Methodological Issues in the Meta-Analysis of Observational Studies: Discussion

David A. Binder
[1]Statistics Canada (retired) 49 Bertona Street, Nepean, Ontario, Canada K2G 4G7

Abstract
Papers by Lauren Griffith, George Wells, and Karla Fox are discussed. Emphasis is given to the reasons for conducting meta-analyses of observational studies and some general concerns when conducting such analyses.

Key Words: Comparability, Modeling, Causality, Scientific inferences

1. Overview – Comparability and Modeling

In this discussion, rather than focus on the details of each of the presentations, I consider the reasons for conducting meta-analyses of observational studies and some concerns one should have with such analyses. Recently, Arlene Swern (2010) mentioned to me that, "People are under the impression that meta-analysis is some magic box into which they can throw vastly disparate studies and get the 'true' answer." The availability of software to perform meta-analyses can lead many to use this tool without thinking about the statistical underpinnings behind the methods. It is true that meta-analyses can yield useful results, but only in cases that satisfy conditions of comparability, and where appropriate modeling has taken place.

As Lauren Griffith pointed out, systematic reviews and meta-analyses of observational studies are more common now. However, one should ask what, in general, are the reasons for combining the studies? Obviously increasing sample size is a goal, in order to obtain sharper inferences. However, if biases are introduced through the use of inappropriate methods, the conclusions can be misleading. This was emphasized by Karla Fox in her discussion of using model-based randomization assumptions when combining data from complex surveys.

I would maintain that the underlying goal of meta-analyses is usually to examine the possibility of *causal relationships*. This is true both for combining observational studies and for combining randomized studies. None of the authors has addressed this issue explicitly. However, it is clear that showing causality is often an implicit goal. However, the use of observational studies to infer causality is somewhat controversial. This is discussed this further in Section 2.

Comparability of the data and appropriate modeling are critical components for inferences using data arising from a variety of sources. This was elucidated nicely in Schenker and Raghunathan (2007). To achieve comparability, Lauren Griffith describes the process of conducting systematic reviews of the individual studies. Clearly, this time-consuming process is important to ensure that the data are being appropriately analyzed.

Lauren Griffith pointed out that combining observational studies is useful for situations where one needs to synthesize the evidence in areas of research that are not amenable to randomized controlled trials.

For modeling the data, Karla Fox addresses some of the issues when combining several observational studies from survey data. She correctly points out that one must be careful to account for differences among surveys, even when variables are comparable. Differences can be due to difference in the target populations associated with the individual surveys, as well as differences in the mode of data collection. Such differences apply not just to survey data, but to any observational study, or non-randomized study, as mentioned by George Wells.

Figure 1 shows a randomization framework for combining data from several surveys. Some discussion of this framework is given in Roberts and Binder (2009).

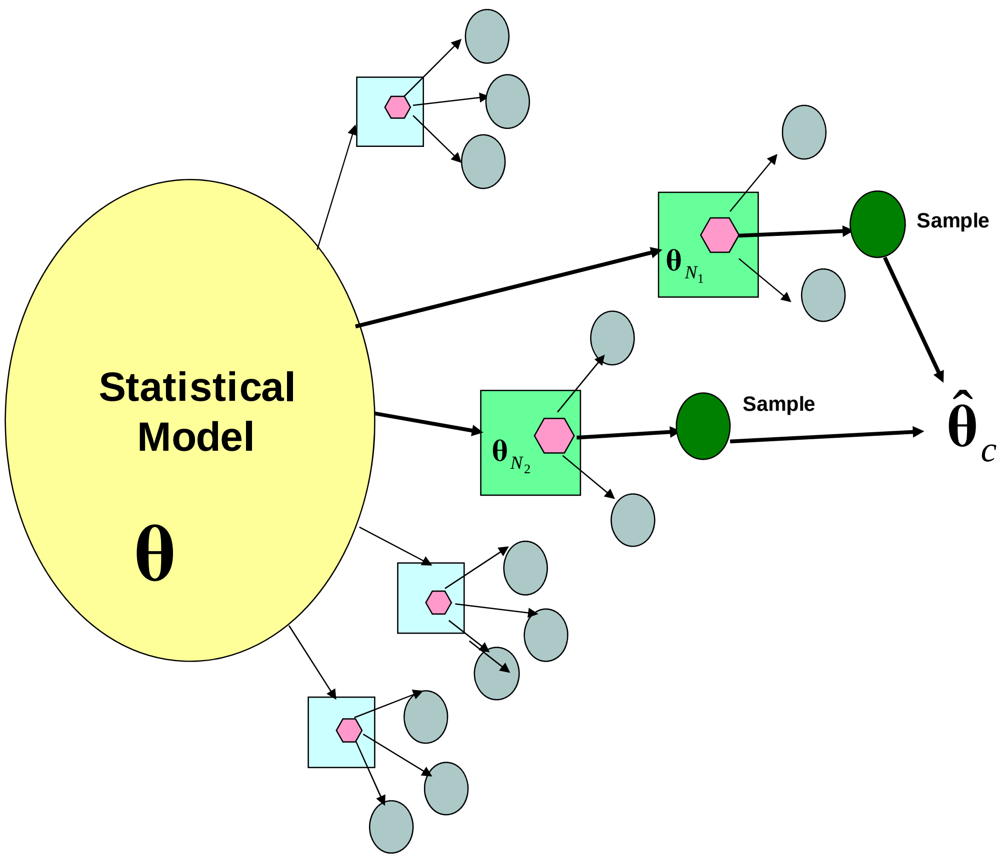

Figure 1: The model-design-based framework for combining data from multiple surveys

In Figure 1, for the first randomization stage, a statistical model with parameter $\boldsymbol{\theta}$ gives rise to the values of the finite population characteristics. Conceptually, there are finite populations that could have been generated, but are not included in the study. Comparability among finite populations is achieved by including in the model any effects due to differences in the surveys (such as the effect due to mode of collection). For the ith finite population, there is a finite population quantity $\boldsymbol{\theta}_{N_i}$, corresponding to an estimator of $\boldsymbol{\theta}$ based on the values for all the finite population units.

The second randomization stage involves taking samples from some of the finite populations using a well-defined, possibly complex, sample design. Again, conceptually, other possible samples could have been selected other than the observed ones. The observed combined sample data is then used to obtain an estimate of $\boldsymbol{\theta}$, given by $\widehat{\boldsymbol{\theta}}_{\boldsymbol{c}}$. It can be seen that in this framework both comparability and modeling have roles to play.

To ensure that a researcher can assess the applicability of an observational study for the research topic being considered, it is important to have available suitable documentation on the studies being reviewed. George Wells reports on the efforts being undertaken to ensure certain protocols for documentation are being followed. He explains what is needed to conduct a systematic reviews of non-randomized studies before engaging in the meta-analysis. These include several details on how the data were obtained, possible sources of bias, and how the data were analyzed.

However, meeting all the requirements for reporting on how the data were collected and analyzed does NOT imply that the conclusions have scientific validity. Such conclusions become valid only when the assumptions for causality are correct.

2. Inferring Causality from Observational Studies

As Freedman (1999) points out, in many situations randomized experiments are impractical or unethical. Most of what we know about causation in such contexts is derived from observational studies. Delicate judgments are required to assess the probable impact of confounders (measured and unmeasured), other sources of bias, and the adequacy of the statistical models used to make adjustments.

Holland (1986) refers to the “Fundamental Problem of Causal Inference”; that is, it is impossible to observe the value of a treatment and a control on the same unit and, therefore, it is impossible to observe the effect of the treatment. This is true even with randomized controlled trials. In Cox and Wermuth (2004), the variables included in a study of causality can include primary responses, potential causes, background variables, and intermediate variables. The classification of these variables and how they affect the analysis can be based on subjective knowledge, or can be based on what is known about the variables. As Pearl (2009) states, *associational assumptions* are testable in principle, given sufficiently large sample and sufficiently fine measurements; however, *causal assumptions* cannot be verified even in principle, unless one resorts to experimental control. Pearl advocates the use of Structural Causal Model diagrams to clarify role of the variables. These can prove quite useful for understanding the processes behind what is being studied.

A more cautious viewpoint was discussed by Thompson (2006), where she stated that careful interpretation of associations can lead to understanding the *glimpses of causality*. This seems to be what George Wells is also advocating; namely that observational studies can lead to useful speculations as to the underlying causes.

On the other hand, some statisticians are more forceful in their rejection of causal models. Rogosa (1987) advocates that a critical distinction be made (a) building statistical models for the processes that generate the social science data and (b) tossing the data at available statistical methods. Although his comments were directed more to the case of analyzing social science data, Rogosa’s perspective is worth paying attention to even in biostatistical contexts. Rogosa's concerns would be relevant for the analysis of survey data from population-based surveys that are discussed by Karla Fox.

My own perspective here is that when causality is inferred from observational data, the assumptions being made for the validity of any conclusions must be clearly stated in terms that the audience for the study can understand. I do agree with Thompson's (2006) phrase that some studies can provide a "glimpse of causality". If it feasible to obtain data from randomized trials to confirm the conclusions, these should be conducted.

3. Some Additional Remarks

Pearl's Structural Causal Model diagrams can be very useful in the understanding of the role of the variables used in the analysis. Would this lead to a clearer understanding of the models used by Lauren Griffith?

No mention is made of the possibility of using cross-validation methods in the context of analyzing several studies in this meta-analysis framework. Is there scope for applying such techniques here? This seems like a logical possibility given that the data are naturally grouped into the studies. But, other possibilities may also be considered. For example one could take training samples and validations samples that cross-cut each of the individual studies. For some discussion of how survey bootstrap replicates have been used for model-building and cross-validation, see Rowe and Binder (2008).

References

Cox, D. R. and N. Wermuth. (2004). Causality: A statistical view. *International Statistical Review*, 72. 285-305.

Freedman, D. 1999. From association to causation: Some remarks on the history of statistics. *Statistical Science*, 14. 243-258.

Holland, P. 1986. Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81, 945-970.

Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3, 96-146.

Roberts, G. and D. Binder. 2009. Analyses based on combining similar information from multiple surveys. In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 2138-2147.

Rogosa, D. 1987. Casual models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational Statistics*, 12. 185-195.

Rowe, G. and D. Binder. 2008. Can survey bootstrap replicates be used for cross-validation? In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 1430-1437.

Schenker, N. and T. E. Raghunathan. 2007. Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, 26. 1802-1811.

Swern, Arlene S. 2010. Private communication.

Thompson, M. E. 2006. SSC Gold Medal address, unpublished.