J. R. Statist. Soc. A (2016)
179, *Part* 2, *pp.* 1–28

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 'get over 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

 0964–1998/16/179000

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 Bareinbiom (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* that was introduced by Rothman, Gallaches and Hatch (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—'Snart-Gravid'

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 (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, Wise, Sørensen, Riis, Mikkelsen and Hatch, 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 most probably would 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, Wise, Sørensen, Riis, Mikkelsen and Hatch (2013) concerned 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, Wise, Sørensen, Riis, Mikkelsen and Hatch (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, Gallacher and Hatch (2013b) in controlling for differential health awareness by using a sampling frame.

As a postscript, we have been informed through personal communication 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 time to pregnancy, 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 nonparticipation 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 nonparticipation.'

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 Barenboim (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 pair wise) 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, Gallacher and Hatch (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) crystalized 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 relationship 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 relationship 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 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 jury is still out on 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 (1926), 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 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, Gallacher and Hatch (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, Gallacher and Hatch (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, Gallacher and Hatch (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, Gallacher and Hatch (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, Gallacher and Hatch (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) provides 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 particular 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, 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 researches 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 (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 Bareinbiom (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 jury is still out on 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 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 MSE = variance + 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).

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* **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.

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, Gallacher and Hatch (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.*, doi 10.1093/aje/kwv007.

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) Report of the AAPOR task force on non-probability sampling.

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. *et al.* (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.*, doi `10.1093/aje/kwv006`.
Harris, M. L., Luxton, D., Wigginton, B. and Lucke, J. C. (2015) Harris *et al.* respond to "social media recruitment". *Am. J. Epidem.*, doi `10.1093/aje/kwv008`.
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, E. M., Mikkelsen, T., Christensen, A. H., Riis, H. T., Sørensen, K. J. 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 *et al.* (2015)Aapor report on big data. *Technical Report*. 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*, 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. (1926) Observations et expériences concernant des dénombrements représentatifs. *Bull. Int. Statist. Inst.*, **22**, 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**, 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. Met.*, **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: and 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-sant cohort study. *J. Med. Internet Res.*, **16**, article e189.
Miettinen, O. S. (1985) *Theoretical Epidemiology.* New York: Wiley.
Mikkelsen, E. M., H, A., 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.
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.
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**, 899–911.
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.e2.
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. *Pub. 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 (1986) *Modern Epidemiology*. 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. 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. *Sur. 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. 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.*, **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. *PlOsOne*, **7**, article e31396.
Toh, S., Hernádez-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*, no. 1, **12**, 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. (Available from `http://nyti.ms/1H00TPy`.)