Battelle
The Business of Innovation

Development of cost effective statistical sampling strategies and optimal design considerations for exposure assessment as part of the National Children's Study

Environmental Exposure & Health 2005

Warren Strauss, Battelle

October 5, 2005

Collaborators/Sponsors

Battelle: Warren Strauss, Michele Morora, Mark Davis, Nicole Iroz-Elardo, John Menkedick, Tim Pivetz, Jeff Lehman

Harvard: Louise Ryan, Andy Houseman, Kevin Roberts, Sohee Park, Cassandra Arroyo, David Loecke

NICHD: Peter Scheidt, Jim Quackenboss, Ruth Brenner, Warren Galke, Terry Dwyer

EPA: Haluk Ozkaynak, Tom McCurdy

About the National Children's Study

Executive Order 13045 (April 1997) → Directed Federal agencies to make it a priority to **identify, assess, and address children's health** and safety risks

Authorized NICHD to conduct a **national longitudinal study of environmental influences** (including physical, chemical, biological, and psychosocial) on children's health and development ← Children's Health Act of 2000

"The Director of NICHD shall….. (1) plan, develop, and implement **a prospective cohort study**, from birth to adulthood, to evaluate the effects of both chronic and intermittent exposures on child health and human development; and (2) investigate basic mechanisms of developmental disorders and environmental factors, both risk and protective, that influence health and developmental processes."

Track 100,000 Births → Pre-conception → Pregnancy → Birth → 21 Years Old

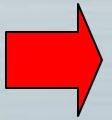
Projected $2.7B Budget

Primary Objectives of the NCS

- A National Longitudinal Study of Environmental Influences on Children's Health and Development
  - Environment defined broadly to include physical, chemical, biological and psychosocial factors
- To investigate basic mechanisms of developmental disorders and environmental factors, both risk and protective
- Determine the presence or absence of the effects of chemical, physical, and social exposures in children's environments
- Determine cause and severity of specific conditions of children that are related to environmental exposures
- Create a national resource for future studies of child health and development

Priority Health Outcomes/Exposures

Priority Exposures	Examples
Physical Environment	Housing quality, neighborhood
Chemical Exposures	Pesticides, phthalates, heavy metals
Biologic Environment	Infectious agents, endotoxins, diet
Genetics	Interaction between environmental factors and genes
Psychosocial milieu	Families, SES, institutions, social networks



Priority Health Outcomes	Examples
Pregnancy Outcomes	Preterm, Birth defects
Neurodevelopment & Behavior	Autism, schizophrenia, learning disabilities
Injury	Head trauma, Injuries requiring hospitalizations
Asthma	Asthma incidence and exacerbation
Obesity & Physical Development	Obesity, Diabetes, altered puberty

See http://www.nationalchildrensstudy.gov/research/hypotheses/ for more information on specific research hypotheses for the NCS

Why a Longitudinal Cohort Study?

- Links between many exposures and children's health not adequately investigated (esp. mixtures)
- Life - Stage effort
  - Timing of exposures
  - Timing of outcomes
- Typical studies limited in size and scope
- This study will be a national resource for other studies

Why Now?

- Environment today vastly different than two generations ago: new envirotoxins, changes in social structure, diet, behavior, and other factors potentially compromising health
- Increases in severe chronic diseases and developmental disorders: relative genetic and environmental etiologies for >80% are unknown
- No large longitudinal cohort study of development in the US since 1960s
- Recent (and future) developments in biomarkers, biotechnology, informatics, and genetics.

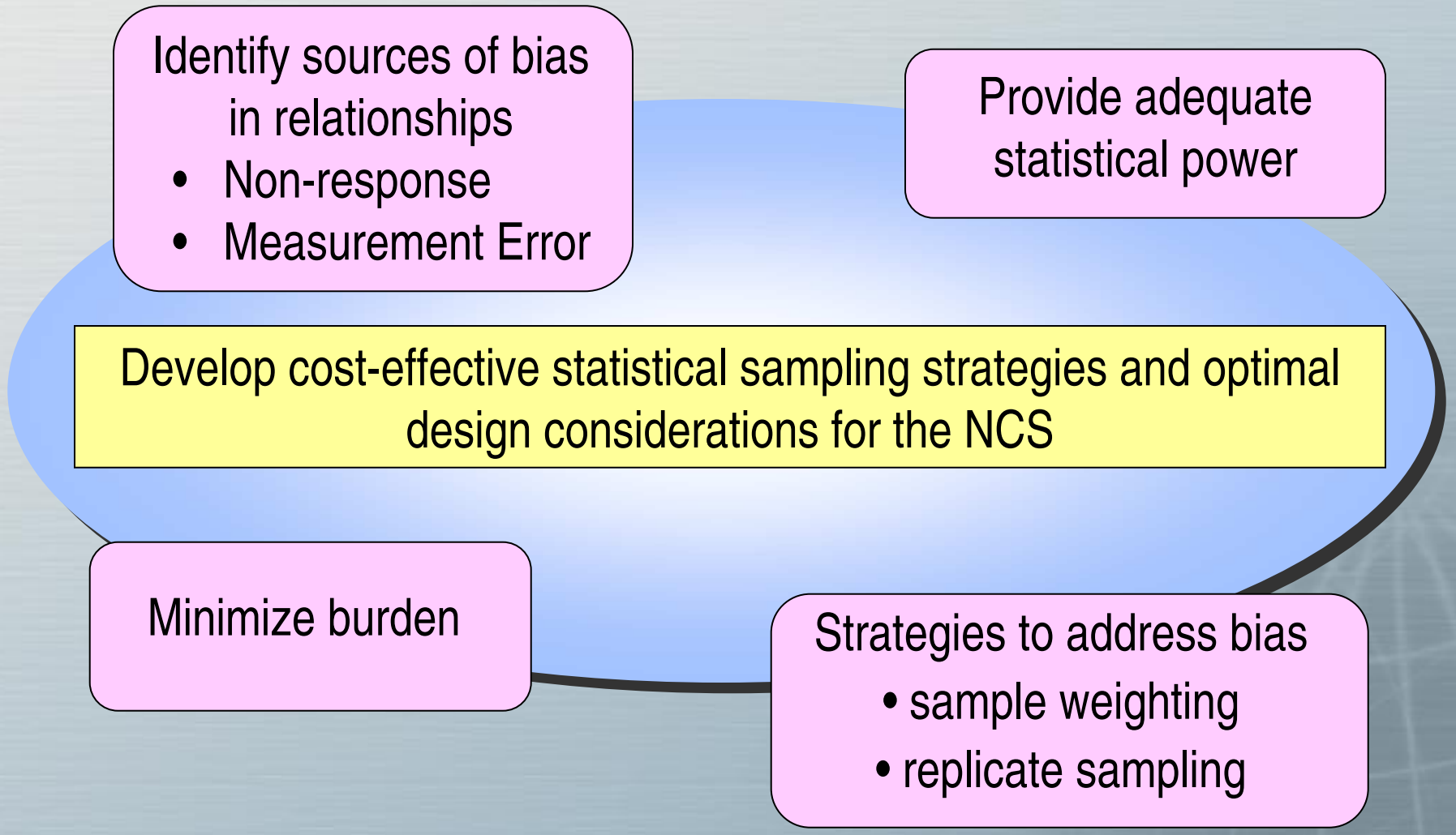
Challenges

- Long-term funding
- Mobility versus long-term follow-up
- Representativeness & diversity

- Develop and use emerging technologies
  - Techniques for specimen & data collection, analysis, and archiving
  - Information technology

- Coordination among centers
- Ethical issues
- Cost and burden associated with exposure and outcome assessments over 20+ year period of follow-up
- How should the study efficiently conduct exposure assessments?
  - The study cannot afford to measure everything on every subject

Exposure Assessment Design



Validation Samples

- A small sample that is designed in a purposive manner to provide information related to the bias or error introduced into the main study cohort by the nature of the design
- The information gathered from the validation sample designed to allow for appropriate statistical adjustments to the data collected in the larger cohort to address bias and error

Example Uses of Validation Samples within the NCS

Larger cohort	Collect exposure information using low-cost, low-precision methods across the cohort	Use of a single biomarker for most respondents, even when there is an anticipated high degree of within-person temporal variability
Validation Samples	Undergoes detailed environmental assessment	Contains people w/ multiple measures over time

Implications for Design

- Cost savings
- Potential to minimize burden
- Possible use of a smaller pre-conception validation sample
  - Use of retrospective measures of exposure for main cohort
  - Corrections for temporal variability
- Careful planning in the study design to ensure that appropriate relationships between measurements are captured

Conceptual Model for Validation Samples

- Let….. Y be the Health Outcome of Interest,
  X be the "gold-standard" measure of exposure, and
  Z be a less precise measure of exposure

- X is measured on a small subset of the cohort, whereas Y and Z are measured on the entire cohort

- **Idea is to leverage the information contained in the small validation sample that contains X to draw inferences on the effect of "true" exposure (X) on outcome (Y).**

- There are three general methods for selecting the subset of study participants that are in the validation sample (i.e., have X):

1. Outcome Dependent Sampling (Depending on Y)
2. Covariate Dependent Sampling (Depending on Z)
3. Random Sampling (No Information)

Utility of Validation Samples in NCS

- Provides a statistical basis to correct for bias and error in exposure assessment when investigating relationships
- Allows NCS to less detailed measures of exposure for the majority of the cohort, while preserving the ability to assess the impact of "true" exposure on disease
  - Assumes "true" exposure can be measured on subset
- Potential for
  - **Substantial cost savings** when detailed exposure assessment is expensive and reasonable surrogate measures can be implemented at a lower cost
    - e.g., passive air sampler and questionnaire vs. continuous/active samplers for pesticides in indoor air
  - **A more feasible study** – especially when applied to pre-conception or peri-conception exposures
    - Temporal variability/bias from early gestation to later in pregnancy is just another source of error that can be addressed using this methodology

Results from Previous Work on Validation Samples (Battelle/Harvard/EPA)

Optimal Designs Depend on Characteristics of Exposure and Health Outcome of Interest

General Conclusions

- Difficult to identify a single "optimal" design
- For each hypothesis of interest, optimally designed sub-studies should be considered and investigated
- Well designed sub-studies can efficiently characterize exposure/health outcome relationships using a fraction of the NCS cohort
- "Detailed" exposure information for all subjects may not be necessary to characterize effect of exposure on health outcome.

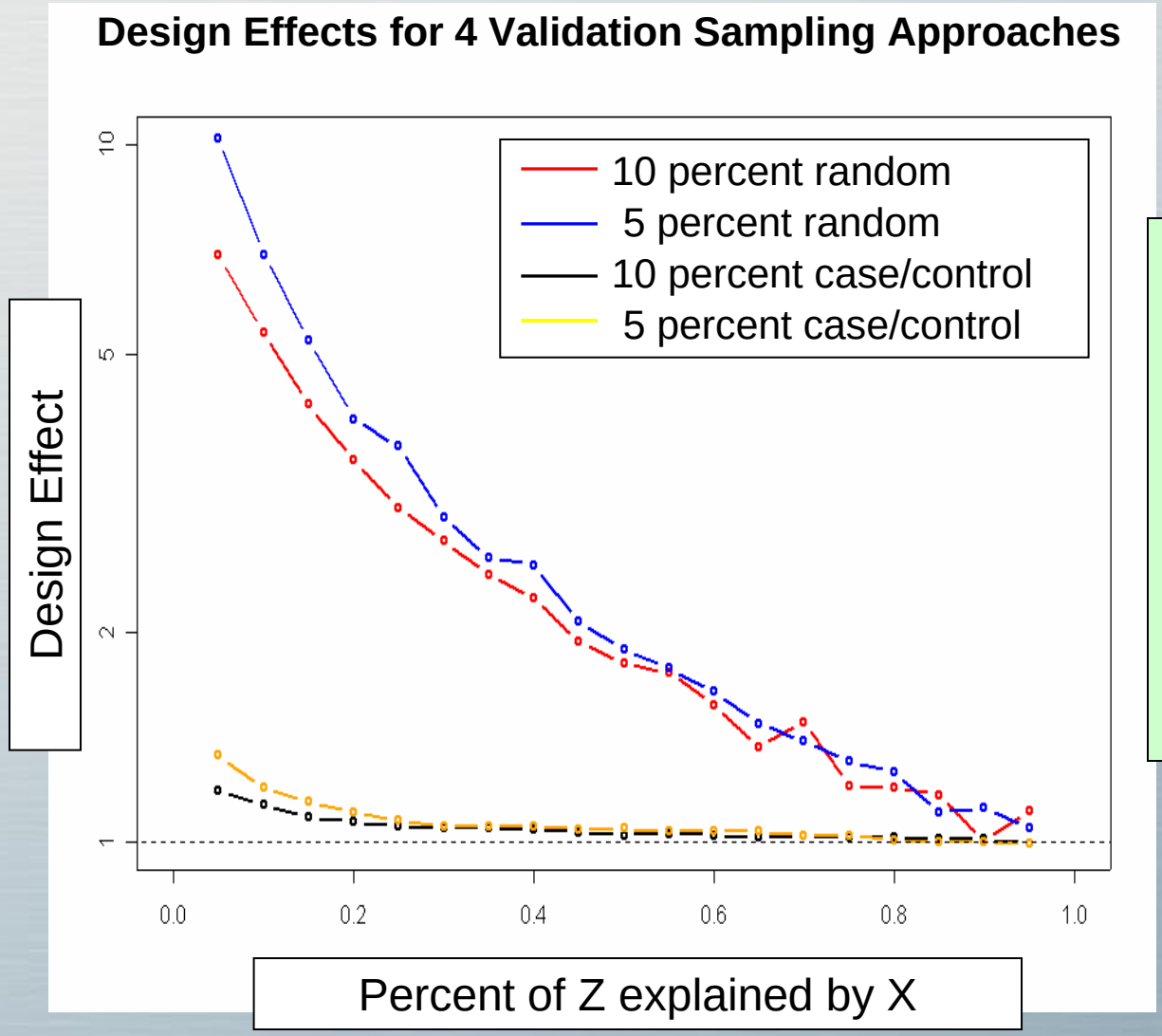
Factors To Consider

- Exposure period
- Pathways/measurements
- Variability of exposure measurements
- Exposure metric related to disease (average, acute, etc.)
- Continuous or binary outcome
- Longitudinal outcomes
- Prevalence of outcome
- Variability of outcome measurements
- Strength of exposure/outcome relationship

Questions on Validation Samples

- What is the sample size of the validation sample that is necessary to support NCS inferences?
  - Is this a function of the size of the NCS cohort (i.e., a 10% sample)?
  - or the number of (X,Z) pairs necessary to accurately establish the relationship between "true" and "surrogate" measures of exposure to utilize in the statistical errors-in-variables correction?
- How is the power to assess relationships between exposure and health outcome affected by the strength of the relationship between X and Z?
- How does the approach to selecting participants into the validation sample affect the above two questions?
- Assess answers to these questions using Design Effects as a function of $\rho_{X,Z}$, n_X, and X sampling strategy
  - Reference Design is one where entire cohort has X measures
  - Simulations correspond to a cohort size of 10,000 with a disease prevalence of 0.025

Design Effects from Validation Sampling



Validation Sampling

- X is a very good measure of exposure, while Z is a less costly, less accurate measure
- Measure Z on everyone and X on a chosen few vs. measuring X on everyone
- Define the portion of variability in Z explained by X assuming, $Z=X+error \sim R^2$

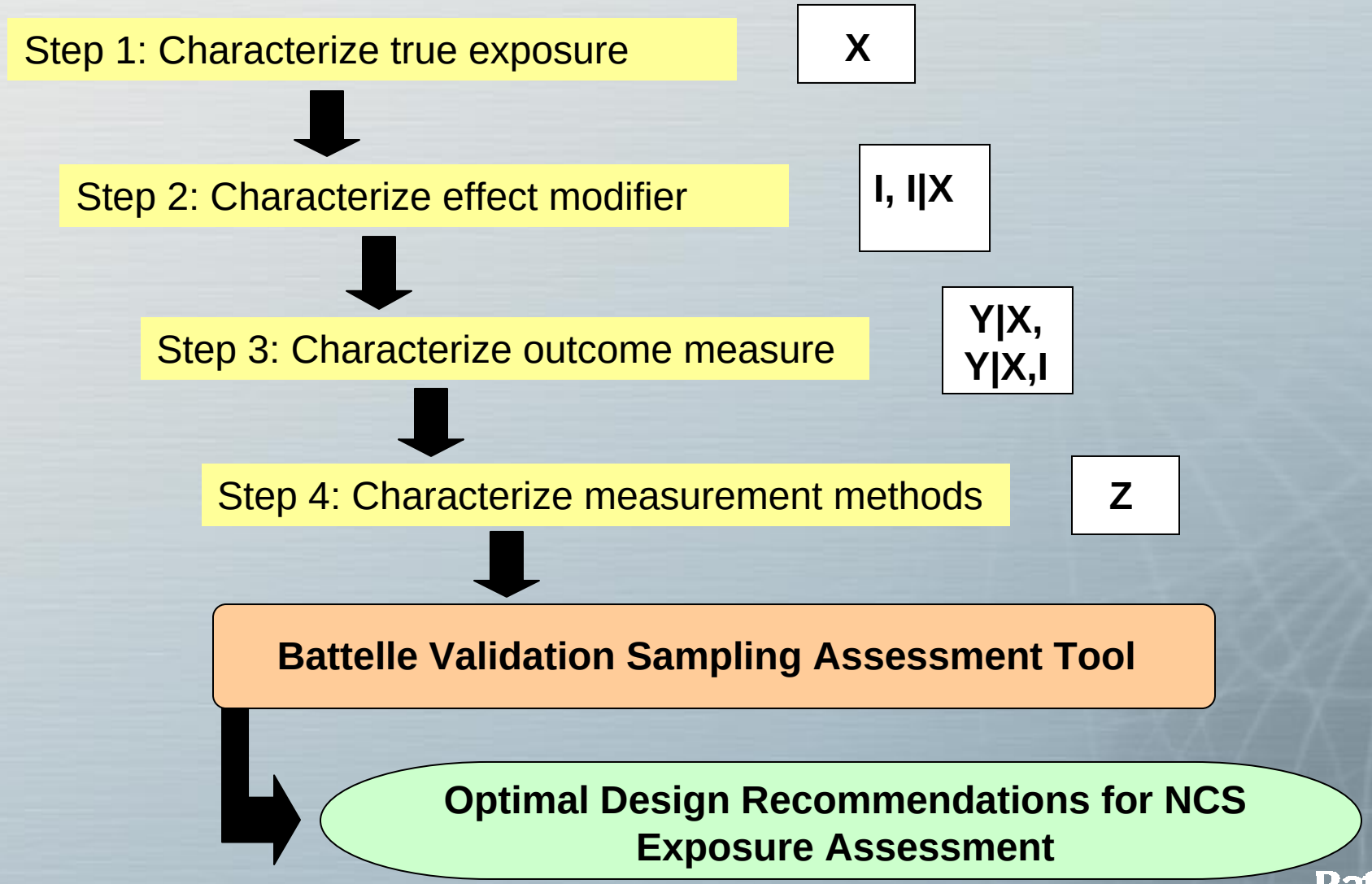
General Conclusions from Previous Work

- Size of validation sample is dictated by the strength of the relationship between X and Z
- Use of an Outcome Dependent Design for the selection of subjects into the validation sample leads to extremely efficient analysis
  - Design effects below 2 – even when $\rho_{X,Z}$ is small (<.3)
- Use of Random Sampling for the selection of subjects into the validation sample leads to slightly less efficient designs
  - Design effects still below 2 when $\rho_{X,Z}$ is reasonably large (above 0.5)

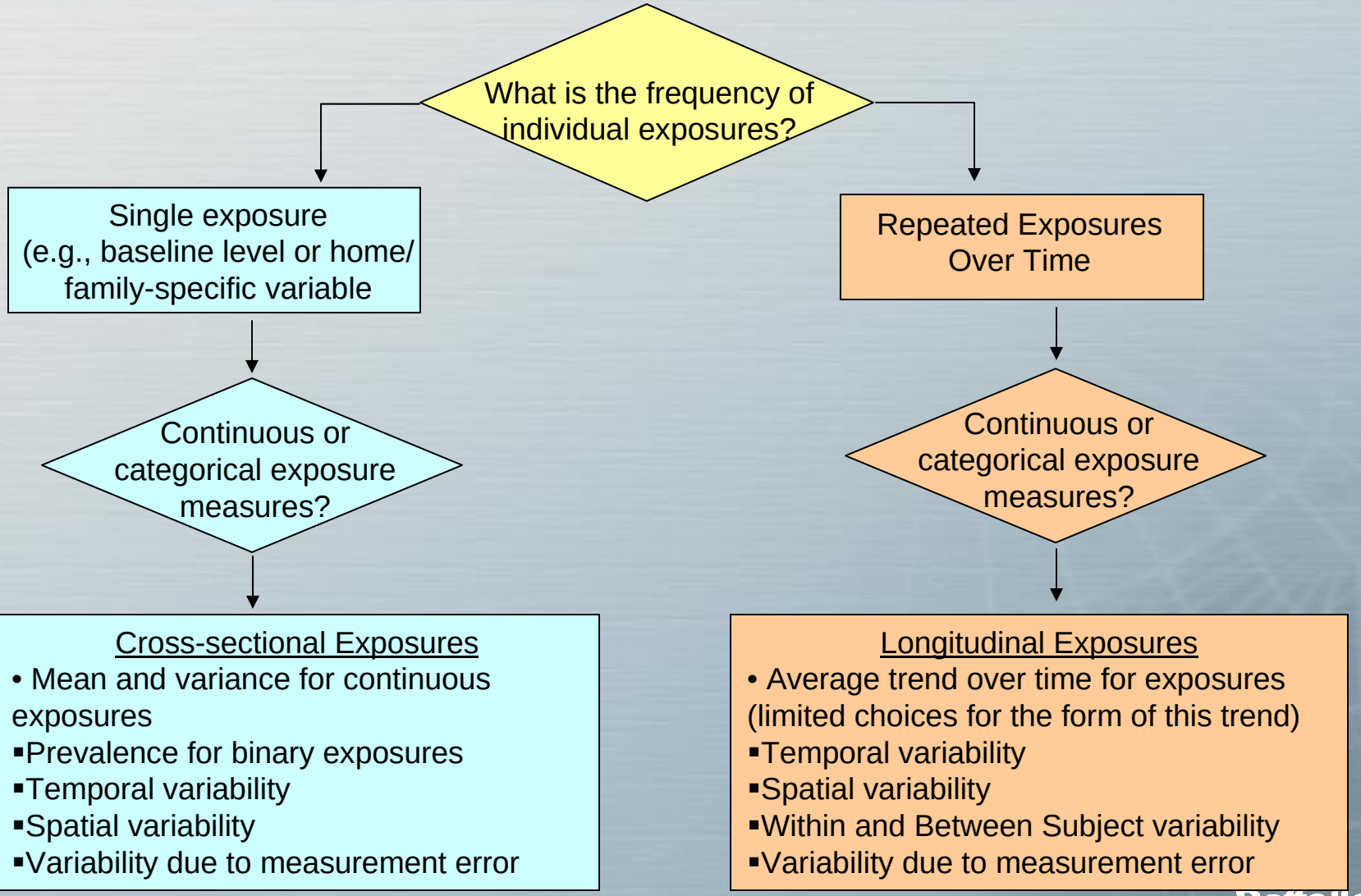
New Work Developing Software for Exposure Assessment Design

- Battelle and Harvard are jointly developing a prototype software tool to allow NCS study planners to research benefits and limitations of utilizing study designs that employ validation sampling techniques
- Tool contains interface that sequentially interviews the user on critical design input regarding the health outcome, exposure, type of relationship between exposure and outcome, potential measurement methods for exposure, sample size, and resource constraints
- Output provided on cost, sample size, and power across a range of designs

Decision Pathway for NCS Environmental Exposure Assessment



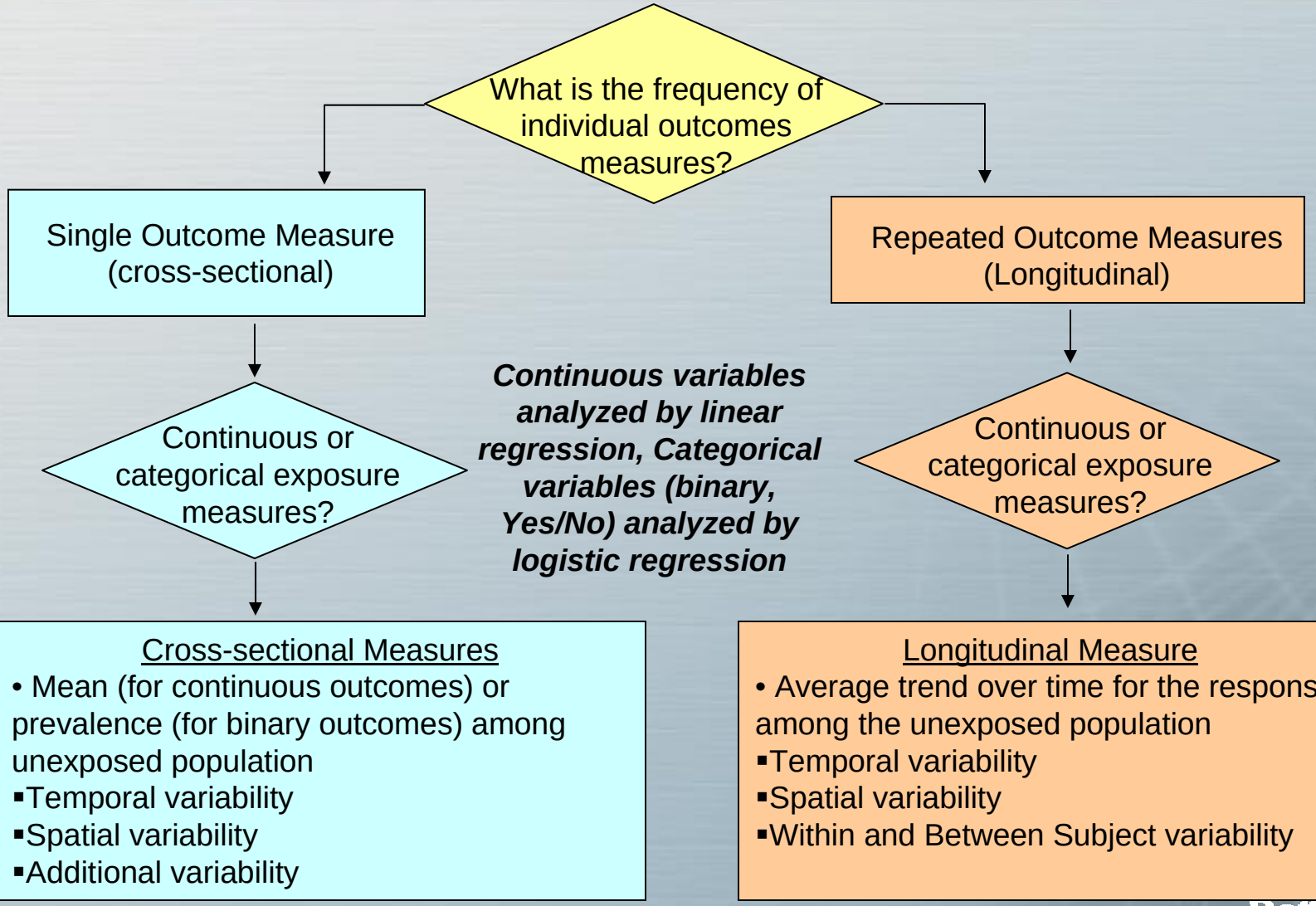
Step 1: Characterize True Exposure, X



Step 2: Characterize Effect Modifier

- Effect modifier (I) may be factors such as genetic predisposition to disease
- Define whether effect modifier is continuous or categorical
- Define the relationship between effect modifier and true exposure X (I|X)
  - In many cases, we would expect exposure and the effect modifier to be independent
  - Example of where there is a relationship:
    - I: Allergic sensitivity to cat dander (effect modifier)
    - X: Exposure to cat allergens
    - Likely a negative association between I and X

Step 3: Characterize Outcome Measures



Step 3: Characterize Outcome Measures

	Categorical Outcome Measure	Continuous Outcome Measure
No Effect Modifier	Logisitic Regression $Logit(\pi) = \beta_0 + \beta_1 \cdot X$ Where π = Pr(Y=1)	Linear Regression $Y = \beta_0 + \beta_1 X + \varepsilon$
With Effect Modifier	Logisitic Regression $Logit(\pi) = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot I + \beta_3 \cdot I \cdot X$ Where π = Pr(Y=1)	Linear Regression $Y = \beta_0 + \beta_1 X + \beta_2 I + \beta_3 IX + \varepsilon$
Longitudinal	Use Generalized Estimating Equations approaches to account for positive correlation among repeated measures	Use Mixed Models Analysis of Variances approaches to account for positive correlation among repeated measures

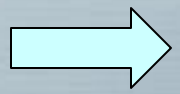
Step 4: Characterize Measurement Methods

For each measure of exposure assessment (Z_j) – the user will provide input on:

- Relationship with X
- Costs (sampling, storage, chemical analysis)
- Detection limits
- Whether it could be archived for future analysis
- Whether it could be collected following another exposure measure

Examples

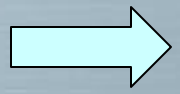
A gold standard measure is available for X, but it is too complicated and costly. Z is a surrogate measure related to the gold standard, assumed to be cheap and easy enough to be measured at all timepoints for every participant.



Z = X + error

Relationship between Z and X defined by $\rho_{x,z}$ (correlation coefficient)

What if there are multiple surrogate measures available? We assume that at least one of the Z's will be measured for everyone.



$Z_1, Z_2, \ldots\ldots, Z_k$

Example 2-Stage Sampling Equations

Random Sampling

Stage 1: $\text{logit}(\pi_1) = \alpha_{10}$

Stage 2: $\text{logit}(\pi_2) = \alpha_{20}$

Covariate Dependent Sampling

Stage 1: $\text{logit}(\pi_1) = \alpha_{10}$

Stage 2: $\text{logit}(\pi_2) = \alpha_{20} + \alpha_{21} \cdot Z$

Outcome Dependent Sampling

Stage 1: $\text{logit}(\pi_1) = \alpha_{10}$

Stage 2: $\text{logit}(\pi_2) = \alpha_{20} + \alpha_{21} \cdot Y$

π_1=Probability of Stage1 Selection (Y,Z)
π_2=Probability of Stage2 Selection (X), given that you were sampled in Stage 1

Two Types of Constrained Optimization

- Constrained Budget: Find α's that minimize the variance of β
- Constrained Variance: Find α's that minimize the cost of sampling

Case Example 1

Cross-sectional Investigation of Autism – Focus on Costs of Exposure Assessment

$Y \sim Bin(P_Y = 0.003)$ Cost associated with measuring Y is $20

$X \sim N(0,1)$ Cost for exposure assessment = $1000

$\Psi_{Y,X} = 2.0$ Odds ratio between X and Y

Total Cohort Size = 100,000

4 Potential Surrogate Measures

Surrogate Measure	Cost	Correlation with X
Z_1	$10	$\rho_{X,Z_1} = 0.3$
Z_2	$50	$\rho_{X,Z_2} = 0.5$
Z_3	$100	$\rho_{X,Z_3} = 0.7$
Z_3	$200	$\rho_{X,Z_3} = 0.9$

Case Example 1: Results for 2-Stage Sampling (X observable)

- Designs that allow for a two-sided test of the Odds Ratio between Y and X with size (α=0.05) and power (1-β = 0.80)

Design	Random Sample		Covariate Dependent Sample (for X)		Outcome Dependent Sample (for X)	
	Cost	N	Cost	N	Cost	N
Classic X	Cost = \$5,606,940 n =5,497					
Z_1/X ρ_{X,Z_1} = 0.3	\$1,920,000 (34%)	n_Y=61,000 n_Z=61,000 n_X=91	\$1,850,000 (33%)	n_Y=61,000 n_Z=61,000 n_X=20	\$183,500 (3.3%)	n_Y=5,550 n_Z=5,550 n_X=17
Z_2/X ρ_{X,Z_2} = 0.5	\$1,600,000 (29%)	n_Y=21,997 n_Z=21,997 n_X=61	\$1,560,000 (28%)	n_Y=22,000 n_Z=22,000 n_X=20	\$405,500 (7.2%)	n_Y=5,550 n_Z=5,550 n_X=17
Z_3/X ρ_{X,Z_3} = 0.7	\$1,400,000 (25%)	n_Y=11,164 n_Z=11,164 n_X=61	\$1,370,000 (24%)	n_Y=11,250 n_Z=11,250 n_X=20	\$683,000 (12.2%)	n_Y=5,550 n_Z=5,550 n_X=17
Z_4/X ρ_{X,Z_4} = 0.9	\$1,550,000 (28%)	n_Y=6800 n_Z=6,800 n_X=55	\$1,510,000 (27%)	n_Y=6,800 n_Z=6,800 n_X=15	\$1,238,000 (22%)	n_Y=5,550 n_Z=5,550 n_X=17

Validation designs

Results produced by Beta version of Battelle's Optimal Design for Exposure Assessment tool

Case Example 1: Results for 2-Stage Sampling
(Consider X to be unobservable / Measure Z_4 at 2nd stage)

- Designs that allow for a two-sided test of the Odds Ratio between Y and X with size (α=0.05) and power (1-β = 0.80)

Design	Random Sample		Outcome Dependent Sample	
	Cost	N	Cost	N
Classic X	Cost = $5,606,940 n =5,497			
Z_1/Z_4	$1,573,780 (28%)	n_Y=9,586	$208,440 (3.7%)	n_Y=6,808
		n_Z=9,586		n_Z=6,808
		n_X=6,431		n_X=21
Z_2/Z_4	$1,571,700 (28%)	n_Y=22,110	$475,790 (8.5%)	n_Y=6,737
		n_Z=22,110		n_Z=6,737
		n_X=120		n_X=21
Z_3/Z_4	$1,365,000 (24%)	n_Y=11,275	$794,320 (14.2%)	n_Y=6,586
		n_Z=11,275		n_Z=6,586
		n_X=60		n_X=20

Validation designs

Results produced by Beta version of Battelle's Optimal Design for Exposure Assessment tool

Case Example 2

Cross-sectional Investigation of maternal exposure to non-persistent pesticides and subtle neuro-cognitive deficits in children

$Y \sim N(100,100)$ — **Cost associated with measuring IQ = $30**

$X \sim LN(4.76,2.34)$ — **Cost for aggregate exposure assessment = $1800**

$\beta = 0.5$ — **Slope explaining IQ as linear function of exposure**

Total Cohort Size = 100,000

2 Potential Surrogate Measures

Surrogate Measure	Cost	Correlation with X
Z_1 (Questionnaire)	$10	$\rho_{X,Z_1} = 0.288$
Z_2 (Solid Food Sample)	$450	$\rho_{X,Z_2} = 0.892$

Case Example 2: Results

- Designs that allow for a two-sided test of the slope between Y and X with size (α=0.05) and power (1-β = 0.80)

Design	Random Sample				
	Cost	n_Y	n_{Z1}	n_{Z2}	n_X
Classic Design	$7,938,540	4,335			4,335
Two Stage (Z_1/X)	$3,016,720	57,508	57,508		398
Two Stage (Z_2/X)	$2,681,880	5,516		5,516	19
Three Stage (Z_1/Z_2/X)	$2,354,080	37,747	37,747	1,848	7

Validation designs

Conclusions

- Validation sampling may allow NCS to conduct more cost-effective data collection while still retaining necessary power to make conclusions about study hypotheses
- Battelle automated tool for considering various sampling design considerations will allow NCS planners and protocol developers to identify optimal sampling strategies using validation studies
- Designs are highly sensitive to design input
    - Pilot studies to identify appropriate surrogates (and relationship with true exposure) will be key