Executive Summary: Discussion on Utility of Validation Samples for the NCS

by

Warren Strauss, Louise Ryan, Jeff Lehman

For longitudinal exposure studies like the NCS, one of the important considerations when planning and designing the study is the need to introduce resource efficiency in the data collection effort while maintaining data quality. To satisfy the scientific objectives of the study, such as collecting sufficient data to adequately assess study hypotheses, data quality is a primary concern; however, to satisfy the resource limitations of the study and maintain the feasibility of the study, efficient data collection is necessary and may often be in opposition to the need to collect detailed (and expensive) study subject information. In this report we discuss and illustrate the use of validation samples for introducing efficiency in the data collection effort.

A validation sample is a small sample that is designed to provide information related to the bias or error introduced by using alternative measures of exposure. The basic idea is that in cases where a reasonable surrogate measure of exposure, such as a lower cost, less detailed, or less accurate measure, is available, it may not be necessary to collect the ”ideal” exposure information for the entire cohort resulting in reduced costs, reduced subject burden, and increased study feasibility. Instead, validation samples, in which both the surrogate measure (Z) and “ideal” measure (X) are collected for a small portion of the cohort, can be used to estimate the relationship between the true measure of exposure (X) and the surrogate measure of exposure (Z). (Note that by “ideal” we are referring to a gold-standard measure of exposure that is of interest in explaining the health outcome but is presumed to be expensive or difficult to collect across the entire cohort.) By capitalizing on the relationship between these alternative measures of exposure, statistical methods can be applied to correct for bias and error when estimating the relationship between a health outcome of interest (Y) and the true measure of exposure. Thus, validation samples have the potential to allow the NCS to capitalize on less precise/accurate measures of exposure for the majority of the cohort while still preserving the ability to assess the impact of “true” exposure on the health outcomes of interest (assuming that true exposure can be assessed on a subset of the cohort). This can lead to significant cost savings when accurate or precise exposure assessment is very expensive and when reasonable, less expensive, surrogate measures are available.

To measure the loss of statistical efficiency as a result of using the validation sampling approach, we compute a design effect that is the ratio of the variance of the estimate of the relationship between Y and X under the validation sampling approach versus the corresponding variance under an approach that measures Y and X on the entire cohort. The magnitude of this design effect allows us to assess the magnitude of the loss of information resulting from the use of a validation sample as opposed to collecting the true exposure information for the entire cohort. Of course, the degree of this loss depends on a number of factors, including: the availability and accuracy of a surrogate (less expensive/detailed) measure of exposure, the strength of the exposure/outcome relationship, the methods used in selecting the validation sample, and the size of the validation sample. Certainly, from a purely statistical perspective the use of a validation sample is less optimal than collecting detailed/burdensome/expensive data for the entire cohort (all design effects are greater than 1 indicating some loss of efficiency); however, from a feasibility and resource efficiency perspective, the use of validation samples can play an important role in allowing the NCS to collect information that will allow adequate assessment of the study hypotheses while maintaining cost efficiency and study feasibility.

As an example of the results presented in this report, Figure 1 displays design effects as a function of the strength of the relationship between *X* and *Z* for validation samples selected at random and as an outcome dependent sample (where subjects are selected into the validation sample based on their observed health outcome), with a validation sample size that is fixed at 10 percent and 5 percent of the original cohort size. To evaluate the effectiveness of the validation sample, consider that if only the subjects in the validation sample were included in the analysis, the resulting design effects would be 10 for the validation sample of size 10 percent of the original cohort size and 20 for the validation sample of size 5 percent of the original cohort size (i.e., the ratio of the full cohort size to the subsample size). As seen in the figure, provided there is some reasonable surrogate measure for the exposure of interest, there can be relatively little loss of statistical efficiency (e.g., design effects less than 2.0 when the portion of the variability in *Z* that is explained by *X* is greater than 0.50).

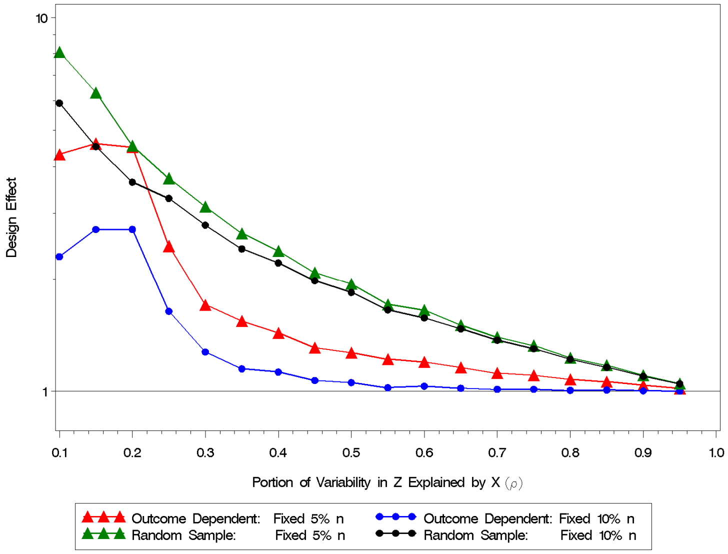

Figure 1. Design effects for outcome dependent and randomly selected validation samples of fixed size (n=500 and n=1000).

To evaluate a more concrete example, suppose that the study collects some expensive exposure measure (*X*) for approximately 10 percent of the cohort, and suppose that there exists a less expensive alternative measure of exposure (*Z*) that is collected across the entire cohort and that has 50 percent of its variability explained by the expensive exposure measure. Figure 1 suggests that even if the subjects for which *X* is measured are selected at random, if we appropriately combine the information contained in *X* and *Z* and relate that information to the health outcome of interest, the result of the validation sample approach would be a design effect on the order of 2.0. This implies that the information available when using the validation approach is equivalent to the information that would be collected if *X* were measured for half (inverse of the design effect) of the original cohort. Assuming it costs one financial unit (e.g., one financial unit may be 1 million dollars) to collect the surrogate measure of exposure across the entire cohort and ten financial units to collect the “true” measure of exposure across the entire cohort, the 10 percent validation sample approach would result in a total cost of two financial units. Comparing this to the five financial units that would be needed to collect both X and Y for half of the original cohort, thereby obtaining the same amount of information as the validation approach (statistically speaking), the potential to reduce costs is significant (costs are 60 percent less for the same amount of information). The question then becomes: is a design effect of 2.0 acceptable given the financial savings that could be realized in using the validation sample approach? In some cases, the cost savings may clearly outweigh the loss of statistical efficiency while in other cases the choice may not be as apparent, requiring further consideration of the size of the validation sample that is needed and/or the accuracy of the surrogate measure of exposure.

Validation samples can also assist study planners in maintaining the feasibility of the study by minimizing study subject burden. For example, when appropriate and acceptable in terms of the loss in statistical efficiency, the validation sample methodology can be used to limit the burden on a large portion of the cohort by only requiring detailed (i.e., burdensome) data collection for a relatively small portion of the cohort (depending on the factors mentioned previously). Additionally, since there are likely many study hypotheses that involve assessment of the relationship between a health outcome and a detailed measure of exposure, for those situations where validation samples are acceptable, we can “spread” the subject burden over the cohort by selecting different subjects in the validation sample for different hypotheses.

Finally, the need for pre/peri-conception exposure information, and the costs/difficulties associated with recruiting and retaining subjects identified in the pre-conception period, can also be addressed by validation samples. In this case, retrospective measures of exposure could be considered the surrogate measure of exposure and could therefore eliminate the need to recruit the entire cohort in the pre-conception period. For example, a small portion of the cohort could be recruited as the pre-conception validation sample that undergoes all of the desired pre/peri-conception data collection, and the remainder of the cohort could be recruited at a later time (e.g., sometime during pregnancy) with their pre/peri-conception exposure information assessed retrospectively through combinations of questionnaires and other retrospective measures. This would allow the NCS to avoid the cost inefficiencies associated with following a large number of women that fail to become pregnant during the study recruitment period, and to utilize other/more efficient sampling strategies (e.g., sampling through OB/GYN offices) to recruit the majority of study subjects.

Of course, there remain several areas for further work in order to better integrate the concept of validation samples into the NCS data collection protocol. Possible areas are: (1) further development of statistical methods and calculations to assess the impact of validation samples for a variety of settings (including longitudinal data measurement), (2) investigation and determination of suitable surrogate measures of exposure, and (3) development of tools, such as software, that would more easily enable NICHD to integrate validation samples into the data collection protocol. By further investigating these areas, utilization of validation samples can be made more practical for study planners, and can ultimately lead to a more feasible and more cost efficient study.