Discussion on Utility of Validation Samples for the NCS

by

Warren Strauss, Louise Ryan, Jeff Lehman

1. Introduction

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. Certainly, in order to satisfy the scientific objectives of the study, such as collecting sufficient data to adequately assess study hypotheses, data quality is a primary concern. On the other hand, in order to satisfy the resource limitations of the study and maintain the feasibility of the study, efficient data collection is also necessary. In many cases, these two study components can represent competing objectives, making determination of the appropriate balance between efficient data collection and sufficient data collection a difficult decision. For the NCS, where study hypotheses range over a large number of exposures and health outcomes, detailed, and perhaps expensive, data collection is a necessity. Given the size of the cohort envisioned for the NCS along with this need for detailed information, efficient and sufficient data collection becomes an even more critical component for controlling the financial costs of the study, minimizing study subject burden (e.g., minimizing the amount of time/effort required of study subjects), and maintaining study feasibility. In this paper we discuss the concept of validation samples for introducing efficiency in the data collection effort, and illustrate their usefulness in this setting.

In general, we define a validation sample as 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 and “ideal” measure are collected for a small portion of the cohort, can be used to estimate the relationship between the true measure of exposure and the surrogate measure of exposure. (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 and the true measure of exposure. Thus, validation samples would 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).

It is for this reason that validation samples can lead to significant cost savings when accurate or precise exposure assessment is very expensive and when reasonable, less expensive, surrogate measures are available. For example, assessing aggregate pesticide exposure through multiple pathways may be an expensive endeavor, whereas, collecting questionnaires about diet and consumer product use might offer a surrogate measure of this aggregate pesticide exposure at a much lower cost. Alternatively, validation samples can lead to a much more feasible study, particularly when it comes to the notion of a pre-conception or peri-conception sample. In this case, we might consider the true exposure measure to be an exposure measure during the peri-conception period and the surrogate measure might be a retrospective exposure measure that is assessed later during pregnancy. Provided this retrospective measure acts as a reasonable surrogate for the true exposure measure, the validation sample methodology would then only require that a small portion of the NCS cohort be sampled in the pre/peri-conception period with the remainder of the cohort identified at a later period (e.g., during the second trimester of pregnancy).

In previous work with EPA, Battelle investigated the use of validation samples for the acquisition of exposure data in longitudinal studies. The specific objective of the project was to develop cost-effective statistical sampling strategies and optimal design considerations for the collection of exposure data in the NCS. In terms of validation samples, the main conclusion of this work was that detailed exposure assessment/information on all study subjects was likely not necessary to characterize the effect of exposure on the health outcome. Instead, carefully designed validation samples (or sub-studies) should be considered as a cost effective strategy to collecting the necessary data. Of course, the work identified a number of factors that must be considered when evaluating the feasibility of validation samples. These relevant factors include: exposure period of interest, pathways of exposure, sources and magnitudes of exposure variability, exposure metric related to disease, strength of exposure/outcome relationship, distribution of exposure and health outcome, and the availability of surrogate measures of exposure. We refer the reader to Strauss et al., 2003 for further information on this previous work.

The goal of this paper is to provide further guidance on the utility of validation samples for the NCS. In particular, we seek to provide guidance on the following questions:

- What is the size of the validation sample that is necessary to effectively support NCS inferences?

    - How does the sample size of the validation subset relate to the size of the cohort?
    - How does the sample size of the validation subset relate to the strength of the relationship between the ideal and surrogate measures of exposure?
    - How does the sample size of the validation subset relate to the strength of the relationship between exposure and health outcome?

- What are the different approaches to selecting the subjects included in the validation sample, and how do these different sampling approaches impact the answers to the above questions?

- What future work is necessary in order to develop tools that would allow NICHD to integrate validation samples into the data collection protocol?

To this end, the remainder of this report is organized in the following way. Section 2 provides an illustration on the use of validation samples when a surrogate measure of exposure is available, and Section 3 discusses the implications of these results and the advantages and limitations of relying on validation samples in the NCS. More specifically, Section 3 discusses the use of validation samples in covering pre-conception and peri-conception hypotheses – both in the case where an outcome dependent design can be assumed (e.g., utilize archived samples) and in the case where the validation sample must be identified prior to knowledge about health outcomes and/or exposure factors. Additionally, Section 3 provides guidance on how low cost (less detailed) exposure measures can be used across the NCS cohort, with a small validation sample undergoing additional detailed exposure assessment. Finally, Section 4 discusses areas for future (more extensive) work to provide tools (e.g., software) that would more easily enable NICHD to integrate validation samples into the data collection protocol.

2. Illustration of Validation Samples

Suppose we are interested in the relationship between an outcome variable, Y, and some measure of exposure X. Letting X_i be the true exposure for individual i, and Y_i this individual's corresponding response (assume the response is binary for our discussion here), a logistic regression model relating Y and X is:

$$\text{Logit}[\Pr(Y_i=1|X_i)] = \text{Logit}(\mu_i) = \beta_0 + \beta_1 X_i \quad \text{for } i=1,\ldots,n,$$

with β_1 representing the relationship between the exposure and the health outcome. The significance of the relationship between X and Y is assessed by evaluating the significance of the parameter β_1, which is the natural log of the odds ratio (the ratio of the odds of the outcome for a unit increase in X). For the case where X is easily and inexpensively measured, there is little need for a validation sample since X can be measured for each study subject and the logistic model straightforwardly applied to the resulting data. However, when X is expensive and/or difficult to measure, such as is the case when X represents a detailed environmental measure or an exposure measure in the periconception period, it may be more optimal to introduce a less expensive or more feasible surrogate measure of exposure (Z) to be measured across the entire study population with a smaller representative sample (i.e., the validation sample) of study participants selected to have both X and Z assessed. In other words, suppose that for reasons of budget and/or logistical constraints, the study can only afford to measure the covariate X on a subset of all study subjects; whereas Z can be measured for all study subjects. (Note that in practice both X and Z are likely to be vectors; however, for purposes of our discussion here we assume that X and Z are univariate.)

In general, selection of the subset of individuals for which X is measured (i.e., the subjects in the validation sample) can be done randomly or can depend on the outcome or some other covariate of interest (e.g., the surrogate measure of exposure). For example, an outcome dependent design may call for measurement of X for all subjects with the outcome of interest and a random selection of subjects without the outcome of interest. Alternatively, a design that depends on the surrogate measure of exposure, a covariate-dependent design, may call for measurement of *X* for all subjects for which *Z* exceeds some threshold value. Depending on the logistics of the health outcome and exposure measurements of interest, each of these strategies for selection of the validation sample may be needed, and each strategy will be associated with different impacts when estimating the parameters of interest (i.e., the relationship between the outcome and the exposure). For the case where *X* can be measured after observing the health outcome, such as if *X* represents an exposure that is measured in archived blood samples, an outcome dependent design may be most promising. On the other hand, other situations may only accommodate a randomly selected validation sample, such as when *X* represents an exposure measure during the peri-conception period and only a random segment of the cohort is recruited during the peri-conception period.

Detailed statistical calculations that allow computation of the impact of introducing the validation paradigm into the data collection effort, in terms of the variance of parameter estimates, are currently being developed; however, in this paper, we provide a simple illustration of the utility of validation samples using a simulation based approach in which maximum likelihood estimation is used to correct for bias and error introduced in estimation of the relationship between *X* and *Y* when only observing *X* for a subset of individuals and observing *Z* for all individuals. To this end, assume the surrogate measure of exposure, *Z*, is related to *X* in the following manner:

$$Z = X + \varepsilon,$$

where the random variable ε represents a random deviation of the value of *Z* from the value of *X* (i.e., *Z* is considered an error-prone version of *X*). Letting σ_x^2 be the variability of *X* and σ_ε^2 be the variability of ε, we define the strength of the relationship between *X* and *Z* as $\rho = \dfrac{\sigma_X^2}{\sigma_X^2 + \sigma_\varepsilon^2}$, the portion of the variability in *Z* that is explained by *X*.

In this case, since *Z* is measured for the entire cohort and *X* is measured for a subset of subjects, maximum likelihood methods are used to estimate the relationship between *X* and *Y* for the entire cohort (not just the subjects for which both *Y* and *X* are measured). Heuristically, the analysis estimates the relationship between *X* and *Z* based on those subjects for which both measurements are available, and then estimates the relationship between *X* and *Y* for the entire cohort by utilizing the relationship between *X* and *Z* and treating *Z* as an error-prone measure of *X* for those subjects for which *X* is not available.

To demonstrate the utility of validation samples we simulate example realizations of the data resulting from the validation approach, estimate the parameters of interest using maximum likelihood, and evaluate the uncertainty associated with these parameter estimates as a function of a variety of important factors. This evaluation is used to demonstrate the potential that validation samples have to introduce efficiency into the data collection protocol, while minimizing the loss of information that may result. Assume the following:

- *X* is distributed normal with mean zero and standard deviation 1

- ε is distributed normal with mean zero and standard deviation determined by the portion of variability in *Z* that is explained by *X* (the parameter ρ).
- The parameter ρ takes on values ranging from 0.10 to 0.95 to demonstrate the impact of having strong and weak relationships between *X* and *Z*.
- The total cohort size is defined to be n=10,000. (Note that this is done in order to limit the computational time needed to apply these models for a large number of simulations. Similar results apply to the case where n=100,000)
- The strength of the relationship (defined by β_1) takes on a value such that under a design for which *X* is collected for all individuals there is approximately 80 percent power to detect the relationship.
- The probability of *Y* given $X = 0$ is 0.025. In other words, for individuals experiencing average exposure, the probability of disease is 0.025.

Using these assumptions, we simulate realizations of the study data (i.e., simulate *Y*, *X*, and *Z* for 10,000 individuals), apply the validation sampling approach (i.e., select those individuals for which *X* is observed), and fit appropriate statistical models relating *Y* to *X*. 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 log odds ratio estimate under the validation sampling approach versus the variance of the log odds ratio estimate under an approach that measures *Y* and *X* on the entire cohort (in this case the surrogate measure *Z* is of no use when applying the model).

In the following figures, we display design effects associated with validation samples that are a fixed proportion of the initial cohort size: 10 percent and 5 percent yielding 1000 and 500 subjects in the validation sample, respectively. Additionally, we evaluate validation samples where the number of individuals in the validation sample depends on the strength of the relationship between *X* and *Z*. This may indicate whether the size of the validation sample is related to the number of (*X*, *Z*) pairs that are necessary to accurately establish the relationship between "true" and "surrogate" measures of exposure. To do this, we fix the variance of the sample standard deviation of ($X - Z$) at a constant that is defined as the value for which the validation sample size is 1000 or 500 when ρ=0.50. In other words, for the validation samples with a variable sample size (labeled "*variable n*" in the figures), the size of the validation sample is 1000 or 500 when ρ=0.50 and the size decreases (increases) as the strength of the relationship between *X* and *Z* increases (decreases). Finally, we investigate the impact of both random sampling of individuals for the validation sample, as well as outcome dependent sampling (using a 2 to 1 proportion of non-diseased individuals to diseased individuals).

Figures 1 and 2 display 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. In particular, Figure 1 displays design effects when the size of the validation sample is fixed at 10 percent (1,000 subjects) and 5 percent (500 subjects) of the original cohort size for both an outcome dependent sampling approach and a random sampling approach. 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 1,000 subjects and 20 for the validation sample of size 500 (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 ρ>0.50). This is especially the case when an outcome dependent sampling design can be utilized.

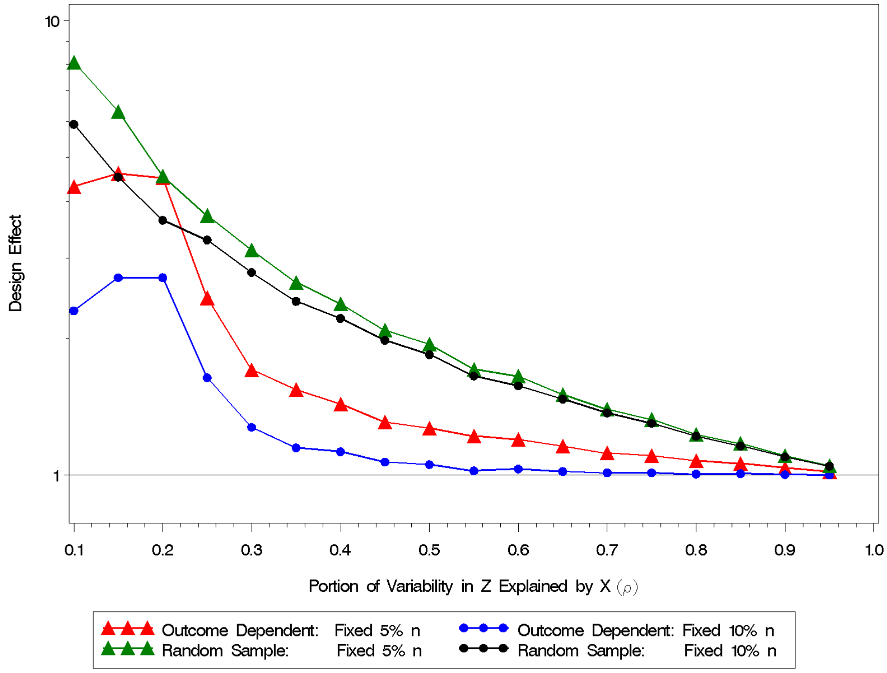

Figure 1. Design effects for outcome dependent and randomly selected validation samples of fixed size (n=500 and n=1000).

Figure 2, on the other hand, displays design effects when the size of the validation sample is allowed to vary depending on the strength of the relationship between *X* and *Z*. In particular, when ρ=0.50, the size of the validation sample is 5 percent or 10 percent of the original cohort size, resulting in 500 or 1,000 subjects in the validation sample. In this case, note that since we allow the number of subjects in the validation sample to vary as a function of the strength of the relationship between *X* and *Z*, the impact of the validation sample in terms of statistical efficiency is even smaller. Of course, since the size of the validation sample increases as ρ decreases, the number of subjects in the validation sample eventually includes the entire cohort (i.e., all 10,000 individuals) which would result in no cost savings in the data collection effort.

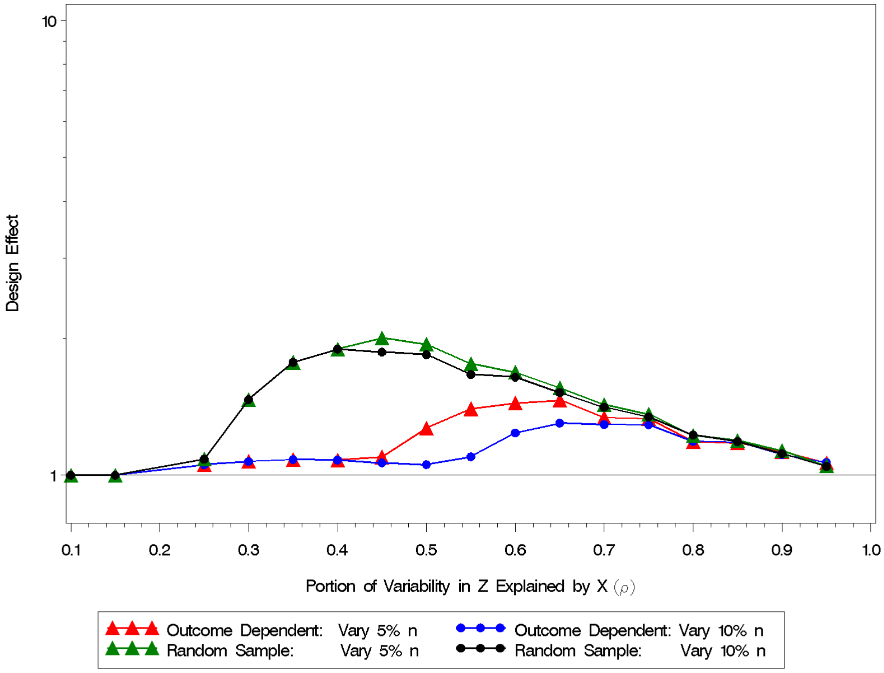

Figure 2. Design effects for outcome dependent and randomly selected validation samples of variable sample size (n=500 and n=1000 when ρ=0.50).

3. Discussion

The results presented in Section 2 illustrate the use of the validation sample paradigm, and indicate that it offers a promising means of introducing efficient data collection into the NCS without losing the ability to assess relationships between exposure measures (even expensive exposure measures) and health outcomes of interest for the entire cohort. In particular, the results 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. The degree of this loss (and therefore the answers to the questions posed in Section 1) 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. As demonstrated and as expected, from a purely statistical perspective the use of a validation sample is less optimal than collecting detailed, burdensome, and/or expensive data for the entire cohort (all design effects are greater than 1 indicating some loss of efficiency); however, from a feasibility, burden, and resource efficiency perspective, the use of validation samples may play a very 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 a concrete example of the potential for cost savings resulting from validation samples, 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. Even if the subjects for which *X* is measured are selected at random, the results in Section 2 suggest that by appropriately combining the information contained in *X* and *Z* and relating that information to the health outcome of interest, the impact of introducing the validation sample 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 necessary and/or the accuracy of the surrogate measure of exposure.

In terms of maintaining study feasibility, validation samples can help study planners to minimize 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.

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.

We would also like to highlight the relatively small loss of statistical efficiency when utilizing outcome dependent designs in combination with validation samples. If the exposure of interest lends itself to selection of the validation sample after the health outcome has been observed, such as may be the case when the exposure can be measured in archived blood samples, then the use of outcome dependent designs offers the potential for very little loss of statistical efficiency with very large gains in resource efficiency. That said, even in the case where outcome dependent designs are not available, randomly selected validation samples, or validation samples that are selected based on the value of some covariate that is related to the exposure or health outcome of interest, will also result in a relatively small loss of information (note that the latter approach, covariate-dependent designs, have not been demonstrated in the examples of Section 2).

Finally, we would like to inform the reader that there are several alternative approaches to analyzing data of the form described here (namely, data subject to possibly outcome dependent missingness). In the above examples, maximum likelihood estimation (MLE) approach was utilized to assess the relationship between the exposures and outcomes of interest. Alternatively, a weighted estimating equations (WEE) approach could be utilized. As has been discussed by a number of authors, there are advantages and disadvantages to both approaches. The MLE approach requires that distributional properties be specified for the covariate *X*, as well as for the outcome of interest *Y*; whereas the WEE approach removes this assumption in that it does not require any distributional assumptions on *X*. So long as the assumed model is correct (i.e., the assumed distributions are correct), the MLE method will have the highest degree of efficiency; however, the disadvantage of the MLE approach is that it may lead to invalid conclusions when the model has been mispecified. The WEE approach solves the robustness problems of the MLE approach in that it will lead to valid inference regardless of the distribution of *X*, but it can be quite inefficient, especially when only a small proportion of the subjects are sampled on *X*. As suggested in Section 4, further investigation of the advantages and disadvantages of these two approaches is a promising area for further research.

4. Future Work

In order to better integrate the concept of validation samples into the NCS data collection protocol, several areas for future (and more extensive) work are needed. Possible areas are as follows:

- Further development of detailed statistical calculations that allow computation of the impact of introducing the validation paradigm into the data collection effort. These calculations are currently under development, and have the potential to allow a more straightforward assessment of the statistical efficiency associated with validation samples under a variety of settings (e.g., outcome dependent sampling, covariate dependent sampling, alternative sample sizes for the validation sample, MLE and WEE analysis techniques, etc.).

- Further methods development work in order to more extensively investigate the impacts of validation samples under more complex settings. In particular, we highlight the need to better consider validation samples in the longitudinal setting and the need to develop examples of the use of covariate dependent sampling (e.g., using a cheap/easy exposure measure to determine whether the expensive exposure measure is collected).

- Investigation, perhaps through pilot studies, and determination of measures of exposure that are suitable surrogates for the detailed measures of exposure that are difficult and/or expensive to collect, and for which it might be acceptable to use the validation sample methodology.

- 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.

5. References

Strauss, W.J., Ryan, L. Lehman J., et. al., “Development of Exposure Assessment Study Design for the National Children’s Study” A Series of Technical Reports submitted by Battelle to U.S. Environmental Protection Agency, National Exposure Research Laboratory under Task Order 19 on Contract 68-D-99-011, May-June 2003.