Published in final edited form as:
*Stat Med*. 2010 June 15; 29(13): 1377–1387. doi:10.1002/sim.3892.

# Improving cost-effectiveness of epidemiological studies via designed missingness strategies

**Warren Strauss**, **Louise Ryan**, **Michele Morara**, **Nicole Iroz-Elardo**, **Mark Davis**, **Matt Cupp**, **Marcia Nishioka**, **James Quackenboss**, **Warren Galke**, **Haluk Özkaynak**, and **Peter Scheidt**
Mr. Strauss, Mr. Morara, Ms. Iroz-Elardo, Mr. Davis, Mr. Cupp and Ms. Nishioka are from Battelle. Dr. Ryan is Professor and Chair of Biostatistics at Harvard University. Mr. Quackenboss and Dr. Özkaynak are members of the Environmental Protection Agency's National Exposure Research Laboratory. Mr. Quackenboss, and Dr. Scheidt are members of the National Children's Study Program Office at the National Institute of Child Health and Human Development within the National Institutes of Health. The late Dr. Galke was a member of the National Children's Study Program Office at the National Institute of Child Health and Human Development within the National Institutes of Health

## Abstract

Modern epidemiological studies face opportunities and challenges posed by an ever-expanding capacity to measure a wide range of environmental exposures, along with sophisticated biomarkers of exposure and response at the individual level. The challenge of deciding what to measure is further complicated for longitudinal studies, where logistical and cost constraints preclude the collection of all possible measurements on all participants at every follow-up time. This is true for the National Children's Study (NCS), a large scale longitudinal study that will enroll children starting in early pregnancy and gather information on their development and environment through early adulthood. The success of the NCS will significantly depend on the accurate, yet cost-effective, characterization of environmental exposures thought to be related to the health outcomes of interest. The purpose of this paper is to explore the use of cost saving, yet valid and adequately powered statistical approaches for gathering exposure information within epidemiological cohort studies. The proposed approach involves the collection of detailed exposure assessment information on a specially selected subset of the study population, and collection of less-costly, and presumably less-detailed and less-burdensome, surrogate measures across the entire cohort. We show that large scale efficiency in costs and burden may be achieved without making substantive sacrifices on the ability to draw reliable inferences concerning the relationship between exposure and health outcome. Several detailed scenarios are provided that document how the targeted sub-sampling design strategy can benefit large cohort studies such as the NCS, as well as other more focused environmental epidemiologic studies.

### Keywords

National Children's Study; Sampling design; Validation Sampling; Measurement Error Adjustment

Please send page proofs to Warren Strauss at: 505 King Ave., Columbus, OH 43201, 614-424-4275 (Telephone), 614-458-4275 (FAX), strauss@battelle.org.

The work was performed at the following locations: Battelle Memorial Institute, 505 King Avenue, Columbus, Ohio 43201 and the National Institute of Child Health and Human Development, 6100 Executive Blvd - 5C01, Rockville, MD 20892.

## Introduction

Modern epidemiological studies face the opportunities and challenges posed by a seemingly ever-expanding capacity to measure a wide range of environmental exposures, along with sophisticated biomarkers of exposure and response at the individual level. The challenge of deciding what to measure is further complicated for longitudinal studies that involve following participants over time. In practice, logistical and cost constraints preclude the collection of all possible measurements on every study subject at every follow-up time. These challenges already face the National Children's Study (NCS), an ambitious longitudinal study to examine the effects of environmental exposures on the health and development of children during their formative years. Key to the success of the NCS will be the detailed characterization of multiple environmental exposure parameters over time. Yet, to be successful, this must happen in a cost-effective manner. We describe here a statistical design framework that uses carefully planned sub-sampling and designed missingness to reduce study cost, while retaining high power to estimate effects of interest.

Environmental exposure assessments often use physical and biological measurements combined with time-activity information to estimate aggregate exposures. The NCS requires a means of identifying those exposure parameters that relate to the targeted health outcomes. The large number of participants in the NCS, and observations of them for up to 21 years, constrains the number and timing of measurements of detailed physical and biological measurements that can be obtained for each individual throughout the study due to the limitations of respondent burden and cost. This will inevitably lead to some degree of missing and potentially error-prone exposure data. The NCS will therefore require statistical tools which allow for efficient collection of exposure information while still preserving the ability to detect key dose-response relationships in the face of such challenges. More importantly, the results from the NCS must be of high scientific quality, and derived with sufficient statistical power to properly characterize the relationship between exposure and health outcome, and to serve as a resource for additional scientific investigations as new information becomes available.

In support of the NCS, we have applied innovative statistical techniques and developed supporting software for multi-stage targeted sub-sampling designs in which precise exposure assessment measures are collected on a carefully selected subset of the study population while less expensive (and presumably less detailed/comprehensive and less burdensome) surrogate measures for the same exposure are collected across the entire cohort [1] This sampling approach builds on well-developed statistical methodology, and integrates the concepts of epidemiological study design and analysis of data with missing covariates. Since the early work of Cornfield [2], the case-control study design has become a critically important tool in epidemiology. Conceptually, one may think of the case/control study as selecting from a large cohort (such as the NCS), with subjects who experience the event having a higher likelihood of selection than those who do not. The case-control design is one of the simplest and most effective examples of a much broader class of outcome dependent designs [3]. Like the case-control study design, outcome dependent sampling can substantially reduce the cost of a study while maintaining high statistical efficiency by selectively observing the exposure variable in a way that depends on the outcome status and other confounding covariates. There are several well-studied variations on the case-control design which exploit the idea of outcome-dependent sampling. White [4] proposed a two-stage case-control design where one starts with a traditional case-control design. Some expensive or difficult to measure covariates are assessed only on a subset of the study subjects, with selection probabilities that depend on other covariates of interest. Weaver and Zhou [5] proposed a two-stage outcome dependent sampling design for continuous

outcomes, followed by Wang and Zhou [6] who proposed a two-stage outcome dependent sampling design with discrete response and auxiliary covariates.

Our approach relies on multiple nested stages of sampling for exposure information, where each successive stage of sampling is pursued using a subset of study participants selected in the previous stage. Each stage represents an increasing level of detail (e.g., the accuracy, completeness, and precision of measurements; the number of locations and media sampled; and the number of observations over time) and cost with respect to exposure assessment (e.g. ranging from routine community-level environmental monitoring and simple questionnaire information, to environmental measures, to biomarker data, to aggregate exposure assessment information, to repeated measurement of multimedia multipathway or aggregate exposure information). The goal of the statistical sampling approach is to determine what fraction, and which specific members, of a study population should be assessed at each stage to allow investigators to determine the most cost efficient (or burden efficient) manner to design the study so that the relationship between disease and exposure can be appropriately characterized.

Our statistical approach is based on the application of well accepted approaches to handling missing covariate data. Because the selection probabilities used to define our assumed staged sampling strategy always depend on data observed at the previous stage, we are in the setting where data can be viewed as missing at random (MAR) [7]. While a number of different approaches are possible, we take a likelihood-based approach based on an assumed parametric model for the joint distribution of the outcome and covariates of interest. Through construction of the observed data likelihood by integrating over the distribution of missing data, we can characterize the relationship between the health outcome and exposure of interest by leveraging the joint exposure/surrogate information collected in the highest stage of sampling with a statistical measurement-error-type adjustment applied to the remainder of the cohort. Computer simulations indicate the potential for large scale efficiencies in sampling and analysis costs and study subject burden associated with the targeted sub-sampling approach compared to traditional study designs in which detailed exposure information is collected on all study participants [8]. These efficiency gains can be realized without making any substantive sacrifices on the ability to draw unbiased inferences on the research problem of characterizing the relationship between exposure and health outcome.

Despite the relatively large body of literature on methods for handling missing covariates, relatively little has been done on optimal design in this setting. Reilly and Pepe [9] use a mean score formulation to address optimal design for settings where covariates cannot be measured on all study subjects for reasons of cost or practicality. Our targeted sub-sampling design approach builds on these principles, using a maximum likelihood approach for staged sampling in a manner that allows flexibility in terms of the research problems that it will support, including: the inclusion of effect modifiers in the model between exposure and health outcome, multiple distributional assumptions for the outcome and exposure variables, and the possibility of repeated measures for the health outcome data in a longitudinal model.

Through careful consideration of these issues in the context of providing the data needed for the various exposure assessment analyses planned for the NCS, we demonstrate via three examples how the proposed targeted sub-sampling methodology can be used to identify cost-effective and efficient sampling schemes for collecting the desired environmental exposure data. While our examples focus on cost characterized in terms of monetary considerations, cost could just as easily be interpreted in terms of respondent burden so long as investigators can quantify the differential burden associated with each measurement.

## Targeted sub-sampling

We assume a simple logistic regression model for characterizing the probability of an adverse health outcome as a function of exposure:

$$\text{logit}(\Pr(Y=1|X))=\beta_0+\beta_1 X, \quad (1)$$

where Y is the binary outcome of interest (e.g. whether a child develops autism by age 6) and X is the exposure variable (e.g. in-utero or post natal pesticide exposure). Targeted sub-sampling assumes that if X can be predicted well by one or more surrogate measures of exposure, then a study can still maintain excellent power for fitting model (1) even if X is measured on only a relatively small sub-sample of the study population. We have developed rigorous optimal design guidelines that characterize the sub-sampling strategies that either maximize study power subject to a cost constraint or minimize cost, subject to a desired power [1]. As discussed later, "cost" may reflect monetary or respondent burden considerations. Our sampling strategy is conducted in stages, with the health outcome (Y) being assessed in the first and largest sample of participants within the cohort. A subset of those participants would have the first-stage surrogate measure ($Z_1$). If multiple stages of surrogate measures are being planned ($Z_2,\ldots, Z_n$), each successive stage represents a smaller subset of the previous stage. The last and smallest stage of sampling assesses the true exposure of interest (X). Our approach allows for the possibility that X is not directly measurable, but rather represents a hypothetical true exposure that can, at best, be approximated. For this case, we assume a latent variable approach where all the various observed surrogates (the Z's) are viewed as providing indirect information about the true exposure of interest. It should be noted that the design's stages of sampling need not correspond to a temporal ordering for sample collections. Rather, the stages correspond to an assumed hierarchy of sampling in which Stage 0 represents a sub-sample of study participants in the NCS cohort that will have the health outcome measured, and each successive stage represents a smaller subset of study participants from the previous stage.

Our design is characterized by a series of conditional probabilities that characterize an individual's chance of being sampled at successive stages in the hierarchy, conditional on data that were observed at previous stages. In a manner similar to that reported by Robins et al. [10], we represent these sampling probabilities through a series of logistic regression equations. For example, the equations characterizing a design involving a two surrogate measures would be:

$$\begin{aligned}\text{logit}(\gamma_0)&=\alpha_{00}\\ \text{logit}(\gamma_1)&=\alpha_{10}+\alpha_{11}\cdot Y\\ \text{logit}(\gamma_2)&=\alpha_{20}+\alpha_{21}\cdot Y+\alpha_{22}\cdot Z_1\\ \text{logit}(\gamma_3)&=\alpha_{30}+\alpha_{31}\cdot Y+\alpha_{32}\cdot Z_1+\alpha_{33}\cdot Z_2,\end{aligned}$$

where $\gamma_0$ represents the probability that a study subject is sampled at stage 0 for the health outcome (Y); $\gamma_1$ represents the conditional probability that a study subject is sampled at stage 1 (for the surrogate measure $Z_1$), given that (s)he was already selected within stage 0 and given the observed value of Y at stage 0; $\gamma_2$ represents the conditional probability that a study subject is sampled at stage 2 (for the surrogate measure $Z_2$), given that (s)he was already selected within stage 1 and given the data observed at stages 0 and 1, and ; $\gamma_3$ represents the conditional probability that a study subject is sampled at stage 3 (for the exposure variable X), given that (s)he was already selected within stage 2 and given the data observed at stages 0, 1 and 2.

The intercept terms ($\alpha_{00}$, $\alpha_{10}$, $\alpha_{20}$, and $\alpha_{30}$) are included in all designs with multiple stages of surrogate measures and a measurable exposure variable. The slope terms $\alpha_{11}$, $\alpha_{21}$, and $\alpha_{31}$ represent parameters associated with outcome dependent sampling in stages 1, 2, and 3; and $\alpha_{22}$, $\alpha_{32}$, and $\alpha_{33}$ represent parameters associated with covariate dependent sampling in stages 2 and 3. Thus, a design with no opportunity for covariate-dependent or outcome-dependent sampling would constrain these slope terms to zero.

Due to the hierarchical nature of the staged sampling approach, the probability of sampling at any stage is a product of the $\gamma$ probabilities from the current and all previous stages of sampling. Thus, the overall probability of being sampled for X in stage 3 (as a function of $\alpha$,Y, $Z_1$ and $Z_2$) is $\pi_3 = \gamma_0 \cdot \gamma_1 \cdot \gamma_2 \cdot \gamma_3$.

We use a maximum likelihood formulation for optimal design, which means statistical distributional assumptions are needed for the marginal distribution of X, and the conditional distributions of Y|X and $Z_1|X,\ldots.Z_P|X$, in to allow proper construction of the joint likelihood which is necessary to support the design. Numerical constrained optimization is used to determine optimal selection probabilities (that is, the set of $\alpha$'s needed to characterize the sampling probabilities) at each model stage, subject to specified constraints related to budget, respondent burden or required levels of precision. For this last optimization goal, the target standard error is calculated as a function of desired statistical power, size of the study, and a choice between a one-sided or two-sided test of the hypothesis.

For the relationship between health outcome and exposure, our approach allows for logistic regression models for binary health outcomes and linear regression models for continuous outcomes, in both the cross-sectional and longitudinal setting. Our approach also allows investigators to develop designs that include an effect modifier (E), which when included in the design, is assumed to be assessed at the second stage of the model among a subset of participants that are observed with the health outcome. Additional detail on the statistical methodology for these designs, including information about the objective function and optimization strategy supporting both cross-sectional and longitudinal designs can be found in our earlier work [1].

## Partitioning and Attributing Cost and Burden to Appropriate Stages

Each variable (Y, E, $Z_j$, and X) that is included in the staged targeted sub-sampling approach may have two different types of associated costs and burden: specimen collection and analytical measurement. Specimen collection costs cover the physical collection of the biological or environmental specimen and the materials and activities that are required prior to specimen analysis (e.g. supplies, equipment, shipment, storage in a repository, etc.). The respondent burden associated with specimen collection may be significant. For example, parents are typically reluctant to agree to blood samples being drawn from infants and young children unless medically necessary. Analytical costs are designed to cover specimen analysis and transferring the results to the study database. There is generally no respondent burden associated with analytical measurement.

We assume that the analytical costs for a variable are always attributable to the stage of sampling associated with that variable but that the specimen collection costs for a variable may be attributable to that variable's stage or any previous stage of sampling. Assignment of specimen collection costs to a previous stage of sampling becomes necessary when the decision of whether or not to measure a variable is a function of information from a previous stage within the targeted sub-sampling paradigm (i.e. covariate or outcome dependent sampling). There are also instances in which a fraction of the specimen collection and/or analytical costs associated with a particular variable might already be covered as part of a

previous stage of sampling, and the costs associated with sampling and measuring that fractional component must be deleted from subsequent stages.

## Case Studies

We present three case studies focused on the design of a hypothetical study relating neurobehavioral and cognitive development in children with exposures to chlorpyrifos, a specific non-persistent pesticide and the potential interaction with decreased paraoxonase activity. Each case explores multiple study designs, including a classic design in which the health outcome, effect modifier, and true measure of exposure is measured on all study subjects (used as a basis of comparison), and two-stage targeted sub-sampling designs in which the health outcome, effect modifier, and a surrogate measure of exposure (either $Z_{SF}$ or $Z_{Quest}$) are measured in the first stage of sampling, and then the true exposure (X) is measured in the second stage of sampling. For the targeted sub-sampling designs, we investigated three methods for selecting study participants for true exposure assessment: random sampling, covariate dependent sampling, in which participants are selected based on a previously obtained value of a surrogate measure, and outcome dependent sampling, in which participants are selected based on a previously obtained health outcome measure. For simplicity, our case examples are based on determining optimal design strategies based on monetary cost, though respondent burden would be important to consider as well.

### Data Sources for Case Studies

The assumed outcome for our case studies is the Peabody Picture Verbal Test – Revised (PPVT-R) scores observed sequentially on children over a 4 year period of time (ages 36, 60, and 84 months) from National Longitudinal Survey of Youth 1979 [11]. The first two case studies focus on a cross-sectional investigation of whether the child was in the lowest $5^{th}$ percentile of scores at 84 months. Case Study 3 focuses on a longitudinal assessment of repeated measures of the continuous PPVT-R scores. We assume a $20 administration cost and a $5 analysis cost for the Peabody exam.

The best current methods for estimating true exposure to chlorpyrifos may rely on a detailed and accurate aggregate exposure analysis combining information about the food, dust, and air in the child's environment with activity patterns during key lifestages [12,13]. Aggregate exposure analysis is costly in terms of both study budget and study subject burden and thus an ideal candidate for consideration in a targeted sub-sampling approach.

We utilized data on childhood chlorpyrifos exposures from the EPA study of Children's Total Exposure to Persistent Pesticides and Other Persistent Organic Pollutants (CTEPP) as the basis for this case study example [14,15]. CTEPP is a recent aggregate exposure study of 257 preschool age children (ages 2 to 5) conducted in North Carolina and Ohio in rural and urban settings with sampling schemes chosen for diversity in economic and daycare status. CTEPP monitored exposures over a 48 hour period at the children's home and day care center (if applicable) by collecting environmental samples(air, dust and soil), personal (hand wipes, diet, water, and urine) samples and questionnaire information (housing characteristics, pesticide products used, and general activity patterns of the participants). Chlorpyrifos data from 127 households in the six Ohio counties was used for this case example.

Aggregate exposure to chlorpyrifos for each participant in CTEPP was estimated using the time-weighted average of concentrations [16] and serves as the gold standard (X). Various candidate surrogate measures ($Z_j$) were calculated by setting all but the component of interest (diet, inhalation, dust, etc) to the mean exposure. The correlations, ρ, between X and $Z_j$ were then calculated by performing a simple linear regression, allowing for a non-zero

intercept term, and taking the square root of the $R^2$ coefficient. Because these analyses clearly suggested that dietary exposure from solid food explained much of the variability in aggregate exposure, we further parceled the dietary surrogate into solid and liquid food. The distributions and costs associated with exposure and two of the surrogate measures are detailed below:

- **Exposure (X)**: The aggregate exposure per day to chlorpyrifos is log normally distributed with a geometric mean of 4.76 and a geometric standard deviation of 1.28. We assumed that it would cost $350 to sample and $1800 to analyze the aggregate exposure data for each child based on costs from the CTEPP Study.
- **Surrogate ($Z_{SF}$)**: The solid food (SF) portion of the duplicate plate analysis was highly correlated ($\rho$=0.89) with aggregate exposure. $Z_{SF}$ was log normally distributed with a geometric mean of 4.62 and standard deviation of 2.12 and would cost $50 to sample and $450 to analyze. The cost of exposure (X) when using Z_SF was adjusted to the additional cost of sampling ($300) and analysis ($1350) exposure.

Since questionnaire data (aside from time activity patterns) is not part of the aggregate exposure calculations, a second surrogate measure ($Z_{Quest}$) was calculated by using four pesticide use and diet questions to predict aggregate exposure in a linear model. Distributions were then calculated on the predicted points and correlated with X as shown below:

- **Surrogate ($Z_{Quest}$)**: The surrogate from the questionnaire data was log normally distributed with a geometric mean of 4.75 and a standard deviation of 1.28. We assumed that administering the questionnaire would cost $5 and analysis would cost $1. The costs of obtaining the questionnaire surrogate ($Z_{Quest}$) were assumed to be independent from the other measures of exposure (X and $Z_{SF}$) for the case studies.

Our case studies also incorporate genetic susceptibility to decreased paraoxonase activity as an interaction between genetics and environment. Preliminary studies have shown that paraoxonase (PON1), an enzyme made in the liver and carried on HDL, may protect against the adverse affects of pesticides/insecticides: the higher the level of paraoxonase, the lower the toxicity effects of chlorpyrifos. Infants produce less paraoxonase than their mothers and thus may be at greater risk for adverse neurocognitive health effects from pesticide exposure [17]. We assumed that approximately 15% of the population has the genetic form of PON1 which corresponds to decreased paraoxonase activity [17,18]. We also assumed that blood will be routinely archived for all NCS participants (thus a $0 sample collection cost) and will cost $40 to genotype PON1.

## Case Study 1

Our first case study is a cross-sectional example relating exposure to chlorpyrifos and decreased paraoxonase activity to the probability of being in the lowest 5% of a standardized neuro-cognitive test, as captured by the following logistic regression:

$$\text{Logit}(\Pr(Y=1))=\beta_0+\beta_1 \cdot E+\beta_2 \cdot X+\beta_3 \cdot E \cdot X.$$

We assume $\beta_1=\beta_2=\beta_3=0.4045$ (corresponding to an odds-ratio of 1.5), and that the optimization goal was to identify the lowest cost sampling design that allows us to test the hypothesis $\beta_3=0$ based on a two-sided test with size ($\alpha$=0.05) and power ($1-\beta=0.8$). Since $\beta_3$ corresponds to the interaction between exposure and the effect modifier, this design essentially is optimized to detect a 50% increase in the odds of being in the lower 5 percent

of this cognitive test associated with a one standard-deviation increase in exposure for study participants with decreased paraoxonase activity ($\psi_{YX} = \exp(\beta_3)=1.5$). The results are provided in the top third of Table 1.

The classic design requires the neuro-cognitive assessment (Y), aggregate exposure (X), and genotyping of PON1 (E) to be measured for 7,210 study participants with an associated total cost of approximately $16M. Utilizing a two-stage targeted sub-sampling design with a solid food surrogate (sampling Y, E, and $Z_{SF}$ jointly in the first stage), and a completely random sampling scheme for aggregate exposure assessment in the second stage lowers the cost to $5.5M, or 35% of the classic design. This is accomplished by increasing those that are jointly sampled for Y, E, and $Z_{SF}$ in the first stage to 9,505 children, with 90 of those children completing the aggregate exposure analysis (X) in the second stage of the design.

While none of the covariate or outcome dependent sampling schemes have a lower total cost than the solid food surrogate with random sampling, it is interesting to note that the outcome dependent sampling scheme for the questionnaire surrogate costs about $1.4M less than completely random sampling designs.

## Case Study 2

For this case study, we build upon the first case study with two important changes:

1. We assume that the specimens associated with exposure (X) were already collected and stored in a repository as part of the NCS core data collection protocol, and are available for future analysis to support this hypothesis. Thus, we assume no additional costs associated with specimen collection for exposure assessment in this Case Study. However, we assume that there will be costs incurred from the chemical analysis of samples from the repository associated with those study participants selected for detailed exposure assessment as part of this design.
2. Instead of conducting the covariate dependent sampling as a function of the continuous surrogate exposure measure (Z), we allow the decision of whether to analyze the full complement of samples associated with true exposure (air, dust, food and soil) to be dependent on the binary effect modifier (E).

We again assume ($\beta_1= \beta_2= \beta_3= 0.41$) and that the optimization goal is to identify the lowest cost design that allows us to reject $H_0$: $\beta_3 =0$ based on a two-sided test with size ($\alpha=0.05$) and power ($1- \beta =0.8$). The main difference from Case Study 1 is that there are no additional costs associated with specimen collection for assessing exposure, since these costs were already included in Stage 0, as described above in Change 1.

The middle section of Table 1 provides the results for Case Study 2. The classic design still requires joint sampling of Y, E and X for 7,210 study participants at a cost of $13.4M. The approximate $2.5M reduction in the cost of the classic design (compared to Case Study 1) is entirely attributable to the assumption that the environmental samples were already collected and available from the repository for chemical analysis. The use of a solid food surrogate in a random sampling scheme lowers the cost to $5M, or 37% ofthe classic design by increasing the Y, E, and $Z_{SF}$ sample size to 9,489 children. Only 96 of those children would need aggregate exposure media analysis.

The above result parallels the results of Case Study 1 random sampling schemes. However, unlike in Case Study 1, additional efficiency gains can be gained by the covariate and outcome dependent targeted sub-sampling designs. This result is largely attributable to the fact that these designs no longer carry the expensive specimen (air, dust, food, and soil) collection costs that were previously attributed to the Stage 1 sample.

When exploring targeted sub-sampling designs that utilize the lowest cost and precision questionnaire-based surrogate measure (*Z_Quest*), the random sampling design would sample 93,266 study participants in the first stage and pursue chemical analysis on 991 randomly selected study participants (i.e., about 1% of the cohort) in the second stage at a total cost of $8.41M. The covariate dependent design where sampling for X is dependent on E would sample 53,382 study participants for *Y* and *Z_Quest* in the first stage and pursue chemical analysis on 1029 study participants in the second stage at a total cost of $5.64M. The covariate dependent sampling designs result in sampling equations for the second stage chemical analysis that result in over-sampling the study participants with decreased paraoxonase activity (as expected). The questionnaire-based targeted sub-sampling design with selection of Stage 2 participants based on the value of the effect modifier offers a significant improvement in cost efficiency. The efficiency gain is achieved due to two factors: (1) Only those study participants with decreased paraoxonase activity (when E=1) contribute information to the target parameter ($\beta_3$) whose variance is being minimized by the constrained optimization, and (2) the questionnaire-based design does not require the higher cost of chemical analysis of food samples in the first stage of the design.

The outcome dependent design would sample 16,992 study participants for *Y* and $Z_{Quest}$ in the first stage and pursue chemical analysis on 721 study participants in the second stage at a total cost of $2.5M. This outcome dependent sampling design, as well as the design using $Z_{SF,}$ resulted in sampling equations for the second stage chemical analysis that leads to over-sampling the study participants with low performance on the neuro-cognitive assessment. The results in the middle section of Table 1 demonstrate that the outcome dependent design with the lowest cost/lowest precision surrogate measure ($Z_{Quest}$) leads to the most efficient sampling design. This result is expected to generally hold true for most instances in which a reliable surrogate measure (Z) can be identified for an outcome dependent design where there are no sample collection costs necessary for obtaining true exposure (*X*) measurements during the first stage of sampling. Of course, this design approach relies heavily on the use of archived specimens, and additional work may be necessary to determine whether these specimens can be stored long-term and still yield accurate analytical results.

## Case Study 3

Ideally, the NCS will be able to detect differences in children's developmental trajectories over time attributable to exposure. The longitudinal example presented here examines study designs needed to understand the effect of chlorpyrifos on Peabody Picture Verbal Test – Revised (PPVT-R) scores over a 4 year period of time (ages 36 months, 60 months, and 84 months). The following model was used as the basis for this longitudinal case-study example:

$$Y_{ij}=\beta_0+\beta_1\, ln(X_i)+\beta_2 Age_{ij}+\beta_3\, ln(X_i)Age_{ij}+R_{0i}+R_{1i}Age_{ij}+\varepsilon_{ij}$$

where $Y_{ij}$ is the PPVT-R test score for the ith child at follow-up time j, $Age_{ij}$ is the age of this child at that time, $X_i$ represents the true pesticide exposure for the ith child, $\beta_0$ to $\beta_3$ are fixed effects parameters that describe the interactive effects of age and exposure on test scores across the population, $R_{0i}$ and $R_{1i}$ are random effects allowing each child to deviate from population average, and $\varepsilon_{ij}$ is the error not explained by the model. It is assumed that the random effects are multivariate normal

$$\begin{pmatrix} R_{0i} \\ R_{1i} \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{R_0} & \sigma^2_{R_0R_1} \\ \sigma^2_{R_0R_1} & \sigma^2_{R_1} \end{pmatrix}\right]$$

and that the error term is normal $\varepsilon_{ij} \sim N(0, \sigma^2_{error})$. It should be noted that the exposure variable ($X_i$) in the above model is not time varying (i.e. it represents a pesticide exposure at a single moment in time for each child) and is assumed to temporally precede health outcome measures ($Y_{ij}$).

Based on an analysis of the NLSY data, we estimated that $\beta_0 = -2.77$, $\beta_2 = 0.834$, $\sigma^2_{R_0} = 160.4$, $\sigma^2_{R_0R_1} = -0.453$, and $\sigma^2_{R_1} = 0.0115$, and $\sigma^2_{error} = 107$. For the purposes of this case study, we assume that exposure to chlorpyrifos, ln(X), will have a negative impact on PPVT-R scores with $\beta_1 = -0.5$. Similarly, we hypothesize that the negative effects of exposure will also diminish cognitive development in children over time, justifying the choice of a negative value for $\beta_3 = -0.01$ which is the parameter whose standard error is being minimized by the sampling design.

The classic design in the longitudinal example as shown in the last section of Table 1 involves participation of 11,321 children in a full aggregate exposure analysis and three waves of PPVT-R testing, costing $25.2M. A random two-stage design utilizing solid food as a surrogate greatly improves upon the classic design. By increasing the number of children to 14,483 that participate in the three waves of PPVT-R testing and the solid food portion of the aggregate exposure study, the number of children needing a full aggregate exposure study decreases to 107. Consequently, a design utilizing the solid food surrogate would only cost $8.5M, or 34% of the classic design. Note that the outcome dependent sampling design presented in Table 2 generated sampling equations based on the average of the three repeated measures on each individual – but that our design also allows for outcome dependent sampling in the longitudinal case based on the minimum or maximum of the repeated measures.

## Aggregate Exposure Study

Many large cohort studies such as the NCS have multiple scientific objectives, making it difficult to identify a single efficient design under the designed missingness paradigm. One possible strategy to deal with the problem of multiple and varied scientific objectives is to implement a global aggregate exposure study as a 2nd stage random sample from within the larger cohort, designed to provide detailed exposure results across a range of hypotheses to be investigated. Using this approach, we assume that information related to major health outcomes of interest and limited risk factor information (surrogate measures of exposure) on all study cohort members are collected in the first stage of the design. The second stage (aggregate exposure study) sample helps characterize the relationship between the risk factors explored in the first stage and more definitive and direct measures of exposure, thereby allowing key exposure/outcome relationships to be explored across the larger cohort using analysis methods that correct for the missing exposure information among study participants that are not involved in the aggregate exposure study. The fraction of study participants selected in the 2nd stage is dependent on a set of parameters related to the strength of exposure/outcome relationships that are being characterized in the study, as well as the strength of the relationship between the surrogate and improved measures of exposure. To illustrate, we investigated a series of cross-sectional sampling designs in which the health outcome (Y) and the risk factor information (Z) were constrained to be assessed

on the entire cohort, and that exposure (X) is measured at random on a smaller subset of the study population.

Since measures of exposure are often continuous, we assume that true exposure follows a standard normal distribution (i.e. X ~N(0,1)), that the risk factor information (Z) is also continuous, that the health outcome (Y) represents the presence or absence of disease, and that the relationship between Y and X is defined by a simple logistic regression model (i.e. $\text{logit}(\Pr(Y=1)) = \beta_0 + \beta_1 \cdot X$).

Table 2 provides sample sizes needed for random 2nd stage sample for exposure within a cohort of similar size to the NCS as a function of the odds ratio between the health outcome and exposure ($\psi_{YX}$), the correlation between aggregate exposure and the risk factor information that is measured across the entire cohort ($\rho_{XZ}$), the prevalence of the health outcome ($P_Y$), and the cohort size. The sample sizes correspond to the number of study participants that are needed in order to detect $\psi_{YX}$ with sufficient precision ($\alpha$=0.05) and power ($1-\beta$=0.8). Because exposure (X) was assumed to follow a standard normal distribution, the odds ratio ($\psi_{YX}$) can be interpreted as the increased odds of the health outcome associated with a one standard-deviation increase in exposure.

The shaded rows in Table 2 correspond to $\rho_{XZ}$=0, and essentially identify the sample size necessary in a traditional design in which the health outcome Y and the aggregate exposure X are assessed on all study participants. For example, to detect an odds ratio of 1.25 ($\psi_{YX}$=1.25) when the prevalence of the health outcome is 1% ($P_Y$=0.01), a traditional design would require 15,947 study participants with Y and X observed. Green shaded areas in Table 2 identify odds ratios ($\psi_{YX}$) that cannot be detected with sufficient precision and power within the sample size constraints under the traditional design.

Continuing with this example of detecting an odds ratio of 1.25 when the prevalence of Y is 1 percent, if we assume a cohort size of 100,000 study participants and a correlation between X and Z of 30 percent (similar to the correlation between questionnaire information and measures of aggregate exposure assessment for nonpersistent pesticides as discussed earlier), use of a random 2nd stage sample for exposure involves 7,696 study participants (compared to 15,947).

The strength of utilizing designed missingness concepts for exposure assessment within a large cohort study is based on leveraging information from Y and Z among a large number of study participants who do not participate in the 2nd stage sample. The size of the total cohort (i.e. those with Y and Z assessed) can be critical – as seen in comparisons between a cohort of size 100,000 and 75,000. For the example of trying to detect an odds-ratio of 1.25 when the prevalence of Y is 1%, the traditional design requires 15,947 study participants. If there is an available surrogate metric of exposure that has a 40 percent correlation with true exposure, then the analysis indicates that only 840 study participants are required in the 2nd stage exposure study if the cohort size were 100,000. However, to characterize the same relationship from a cohort of 75,000 participants would require 4,797 participants in the 2nd stage exposure study. While inclusion of 4,797 study participants still represents a large reduction compared to the traditional design, the loss of 25,000 study participants with Y and Z information has a large impact (making it necessary to include 4,797 rather than 840 study participants in the 2nd stage sample in order to characterize the same relationship).

Conceptually, the aggregate exposure study as described above represents a random 2nd stage sample from the larger cohort, where more detailed and burdensome exposure samples are collected and analyzed in real time to provide coverage across a range of potential exposure/outcome hypotheses. Additional nested outcome-dependent designs should also be explored to support hypotheses in which (1) specimen collection costs and burden for

exposure are minimal and (2) the samples can be archived in a repository for future chemical analysis. However, these investigations are likely to be hypothesis specific.

Finally, it should be noted that since the designs illustrated in Table 2 are based on a maximum likelihood approach, it is understood that the results are based on large sample approximations. Thus, we cannot assume that a resulting design will sufficiently characterize the relationship between X and Z to allow for calibration of the relationship between Y and X with fewer than 20 study participants included in the $2^{nd}$ stage sample.

## Conclusions

The multi-stage targeted sampling design methodology, alternatively referred to as designed missingness, offers an appealing degree of flexibility for developing strategies for logistically feasible and cost effective collection of environmental exposure information for epidemiological cohort studies. These design strategies require study planners to identify a reasonable proxy for exposure that can be assessed as part of the core data collection protocol, while the more burdensome and expensive environmental and exposure monitoring and chemical analysis corresponding to a more direct and comprehensive exposure assessment is pursued on a much smaller, and sometimes targeted, segment of the study population. In comparison to classic design strategies that require a complete matrix of data on all study participants needed for a given hypothesis, the case studies investigated here suggest that significant efficiencies can be gained by employing multi-stage targeted sub-sampling strategies without sacrificing the ability to draw unbiased inferences regarding the relationship between health effects and exposures. While our case example were formulated in terms of monetary cost, the results are relevant to considerations related to respondent burden as well. The “gold standard” exposure measurement in our case studies corresponded to an aggregate exposure analysis reflecting the food, dust and air in each child’s environment, while the surrogate measures focused on the use of only partial assessments. Hence, the cost of measurement was proportional to the associated respondent burden.

We have implemented the proposed design methodology through a windows-based software tool that allows study planners to properly account for the costs associated with employing these design strategies. Users can partition the sampling costs to previous design stages for outcome and covariate dependent designs and allocate fractional costs when the surrogate measure in an earlier stage is a component of the more accurate measure of true exposure. While our design methodology allows substitution of sampling and analytical costs with estimated units of burden as described above, it does not currently account for potential item non-response associated with the different types of measures incorporated into the design. In future work, we plan on developing optimal design strategies that allow incorporate item non-response associated with measures at each stage of the sampling paradigm – thereby allowing researchers to develop designs that are optimal with respect to both costs and the potential effects of non-response due to burdensome data collection. We also plan on developing simulation and analysis tools that will allow us to assess more complex designs, such as those that account for spatial correlation or other effects of clustering that are integrated into studies such as the NCS.

Our analysis within the case studies suggest that when the complexities and costs associated with the collection of physical samples (e.g. air, dust, soil, food) for an exposure assessment are appropriately taken into consideration, multi-stage sampling strategies that select study participants for detailed exposure assessment at random tend to be more cost efficient than outcome or covariate dependent design strategies, unless:

1. The value of the surrogate measure can be identified in advance of specimen collection for more detailed aggregate exposure assessment, in which case

covariate dependent sampling designs can lead to greater cost/burden efficiencies, or

2. Detailed exposure specimens have already been collected across the entire cohort and stored within a repository for future chemical analysis as part of the core data collection protocol. When this is the case, covariate and outcome dependent sampling designs will almost always outperform the random sampling design because the costs associated with specimen collection are removed from consideration and the covariate and/or outcome information can be used to better target where resources are spent for exposure assessment laboratory analyses. While not presented in this manuscript, designs that are dependent on both outcome and covariate information, such as those proposed by Breslow and Cain [19], are also expected to perform well. These designs are also well supported by our methodology and software tools.

When samples that support the true exposure are available from within the study repository, our analyses suggest that low-cost/low-detail/low-burden surrogate measures can be exploited by an outcome dependent sampling design that mirrors the classic nested case/control study design, offering the greatest cost efficiencies.

The multistage targeted sub-sampling approach and software may be used for the NCS in multiple ways, including:

- The design and implementation of one or more global aggregate exposure studies nested within the NCS cohort to serve as a resource for current NCS scientific goals and any future investigations that focus on chemical or biological exposures;
- Determination of appropriately sized fractions of the study population that must undergo both health outcome and exposure assessments to support key hypotheses (it may not always be necessary to measure the health outcome across the entire study cohort – which can also lead to cost and burden efficiencies); and
- The design and implementation of any number of nested studies, including nested case/control studies, from within the NCS cohort that efficiently targets samples from the NCS repository for chemical analysis in support of specific hypotheses.

The identification of appropriate surrogate measures of exposure will be a critical part of capitalizing on these design strategies. This may entail careful scrutiny of existing data sources (such as the ones utilized in developing the case-studies for this report), and/or conducting pilot field and laboratory studies that assess the relative costs and performance of candidate exposure metrics. This type of research, coupled with use of the design methodology and software, can be used to systematically evaluate whether it is worthwhile including additional or more personal exposure assessment methods, or more precise chemical analysis methods that have higher associated costs at varying stages of the design.

Finally, it should be noted that the results presented in this paper demonstrate efficiencies gained from power studies that assume a fully parametric likelihood-based model. Often, following study design, analysts will utilize models different to those assumed at the design phase (e.g.semiparametric analysis techniques) to establish the relationship between exposures and outcome. Depending on the nature of the analysis model, such a discrepancy has the potential to result in bias and/or efficiency loss. We performed preliminary investigations to assess how the misspecification of a model between Y and X in the design phase will affect the estimation of the parameters in the analysis phase. In particular we considered cases where the true logistic model Y|E, X included a non-linear term in X, but where the design phase was based on the assumption of a linear logistic model. Assuming a likelihood-based analysis fitted with MCMC-based approaches, we detected a modest loss in

power depending on the magnitude of the coefficient associated with the non-linear term, but we did not observe a statistically significant bias in the parameter estimations. Additional research needs to be done to fully assess the performance of these designs in the presence of misspecification, as well as extension of these methods to support non-linear functions of exposure.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This work was partially supported by the National Children's Study Program Office at the National Institute of Child Health and Human Development under Contract 282-98-0019, Task Order 16, and by the National Exposure Research Laboratory (NERL) at the U.S. Environmental Protection Agency under Contract 68-D-99-011, Task Order 0019. The authors would also like to acknowledge Dr. Marsha Morgan at EPA-NERL for her contribution of childhood pesticide exposure data and critical review and input for this manuscript. We are grateful to Jonas Ellenberg for the opportunity to include our work in this special collection of papers related to the NCS. Finally, we would like to dedicate this paper to the memory of our friend, colleague and coauthor, Warren Galke.

## References

1. Morara M, Ryan L, Houseman A, Strauss W. Optimal design for epidemiological studies subject to designed missingness. Lifetime Data Analysis. 2007; 13:583–605. [PubMed: 18080755]
2. Cornfield J. A method of estimating comparative rates from clinical data: applications to cancer of lung, breast and cervix. Journal of the National Cancer Institute. 1951; 11:1269–1275. [PubMed: 14861651]
3. Zhou H, Weaver MA. Outcome dependent selection models. Encyclopedia of Environmetrics. 2001; 3:1499–1502.
4. Weaver MA, Zhou H. An Estimated Likelihood Method for Continuous Outcome Regression Models With Outcome-Dependent Sampling. Journal of the American Statistical Association. 2005; 100(470):459–469.
5. Wang X, Zhou H. A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates. Biometrics. 2006; 62(4):1149–1160. [PubMed: 17156290]
6. White JE. A two-stage design for the study of the relationship between a rare exposure and a rare disease. American Journal of Epidemiology. 1982; 115(1):119–128. [PubMed: 7055123]
7. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. John Wiley and Sons; New York: 1987.
8. Strauss, W.; Ryan, L.; Lehman, J. Discussion on Utility of Validation Samples for the NCS. Technical Report submitted by Battelle to NICHD, National Children's Study Program Office under Work Assignment 9 of Task Order 11 on Contract 282-98-0019. 2004 [accessed September 23, 2008]. http://www.nationalchildrensstudy.gov/research/reviewsreports/Pages/Discussion-on-Utility-of-Validation-Samples-for-the-National-Children-s-Study.pdf
9. Reilly M, Pepe M. A mean score method for missing and auxiliary covariate data in regression models. Biometrika. 1995; 82:299–314.
10. Robins J, Rotnitzky A, Zhao LP. Estimation of regression coeffcients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89:846–866.
11. NLSY Child & Young Adult Data Users Guide: A Guide to the National Longitudinal Survey of Youth 1979; 1986–2002 Child Data; 1994–2002 Young Adult Data. Center for Human Resource Research. The Ohio State University; October2004 [Accessed on February 1, 2006]. http://www.bls.gov/nls/nlsy79ch.htm
12. Özkaynak H, Whyatt L, Needham LL, Akland G, Quackenboss J. Exposure Assessment Implications for the Design and Implementation of the National Children's Study. Environmental Health Perspectives. 2005; 113:1108–1115. [PubMed: 16079086]

13. Bradman A, Whyatt RM. Characterizing Exposures to Nonpersistent Pesticides During Pregnancy and Early Childhood in the National Children's Study: A Review of Monitoring and Measurement Methodologies. Environmental Health Perspectives. 2005; 113:1092–1107. [PubMed: 16079084]

14. Morgan MK, Sheldon LS, Croghan CW, Jones PA, Robertson G, Chuang JC, Wilson NK, Lyu C. Exposures of preschool children to chlorpyrifos and its degradation product 3,5,6-trichloro-2-pyridinol in their everyday environments. Journal of Exposure Analysis and Environmental Epidemiology. 2005b; 15:297–309. [PubMed: 15367928]

15. Wilson NK, Chuang JC, Iachan R, Lyu C, Gorgon SM, Morgan MK, Ozkaynak H, Sheldon LS. Design and sampling methodology for a large study of preschool children's aggregate exposures to persistent organic pollutants in their everyday environments. Journal of Exposure Analysis and Environmental Epidemiology. 2004; 14:260–274. [PubMed: 15141155]

16. Morgan, MK.; Sheldon, LS.; Crogham, CW.; Chuang, JC.; Lordo, RA.; Wilson, NK.; Lyu, C.; Brinkman, M.; Morse, N.; Chou, YL.; Hamilton; Finegold, JK.; Hand, K.; Grodon, SM. A Pilot Study of Children's Total Exposure to Persistent Pesticides and other Persistent Organic Pollutants (CTEPP). Vol. 1. US EPA; 2005a [Accessed on September 23, 2008]. http://www.epa.gov/heasd/ctepp/ctepp_report.pdf

17. Chen J, Kumar M, Chan W, Berkowitz G, Wetmur JG. Increased Influence of Genetic Variation on PON1 Activity in Neonates. Environmental Health Perspectives. 2003; 111(11):1403–1409. [PubMed: 12928148]

18. McKeown-Eyssen F, Baines C, Cole DEC, Riley N, Tyndale RF, Marshall L, Jazmaji V. Case-control study of genotypes in multiple chemical sensitivity: CYP2D6, NAT1, NAT2, PON1, PON2 and MTHFR. International Journal of Epidemiology. 2004; 33:1–8.

19. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika. 1988; 75(1):11–20.

**Table 1**

Candidate Sampling Designs from Case Study Examples

| **Case Study 1 (Cross-Sectional): $\psi$ = 1.5 ($\beta_3$=0.4055) 2-sided test with size ($\alpha$=0.05) and power (1−$\beta$=0.80)** | | | | | | |
|---|---|---|---|---|---|---|
| **Design** | **Random Sampling** | | **Covariate Dependent Sampling (Dependent on Z)** | | **Outcome Dependent Sampling** | |
| | **Cost** | **Sample Sizes** | **Cost** | **Sample Sizes** | **Cost** | **Sample Sizes** |
| ***Classic X*** | *$15,970,150; n = 7210* | | | | | |
| **2-Stage Z_Quest/X ρ=0.288** | $8,730,190 (54.7%) | $n_0$ = 96646<br>$n_1$= 869 | $8,710,604 (54.5%) | $n_0$= 92734<br>$n_1$=990 | $7,350,710 (46.0%) | $n_0$= 11751<br>$n_1$= 1335 |
| **2-Stage Z_SF/X ρ=0.892** | $5,518,340 (34.6%) | $n_0$= 9505<br>$n_1$= 90 | $8,305,180 (52.0%) | $n_0$=9390<br>$n_1$=135 | $7,082,560 (44.3%) | $n_0$= 7595<br>$n_1$= 380 |
| **Case Study 2 (Cross-Sectional): $\psi$ = 1.5 ($\beta$3=0.4055) 2-sided test with size ($\alpha$ =0.05) and power (1−$\beta$ =0.80)** | | | | | | |
| **Design** | **Random Sampling** | | **Covariate Dependent Sampling (Dependent on E)** | | **Outcome Dependent Sampling** | |
| | **Cost** | **Sample Sizes** | **Cost** | **Sample Sizes** | **Cost** | **Sample Sizes** |
| ***Classic X*** | $13,446,000; *n = 7210* | | | | | |
| **2-Stage Z_Quest/X ρ=0.288** | $8,405,620 (62.5%) | $n_0$= 93266<br>$n_1$= 991 | $5,641,020 (42.0%) | $n_0$= 53382<br>$n_1$=1029 | $2,503,640 (18.6%) | $n_0$ = 16992<br>$n_1$ = 721 |
| **2-Stage Z_SF/X ρ=0.892** | $5,015,670 (37.3%) | $n_0$ = 9489<br>$n_1$= 96 | $4,918,830 (36.6%) | $n_0$ = 9218<br>$n_1$ =128 | $4,424,380 (32.9%) | $n_0$= 7595<br>$n_1$ = 380 |
| **Case Study 3 (Longitudinal): $\beta_3$=−0.01 2-sided test with size ($\alpha$=0.05) and power (1−$\beta$=0.80)** | | | | | | |
| **Design** | **Random Sampling** | | **Covariate Dependent Sampling (Dependent on Z)** | | **Outcome Dependent Sampling** | |
| | **Cost** | **Sample Sizes** | **Cost** | **Sample Sizes** | **Cost** | **Sample Sizes** |
| ***Classic X*** | *$25,189,225; n = 11,321* | | | | | |
| **2-Stage Z_Quest/X ρ=0.288** | $15,477,600 (61.4%) | $n_0$=100,000<br>$n_1$=3,431 | $15,491,200 (61.5%) | $n_0$=99,998<br>$n_1$ =3,438 | $25,443,626 (101.1%) | $n_0$ =12,254<br>$n_1$ =11,202 |
| **2-Stage Z_SF/X ρ=0.892** | $8,503,830 (33.8%) | $n_0$ =14,483<br>$n_1$=107 | $12,795,000 (50.8%) | $n_0$=14,385<br>$n_1$=155 | $12,794,824 (50.8%) | $n_0$=14,389<br>$n_1$=152 |

**Table 2**

Sample Size Requirements for a Random 2nd Stage Exposure Sample within a Large Cohort Study.

| $\psi_{YX}$ ($\psi$ Y(X95/X50)) | $\rho_{XZ}$ | 100,000 Cohort Size | | | | | 75,000 Cohort Size | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $P_Y$ = 0.0025 | $P_Y$ = 0.005 | $P_Y$ = 0.01 | $P_Y$ = 0.025 | $P_Y$ = 0.05 | $P_Y$ = 0.0025 | $P_Y$ = 0.005 | $P_Y$ = 0.01 | $P_Y$ = 0.025 | $P_Y$ = 0.05 |
| 1.05 (1.10) | 0.00 | 1322202 | 662777 | 333075 | 135294 | 69439 | 1322202 | 662777 | 333075 | 135294 | 69439 |
| | 0.10 | | | | | 69104 | | | | | 69383 |
| | 0.30 | | | | | 66386 | | | | | 68889 |
| | 0.50 | | | | | 59209 | | | | | 67586 |
| | 0.70 | | | | | 39990 | | | | | 64097 |
| | 0.90 | | | | | 13 | | | | | 45737 |
| 1.10 (1.205) | 0.00 | 346504 | 173699 | 87300 | 35472 | 18214 | 346504 | 173699 | 87300 | 35472 | 18214 |
| | 0.10 | | | 87169 | 34820 | 17391 | | | | 35072 | 17642 |
| | 0.30 | | | 86041 | 29097 | 10167 | | | | 31565 | 12617 |
| | 0.50 | | | 83063 | 14005 | 86 | | | | 22309 | 485 |
| | 0.70 | | | 75091 | 30 | 14 | | | | 217 | 17 |
| | 0.90 | | | 33104 | 4 | 3 | | | | 5 | 3 |
| 1.25 (1.549) | 0.00 | 63232 | 31709 | 15947 | 6492 | 3343 | 63232 | 31709 | 15947 | 6492 | 3343 |
| | 0.10 | 62866 | 31022 | 15102 | 5570 | 2457 | 63118 | 31274 | 15353 | 5812 | 2668 |
| | 0.30 | 59603 | 24967 | 7696 | 254 | 121 | 62074 | 27434 | 10135 | 610 | 147 |
| | 0.50 | 50989 | 9021 | 65 | 32 | 28 | 59317 | 17304 | 151 | 36 | 29 |
| | 0.70 | 27940 | 24 | 13 | 10 | 9 | 51940 | 60 | 15 | 10 | 10 |
| | 0.90 | 9 | 4 | 3 | 3 | 2 | 13134 | 4 | 3 | 3 | 2 |
| 1.50 (2.214) | 0.00 | 19169 | 9622 | 4849 | 1985 | 1032 | 19169 | 9622 | 4849 | 1985 | 1032 |
| | 0.10 | 18355 | 8720 | 3930 | 1243 | 602 | 18606 | 8968 | 4163 | 1385 | 645 |
| | 0.30 | 11212 | 1249 | 163 | 98 | 85 | 13665 | 3287 | 244 | 107 | 88 |
| | 0.50 | 100 | 39 | 30 | 26 | 25 | 1088 | 49 | 32 | 27 | 25 |
| | 0.70 | 14 | 11 | 10 | 9 | 9 | 18 | 12 | 10 | 9 | 9 |
| | 0.90 | 3 | 3 | 2 | 2 | 2 | 3 | 3 | 3 | 2 | 2 |

$\psi Y(X_{95}/X_{50})$ corresponds to the increased odds of the health outcome based on a comparison of the 95th percentile of exposure to the 50th percentile of exposure.