Executive Summary - Efficiency Considerations for Clustering the National Children's Study

by

Jeff Lehman, Warren Strauss, Louise Ryan, Steve Rust

Since one of the NCS givens calls for the study to "… include (geographic) clustering of samples to allow for efficient collection of exposure and outcome measures, and measurement of context (physical and social)", an important question that must be considered when designing the study is what degree of clustering should be employed. In other words, how many clusters are needed for the NCS to efficiently collect the necessary data while maintaining the ability to powerfully assess the hypotheses of interest in the NCS? The answer to this question depends on a number of factors (such as data collection costs, within versus between cluster variability, data availability considerations, etc.) related to the characteristics of the exposures and the health outcomes considered in the NCS hypotheses. In this paper, we evaluate the loss/gain in statistical efficiency as a function of the number of clusters by comparing designs with 25, 50, 100, 250, and 500 clusters (each assuming a total sample size of 100,000 individuals) in terms of their ability to estimate/detect relationships between continuous outcomes and continuous exposures and between binary outcomes and binary exposures. These statistical comparisons depend on the sources of variability/covariability in the exposures and the health outcomes of interest. In particular, the amount of within cluster correlation in the health outcome (denoted as ρ) along with the amount of within cluster correlation in the exposure factor (denoted as λ) has significant influence on the loss/gain in statistical efficiency resulting from clustering the design. Generally speaking, when these correlations are small, so that the health outcome and the exposure factor vary significantly within cluster, there are little to no differences between designs with 500 clusters of 200 individuals and designs with 25 clusters of 4000 individuals (i.e., there is little loss of information resulting from clustering the design). On the other hand, for cluster-specific exposure factors (i.e., an exposure factor that only varies between clusters and is constant within clusters) and for larger values of the within cluster correlation in the health outcome, the impacts of design clustering are less trivial.

Figure 1 and Table 1 display example comparisons of the impact of clustering the NCS when estimating relationships across the entire cohort. In particular, Figure 1 displays relative design effects and power for estimating/detecting a relationship between a continuous outcome and a continuous exposure when the within cluster correlation in the health outcome is 0.01 and as the within cluster correlation in the exposure varies between 0 and 1. In this figure, the relative design effect represents the variance of the parameter estimate under the selected design to the corresponding variance under a design with 250 clusters of size 400 individuals and the displayed power represents the power to detect a relationship that has a magnitude which is detectable with 80% power under a design with 250 clusters of 400 individuals (note that this magnitude may change as a function of ρ and λ). By so doing, the figure indicates the loss of (or gain in) efficiency when attempting to estimate a relationship under the selected design versus

adopting a design with 250 clusters (i.e., treating the 250 cluster design as a reference design). For this case, the figure demonstrates that there is very little loss of efficiency/power as a result of design clustering when the within cluster correlation in the health outcome is relatively small (i.e., ρ=0/.01) and the within cluster correlation in the exposure factor is small (e.g., λ less than 0.5). However, note that when assessing relationships between health outcomes and cluster-specific exposure factors (λ=1), there may be a substantial loss in power when going from a 250 cluster design to a 25 cluster design (e.g., from 80% power to approximately 20% power).

Table 1, on the other hand, summarizes the comparison of these designs when evaluating the relationship between a binary outcome and a binary exposure. In particular, the table displays the odds ratios that are detectable with at least 80% power (increments of 0.1) under the different designs and under different degrees of clustering in exposure and health outcome. As above, note that there is very little difference between the five designs if the within cluster correlation in the health outcome and the within cluster correlation in exposure are small (e.g., λ=0.01 and ρ=0.001 or 0.01); whereas, when ρ and λ increase selecting fewer clusters results in a loss of power (as demonstrated by the larger odds ratios detectable with 80% power).

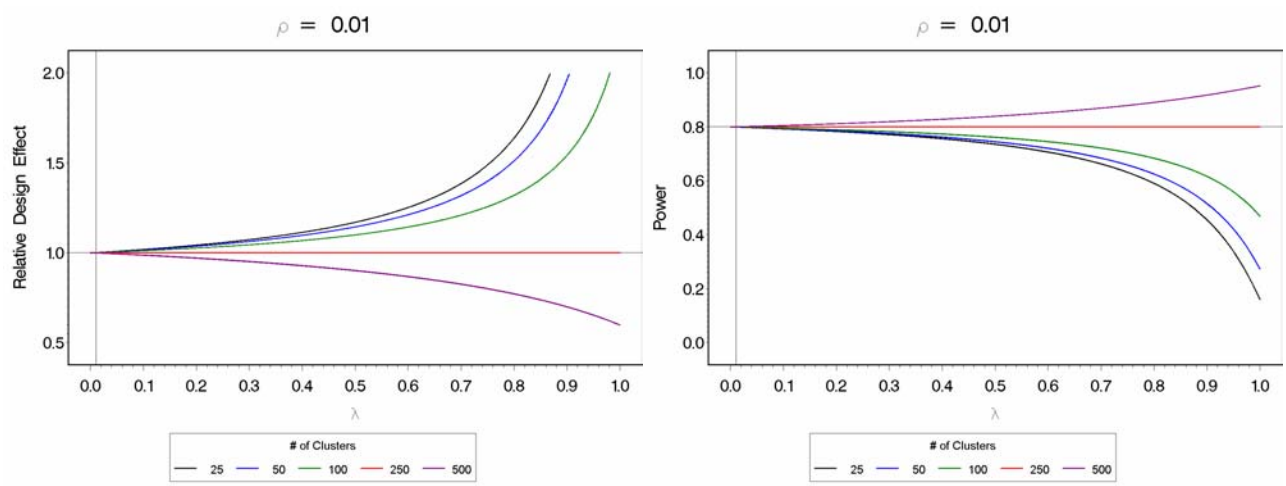

Figure 1. Relative design effects (with respect to a 250 cluster design) and comparison of power for estimating/detecting relationships across the entire cohort.

As demonstrated in these examples, in order to quantify the impacts of clustering and to indicate the tradeoffs inherent in selecting a design with a larger or smaller number of clusters, we must consider reasonable values for the within and between cluster variability in exposures and health outcomes of interest in the NCS. The NCS primary health outcomes (asthma, obesity, pregnancy outcomes, neurodevelopmental and behavioral outcomes, injury outcomes), are evaluated on a subject-specific level and will arguably have a relatively small degree of geographic clustering (i.e., small amount of within cluster correlation in health outcomes). For exposure factors, on the other hand, there are certainly exposure factors that vary primarily within cluster (e.g., personal activity levels, exposure to pesticides, exposure to mediators of inflammation, etc.) as well as factors that vary primarily, or only, between clusters (e.g., community-level $PM_{2.5}$ concentrations, community-level housing variables, etc.); however, since most of the NCS core hypotheses call for subject-specific analyses, and since many of the

variables that vary only between clusters may be interacted with subject-specific variables to form an "exposure" variable (e.g., interaction of community-level $PM_{2.5}$ concentrations and subject-specific activity patterns to form a $PM_{2.5}$ exposure metric), it may be reasonable to assume that most of the primary exposure covariates will have significant within-cluster variability.

Table 1. Odds ratios detectable with 80% power for a common health outcome (probability of disease for unexposed individuals of 0.05) and an exposure factor that occurs in 10% of the population.

ρ	λ	Odds Ratio Detectable with 80% Power				
		25 Clusters	50 Clusters	100 Clusters	250 Clusters	500 Clusters
0.001	0	1.2	1.2	1.2	1.2	1.2
	0.01	1.2	1.2	1.2	1.2	1.2
	0.1	1.2	1.2	1.2	1.2	1.2
	0.5	1.3	1.2	1.2	1.2	1.2
	1	1.4	1.3	1.2	1.2	1.2
0.01	0	1.2	1.2	1.2	1.2	1.2
	0.01	1.2	1.2	1.2	1.2	1.2
	0.1	1.4	1.3	1.2	1.2	1.2
	0.5	1.7	1.5	1.4	1.3	1.2
	1	2.0	1.7	1.5	1.4	1.3
0.1	0	1.2	1.2	1.2	1.2	1.2
	0.01	1.4	1.3	1.2	1.2	1.2
	0.1	2.1	1.7	1.5	1.4	1.3
	0.5	3.6	2.7	2.1	1.7	1.5
	1	5.1	3.5	2.7	2.0	1.7
0.5	0	1.3	1.2	1.2	1.2	1.2
	0.01	2.0	1.6	1.4	1.3	1.2
	0.1	4.2	2.9	2.2	1.7	1.5
	0.5	>10	6.1	4.0	2.7	2.1
	1	>10	>10	5.9	3.5	2.7

Thus, under the assumption that there is a relatively small degree of within cluster correlation in the health outcomes and exposure factors of primary interest in the NCS (i.e., a significant portion of the variability in the health outcome and the exposure occurs within a cluster), there appears to be little loss of statistical efficiency in estimating relationships across the entire cohort (small relative design effects and little loss of power) when comparing designs with 50 or 100 clusters to designs with 250 or 500 clusters. Balanced against other considerations related to sample clustering, such as the financial efficiencies of data collection in a smaller number of geographic regions, this suggests little need for selecting a large number of clusters in the NCS. This is not meant to imply that there is no advantage to selecting a design with a larger number of clusters since from a purely statistical view selecting more clusters offers clear advantages (e.g., when evaluating relationships between health outcomes and/or exposure factors that are cluster-specific, such as average housing age or median household income, or vary primarily between clusters). Rather we are suggesting that there may not be a *significant* statistical advantage to selecting a design with 250 or 500 clusters as compared to a design with 50 or 100 clusters when considering other design factors, such as cost considerations, data availability considerations, etc.