Commentary

Statistics in Medicine

(www.interscience.wiley.com) DOI: 10.1002/sim.3888

Published online in Wiley Interscience

Discussion

Roderick J. Little[*†]

The National Children's Study is a massive and important undertaking, and I appreciate the opportunity to discuss this varied set of articles on statistical aspects of the study. My own connection with the NCS was as a member of the Federal Advisory Committee during the formative years, where the basic design of the study was debated. As the only statistician on the committee, I was a strong supporter of a national probability sample design, rather than the 'medical center model', which was the default position for some proponents of the study. I liked to cite Sir Maurice Kendall, Director of the World Fertility Survey [1], a large study that conducted national probability sample surveys on human fertility in over 40 developing countries in the 1970s. Sir Maurice argued that probability sampling was essential because it was the only *scientific* way to select the sample. The economist Robert Michael was also a strong advocate of probability sampling on the Federal Advisory Committee of the NCS, and I recommend his articulate discussion of the issues with Colm O'Muircheartaigh [2].

The question of randomization arises in medical studies both for selection of a sample and for treatment allocation. In clinical trials the focus is on random treatment allocation, with the sample usually being one of convenience. In the NCS the selection is random, and the assignment of treatments or causal agents is not. Randomization in treatment allocation is a tool for protecting *internal* validity, helping to ensure that estimates of causal effects are valid for the sample at hand. Randomization in selection is a tool for protecting *external* validity, meaning that estimates based on the sample can be applied to the target population of interest. When the goals of a study are primarily descriptive (as with the World Fertility Survey), external validity is paramount. When the goals of a study are primarily to discover and disentangle causal effects (as with the NCS), internal validity is job 1—good external validity is no help if internal validity is compromised.

Other major practical issues were involved—coverage, what is measured and where, attrition rates—but I saw the central debate between household probability sampling and medical center models for the NCS as a tussle between how to balance internal and external validity. Nobody was opposed to the idea of probability sampling; the argument was primarily that it would unnecessarily divert resources needed to ensure internal validity, by careful measurement of causal factors, outcomes, and confounders. The latter is particularly important here since the NCS is an observational study, so internal validity is not protected by randomized assignment of the causal agents.

The key to the interplay between internal and external validity is effect modification, that is, interactions between causal agents and baseline characteristics. If causal effects are the same for everyone, or the god(s) of fate randomly distribute effect sizes over the population, then selection of a random sample is unnecessary. In practice, however, effects of causal agents always vary depending on subject characteristics Z, observed or unobserved. In other words, effect modification is close to universal. Given this, the average causal effect can be greatly distorted if the distribution of Z differs markedly for the sample and the target population; random selection limits this difference to sampling error.

This role of randomized selection for valid causal inference is underlined by Ellenberg, who provides a useful concrete example where the utility of a treatment is distorted by studies directed at a highly selected sample. Let me reiterate Ellenberg's example in more abstract language. The treatment (phenobarbital or diazepam) has a potential benefit, the avoidance of future febrile seizures, and a potential drawback, potential negative effects on cognitive function. Let Y be an outcome that weighs these two outcomes in some rational way, with higher values for better outcomes. For example, $Y=1$ if there is no future seizure and no adverse effect on cognition, $Y=0$ if there is a future seizure and adverse effects on cognition, and $0<Y<1$ if there are future seizures with no adverse effect on cognition, or no future seizures with adverse effects on cognition. Following Rubin [3, 4] and others, define the causal effect of the treatment for an individual as the difference D in Y if that individual was assigned the treatment and if that individual was assigned the control. The distribution of D depends on a baseline factor Z in the target population—one such factor is the propensity for future seizures, since if the treatment is effective in preventing seizures, the expected value of D will tend to be high when the propensity for seizures is high, and low if the propensity is low, other things being equal. It is clear that the average treatment effect could be positive if the sample is concentrated on individuals with high propensities for seizure, such as might arise in a sample recruited from medical centers, but negative in the general population, where the distribution of propensities of seizure is lower. Biased selection of the sample thus distorts the estimated average treatment effect in the target population, to the extent that even the sign might be wrong.

Department of Biostatistics, University of Michigan, MI, U.S.A.

[*]*Correspondence to: Roderick J. Little, Department of Biostatistics, University of Michigan, 1420 Washington Hgts., Ann Arbor MI 48109, U.S.A.*

[†]*E-mail: rlittle@umich.edu*

Copyright © 2010 John Wiley & Sons, Ltd.

One is tempted to view the central role of effect modification in this argument as limiting the priority given to random sampling, because (a) effect modifiers are a form of interaction, and most of the emphasis in medical research is on main effects; (b) often studies have very limited power to detect interactions, and they may be omitted from models on grounds of parsimony—indeed the search for effect modifiers is frowned on as data mining; and (c) interactions with unmeasured variables are off the radar screen altogether. However, the fact that effect modification is often not well measured does not mean that it is not important. With some reflection, it is clear that the causal effects studied in the NCS are all subject to modification by baseline factors, to a greater or lesser degree. Given this, and the fact that the method of sample selection pervades all of the NCS study aims, the need to aim for a random sample becomes clear (at least to this observer).

As Montaquilla *et al.* note, most of the analysis of NCS data will be model-based, to deal with confounding factors and repeated measures. I am a strong believer in randomization for selection, even though I am a Bayesian and think that survey analysis should be model rather than design-based [5, 6]. Some Bayesians are lukewarm about randomization, because it is not the source of the inference. The importance of randomization in model-based inference was clarified formally by Rubin [7], who formulated joint models for the selection of the sample, the allocation of treatments, and outcome measures, and clarified that randomization ensures that the selection and allocation processes are ignorable. For non-randomized forms of selection or allocation, the ignorability is an assumption, and often a questionable one at that. Non-ignorable models can be formulated, but they inevitably involve subjective elements and should be avoided if at all possible.

There was an impressive effort to involve scientists with diverse interests in children's health to develop hypotheses for the study; this very democratic approach had the advantage of minimizing serious gaps and creating buy-in for the considerable expenditures involved. On the other hand, the scientists involved all had an interest in advancing their research priorities, and it is easier to say 'yes' than 'no' to any particular topic. As a result, the NCS is extremely inclusive and broad, and faced with serious issues with respondent burden. My own sense was that the planning process might have spent more time on what needs to be measured, rather than on specific hypotheses, since the latter presumably evolve over time and are subject to the make-up of the individuals in the study groups and the 'hot topics' of the day.

I now make some brief comments on the other articles in this set. The Montaqilla *et al.* article provides a clear description of the sample design, and issues involved, particularly the formidable challenge of recruiting women prior to delivery. The general parameters of the multi-stage cluster design seem appropriate, and a lot of thought has gone into the devilish details. The decision to adopt an equal probability design deserves some comment. Broadly speaking I agree that equal probability makes sense in a broad study like this with multiple objectives, since oversampling for one question leads to inefficiencies for others. The avoidance of sampling weights is a useful simplification, although weights still arise in the context of adjustments for unit nonresponse or post-stratification. My one reservation about the clustered equal-probability design is whether it might miss some areas with high levels of environmental pollutants, which we might like to have in the sample. Thus, reserving some modest fraction of the sample for 'interesting' areas seems worth serious consideration; defining 'interesting' is an 'interesting' problem.

One minor comment is that I disagree with their statement that 'sampling weights are not used in model-based analysis'. My own view is that sampling weights are important for robust modeling, although more as covariates than for weighting the data. For more discussion of the role of weights in model-based survey inference, see [5].

The work reported by Strauss *et al.* on designed missing data strategies is most welcome, given the wide range of information being collected and high associated respondent burden. One way to reduce burden would be conceive of a rather limited core set of measurements for the entire sample, and then include modules that are applied to smaller subsamples, yielding a matrix sampling design [8]. Such a design is used to reduce burden in the National Assessment of Educational Progress [9]. Another example of a study employing a designed missing data strategy, more similar to that discussed by Strauss *et al.*, is the Aging, Demographics, and Memory Study [10], a supplement to the Health and Retirement Study (HRS) that conducts in-person clinical assessments for dementia on selected HRS respondents in order to gather information on their cognitive status. A simple questionnaire measure of cognitive impairment is included in another component of the HRS, the Study of Assets and Health Dynamics Among the Oldest Old [11]. The result is a brief measure of cognitive impairment for the full sample, and a detailed assessment of cognitive assessment for a subsample. The detailed measure can be predicted for the cases not in the subsample, with gains in efficiency. Multiple imputation [8, 12] provides a convenient methodology for analyzing the data. These ideas clearly have applications to the NCS, as the authors' simulations demonstrate. I look forward to further refinement of these methods as the NCS develops.

References

1. Cleland J, Hobcraft J. *Reproductive Change in Developing Countries*: *Insights from the World Fertility Survey*. Oxford University: Oxford, 1985.
2. Michael RT, O'Muircheartaigh CA. Design priorities and disciplinary perspectives: The case of the US National Children's Study. *Journal of the Royal Statistical Society*: *Series A* 2008; **171**(2):465–480.
3. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.
4. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Reviews of Public Health* 2000; **21**:121–145.
5. Little RJA. To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* 2004; **99**:546–556.
6. Little RJA. Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician* 2006; **60**(3):213–223.
7. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* 1978; **7**:34–58.

Copyright © 2010 John Wiley & Sons, Ltd.

8. Raghunathan TE, Grizzle JE. A split questionnaire design. *Journal of the American Statistical Association* 1995; **90**:55–63.
9. Mislevy RJ, Beaton AE, Kaplan B, Sheehan KM. Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement* 1992; **29**(2):133–161.
10. Heeringa SG. Technical description of the Asset and Health Dynamics Among the Oldest Old (AHEAD) Study sample design. *HRS Documentation Report DR-003*, Institute for Social Research, University of Michigan, 1995.
11. Langa K, Plassman B, Wallace R, Herzog AR, Heeringa S, Ofstedal MB, Burke J, Fisher G, Fultz N, Hurd M, Potter G, Rodgers W, Steffens D, Weir D, Willis R. The aging, demographics and memory study: study design and methods. *Neuroepidemiology* 2005; **25**:181–191.
12. Rubin DB. *Multiple Imputation for Nonresponse in Censuses and Surveys*. Wiley: New York, 1978.

Copyright © 2010 John Wiley & Sons, Ltd.