability outward from the observational units in the data set employed in this paper?

The obvious glib answer to this question—all units similar in all relevant ways to the units on which this paper is based—is inadequate until the authors take explicit positions on the meaning of *similar*

Table 3. Types of inference supported by various sampling and design assumptions, from Draper (1995)

		Strong ignorability of treatment assignment	
		Difficult to justify	*Justifiable*
Exchangeability of sampled and unsampled units in target population	Difficult to justify	Calibration inference	Specific causal inference
	Justifiable	Sampling inference	General causal inference

and *relevant*. Fisher (1956), who was right on so many things, was wrong on this point, when he encouraged investigators to generalize inferentially to the population of all possible data sets that might have been generated by the investigator's data collection method; this is essentially just the glib answer to the above question and is insufficiently responsive to problem context.

Michael Elliott (*University of Michigan, Ann Arbor*)

Keiding and Louis have provided an excellent discussion of the issues regarding the use of convenience or more generally non-probability samples in epidemiological research. My comments focus on the point that randomization negates the influence of *unobserved confounders*, whereas representative sampling negates the influence of unobserved *effect modifiers*. To illustrate, suppose that we have an outcome Y, an exposure A and an *unobserved* U that is both a common cause of A and Y (confounder) and modifies the effect of A on Y (interaction), and the true model generating the data is of the simple linear form $E(Y|U, A) = \beta_0 + \beta_1 A + \beta_2 U + \beta_3 AU$.

Suppose further in specific population Θ that we have

$$\begin{pmatrix} U \\ A \end{pmatrix} \sim N\begin{pmatrix} \mu_U & \sigma_U^2 & \sigma_{UA} \\ \mu_A & \sigma_{UA} & \sigma_A^2 \end{pmatrix}.$$

If we correctly estimate $E(Y|A)$ by using a simple random sample from Θ using a naive regression model, we find that Y is actually non-linear in A:

$$E(Y|A) = \beta_0 + \alpha_0\beta_2 + \{\beta_1 + (\alpha_0 + \alpha_1)\beta_3\}A + \alpha_1\beta_3 A^2,$$

for $\alpha_0 = \mu_U - (\sigma_{UA}/\sigma_A^2)\mu_A$ and $\alpha_1 = (\sigma_{UA}/\sigma_A^2)A$. If we are interested only in describing associations between Y and A in Θ, we could stop here. But, if we are interested in the *causal effect* of A on Y, then, in the language of counterfactuals,

$$E\{Y(A = a_2) - Y(A = a_1)\} = E_U[E\{Y(A = a_2) - Y(A = a_1)|U = u\}] = \{\beta_1 + \beta_3 E(U)\}(a_2 - a_1). \quad (2)$$

If we return to our sample from Θ and assume that we can break the association between U and A, by using a set of covariates X so that then $\alpha_{0|X} = \mu_U$, $\alpha_{1|X} = 0$, and the coefficient in the marginal model associated with A becomes $\beta_1 + \mu_U\beta_3$. The question then becomes what the correct value for $E(U)$ in equation (2) is. If our target is the expected value of U in the population Θ, then an alternative sampling scheme that does not yield an unbiased estimator of U will not yield a correct estimate of equation (2) even if the unobserved confounding is accounted for by X.

Thus we need both the internal validity of randomization and the external validity of representative sampling to estimate equation (2) correctly. As noted by Keiding and Louis, Rothman *et al.* (2012) made a vocal case that equation (2) is *not* a sensible scientific target. Instead, we need to unearth U and to make it observed, so that the correct inferential target is $E\{Y(A = a_2) - Y(A = a_1)U = u\} = (\beta_1 + \beta_3 u)(a_2 - a_1)$. But is that always possible? If we have X that are sufficient to break the association between A and U, should we focus on assessing interactions between A and X as an approximation to the interactions between A and U? If we are interested in a causal effect across a well-defined population, then using probability sampling to obtain a consistent estimator of equation (2) seems more appropriate.

Colin B. Fogarty and Dylan S. Small (*University of Pennsylvania, Philadelphia*) **and Joseph L. Gastwirth** (*George Washington University, Washington DC*)

This important paper discusses the conditions under which the findings of an epidemiological study have external validity for the population from which subjects were taken or preferably other populations. A first step is showing internal validity, i.e. inferences are unbiased with respect to the initial study group. Randomized experiments are internally valid because randomization removes any association between the assigned treatment and the potential outcomes. Analysing non-randomized observational studies requires procedures that control for confounding variables.

Self-selection often creates noticeably different covariate distributions in the two groups being compared.

Statistical methods adjusting for covariate imbalances require a substantial overlap in these two distributions; otherwise inferring that the treatment effect applies to the entire study group requires extrapolation of counterfactuals where common support does not exist. To preserve internal validity in the absence of covariate overlap, one should make inferences for a subset which has sufficient overlap. For internally valid inferences to be transportable to populations of interest, it is useful for this subset to be well defined and readily interpretable. Fogarty *et al.* (2015) used the maximal box problem to define a study group which

exhibits verifiable overlap on important covariates while being easily understood in terms of covariates themselves. This allows for internally valid findings with respect to a subset of the data, which may lead to external validity for some, but not all, larger populations.

Section 7 describes the purpose of randomization in treatment assignment for an experiment and in recruitment of survey participants. The authors cite Wirth and Tchetgen Tchetgen (2014) who demonstrated the need to account for survey design to reduce selection bias, noting that a failure to do so may lead to imbalances that are associated with the exposure or the outcome, thereby precluding external validity.

In non-randomized studies, sensitivity analyses are commonly used to assess the internal validity of a study's findings to unmeasured confounding; see Cornfield *et al*. (1959), Rosenbaum (2002), Yu and Gastwirth (2005), Wang and Krieger (2006) and Ding and Vanderweele (2014). Similar methods can be used to assess the potential influence of non-random selection on external validity. Gastwirth (2003) developed sensitivity analysis for mall intercept surveys used to demonstrate trademark infringement or misleading advertising, where the relevant population is potential consumers. Although prescreening occurs, inevitably some individuals who are not potential purchasers are included in the survey. Cornfield's inequality was modified to assess the fraction of the sample comprised of inappropriate respondents and the increased probability that they have to being misled required to reduce an observed degree of confusion to a legally acceptable level (say, less than 10%).

Joel B. Greenhouse (*Carnegie Mellon University, Pittsburgh*)

Keiding and Louis provide a comprehensive review of current issues related to self-selection and generalizability in epidemiological studies and surveys. The central issue, of course, is not advocacy for internally *versus* externally valid studies but rather for research programmes that advance scientific discovery and bring differing opinions concerning causal effects to consensus. Experience suggests evidence from both types of studies is necessary (see, for example, Cornfield (1959) and Cornfield *et al*. (1959)).

The field of comparative effectiveness research provides a good illustration. A goal of comparative effectiveness research is to help stakeholders to reach consensus about the benefits and harms of medical interventions, both at the individual and at the population levels. Questions that are key to effectiveness research include whether a treatment effect is homogeneous and replicable in a specific population and whether it is generalizable to other populations and setting. Since a single study is rarely sufficient for answering questions such as these, we have argued that it is necessary to generate, synthesize and weigh evidence from multiple sources of data, including experimental and observational studies (Greenhouse and Kelleher, 2005). To this end, my colleagues and I have been developing methods for combining and interpreting data from multiple studies that both capture the strengths and minimize the weaknesses of different study designs (see, for example, Greenhouse *et al*. (2015) and Kaizar (2015)). Much more work needs to be done and I welcome Keiding and Louis's call for statisticians, epidemiologists, survey researchers and, I would add, subject matter specialists to continue to work together to understand these research challenges better and to develop new methods to address them.

Finally, it is interesting that like Keiding and Louis, who have expertise in both epidemiological and survey research, some of the earlier contributors to modern epidemiological methods also had strong foundations in survey sampling, e.g. W. G. Cochran (Rothamsted), J. Cornfield (Bureau of Labor Statistics) and S. W. Greenhouse (Census Bureau). I suspect that there is a clarity of thinking about issues, such as target and study populations, sampling frames, weighting, clustering and non-response, that provides a useful framework for approaching methodological problems in epidemiology. This paper serves as an excellent reminder of the synergy between applications and methods development and the key role for statistical thinking in what some are now calling the data sciences.

David J. Hand (*Imperial College London*) **and Agnes M. Herzberg** (*Queen's University, Kingston*)

Keiding and Louis provide a timely exploration of the merits and demerits of self-selected entry to epidemiological studies and surveys. Such studies have merits. In particular, the possibility of speed means that they have the potential to give results essentially immediately in situations where classic approaches may take much longer. This has clear importance for economic decision making, social policy and other areas. However, self-selection also means that the potential for bias is greater. As the authors point out, it all boils down to a balance between sources and sizes of different aspects of data quality, especially timeliness, bias and variance. Even nowadays, time pressures are such that in the UK gross domestic product estimates are released 25 days after the quarter, based on only 44% of the data (which explains subsequent revisions as more data arrive).

Since self-selected surveys are here to stay, we believe that more focus should be put on developing

accessible model-based ways of improving self-selected survey results, and of publicizing these strategies, along with publicizing the potential disadvantages if they are not followed. This paper provides an excellent start. Perhaps publication of the invalid and misleading conclusions of some high profile self-selected surveys would be a useful next step.

The growth of interest in 'big data' has aggravated the situation. Often implicit in the notion of big data is the idea that one has 'all' of the data. This is often not so, even in situations involving administrative data.

The authors do not mention another legitimate and valuable use of self-selected surveys, i.e. characterizing the span of possible responses. This is neither conditional nor marginal (provided that there is at least one response in all potential categories), but it is a necessary precursor to other, carefully designed, studies.

We endorse the three summarizing conclusions and would like to congratulate the authors on the timeliness of their paper.

Michael G. Hudgens and Jason P. Fine (*University of North Carolina at Chapel Hill*)

We congratulate Keiding and Louis for this outstanding contribution contrasting inference in epidemiology and survey sampling. The primary focus of survey sampling entails externally valid inference from a study sample to a specified target population. In contrast, Keiding and Louis contend that recent analytic epidemiology emphasizes internal validity, with generalizability addressed informally, leading to unclear external conclusions about ill-defined target populations. They recommend that epidemiology adopts a more rigorous, quantitative approach to externally valid inferences by using recently developed epidemiologic methods rooted in causal inference and survey sampling fundamentals.

Keiding and Louis focus on estimation of relative risks (or risk differences) characterizing associations between outcomes and exposures, and their generalizability to reference populations. Applications in which absolute risks are of interest pose different but related challenges. For example, the Population Surveillance Group at the US National Cancer Institute aims to develop risk calculators for the US population, providing estimated probabilities of death from cancer and death from other causes following diagnosis. Often multiple sources of data, drawn from potentially different populations, may be used in these analyses. Even if the relative risks in such data sets are the same as in the general population, concerns may arise regarding target population validity if baseline risks differ from the target population.

Keiding and Louis mention instrumental variables (IVs) for addressing unmeasured confounding. IV methods have been widely adopted to obtain within-study validity in social sciences. Recently, IVs have received increased attention in epidemiology, particularly in comparative effectiveness studies involving secondary analyses combining multiple population-based data sets. Questions arise in this setting about the study population definition and the precise conditions under which IVs provide valid effect estimates in such populations. A related question is whether IV methods might be adapted for transporting inferences to well-defined target populations.

Traditionally survey sampling has not focused on confounding or, more generally, causal inference. In contrast, a primary focus of epidemiology is determining which exposure(s) cause disease and many statistical methods for causal inference have been developed by epidemiologists. A key assumption in causal inference is no interference, i.e. the treatment or exposure of one individual does not affect other individuals' outcomes. In many scenarios this assumption may not hold. Web-based studies are one setting where interference may be present, e.g. if study participants interact on the same social network Web site. In these settings formal methods for transportability must consider selection bias, confounding and interference.

Jay S. Kaufman (*McGill University, Montreal*)

Keiding and Louis provide an extensive engaging discussion. They mention superpopulation sampling as a basis for statistical inference (Section 4.2), and randomization as another basis (Section 7.1). But what is the justification for statistical inference in a cohort that is neither representatively sampled nor randomized? Papers from the Nurses' Health Study, for example, include p-values, but what exactly do these numbers mean? Greenland suggested some possible interpretations, but these remain qualitative because of the absence of any underlying anchor for formal inference (Greenland, 1990). The dangers inherent in foregoing such an anchor are exemplified by the hormone replacement therapy example (Section 7.5), in which confusion about inferential targets led to accusations that observational studies are inherently suspect. The eventual resolution of this putative conflict between designs vindicates the perspective of Keiding and Louis.

There is a glimmer of logic in the opposing narrative when one asks only whether exposure is a cause of disease. For example, does smoking cause lung cancer? With emphasis only on rejecting the null hypothesis, it may not matter much who is the target because, if smoking causes cancer in anyone, that is a finding. Focus on estimation follows later as a refinement. Sympathy for this perspective evaporates, however, when we consider that, within selected samples, associations may lack causal interpretations even absent confounding. This is the saga behind many fumbles eventually explained by selection bias, including low birth weight and obesity paradoxes (Hernández-Díaz *et al*., 2006; Banack and Kaufman, 2014).

Keiding and Louis review some historical definitions of representativeness (Kruskal and Mosteller, 1979), without offering alternative models. One is that generalizability rests on balance between potential response types (Robins, 1988). For example, consider the latent stratum of people who suffer the outcome if exposed, but not if unexposed. The proportion of such individuals can differ between samples via several mechanisms, including selective exit via competing risks (Flanders and Klein, 2007). It is evident that the causal effect is a function of this proportion in the study, and therefore that its constancy is a critical consideration. This is perhaps more obvious for econometricians because of their focus on 'local' treatment effects that are specific to such latent strata, and it may therefore explain the attention to target populations in economics while this matter is denigrated in the *International Journal of Epidemiology* (Heckman and Vytlacil, 2007; Rothman *et al*., 2013).

The essay by Keiding and Louis is already ample and wide ranging, but it is a shame to posit that conditional effect estimates are more generalizable than marginal estimates (Section 1) without any discussion of collapsibility (Duan *et al*., 2008). Likewise, it seems unfortunate to leave unchallenged the suggestion (Section 7.4.2) that relative effect estimates are more conserved across studies than absolute estimates (Poole *et al*., 2015).

Thomas King (*Newcastle University*)

The nature of a population is unknowable and constantly changing but adjusting sample estimates by stratification relies on good information on the population. This can allow for some problems in coverage; for example household surveys omit those not in standard households just as Web surveys omit those without access. A census can act as a population base but uses independent first-order adjustments (Baffour *et al*., 2013) and estimates a response rate of the order of 95%. Indeed all of the compensation for confounding, post-stratification or weighting relies on linear models.

In epidemiology, most health outcomes obey a social gradient (Wilkinson and Marmot, 2006), which is only partly explained by factors understood as a mechanism. But the need to include social status as a confounder is complicated by the influence of social status on the propensity to participate in research. Social biases in recruited samples should be expected and post-stratification is indicated. However, social status is a difficult-to-measure construct which at best follows an ordinal scale, often captured by composite indicators (Bennett *et al*., 2009), yet ordinal measures miss the detail of confounding, or the full shape of the distribution. So the sample should be designed with external validity in mind when the confounding relationship is complex and not completely understood.

In child development, when we compare between longitudinal analyses, we need to make a judgement about the inferential basis for comparing and contrasting populations, and most analyses cannot decide measurement and cultural explanations for differences (Ermisch *et al*., 2012). The child's environment influences aspects of development differentially, so sophisticated longitudinal data are required (Shonkoff and Phillips, 2000). Advance has been limited by disciplinary focus either on issues of sampling (sociology) or psychometrics (psychology) when what is needed is better design which delivers both (West *et al*., 1998). Longitudinal design for internal validity only will limit reanalysis and realize confounded estimates of unknown or non-stationary mechanisms.

Linear models in general have their limitations, and so quantile estimation is appropriate for some, e.g. psychosocial outcomes (Tzavidis *et al*., 2016). Study populations can omit those at greatest risk such as the poorest or those with unstable housing tenure. An inference implies a population, and unbiased estimation: both assumptions can be further tested and improved, as can the model in general. External validity gives a source of calibration and model criticism which is always needed when further data collection is a prohibitive undertaking.

Han Liu and Yang Ning (*Princeton University*)

We congratulate Keiding and Louis for making a timely and thought-provoking contribution. In particular, we found the 'big-data'-related discussion in Section 7.4 interesting. Here we make two comments along this direction. The first is about a formal framework that highlights the potential benefit of using extra

data to alleviate the problem of self-selection. The second is an idea that brings the experimental design dimension into big data analysis.

Big data may alleviate the problem of self-selection

First, we present a formal framework that illustrates the advantage of big data for inferring the population mean. Assume that we are interested in estimating $\mathbb{E}(Y)$, where Y is randomly sampled from a target population. Let T be a 0–1 random variable denoting whether Y is observed or not, i.e. we observe Y only if $T=1$. If Y and T are not independent, the mean of Y after self-selection is typically not identical to the overall mean, i.e. $\mathbb{E}(Y|T=1) \neq \mathbb{E}(Y)$. To calibrate this bias, the main challenge is that there may not be enough information in self-enrolment studies to adjust for the propensity score. However, big data may alleviate this issue. By exploiting the public data (e.g. social media) that are available on the Internet, we can acquire a massive amount of information for both respondents and non-respondents. We denote these variables by $\mathbf{X}$. Thus, we can impose a parametric (non-parametric or semiparametric) surrogate model $\pi(\mathbf{X})$ to approximate the propensity score $\mathbb{P}(T=1|\mathbf{X})$. Given n independent and identically distributed copies of $(Y,T,\mathbf{X})$, the Horvitz–Thompson estimator (Horvitz and Thompson, 1952) of $\mathbb{E}(Y)$ is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(\mathbf{X}_i)},$$

where $\hat{\pi}(\mathbf{X})$ is an estimator of $\pi(\mathbf{X})$. If $\mathbf{X}$ is of high dimension, it is often reasonable to assume that the propensity score $\mathbb{P}(T=1|\mathbf{X})$ has a sparse representation. For instance, under the logistic model assumption, we can estimate $\hat{\pi}(\mathbf{X})$ by $\pi_{\hat{\boldsymbol{\beta}}}(\mathbf{X})$, where $\hat{\boldsymbol{\beta}}$ is the penalized maximum likelihood estimator

$$\hat{\boldsymbol{\beta}} = \arg\max \frac{1}{n}\sum_{i=1}^{n}\left\{T_i\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{1+\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}\right\} - \mathcal{R}_\lambda(\beta).$$

Here, $\mathcal{R}_\lambda(\boldsymbol{\beta})$ is some penalty function with a tuning parameter λ. Under some regularity conditions, the Horvitz–Thompson estimator with $\pi_{\hat{\boldsymbol{\beta}}}(\mathbf{X})$ has good theoretical properties such as consistency and asymptotic normality.

More sophisticated data acquisition helps big data analysis

Second, we comment that big data analysis can be benefitted by more careful data collection processes. When fitting undirected graphical models, the presence of hidden variables is a threat to valid inference on the graph structure. Tan *et al.* (2015) showed that the effect of hidden variables can be adjusted by collecting multiple measurements for each subject. The main reason is that, with multiple measurements per subject, we can exploit careful conditioning arguments to treat hidden variables as a nuisance. Thus, more sophisticated data acquisition is beneficial to big data analysis.

Xavier de Luna (*Umeå University*)

I enjoyed reading this thought-provoking paper. I believe that some new light can be shed on the important issues discussed by adopting a Popperian standpoint.

Let me start by fully agreeing that surveys based on probability sampling are an invaluable tool to provide predictions of parameters for well-defined populations. Such predictions can in principle be empirically corroborated through independent surveys or administrative registers.

In epidemiology (and other empirical sciences like economics) where causal parameters are often targeted, the predictions of interest are arguably of a different nature. Consider the fictitious example of an observational study yielding a risk of lung cancer which is five times higher when comparing those smoking during 30 years with non-smokers. What are the predictions that we are willing to make from this study based on a given sample? *Internal validity* refers to the prediction that if everyone in the sample would have chosen to smoke during 30 years then we would have observed five times more lung cancer cases than if all these individuals would have chosen not to smoke at all: this prediction cannot be corroborated empirically. Thus, 'internal validity' is a metaphysical concept that is often considered useful for causal reasoning and inference (Neyman, 1990; Rubin, 1974; Dawid, 2004). *External validity* the (representativity version in the paper) is then about the need to extend the latter prediction to a well-defined population.

My Popperian interpretation of the second *external validity* concept discussed in the paper ('generalization to the abstract': Miettinen (1985)) is that 'the abstract' refers to predictions that can be corroborated empirically. From the above lung cancer study, we may predict that taking another sample with similar characteristics, and randomly assigning individuals to smoking (30 years) or not smoking, will yield a larger proportion (closer to five times larger) of lung cancer among smokers. This prediction can be

checked in 30 years from now (although difficult to implement). Another prediction, which is easier to corroborate, is that a higher incidence of lung cancer is expected when comparing smokers with non-smokers in future observational studies conducted on samples with different geographic, genetic, and/or socio-economic characteristics. Thus, when causal parameters are targeted, probability sampling is not paramount (although related issues such as outcome-based sampling and attrition by death may need treatment). What is essential is corroboration of the results of one study by other studies performed under different conditions.

Jorge Mateu and Pau Aragó (*University Jaume I, Castellón*)

Keiding and Louis are to be congratulated on a valuable contribution and thought-provoking paper on this timely topic of surveys based on self-selection of responders through social media tools. This clearly involves a problem related to sampling in the classical fashion confronting this with a more modern vision of sampling, and with the way that information is treated and saved, and further statistically analysed.

Social media (e.g. Facebook, Twitter and Flickr) have become a social fabric of our society. By simplifying the sharing and dissemination of user-generated content, social media have changed the way that individual level information is generated, distributed and exchanged. Massive streams of social media data provide alternatives to traditional data collection approaches for understanding people's opinions and observations and, thus, are increasingly investigated by researchers in many domains.

Massive numbers of social media users are engaged at any moment to view and generate content, and publish where they are along with the content they generate. Hence social media can be regarded as a major public spatiotemporal source of data. And with the social media spatiotemporal characteristics in mind, in a very particular form in epidemiological studies, we have several aspects to comment on.

(a) Social media are a useful platform for rapidly reaching and enrolling large numbers of people who are not reachable via traditional techniques. Social-media-based recruitment has the potential to expand the geographic reach of investigators and to identify potential participants more cheaply than traditional approaches. However, social-media-based recruitment may raise concerns about data accuracy and ethical handling of information.
(b) Space–time data are becoming available in overwhelming volumes and diverse forms as a result of crowd sourcing and citizen science data. The processing of such diverse and massive data poses conceptual, methodological and technical challenges, which are exacerbated by the diversity of data. Sufficiently flexible and powerful solutions that explicitly exploit the space–time ordering are not available to date, because existing methods were not designed for global, high volume, hyperdimensional, heterogeneous and uncertain space–time data.

Paul D. McNicholas (*McMaster University, Hamilton*) **and Sanjeena Subedi** (*University of Guelph*)

We congratulate Keiding and Louis on an interesting and timely contribution. They point out that the availability of 'big data', together with advances in statistics and computer science, will help to address population goals (the first paragraph in Section 7). Big data are quite often defined in terms of three or more Vs, and the authors quote a sentence from Japec *et al*. (2015) where a three-V definition is provided. In the sentence quoted, it is highlighted that big data are often

> 'characterized not just by their large volume, but also by their variety and velocity, the organic way in which they are created, and the new types of processes needed to analyze them and make inference from them'.

The authors give some specific examples of big data, including voting records and a breast cancer database (Section 7.4). When the authors speak to the benefits of big data, are voting records and databases the sort of data they are referring to?

The issue here goes beyond semantics and it is important to understand clearly what the authors mean by big data. This is, perhaps, particularly important as it pertains to a key issue in big data applications: error correction. The authors provide details of case-studies and note the work of Ansolabehere and Hersh (2012), where careful analyses were performed either to validate the findings of surveys or to explore the discrepancies between a survey and actual records from a population of interest. However, what is called big data in these cases seems to be administrative data, e.g. voting records and databases. In a very nice article Puts *et al*. (2015) discuss challenges around finding errors in big data. In doing so, they draw a distinction between big data and administrative data, writing 'Administrative data can be high-volume, but differ from Big Data with respect to velocity and variety'. They later re-enforce the relative simplicity of dealing with administrative data, clarifying that

> 'Having gained such experience editing large administrative data sets, we felt ready to process Big Data. However, we soon found out we were unprepared for the task.'

We wonder whether the authors might comment on big data, as opposed to administrative data, in the context of epidemiological studies.

Fionn Murtagh (*University of Derby and Goldsmiths University of London*)
Interesting perspectives that support Keiding and Louis include Friedman *et al*. (2015), and the following quote, from Laurison and Friedman (2015):

> '... the GBCS [Great British Class Survey] data have three important limitations. First, the GBCS was a self-selecting web-based survey, ... This means it is not possible to make formal inferences. ... the nationally representative nature of the Labour Force Survey (LFS) along with its detailed and accurate measures ... facilitates a much more in-depth investigation. ...'

In a blog posting, Laurison (2015) pointed very clearly to how, just 'Because the GBCS is not a random-sample or representative survey', other ways can and are being found to draw great benefit (`http://www.thesociologicalreview.com/information/blog/three-myths-and-facts-about-the-great-british-class-survey.html`).

Another different study on open, free-text questionnaires (Züll and Scholz (2011); see also Züll and Scholz (2015)) notes selection bias, but also

> 'However, the reasonable use of data always depends on the focus of analyses. So, if the bias is taken into account, then group-specific analyses of open-ended questions data seem appropriate.'

The bridge between the data that are analysed and the calibrating 'big data' is well addressed by the geometry and topology of data. Those form the link between sampled data and the greater cosmos. Eminent quantitative and qualitative sociologist Pierre Bourdieu's concept of field is a prime exemplar. Consider, as noted by Lebaron (2009), how Bourdieu's work involves 'putting his thinking in mathematical terms', and that it 'led him to a conscious and systematic move toward a geometric frame-model'. This is a multi-dimensional, 'structural vision'. Bourdieu's analytics

> 'amounted to the global [hence big data] effects of a complex structure of interrelationships, which is not reducible to the combination of the multiple [... effects] of independent variables'.

The concept of field, here, uses geometric data analysis that is core to the integrated data and methodology approach used in the correspondence analysis platform (Murtagh, 2010).

An approach to drawing benefit from big data is precisely as described by Keiding and Louis. The noting of the need for the 'formulation of abstract laws' that bridge sampled data and calibrating big data can be addressed, for the data analyst and for the application specialist, as geometric and topological.

Ross L. Prentice (*Fred Hutchinson Cancer Research Center and University of Washington, Seattle*)
I also thank Professor Keiding and Professor Louis for their thoughtful and timely paper, and for their generous remarks on some of our Women's Health Initiative analyses. In spite of the typical greater transportability of ratio *versus* absolute risk measures in epidemiology, interactions of exposures (or treatments) under study with study subject characteristics can be anticipated with some frequency even for such relative measures as odds ratios or hazard ratios. After all, statistical interaction is defined in a model-dependent fashion, and lack of interaction on one scale will typically translate to some degree of interaction on another.

Timing issues, such as duration of exposure, or ages at exposure, may also be important determinants of the hazard ratio function: we studied transportability from the Women's Health Initiative randomized hormone therapy (HT) trials to the companion observational cohort study, with women drawn from the same catchment population, followed under a common overarching protocol, over the same time period. Even in this nearly optimal context for transportability, allowing the hazard ratio to depend on the duration of HT was necessary for common cardio-vascular disease findings, as was noted by Keiding and Louis. For breast cancer, however, controlling for duration of use, standard confounding factors and mammography use patterns was insufficient. Another timing variable, years from menopause to first use of HT, seemed important for this (Prentice *et al*., 2008a,b), with higher hazard ratios among women who start HT at or soon after the menopause. This timing variable is contributing to a valuable discussion concerning the biology of HT effects in relation to undiagnosed breast cancers, breast involution following menopause

and competition for breast epithelial cell receptors, illustrating the potential for descriptive analyses to complement related basic science research.

These analyses suggest that ratio measure reporting may need to be broken out by key timing variables to enhance transportability. Additional enhancement may arise from standardization to the distribution of a specific distribution of potentially interacting variables. An important interface between epidemiology and surveys could transpire if survey cohorts, having suitable probability frames and substantial specimen and data repositories, were maintained to provide exposure and confounding factor information on the larger target population, for joint analyses of survey data with corresponding epidemiologic cohort data. I would be interested in Keiding and Louis's thoughts on the potential of this type of research agenda development for ratio measure transportability and, especially, for the burgeoning enterprise of absolute risk calculators.

Kenneth J. Rothman (*Research Triangle Institute and Boston University School of Public Health*), **Elizabeth E. Hatch** (*Boston University School of Public Health*), **Charles Poole** (*University of North Carolina at Chapel Hill*), **Lauren A. Wise** (*Boston University School of Public Health*), **John E. J. Gallacher** (*Cardiff University*), **Timothy L. Lash** (*Emory University, Atlanta*) **and Ellen M. Mikkelsen and Henrik T. Sørensen** (*Aarhus University Hospital*)

If you assume that the world is flat, then flat is what you will find. Keiding and Louis describe survey research as the 'gold standard', an assumption rather than a fact. Their doctrine overlooks the reality that epidemiologic studies have various goals. Some share survey goals of describing a specific population, but others are undertaken to describe associations that would apply to people in general, especially future people, i.e. they have scientific goals (Rothman, 2010).

Keiding and Louis assert that, without a survey sampling frame, 'findings pertain only to the actual participants'. As there is no survey sampling frame for most randomized trials, they imply that the results of these randomized trials pertain only to the study patients. If true, that would be an argument against undertaking such studies, which are designed to produce knowledge about future patients, not just the patients being studied. How does one sample from future patients? Keiding and Louis propose that it is somehow necessary:

> 'Epidemiological and clinical studies that purport to make generalizable conclusions need to operate at least to a degree as a survey'

because in their doctrine external validity comes only from sampling representativeness. They go on to say, effectively, that trials should include, rather than eligible and consenting patients, a random sample of target populations who are then assigned randomly to treatments. Their view stops just short of Miettinen's joke that devotion to representativeness would require studies of laboratory animals to trap rats or mice randomly from the general rodent population rather than using syngenic animals.

Keiding and Louis question the generalizability of the *SnartGravid* cohort because it was recruited via the Internet, but then they praise the study of Schisterman *et al*. (2014) because it

> 'used Facebook as a recruiting mode (E. F. Schisterman, personal communication), but enrolment still depended on qualifying for the study'.

Yet, both these features also applied to *SnartGravid*. They went on to say

> 'This use of the Internet causes little concern over that associated with traditional recruitment methods and may be the most effective way to accrue to well-designed studies'.

So recruiting via the Internet is a bad idea, because it has no identifiable sampling frame, unless one recruits via Facebook.

We agree that selection bias may arise in cohorts that are recruited via the Internet, as in other settings, but lack of sampling representativeness is not the underlying problem (Rothman *et al*., 2013). Indeed, a recent comparison of *SnartGravid* and registry data provided empirical evidence against selection bias (Hatch *et al*., 2015).

Martin Schumacher (*University of Freiburg*), **Jan Beyersmann** (*University of Ulm*) **and Nadine Binder** (*University of Freiburg*)

We congratulate our colleagues Niels Keiding and Tom Louis for their excellent review on the challenges of self-selected entry to epidemiological studies. We appreciate the opportunity to contribute to the discussion

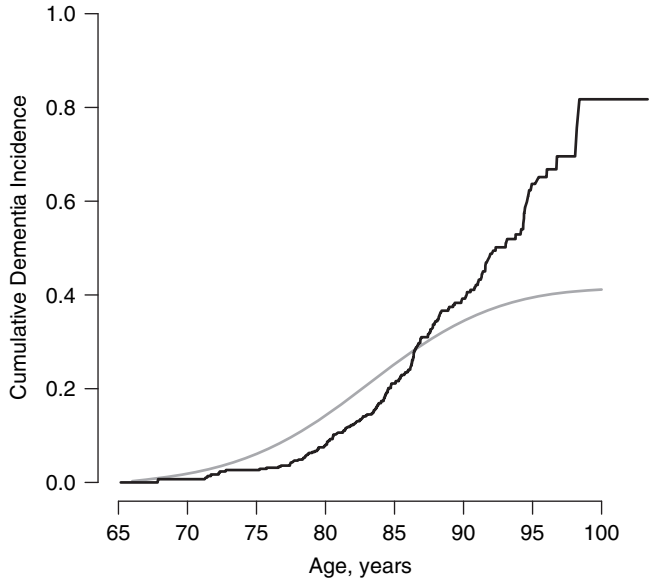

Fig. 2. Comparing two different approaches towards estimating age-specific cumulative dementia incidence: we consider the random sample of the French elderly-population-based PAQUID cohort study of mental and physical aging data provided with the R package SmoothHazard (Touraine *et al.* (2013); for study details see Joly *et al.* (2002)); using the R package we estimate and plot the cumulative dementia incidence model-based results from the illness–death model (———, adequate analysis) and compare them with the 1 minus the Kaplan–Meier estimate (———, inadequate for this kind of data)

by pointing to the further aspect of follow-up within studies where Web-based technologies may also provide an attractive alternative to traditional approaches. Whereas Keiding and Louis briefly strive on that issue (Sections 5.4 and 5.5: 'Are longitudinal analyses protected?' and 'Relationship to missing data'), we shall outline some additional sources of bias that must be considered but can hopefully be avoided. Using multistate models (Andersen and Keiding, 2002; Beyersmann *et al.*, 2012), Grüger *et al.* (1991) provided conditions for 'non-informativeness' of a sampling scheme, i.e. the sequence of examination times where the disease state of an individual is observed. Here, the likelihood that is obtained is proportional to the likelihood that we would obtain if the examination times were fixed in advance, permitting an unbiased analysis. In terms of a tumour marker study, they considered the sampling scheme 'patient self-selection', where the patient decides on the next examination time. This decision may depend on the current disease status and thus a straightforward analysis will give biased results. The bias can only be avoided if the selection mechanism is known or can be estimated such that corresponding weights can be incorporated in the likelihood (Andersen *et al.*, 1993). Follow-up in epidemiological studies is also essential to study incidence of diseases or conditions of interest and to estimate effects of potential risk factors. Without well-organized registries available, the occurrence of disease and corresponding event time information can only be observed in participants who are alive and will be missing in those who died during follow-up, the latter who cannot be considered in the analysis. This is comparable with studies with fixed follow-up visits, often ignoring the death cases in the analysis. In Binder *et al.* (2014) we have highlighted that the bias resulting from such a 'naive' analysis can be large in magnitude and in either direction when estimating hazard ratios corresponding to a risk factor. It can be reduced by applying an adequate illness–death model (Joly *et al.*, 2002; Leffondré *et al.*, 2013) which, however, requires event time information on the death cases. In addition, in many studies, the dates of disease onset are not reported exactly but must be considered as interval censored. By failing to do so, the cumulative incidence of disease will be underestimated (Joly *et al.*, 2002) whereas excluding the death cases will lead to a gross overestimation (Wolkewitz *et al.*, 2014). Fig. 2 illustrates this by using data from a cohort study in elderly patients estimating age-specific cumulative incidence of dementia (Joly *et al.*, 2002; Touraine *et al.*, 2013). Finally, we conclude in accordance with Keiding and Louis that the challenges of self-selection are not only present at study entry but also persist during follow-up of studies.

Alfred Stein (*University of Twente, Enschede*)
This paper is an interesting contribution to address problems and opportunities in epidemiology generated by the advent of the Internet. A major message is that traditional statistical methods are losing ground, whereas new opportunities in the frame of big data are emerging. Some traditional concepts are clearly surviving, like the central idea of collecting representative information from a large amount of data. The current paper makes a clear contribution in this sense. An interesting aspect thus concerns the representativity of a collection of data. Here, the paper builds an analogy between epidemiology and sampling. Representativeness of a sample is then an appealing idea in particular to reconstruct a profile that can serve as a representative profile. Reference is made towards agricultural sampling methods. This analogy falls somewhat short, for several reasons.

(a) During agricultural or environmental sampling (in brief: survey) we are in principle free to select any sampling spot, except if there are clear obstructions. If data are missing because of such an obstruction, then we know what the obstruction was like, and we could for example decide on a proxy location. In a Web-based epidemiological sample, however, we must deal with the data as they come, and if some part of the population is absent then we do *not* know this. The only liberty is to make a selection from the existing data.
(b) In a survey it is common to define a criterion for optimality: there is ample literature on how to do this in particular circumstances. Examples are an equal coverage of a specific area, making the best possible map, or creating the optimal variogram. If data freely float in through the Web, then there is no choice in any optimality criterion.

This brings us to 'big data' developments, where the emphasis is on segmenting and classifying, more than on anything else, i.e. on finding the relevant information for a specific purpose. Clearly, once more the 'purpose' is important, and, unless explicitly stated and *thus filtering the accrual* on this, there is little more than *filtering the data* that can be done. The big advantage, however, is the huge amount of data. Abundance of data may thus be adequate to counterbalance the lack of statistical rigour. So far, no convincing evidence has been provided for this, but with the current paper I have the impression that the first important steps have been set.

Jan P. Vandenbroucke (*University of Leiden and University of Aarhus*)
Any attempt at generalizing about generalization should make many distinctions. Keiding and Louis mainly discuss two extremes: self-selection *versus* random sampling. The original position by Miettinen was not really about self-selection. it was rooted in the experience that the major part of our knowledge about cigarette smoking and lung cancer was based on studies like the British studies about doctors (Doll and Hill, 1954). How should one generalize such results?: to men of the same social class?; to non-British men?; to women?; to non-whites? Do we need a new study for each combination of these groups, or to use a random sample? In a random sample of some community in Britian at the time of the original studies (the 1950s), there might have been too few non-white women to make any judgement. Thus, there was no solution. The largest US study at the time—by the American Cancer Society (Hammond and Horn, 1954)—used volunteers who interviewed between five and 10 middle-aged white men whom they knew and who were judged to be co-operative, about their smoking habits; the volunteers reported after 1 year whether the men had died. The generalization tacitly made in diverse reports on smoking and lung cancer in the 1960s was 'to all human beings'. This was helped by knowledge about animal experiments on the carcinogenicity of tobacco smoke in epithelial cells, the spread of lesions in pathology reports of lungs of smokers and knowledge that soot was a human carcinogen (based on reports from chimney-sweepers) (Cornfield *et al*., 1959). The analogy, which was often made in epidemiologic teaching in the 1980s, was that we know about the working of human neurons by studing giant neurons of squids, and about contractility of the heart from frog preparations—which are even more daring extrapolations of mechanisms across species. Whether or not a generalization can be made on particular studies bypasses statistics, epidemiology and even structural causal modelling. It is done case by case, and next to statistics makes use of biologic and other mechanistic insights, such as sociologic and behavioural. For external generalization, there seems little difference between observational studies and randomized trials, because also in randomized trials external validity is '... a complex reflection in which prior knowledge, statistical considerations, biological plausibility and eligibility criteria all have place' (Dekkers *et al*., 2010). Thus, we agree with Keiding and Louis that 'Justifying epidemiological generalization is difficult concrete work...'.

Herbert I. Weisberg (*Causalytics, Needham*)
I commend Keiding and Louis for tackling a difficult issue of great practical importance, and I heartily

endorse their three main summary points. Drawing on insights from the two cognate disciplines of epidemiology and survey research will be critical as we work through the emerging opportunities and challenges of 'big data'. However, I feel a need to quibble about their specification of the core issue as 'whether the conditional effects that we wish to transport are actually transportable'. As a conditional effect, the authors seem to have in mind a population level parameter such as the relative risk derived from a randomized experiment or observational study.

A causal effect in a given population is a particular statistical summary (e.g. a risk ratio, mean difference or hazard ratio) of individual causal effects in that particular population. Thus, I prefer to reframe the central and root issue as 'how our research methods should take into account the possibility of substantial causal effect heterogeneity'. Statistical summaries of causal effects are necessary and useful, but there is much confusion about the implications of effect variability across individuals. Here follow two important examples.

The transportability perspective implicitly assumes a 'real' causal effect largely independent of individual characteristics and circumstances. Adjustment for confounding presumably reveals this main effect. However, it turns out that confounding and effect heterogeneity actually represent two sides of the same coin: namely, causal effect variability across individuals (see Weisberg (2010)). As a result, if we manage to adjust successfully for confounding by 'controlling' for a certain set of relevant covariates, the resulting main effect may be difficult interpret and unlikely to transport. So, it is better to estimate stratum-specific effects, conditioning on the covariates or associated propensity scores.

Another common misunderstanding pertains to the adoption of strict entry criteria in clinical trials to exclude patients who are deemed at high risk of an adverse event. The result can be a bias of the risk ratio relative to its value in the real population of interest, even to the point of reversing its direction (Weisberg *et al*., 2009). Moreover, the presumed reduction in risk to patients may be illusory, because those excluded tend to be 'doomed' individuals at high risk *with or without* the treatment. In studies that entail self-selection, whether observational or randomized, a similar phenomenon may be at work.

Weixuan Zhu (*Universidad Carlos III de Madrid*) **and Weining Shen** (*University of California at Irvine*)

We congratulate Professor Keiding and Professor Louis for their thought-provoking paper on epidemiological and survey cultures. They discuss the transportability of conditional effects and approaches such as covariate adjustment and instrumental variables. We would like to extend their discussion to the design of clinical trials.

In a recent paper, Shen *et al*. (2015) considered continuous monitoring of early phase clinical trials with low compliance rates. Possible examples include cessation of smoking and Internet-based interventions for cancer-related sexual dysfunction. Shen *et al*. (2015) proposed the use of baseline covariates to predict the potential compliance behaviour of patients under the principal stratification framework and to identify the causal effects to help to make early stopping decisions for the trial. A few related issues were discussed in their paper.

(a) *Validity of key assumptions*: Shen *et al*. (2015) considered a smoking cessation example and focused on a two-arm, placebo-controlled randomized phase II clinical trial. They noted the plausibility of the one-sided access assumption as the patients who were assigned to the control group may not have access to the treatment group given that the new experimental agent is not available in the market. Other assumptions such as exclusion restrictions were also discussed and justified in Shen *et al*. (2015). This example is related to Keiding and Louis's discussion of the transportability of conditional effects for causal inference.

(b) *Effect of model misspecification on stopping rules*: Shen *et al*. (2015) included several sensitivity analyses to evaluate the numerical effect of confounders, model misspecification and assumption violation. It seems that the stopping rules are less sensitive to these problems compared with the parameter estimates. One possible reason is the use of a continuous monitoring design, which allows for sequential updating of the data and makes the early stopping decision robust.

(c) *Comparison with intent-to-treat analysis*: Shen *et al*. (2015) discussed the trade-off that is associated with using the causal effect instead of the intention to treat. Using causal effects may provide greater accuracy in decision making but requires additional assumptions, and the performance relies on correct model specification, which may be difficult to test and verify in practice. It is helpful to consult experts regarding the non-compliance information.

(d) *Targeting future compliant groups*: in addition to stopping rules, Shen *et al*. (2015) also discussed identifying factors that are predictive of non-compliance, and using them to develop strategies to decrease non-compliance and to improve the benefit to the intention-to-treat population. They gave

an example: if depression is found to be a strong predictor of non-compliance, then using a combination therapy of the experimental drug with an antidepressant or another intervention may achieve better treatment performance. This relates to Keiding and Louis's discussion on representativity and generalization.

The **authors** replied later, in writing, as follows.

We thank the discussants for their insightful commentary and suggestions, and we respond by theme.

Representativity and internal–external validity

Representativity is necessary when the object of an investigation is estimating basic population characteristics from a sample. This sample survey goal applies, for example, when assessing population prevalence. It is more delicate and situation dependent to determine the role of representativity in analytic epidemiology and clinical trials, where the target is association between an exposure or treatment and an effect, or other relationships between variables.

Exemplifying the critical attitude on the relevance of representative sampling in analytic epidemiology, *Vandenbroucke* refers to the important epidemiological finding from a narrow study group: the Doll and Hill (1954) study of excess lung cancer incidence among smoking male British doctors. Similarly, *Rothman and his colleagues* recall Miettinen's sarcastic remark about non-representativity of laboratory studies based on highly selected animals. However, *Kaufman* warns that 'within selected samples, associations may lack causal interpretations even absent confounding' and highlights 'many fumbles' in the slow resolution of the so-called low birth weight and obesity paradox.

King's observations that most outcomes have a social gradient and that social status is a confounder and is also associated with participation reinforce our point that context matters, and that very few survey or epidemiological targets of inference are immutable. Furthermore, social status is difficult to measure and adjustments based on it are likely to be inadequate.

Elliott offers a Gaussian linear model example to delineate the roles of randomization and representative sampling. Randomization 'breaks' confounding and representative sampling allows estimation of a causal effect. He makes the important distinction between estimating the causal effect in the study population (under representative sampling it does not require knowing the distribution of the unmeasured confounder U) and generalization to other populations (additional information is needed on U). This point provides a possible explanation for the unwillingness of Rothman and his colleagues to endorse representative sampling (see Section 7.3). We emphasize that, even if the distribution of U is properly specified, Pearl and Bareinboim (2014) showed that all other aspects of the model need to be docked to a reference population for valid transportation.

Little endorses the importance of representation for both descriptive and analytic estimands, and he properly calls for increased attention to estimation goals. His comments on the ferment over probability *versus* non-probability sampling in the National Children's Study are directly relevant. In designing the National Children's Study many argued for a frame-based probability sample; others rejoined that doing so was impossible and also not necessary. We agree that it was impossible to acquire a perfect probability sample, but we offer that the attempt to do so would ensure good demographic, geographic and exposure coverage, and make available approximately correct sampling propensities.

Draper relates representation to exchangeability of sampled and non-sampled units, and to assignment ignorability. We strongly endorse his call for journals to require that authors calibrate and justify the scope of valid inferences.

Rothman and his colleagues misinterpret our primary thesis, e.g. by claiming that we anoint survey research as the gold standard, and that all studies must aspire to survey status. We do not. We use *SnartGravid* as a case-study for its explicit discussion of selection issues and its subject matter results. Internet-based recruiting requires considerable care, and their representativity analysis using Danish administrative data is far more convincing than *ex cathedra* declarations. Schisterman *et al*. (2014) used Faceoook to find participants, but they imposed rigorous eligibility criteria, and we judge that recruitment was equivalent to the use of newspaper advertisements, clinic posters, etc.

The premise that relationships are immutable and so internal validity suffices, although comforting to believers, is absolutely incorrect. Context and conduct matter, as experiments in psychology (see the careful study by Henrich *et al*. (2010), which was quoted by *Brick*) and in cell phone surveys demonstrate. Effect modification is always in play; magnitude is what matters.

Self-enrolment

Fogarty and his colleagues offer sound advice on analysing self-selected studies, including that valid adjust-

ment for covariate imbalance depends on substantial overlap in treatment and control group distributions, so basing inference on the study subset with sufficient overlap is recommended. We endorse this, now common, practice and also their call for sensitivity analyses to assess the potential effect of non-random selection on external validity.

Hand and Herzberg note that self-enrolled studies and surveys have the advantage of relatively rapid conduct to inform policy, e.g. economic decision making. However, they caution that the likelihood of substantial bias is high and (as does *Eltinge*) call for striking a balance between timeliness, bias and variance. They conclude that self-selected surveys and 'big data' are here to stay, and so improved modelling is needed. We endorse their encouragement of innovations to improve quality, and warning that substandard approaches may replace the less timely, but more valid, probability samples.

Bethlehem reminds us that interviewer-assisted surveys are high quality, but expensive; that on-line, general population surveys will have poor coverage unless a large fraction of the population has access, and that probability sampling via the Internet is difficult if not impossible. *Post hoc* weighting may help, but he concludes that self-selected surveys are inappropriate for official statistics. That is possibly true but, with declining response rates for traditional surveys, improved Internet-based approaches need to be developed.

Baffour emphasizes the key role of a sampling frame to generate a representative sample, and the benefits of dual frames (e.g. land-lines and mobile phones) to enhance validity. He notes the close relationship between selection bias and the bias induced by non-response that is associated with the end points of interest. We endorse his call for innovative designs and analyses to mitigate such biases. *Zhu and Shen's* relating of self-selection to the problem of low compliance in early phase randomized clinical trials provides another analogy.

Bordley reports on successful use of an on-line survey by General Motors, but this success depended on the reference population being exactly those who participated. Other contexts with this alignment will be similarly successful.

Meng's example shows that, unless a very large fraction of a reference population participates, a large but self-selected sample's mean-squared error will be much greater than that of a small random sample. This caution also applies to use of organic data. Potential remediation results from using information to model the correlation between participation propensities and population values, and we caution that effectiveness can be end point dependent.

Brick formalizes the evaluation of self-selected samples, addressing representation and measurement issues leading to the modern concept of total survey error. He claims that comparisons and associations may be more transportable than level, echoing the belief in the relative safety of conditional effect measures that we identify in our paper and discuss below.

Treatment effect heterogeneity

Treatment effect heterogeneity is central to representativity in epidemiology. *Kaufman* provides a basic, far-reaching example wherein there is a latent stratum of people who suffer the outcome if exposed, but not if not exposed. Findings from the sample are directly generalizable, if the proportion of such individuals is the same in the study group and target population (and are representative if the relative proportions are known). Similarly, *Weisberg* addresses *causal effect heterogeneity*, recommending focus on conditional rather than marginal measures. (Our use of conditional effect typically *is* what Weisberg calls a stratum-specific effect.) He also comments that strict entry criteria, in particular to clinical trials, may bias the results. In his example, exclusion of 'doomed' patients who will die with or without the treatment might bias the risk ratio estimate relative to the population value, even to the point of reversing its direction. Related, *Ding* mentions his work exploring randomization inference as a basis for assessing treatment effect variation beyond what is explained by observed covariates.

Transportability of conditional and marginal effects; collapsibility

Ding and *Kaufman* each mention our statement that, subject to accounting for relevant confounders, conditional effects, more often than marginal effects, can be transportable. Our statement is not an 'oracle'; we want to document that use of regression analysis and standardization (Keiding and Clayton, 2014) in aetiological research usually has conditional effects as targets, and that these work well in many situations. Conditioning is on observed confounders, i.e. pretreatment variables. *Ding* generalizes this by allowing conditioning on post-treatment surrogates related to the outcome of interest. We support such innovations but caution that they are very model dependent.

Kaufman notes that we did not mention collapsibility. It is important and implicit in our comment just before Section 3.1 that different scales produce different effects. *Hudgens and Fine* make the

related point (which we support) that absolute effects are often more relevant targets than conditional effects.

Statistical inference

In survey analysis, principles of statistical inference including contrasting design-based and model-based approaches are of central interest. Similar issues arise in epidemiology (see Robins (1988), quoted by *Kaufman*), but they have not permeated the discourse. Indeed, we gave them little attention, and we thank several discussants for bringing them up. *Kaufman* acknowledges that sometimes randomization is justified, in a few cases via superpopulation sampling. He questions the justification for the sampling properties of estimates in many observational studies, citing the (American) Nurses' Health Study as an example. Similarly, *Choi and Lai* point out that the confidence intervals in *SnartGravid* cannot be justified, because the study group is not a probability sample from a target population.

We mentioned in passing *instrumental variable analysis* as a candidate tool for confounder control, agreeing with *Hudgens and Fine* and with *le Cessie* that careful evaluation of its role is needed in the epidemiological context.

Empirical validation of aetiological studies and surveys

Several discussants offer the good advice that it is often necessary to collect additional empirical information to validate observed associations in an aetiological study. *le Cessie* mentions use of several control groups in case–control studies, and she reports experience from a Dutch cohort study on pathways to obesity-related diseases that employed three recruitment approaches. *Choi and Lai* remind us that multiple independent studies are often the best way of confirming or questioning results from a study. *Greenhouse* supports this view and cites work on systematic approaches to combining evidence. *Ding* proposes sensitivity analyses and notes that few approaches are available for external validity. *Hong* proposes two empirical checks on transportability that may be helpful in some contexts. *de Luna* concludes his philosophical analysis of key concepts by interpreting Miettinen's 'the abstract' as predictions that can be corroborated in other studies.

The role of big data

In response to *McNicholas and Subedi,* our use of the term 'big data' is broad, including administrative registers and 'organic' information. *Little* observes that surveys must be designed to link to administrative and other big data for continued relevance of probability surveys. The benefits of using administrative data to help to address population goals and to calibrate survey responses are quite direct, but also there is potential for organic data to add value. The challenges are substantial and taking them on is not for the (statistically) faint of heart!

Liu and Ning discuss the role of using big data to help to repair self-selection and they promote designed approaches to collecting such data. Their formulation is similar to Meng's with sample inclusion propensities that depend on the potentially observed data. Big data have the potential to estimate dependences and to improve weighted adjustments; their penalized, doubly robust approach merits evaluation.

Mateu and Aragó explore potentials in the *tsunami* of social media data. These media are effective in contacting and enrolling participants (Schisterman *et al*. (2014) used Facebook) but, when going beyond recruitment, Mateu and Aragó note that data quality and ethical concerns must be addressed. We warn that the effective sample size of social media data is likely to be orders of magnitude less than its apparent size.

Women's Health Initiative

Both *Prentice* and *Anderson* offer important elaborations on our sketch of the Women's Health Initiative (WHI) analyses. Paraphrasing Anderson: the impressive results that were achieved by the WHI are based on careful evaluation of results from the self-selected observational study and self-selected study population for the randomized trials. WHI investigators were primarily concerned with the internal validity, and Anderson argues convincingly that to recruit using a population-based probability sample would have been unrealistic and even counterproductive, at least in the USA. We counter that calibrations of the Nurses' Health Study by Hernán and his colleagues were innovative and effective for assessing external validity. *Prentice* points out that in the careful process that ultimately allowed a full reconciliation of WHI data it was important to involve several *timing* variables, such as duration of use of hormone therapy and years from menopause to initiating hormone therapy.

Timing

Schumacher and his colleagues provide a general treatment of timing variables related to patient self-selection during follow-up, citing the important contribution by Grüger *et al*. (1991). Surprisingly, the need to distinguish between protocol-driven follow-up visits and patient-initiated visits has been studied

very little (see Keiding (2014), section 7.2, for a brief survey, and Chan *et al.* (l998) for a modelling approach). Schumacher and his colleagues also emphasize the need for proper accounting of attrition via competing risks and the requisite data requirements.

Study context

We thank *Hernán* for offering the vote of thanks. He delineates the steps in specializing from the total population ('humankind') to the study sample, and he classifies the relative roles of subject matter insight and statistical issues in these steps. However, his classification does not correspond to our experience in which subject matter insights (from cardiologists, oncologists, social scientists, etc.) need to be combined with methodology (from statisticians, epidemiologists and survey analysts) in all steps. We also disagree with Hernán's implicit assertion that the main issue with representativity in the *SnartGravid* study is whether time to pregnancy depends on use or non-use of the Internet (see also statements that we quoted from *Huybrechts and her colleagues*). Our concern is the more general one for self-selected studies: whether the *propensity to volunteer* is associated with the outcome. Similar issues apply to surveys.

Eltinge embeds the internal–external trade-offs in a total survey error framework. He proposes optimally allocating resources on the basis of (cost, quality, risk) profiles. For example, investing in frame development mitigates the risk of non-representation and selection bias, but it can reduce resources for conducting the study. We endorse his call for identifying the degree of departure from full transportation that is acceptable and his encouragement of cost–benefit analysis.

Straf echoes our principal points and emphasizes their importance in the policy arena. He also proposes the use of a systems approach to address key goals.

Waller partitions an analysis into a model for the underlying process and a measurement or observation model. His partitioning supports combining evidence from multiple sources at a variety of spatiotemporal scales and has the potential to simplify (but by no means making simple!) some evaluations when assessing the degree of transportation.

Seconding the vote of thanks, *Miller* reminds us that non-probability sampling is only one of many factors that affect quality, and that *measurement* is of at least co-equal importance. For example, the survey mode that is used to collect data can have a powerful effect on internal validity, and modes have become more complex with the advent of adaptive design. Miller adds that there is a long history of experiments on surveys, with the attendant internal–external validity trade-offs. He advocates, and we strongly support, development of 'truth benchmarks' to evaluate quality in the survey and epidemiological domains.

Stein disagrees with our claim that the issues that we discuss are directly relevant to agriculture and ecology. We agree that in these contexts a probability sample can be more easily implemented and departures documented. However, the issues of internal–external validity and transportability remain important. His comment on the difficulty of defining optimality for data that 'freely float through the Web' echoes that of other discussants.

Murtagh muses on the geometry and topology of the relationship between the data analysed and big data. Implementing his construct would be challenging, but we thank him for his broad perspective.

Linkage of surveys and epidemiological studies to big data sources

Several commentators point out that systematic approaches are needed to embed the detailed information in 'small' focused studies with specific hypotheses into a population distribution of background and other variables. There are several benefits, among them better confounder control and possibilities for supplementing conditional effect measures with the less easily available absolute effects. These effects are usually necessary when communicating results to patients or participants and informing policy. *Prentice* proposes institutionalizing this process by creating repositories for data from large cohorts, with the goal of improving the potential to transport ratio measures, and perhaps also absolute risk.

Closing remarks

Surveys based on a probability sample *prima facie* target a reference population and produce a representative sample (sample inclusion propensities are available); Web-based and other forms of volunteer sampling degrade and possibly destroy representation, because context matters and causal relationships are not immutable. So, all studies must address representation up front by clarifying goals and if necessary collecting information that can support transportation.

References in the discussion

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993) *Statistical Methods based on Counting Processes*. New York: Springer.

Andersen, R., Kaspar, J. and Frankel, M. (1979) *Total Survey Error*. San Francisco: Jossey-Bass.
Andersen, P. K. and Keiding, N. (2002) Multi-state models for event history analysis. *Statist. Meth. Med. Res.*, **11**, 91–115.
Angelucci, M. and Di Maro, V. (2015) Program evaluation and spillover effects. *Working Paper WPS7243*. Development Research Group, World Bank Group, Washington DC. (Available from `http://documents.worldbank.org/curated/en/2015/04/24407918/program-evaluation-spillover-effects`.)
Ansolabehere, S. and Hersh, E. (2012) Validation: what Big Data reveal about survey misreporting and the real electorate. *Polit. Anal.*, **20**, 437–459.
Aral, S. (2015) Networked experiments. In *The Oxford Handbook on the Economics of Networks*. Oxford: Oxford University Press. To be published.
Arcos, A., Rueda, M. M., Trujillo, M. and Molina, D. (2014) Review of estimation methods for landline and cell phone surveys. *Sociol. Meth. Res.*, **44**, 458–485.
Arrow, K. J. and Fisher, A. C. (1974) Environmental preservation, uncertainty and irreversibility. *Q. J. Econ.*, **88**, 312–319.
Baffour, B., King, T. and Valente, P. (2013) The modern census: evolution, examples and evaluation. *Int. Statist. Rev.*, **81**, 407–425.
Bales, K., Hesketh, O. and Silverman, B. (2015) Modern slavery in the UK: how many victims? *Significance*, **12**, no. 3, 16–21.
Banack, H. R. and Kaufman, J. S. (2014) The obesity paradox: understanding the effect of obesity on mortality among individuals with cardiovascular disease. *Prev. Med.*, **62**, 96–102.
Barr, M. L., Ferguson, R. A., Hughes, P. J. and Steel, D. G. (2014) Developing a weighting strategy to include mobile telephone numbers into an ongoing population health survey using an overlapping dual frame design with limited benchmark information. *BMC Med. Res. Methodol.*, **14**, 102–112.
Bennett, T., Savage, M., Silva, E., Warde, A., Gayo-Cal, M. and Wright, D. (2009) *Culture, Class, Distinction*. London: Routledge.
Bethlehem, J. G. (2010) Selection bias in web surveys. *Int. Statist. Rev.*, **78**, 161–188.
Bethlehem, J. G. (2015) On the quality of web panels. In *Survey Measurements—Techniques, Data Quality and Sources of Error* (ed. U. Engels), pp. 112–129. Frankfurt: Campus.
Bethlehem, J. G. and Callegaro, M. (2014) Introduction to part IV—Weighting adjustment. In *Online Panel Research—a Data Quality Perspective* (eds M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick and P. Lavrakas), pp. 264–272. Chichester: Wiley.
Beyersmann, J., Allignol, A. and Schumacher, M. (2012) *Competing Risks and Multistate Models with R*. New York: Springer.
Biemer, P. P., Trewin, D., Bergdahl, H. and Japec, L. (2014) A system for managing the quality of official statistics (with discussion). *J. Off. Statist.*, **30**, 381–442.
Binder, N. and Schumacher, M. (2014) Missing information caused by death leads to bias in relative risk estimates. *J. Clin. Epidem.*, **67**, 1111–1120.
Blumberg, S. J. and Luke, J. V. (2015) Wireless substitution: early release estimates from the National Health Interview Survey, July-December 2014. National Center for Health Statistics, (Available from `http://www.cdc.gov/nchs/nhis.earlyrelease/wireless201506.pdf`.)
Boef, A. G. C., le Cessie, S., Dekkers, O. M., Frey, P., Kearney, P. M., Kerse, N., Mallen, C. D., McCarthy, V. J. C., Mooijaart, S. P., Muth, C., Rodondi, N., Rosemann, T., Russell, A., Schers, H., Virgini, V., de Waal, M. W. M., Warner, A., Gussekloo, J. and den Elzen, W. P. J. (2016) Physician's prescribing preference as an instrumental variable: exploring assumptions using survey data. *Epidemiology*, to be published.
Brackstone, G. (1999) Managing data quality in a statistical agency. *Surv. Methodol.*, **25**, 139–149.
Breslow, N., Lumley, T., Ballantyne, C., Chambless, L. and Kulich, M. (2009) Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statist. Biosci.*, **1**, 32–49.
Cannell, C. F., Miller, P. V. and Oksenberg, L. (1981) Research on interviewing techniques. In *Sociological Methodology* (ed. S. Leinhardt). San Francisco: Jossey-Bass.
Chan, I., Hillman, D. and Louis T. A. (1998) Treatment comparisons with screenable endpoints. *Computnl Statist. Data Anal.*, **27**, 401–419.
Chen, H., Geng, Z. and Jia, J. (2007) Criteria for surrogate end points. *J. R. Statist. Soc.* B, **69**, 919–932.
Choi, A. L., Mogensen, U. B., Bjerve, K. S., Debes, F., Weihe, P., Grandjean, P. and Budtz-Jørgensen, E. (2014) Negative confounding by essential fatty acids in methylmercury neurotoxicity associations. *Neurtoxicol. Teratol.*, **42**, 85–92.
Cornfield, J. (1959) Principles of research. *Am. J. Mentl Defic.*, **64**, 240–252.
Cornfield, J., Haenszel, W., Hammond, E. D., Lilienfield, A. M., Shimkin, M. B. and Wynder, E. L. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natn. Cancer Inst.*, **22**, 173–203.
Cox, D. R. (1984) Interaction. *Int. Statist. Rev.*, **52**, 1–24.
Dawid, A. P. (2004) Probability, causality and the empirical world: a Bayes-de Finetti-Popper-Borel synthesis. *Statist. Sci.*, **19**, 44–57.

Dekkers, O. M., von Elm, E., Algra, A., Romijn, J. A. and Vandenbroucke, P. (2010) How to assess the external validity of therapeutic trials: a conceptual approach. *Int. J. Epidem.*, **39**, 89–94.
De Mutsert, R., den Heijer, M., Rabelink, T. J., Smit, W. A., Romijin, J. A., Jukema, J. W., de Roos, A., Cobbaert, C. M., Kloppenburg, M., Le Cessie, S., Middeldorp, S. and Rosendaal, F. R. (2013) The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. *Eur. J. Epidem.*, **28**, 513–523.
Dillman, D. A. (1996) Why innovation is difficult in government surveys. *J. Off. Statist.*, **12**, 113–124.
Ding, P., Feller, A. and Miratrix, L. (2016) Randomization inference for treatment effect variation. *J. R. Statist. Soc.* B, **78**, in the press.
Ding, P. and Vanderweele, T. J. (2014) Generalized Cornfield conditions for the risk difference. *Biometrika*, **101**, 971–977.
Doll, R. and Hill, A. B. (1954) The mortality of doctors in relation to their smoking habits: a preliminary report. *Br. Med. J.*, **328**, 1451–1455.
Draper, D. (1995) Inference and hierarchical modeling in the social sciences (with discussion and rejoinder). *J. Educ. Behav. Statist.*, **20**, 115–233.
Duan, N., Meng, X. L., Lin, J. Y., Chen, C. N. and Alegria, M. (2008) Disparities in defining disparities: statistical conceptual frameworks. *Statist. Med.*, **27**, 3941–3956.
Ellenberg, J. H. (2010) The National Children's Study (NCS): establishment and protection of the inferential base *Statist. Med.*, **29**, 1360–1367.
Ermisch, J., Jantti, M. and Smeeding, T. M. (2012) *From Parents to Children: the Intergenerational Transmission of Advantage*. New York: Russell Sage.
Fisher, R. A. (1956) *Statistical Methods and Scientific Inference*. London: Macmillan.
Flanders, W. D. and Klein, M. (2007) Properities of 2 counterfactual effect definitions of a point exposure. *Epidemiology*, **18**, 453–460.
Fogarty, C. B., Mikkelsen, M. E., Gaieski, D. F. and Small, D. S. (2015) Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Am. Statist Ass.,* to be published.
Frangakis, C. E. and Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics*, **58**, 21–29.
Friedman, S., Laurison, D. and Miles, A. (2015) Breaking the 'class' ceiling?: Social mobility into Britain's elite occupations. *Sociol. Rev.*, **63**, 259–289.
Gastwirth, J. L. (2003) Issues arising in using samples as evidence in trademark cases. *J. Econmetr.*, **113**, 69–82.
Grandjean, P., Weihe, P., White, R. F., Debes, F., Araki, S., Yokoyama, K., Murata, K., Sørensen, N., Dahl, R. and Jørgensen, P. J. (1997) Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. *Neurtoxicol. Teratol.*, **19**, 417–428.
Greenhouse, J. B., Anderson, H., Bridge, J. A., Libby, A., Valuck, R. and Kelleher, K. J. (2015) Combining information from multiple data sources for comparative effectiveness research: an introduction to cross-design synthesis with a case study. In *Methods for Comparative Effectiveness Research*. New York: Taylor and Francis. To be published.
Greenhouse, J. B. and Kelleher, K. (2005) Thinking outside the (black) box: antidepressants, suicidality, and research synthesis. *Pediatrics*, **116**, 231–233.
Greeland, S. (1990) Randomization, statistics, and causal inference. *Epidemiology*, **1**, 421–429.
Greenland, S. and Draper, D. (2014) Exchangeability. In *International Encyclopedia of Statistical Science* (ed. M. Lovric), pp. 474–476. New York: Springer.
Groves, R. (1989) *Survey Errors and Survey Costs*. New York: Wiley.
Groves, R. M. and Lyberg, L. E. (2010) Total survey error: past, present and future. *Publ. Opin. Q.*, **74**, 849–879.
Grüger, J., Kay, R. and Schumacher, M. (1991) The validity of inferences based on incomplete observations in disease state models. *Biometrics*, **47**, 595–605.
Gundersen, D. A., ZuWallack, R. S., Dayton, J., Echeverria, S. E. and Delnevo, C. D. (2014) Assessing the feasibility and sample quality of a national random-digit-dialing cellular phone survey of young adults. *Am. J. Epidem.*, **179**, 39–47.
Hamilton, S. F., Sunding, D. L. and Zilberman, D. (2003) Public goods and the value of product quality regulations: the case of food safety. *J. Publ. Econ.*, **87**, 799–817.
Hammond, E. C. and Horn, D. (1955) The relationship between human smoking habits and death rates: a follow-up study of 187,766 men. *J. Am. Med. Ass.*, **155**, 1316–1328.
Hartman, E., Grieve, R., Ramsahai, R. and Sekhon, J. S. (2015) From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *J. R. Statist. Soc.* A, **178**, 757–778.
Haskins, R. and Margolis, G. (2014) *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy*. Washington DC: Brookings Institution Press.
Hatch, E. E., Hahn, K. A., Wise, L. A., Mikkelsen, E. M., Kumar, R., Fox, M. P., Brooks, D. R., Riis, A. H., Sorensen, H. T. and Rothman, K. J. (2015) Evaluation of selection bias in an internet-based study of pregnancy planners. *Epidemiology*, to be published.
Heckman, J. J. and Vytlacil, E. J. (2007) Econometric evaluation of social programs: Part I, Casual models, structural models and econometric policy evaluation. In *Handbook of Econometrics*, vol. 6, pp. 4779–4874.

Henrich, J., Heine, S. and Norenzayan, A. (2010) The weirdest people in the world? *Behav. Brain Sci.*, **33**, 61–83.
Hernández-Díaz, S., Schisterman, E. F. and Hernán, M. A. (2006) The birth weight "paradox" uncovered? *Am. J. Epidem.*, **164**, 1115–1120.
Hess, C. and Ostrom, E. (eds) (2006) *Understanding Knowledge as a Commons*. Cambridge: MIT Press.
Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Ass.*, **47**, 663–685.
Ioannidis, J. P. A. (2005a) Contradicted and initially stronger effects in highly cited clinical research. *J. Am. Med. Ass.*, **294**, 218–228.
Ioannidis, J. P. A. (2005b) Why most published research findings are false. *PLOS Med.*, **2**, no.8, article e124.
Japec, L., Kreuter, F., Bert, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. and Usher, A. (2015) *AAPOR Report on Big Data*. Deerfield: American Association for Public Opinion Research.
Jiang, Z., Ding, P. and Geng, Z. (2015) Qualitative evaluation of associations by the transitivity of the association signs. *Statist. Sin.*, **25**, 1065–1079.
Jiang, Z., Ding, P. and Geng, Z. (2016) Principal causal effect identification and surrogate end point evaluation by multiple trials. *J. R. Statist. Soc.* B, **78**, in the press.
Joly, P., Commenges, D., Helmer, C. and Letenneu, L. (2002) A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*, **3**, 433–443.
Ju, C. and Geng, Z. (2010) Criteria for surrogate end points based on causal distributions. *J. R. Statist. Soc.* B, **72**, 129–142.
Kahneman, D. and Knetsch, J. L. (1992) Valuing public goods: the purchase of moral satisfaction. *J. Environ. Econ. Mangmnt*, **22**, 57–70.
Kaizar, E. E. (2015) Incorporating both randomized and observational data into a single analysis. *A. Rev. Statist. Appl.*, **2**, 49–72.
Kalton, G. and Schuman, H. (1982) The effect of the question on survey responses: a review (with discusion). *J. R. Statist. Soc.* A, **145**, 42–73.
Keiding, N. (2014) Event history analysis. *A. Rev. Statist. Appl.*, **1**, 333–360.
Keiding, N. and Clayton, D. (2014) Standardization and control for confounding in observational studies: a historical perspective. *Statist. Sci.*, **29**, 529–558.
Kenett, R. S. and Shmueli, G. (2014) On information quality (with comments). *J. R. Statist. Soc.* A, **177**, 3–38.
Kish, L. (1959) Some statistical problems in research design. *Am. Sociol. Rev.*, **24**, 328–338.
Kruskal, W. and Mosteller, F. (1979) Representative sampling: lii, The current statistical literature. *Int. Statist. Rev.*, **47**, 245–265.
Lai, T. L. (2014) Statistics in an new era for finance and health care. In *Past, Present and Future of Statistical Science* (eds X. Lin, G. Genest, D. L. Banks, G. Molenberghs, D. W. Scott and J.-L. Wang), pp. 369–379. Boca Raton: Chapman and Hall–CRC.
Lai, T. L. and Lavori, P. W. (2011) Innovative clinical trial designs: toward a 21st-century health care system. *Statist. Biosci.*, **3**, 145–168.
Laurison, D. and Friedman, S. (2015) Introducing the class ceiling: social mobility and Britain's elite occupations. *Working Paper*. Sociology Department, London School of Economics and Political Science, London. (Available from `http://www.lse.ac.uk/sociology/pdf/Working-Paper-Introducing-the-Class-Ceiling.pdf`.)
Lebaron, F. (2009) How Bourdieu 'quantified' Bourdieu: the geometric modelling of data. In *Quantifying Theory: Pierre Bourdieu* (eds K. Robson and C. Sanders). New York: Springer.
Leek, J. T. and Peng, R. D. (2015) What is the question?: Mistaking the type of question being considered is the most common error in data analysis. *Science*, **347**, 1314–1315.
Leffondré, K., Touraine, C., Helmer, C. and Joly, P. (2013) Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model? *Int. J. Epidem.*, **42**, 1177–1186.
Little, R. J. (2010) Discussion. *Statist. Med.*, **29**, 1388–1390.
Livingstone, M., Dietze, P., Ferris, J., Pennay, D., Hayes, L. and Lenton, S. (2013) Surveying alcohol and other drug use through telephone sampling: a comparison of landline and mobile phone samples. *BMC Res. Methodol.*, **13**, 41–48.
Meng, X.-L. (2014) A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it). In *Past, Present, and Future of Statistical Science* (eds X. Lin, G. Genest, D. L. Banks, G. Molenberghs, D. W. Scott and J.-L. Wang), pp. 537–622. Boca Raton: CRC Press.
Meng, X.-L. (2015) Statistical paradises and paradoxes in Big Data. *American Statistical Association Chicago Chapter Meet. on World Statistics, Oct 20th*.
Michael, R. T. and O'Muircheartaigh, C. A. (2008) Design priorities and disciplinary perspectives: the case of the US National Children's Study. *J. R. Statist. Soc.* A, **171**, 465–480.
Miettinen, O. S. (1985) *Theoretical Epidemiology*. New York: Wiley.
Mikkelsen, E. M., Riis, A. H., Wise, L. A., Hatch, E. E., Rothman, K. J. and Sørensen, H. T. (2013) Pre-gravid oral contraceptive use and time to pregnancy: a Danish prospective cohort study. *Hum. Reprodn*, **28**, 1398–1405.
Murtagh, F. (2010) The Correspondence Analysis platform for uncovering deep structure in data and information. *Comput. J.*, **53**, 304–315.

Neyman, J. (1990) On the application of probability theory to agricultural experiments: essay on principles; section 9 (Engl. transl. by D. Dabrowska and T. Speed). *Statist. Sci.*, **5**, 463–472.
Pearl, J. (2015) Generalized experimental findings. *J. Causl Inf.*, **3**, 259–266.
Pearl, J. and Bareinboim, E. (2014) External validity: from do-calculus to transportability across populations. *Statist. Sci.*, **29**, 579–595.
Pew Research Center (2012) Assessing the representativeness of public opinion surveys. Pew Research Center. (Available from `http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/`.)
Pomp, E. R., Van Stralen, K. J., le Cessie, S., Vandenbroucke, J. P., Rosendaal, F. R. and Doggen, C. J. M. (2010) Experience with multiple control groups in a large population-based case-control study on genetic and environmental risk factors. *Eur. J. Epidem.*, **25**, 459–466.
Poole, C., Shrie, I. and VanderWeele, T. J. (2015) Is the risk difference really a more hetergeneous measure? *Epidemiology*, **26**, 714–718.
Prentice, R. L. (1989) Surrogate endpoints in clinical trials: definition and operational criteria. *Statist. Med.*, **8**, 431–440.
Prentice, R. L., Chlebowski, R. T., Stefanick, M. L., Manson, J. E., Langer, R. D., Pettinger, M., Hendrix, S. L., Hubbell, F. A., Kooperberg, C., Kuller, L. H., Lane, D. S., McTiernan, A., O'Sullivan, M. J., Rossouw, J. E. and Anderson, G. L. (2008a) Conjugated equine estrogens and breast cancer risk in the Women's Health Initiative clinical trial and observational study. *Am. J. Epidem.*, **167**, 1407–1415.
Prentice, R. L., Chlebowski, R. T., Stefanick, M. L., Manson, J. E., Pettinger, M., Hendrix, S. L., Hubbell, F. A., Kooperberg, C., Kuller, L. H., Lane, D. S., McTiernan, A., O'Sullivan, M. J., Rossouw, J. E. and Anderson, G. L. (2008b) Estrogen plus progestin therapy and breast cancer in recently postmenopausal women. *Am. J. Epidem.*, **167**, 1207–1216.
Prentice, R. L., Langer, R., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G., Barad, D., Curb, J. D., Kotchen, J., Kuller, L., Limacher, M. and Wactawski-Wende, J. and the Women's Health Initiative Investigators (2005) Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initative clinical trial. *Am. J. Epidem.*, **162**, 404–414.
Prentice, R. L., Langer, R. L., Stefanick, M., Howard, B., Pettinger, M., Anderson, G. L., Barad, D., Curb, J., Kotchen, J., Kuller, L., Limacher, M. and Wactawski-Wende, J., and for the Women's Health Initiative Investigators (2006) Combined analysis of Women's Health Initative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *Am. J. Epidem.*, **163**, 589–599.
Prentice, R. L., Manson, J. E., Langer, R. D., Anderson, G. L., Pettinger, M., Jackson, R. D., Johnson, K. C., Kuller, L. H., Lane, S. D., Wactawski-Wende, J., Brzyski, R., Allison, M., Okene, J., Sarto, G. and Rossow, J. E. (2009) Benefits and risks of postmenopausal hormone therapy when intiated soon after the menopause. *Am. J. Epidem.*, **170**, 12–23.
Price, M., Gohdes, A. and Ball, P. (2015) Documents of war: understanding the Syrian conflict. *Significance*, **12**, no. 2, 14–19.
Puts, M., Dass, P. and de Waal, T. (2015) Finding errors in Big Data. *Significance*, **12**, no. 3, 26–29.
Robins, J. M. (1988) Confidence intervals for causal parameters. *Statist. Med.*, **7**, 773–785.
Rosenbaum, P. R. (2002) *Observational Studies*. New York: Springer.
Rosenbaum, P. R. and Rubins, D. B. (1993) The central role of the prospensity score in observational studies. *Biometrika*, **70**, 41–55.
Rothman, K. J. (2010) Real world data. *Val. Hlth*, **10**, 322–323.
Rothman, K. J., Gallacher, J. E. J. and Hatch, E. E. (2013) Why representativeness should be avoided. *Int. J. Epidem.*, **42**, 1012–1014.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rubin, D. B. (1978) Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, **7**, 34–58.
Samuelson, P. (1954) The pure theory of public expenditure. *Rev. Econ. Statist.*, **36**, 387–389.
Schisterman, E. F., Silver, R. M., Lesher, L. L., Faraggi, D., Wactawski-Wende, J., Townsend, J. M., Lunch, A. M., Perkins, N. J., Mumford, S. L. and Galai, N. (2014) Preconception low-dose aspirin and pregnancy outcomes: results from the EAGeR randomised trial. *Lancet*, **384**, 29–36.
Shen, W., Ning, J. and Yuan, Y. (2015) Bayesian sequential monitoring design for two-arm randomized clinical trials with noncompliance. *Statist. Med.*, **34**, 2104–2115.
Shih, M. C., Turakhia, M. and Lai, T. L. (2015) Innovative designs of point-of-care comparative effectiveness trials. *Contemp. Clin. Trials*, to be published, doi 10.1016/j.cct.2015.06.007.
Shonkoff, J. and Phillips, D. (2000) *From Neurons to Neighborhoods: the Science of Early Child Development*. Washington DC: National Academies Press.
Stuart, E. A., Cole, S. R., Bradshaw, C. P. and Leaf, P. J. (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Statist. Soc.* A, **174**, 369–386.
Swanson, S. A., Miller, M., Robins, J. M. and Hernan, M. A. (2015) Definition and evaluation of the monotonicity condition for preference-based instruments. *Epidemiology*, **26**, 414–420.

Tan, K., Ning, Y., Witten, D. and Liu, H. (2015) Do-over: replicates in high dimensions with applications to latent variable graphical models. *Technical Report*. Princeton University, Princeton.
Touraine, C., Gerds, T. A. and Joly, P. (2013) The SmoothHazard package for R: fitting regression models to interval-censored observations of illness-death models. *Technical Report 12*. Department of Biostatistics, University of Copenhagen, Copenhagen.
Tzavidis, N., Salvati, N., Schmid, T., Flouri, E. and Midouhas, E. (2016) Longitudinal analysis of the strengths and difficulties questionnaire scores of the Millennium Cohort Study children in England using M-quantile random-effects regression. *J. R. Statist. Soc.* A, **179**, 427–452.
United Nations Environment Programme (2009) Legally binding instrument on mercury. United Nations Environment Programme, Geneva. (Available from `www.chem.unep.ch/mercury/OEWG/Meeting.htm`.)
US Environmental Protection Agency (2004) What you need to know about mercury in fish and shellfish. US Environmental Protection Agency, Washington DC. (Available from `http://water.epa.gov/scitech/swguidance/fishshellfish/outreach/advice_index.cfm`.)
Van Den Brakel, J. A., Brüggen, E. and Krosnick, J. (2015) Establishing the accuracy of online panels for survey research. *International Statistical Institute 60th World Statistics Congr., Rio de Janeiro*.
VanderWeele, T. J. (2005) *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.
Wang, L. and Krieger, A. M. (2006) Causal conclusions are most sensitive to unobserved binary covariates. *Statist. Med.*, **25**, 2257–2271.
Weisberg, H. (2010) *Bias and Causation: Models and Judgment for Valid Comparisons*. Hoboken: Wiley.
Weisberg, H. I., Hayden, V. C. and Pontes, V. P. (2009) Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? *Clin. Trials*, **6**, 109–118.
West, K. K., Hauser, R. M. and Scanlan, T. M. (eds) (1998) *Longitudinal Surveys of Children*. Washington DC: National Academies Press.
Wikle, C. K. (2003) Hierarchical models in environmental science. *Int. Statist. Rev.*, **71**, 181–199.
Wilkinson, R. and Marmot, M. G. (2006) *Social Determinants of Health*, 2nd edn. Oxford: Oxford University Press.
Wirth, K. E. and Tchetgen Tchetgen, E. J. (2014) Accounting for selection bias and association studies with complex survey data. *Epidemiology*, **25**, 444–453.
Wolkewitz, M., Cooper, B. S., Bonten, M. J. M., Barnett, A. G. and Schumacher, M. (2014) Interpreting and comparing risks in the presence of competing events. *Br. Med. J.*, **349**, article g5060.
Yu, B. and Gastwirth, J. L. (2005) Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics*, **6**, 201–209.
Züll, C. and Scholz, E. (2011) Who took the burden to answer on the meaning of left and right?: Response behaviour on an open-ended question. In *Public Opinion and the Internet: Proc. 64th A. WAPOR Conf.* (Available from `http://wapor.unl.edu/wp-content/uploads/2011/09/Zuell_Scholz.docx`.)
Züll, C. and Scholz, E. (2015) Who is willing to answer open-ended questions on the meaning of left and right? *Bull. Sociol. Methodol.*, **127**, 26–42.