Measurement of SED in Child Populations: Design and Estimation Considerations in Multi-phase Studies

Steven G. Heeringa
Institute for Social Research
University of Michigan

Presentation to the National Academy of Sciences Workshop on Integrating New Measures of Serious Emotional Disturbance in Children into SAMHSA's Data Collection Programs.

June 11, 2015

Multi-Phase Survey Design Considerations

- Statistical objectives

- Multi-phase design choices
  - General framework
  - Costs (data acquisition, screening survey, classification survey)
  - Errors (variance of estimates, survey bias, misclassification)
  - Optimal design conditional on existing data

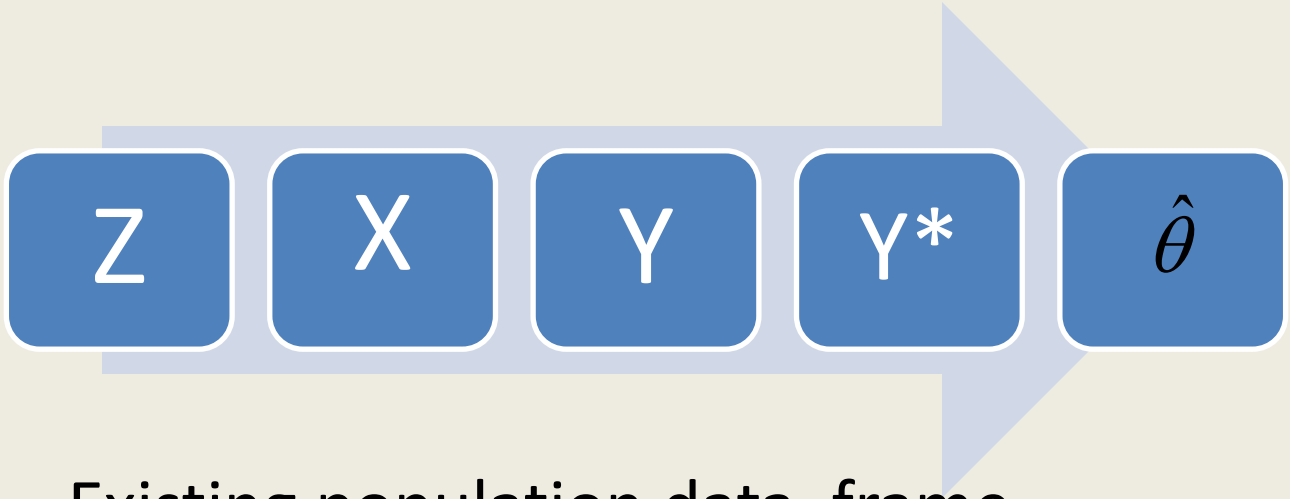
- Measurement error across phases

- Estimation and Inference
  - Direct, design-based methods
  - Model-assisted, model-based

Statistical Objectives of a Screening Study

- Target population
- Estimation of prevalence, population size
- Screening to identify a sample for in-depth subpopulation study
  - Descriptive characteristics, DX types, symptoms
  - Incidence, age of onset
  - Associated factors, causal insights (?)
  - Treatment seeking, treatment compliance

Notation for the Sequence: Design, Observation, Measurement and Estimation



Z – Existing population data, frame

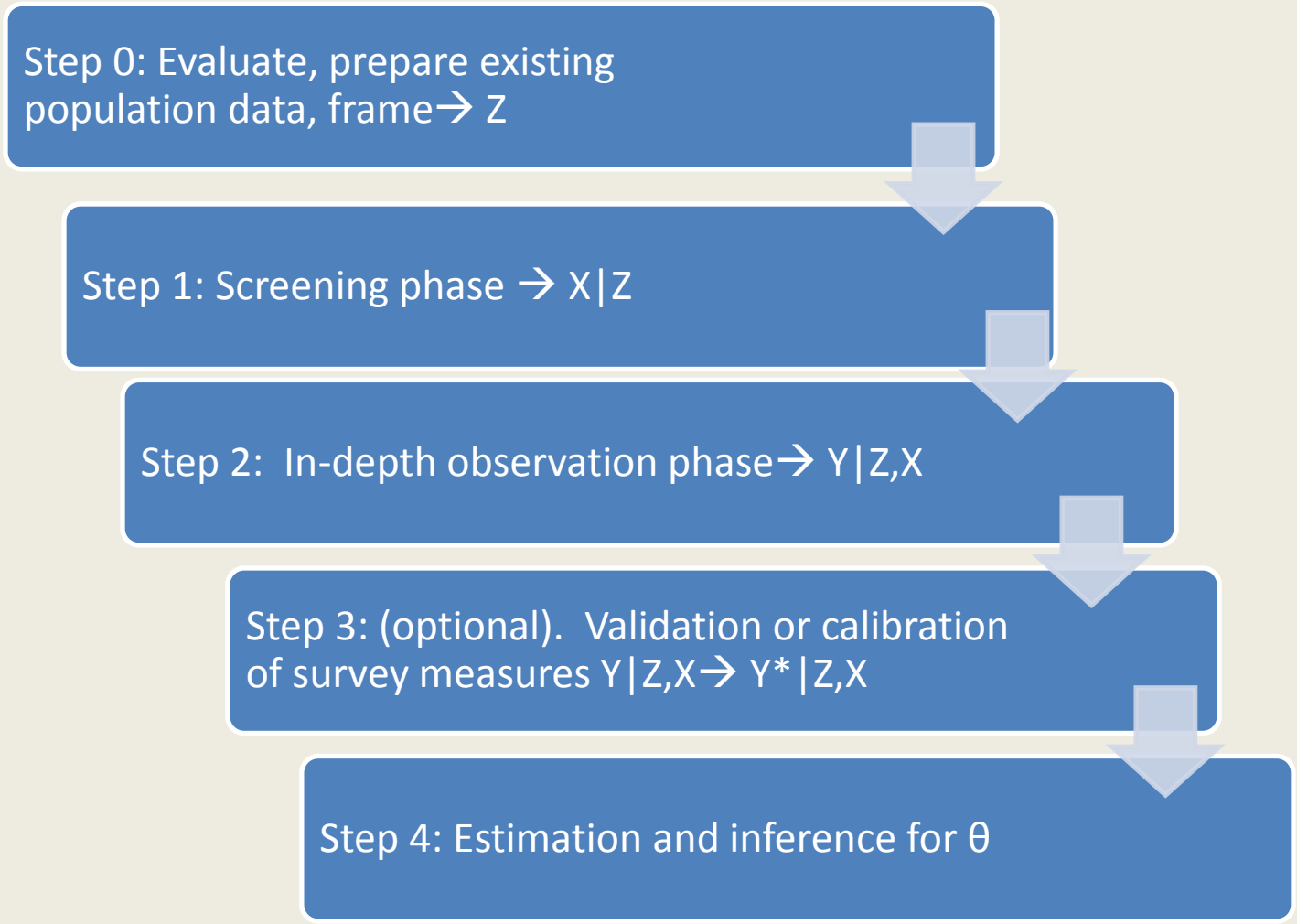
X - Phase 1 Screening data

Y - Phase 2 measurement of outcome of interest

Y*- Validated, calibrated outcome of interest

$\hat{\theta}$ - Estimate of population parameter

Multi-phase Design Framework



Cost Factors in Optimal Multi-phase Design

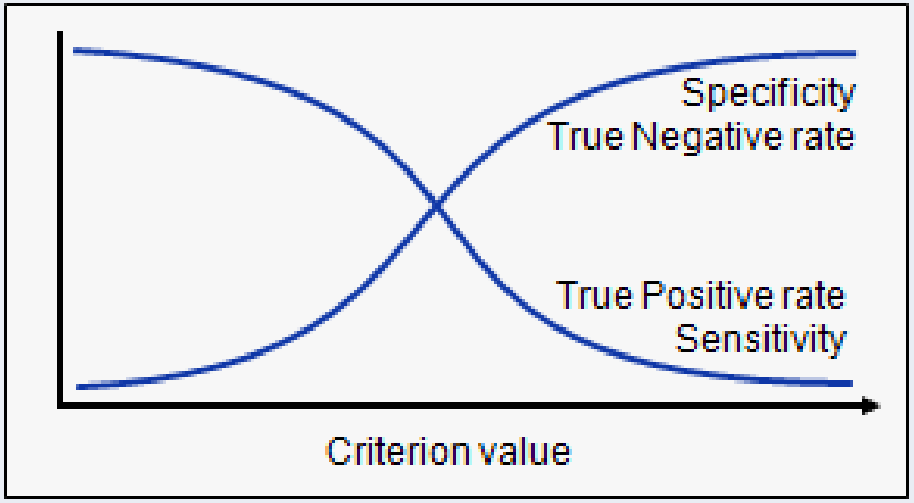
- Prevalence of target population
  - Prevalence estimation, drives n for estimating P
  - Subpop study, drives n to achieve eligible sample of size m
- Need for new Step 1 survey screening (alternative is to assign screener status using existing data source).
- Ratio of phase-specific unit costs: C(2)/C(1)
- Sensitivity of Step 1 screener
  - High false positive rate requires larger Phase 2 follow-up sample size to identify eligible case sample of size m.
- Need for validation, calibration for Y→Y*

Error Factors in Optimal Multi-Phase Design

- Prevalence of target group – drives sampling variance
- Strength of associations: Step 1 (X|Z), Step 2 (Y|Z,X)
- Specificity of screener
  - Coverage of all true cases requires Step 2 subsampling of negative screens
  - High false negative rate based on Step 1 screener implies need for variable weighting of true cases in positive screen and those in subsample of screen negative cases
  - Variable subsampling and weighting of Step 1 +/- screens
    - Increases variance of estimates of population prevalence
    - Inflates variances of estimates for analyses of true subpopulation cases
- Validity of Y for Y*- potential for classification bias

Measurement:
Sensitivity/Specificity of Step 1 Screening

Step 1 Screening or Model Assignment, g(Z,X)	Step 2 Observed Status (Y)	
	NO	YES
NO		
YES		



Measurement Example: Sensitivity/Specificity of Step 1 Screening (Example: true prevalence=.20)

Step 1 Screening or Model Assignment, g(Z,X)	Step 2 Observed Status (Y)		
	NO (0)	YES (1)	Total
NO (0)	P_{00}=.64	P_{01}=.04	P_{0+}=.36
YES (1)	P_{10}=.16	P_{11}=.16	P_{1+}=.64
Total	P_{+0}=.80	P_{+1}=.20	P_{++}=1.00

Screener Sensitivity=P_{11}/P_{+1}=0.8 Screener Specificity=P_{00}/P_{+0}=0.8

Approximate % Increase in Variance of Estimated Prevalence Based on Step 2 Sample

$$L_{percent} \approx \left[\frac{\sum_{r=0}^{1} n_r^{(2)} \cdot W_r^2}{\sum_{r=0}^{1} \left(n_r^{(2)} \cdot W_r \right)^2} \cdot (n^{(2)}) - 1 \right] \cdot 100 = \frac{Var(W_i)}{\bar{W}^2} \cdot 100$$

$=$ Relvariance of Step 2 design weights for all cases in Step 2 sample.

Example of Weighting Loss in Variance of Estimates of Population Prevalence Due to Step 2 Subsampling of Step 1 Negative Screens (true prevalence, P=.20)

f_{pos}	$f_{neg,\ sub}$	% Increase in Var(p)
1.0	0.5	8%
1.0	0.33	22%
1.0	0.25	36%
1.0	0.10	130%

Expected Disposition of Step 2 Eligible Cases in a Two-Phase Design

Step 1 Screening or Model Assignment, g(Z,X)	Step 2 Expected Eligible Cases .	
	Expected sample size. Eligible true cases	Relative Design Weight. Step 1 is epsem.
NO (0)	$E(m_{01}) = \frac{n \cdot (1-P)}{K} \cdot (1-Spec)$	$W_i = K = 1/f_{neg,\ sub}$
YES (1)	$E(m_{11}) = n \cdot P \cdot Sens$	$W_i = 1.0$

Approximate % Increase in Variance of Mean Estimates for Phase 2 Eligible Subpopulation Sample

$$L_{percent} \approx \left[\frac{\sum_{r=0}^{1} m_{r1} \cdot W_r^2}{\sum_{r=0}^{1} \left(m_{r1} \cdot W_r \right)^2} \cdot (m_{+1}) - 1 \right] \cdot 100 = \frac{Var(W_i)}{\bar{W}^2} \cdot 100$$

= Relvariance of Step 2 design weights for true cases in Step 2 sample.

Example of Approximate % Weighting Loss in Variance of Estimated Means for Phase 2 Eligible Subpopulation Sample (true prevalence=.20, sensitivity=0.8)

f_{pos}	$f_{neg,\ sub}$	Step 1 Screen Specificity					
		1.0	0.9	0.8	0.7	0.6	0.5
1.0	0.5	0%	11%	13%	12%	11%	10%
1.0	0.33	0%	30%	34%	33%	30%	20%
1.0	0.25	0%	50%	56%	54%	50%	46%
1.0	0.10	0	180%	203%	194%	180%	265%

Measurement:
Reliability and Validity in True Case Identification

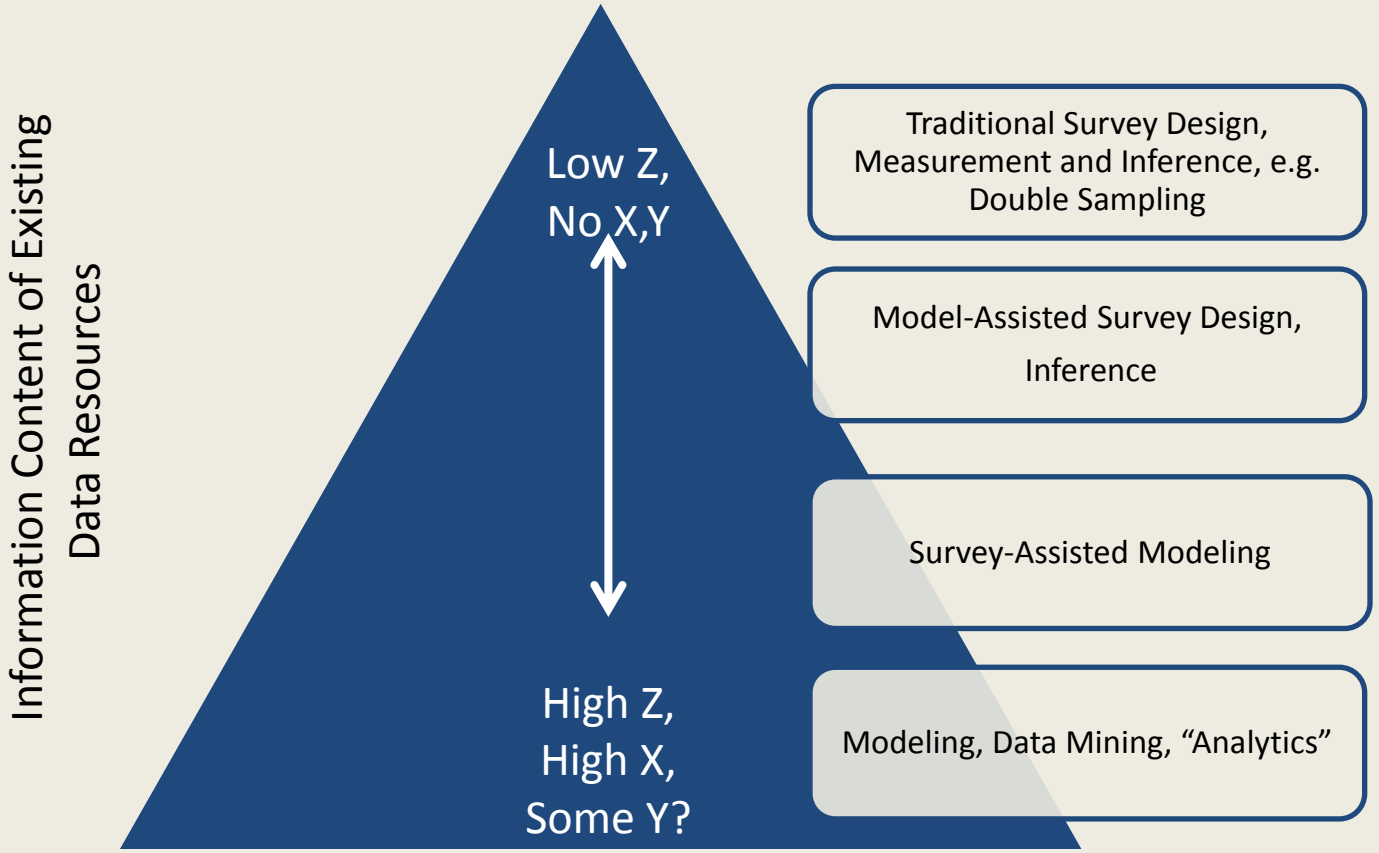
Kessler, et al. (2009). The National Comorbidity Survey Adolescent Supplement (NCS-A): III. Concordance of DS<-IV/CIDI diagnoses with clinical reassessments. *J Am Acad Child Adolescent Psychiatry:* 48(4):386-399

- Any disruptive behavior disorder, AUC = .84

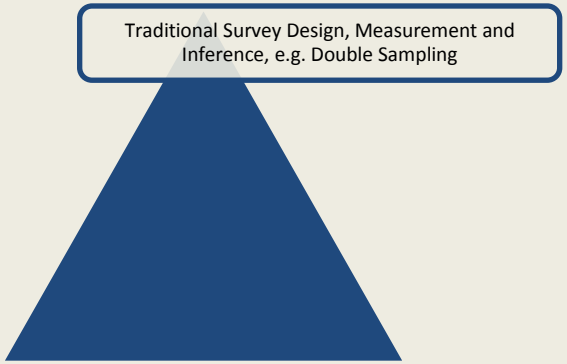
Observed/Assigned Case Status, Y	**True Case Status (Y*)**	
	NO	YES
NO	Specificity: 90.9% Sensitivity: 77.9%	
YES		
Total		

Integrating survey and administrative data.

Adaptation to Information Content of Available Data



Multi-phase Data Collections: Double Sample



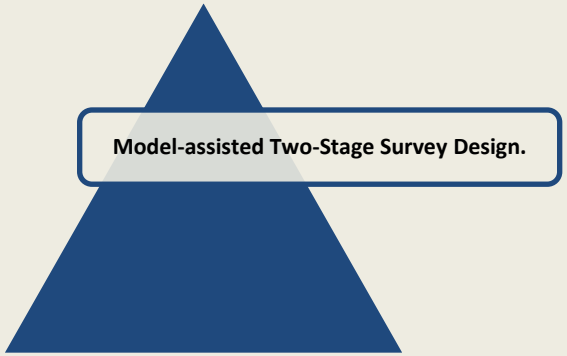
- Probability sample selected on the basis of Z
- Step 1 screening ascertains X for full sample
- Step 2 in depth interview or clinical follow-up for subsample ascertains Y or Y*.
- Optional: Calibration study after Step 2 determines Y→ Y*

Flint Men's Health Study

- Heeringa, SG., Alcser, KH., et al. (2001), "Potential Selection Bias in a Community-Based Study of PSA Levels in African-American Men," *Journal of Clinical Epidemiology*,54(2),142-148.

- Multi-phase design
  - Step 0: Area probability sample frame for Flint, MI disproportionately allocated to efficiently identify African-American households.
  - Step 1: Screening of new household sample to: 1) identify African-American males age 40+, 2)conduct health history interview, 3)obtain blood sample for PSA test (X|Z).
  - Step 2: Sample of Step 1 participants stratified by measured PSA level. Urologist clinical visit for clinical tests and transrectal ultrasound (TRUS) to determine probable cancer→ Y|X,Z.
  - Step 3: Biopsy to confirm cancer in detected growths, Y→Y*

Multi-phase Data Collections: Model-Assisted Survey Design



- Z, X^0 known for the population or existing probability sample
- Model of $f(Y^*|Z,X^0)$ is assumed
- Step 1: Under the assumed model, near optimal sample is selected directly based on $f(Y|Z,X)$ and known values of Z, X^0
- Step 2: In depth interview or clinical follow-up for the subsample ascertains Y or Y*, $f(Y|Z,X)$ is estimated and used in population estimation.
- Optional: Calibration study after Step 2 determines properties of Y→ Y*
- Standard estimation of θ from sample data

Aging Demographics and Memory Study (ADAMS) - Direct Estimation

- Langa, K.M., Plassman, B.L., Wallace, R.B., Herzog, A.R., Heeringa, S.G., Ofstedal, M.B., Burke, J.R., Fisher, G.G., Fultz, N.H., Hurd, M.D., Potter. G.G., Rodgers. W.L., Steffans, D.C., Weir, D.R., Willis, R.J. (2005). "The Aging, Demographics and Memory Study: Study Design and Methods". *Neuroepidemiology, 25,* 181-191.

- Multi-phase design
  - Step 0: Health and Retirement Survey (HRS) longitudinal panel of U.S. adults born prior to 1949. Rich longitudinal data including cognition test measures from HRS 2000, 2002. Ability to estimate a logit model of the probability of dementia from an external data set. Based on existing information in the HRS and a model the HRS panel "frame" was stratified by age, gender and cognitive score.

Dementia Probability Model (VSMA)*

$$logit\left\{ p(dementia \mid X) \right\} = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Educ + \beta_3 \cdot CogScore$$

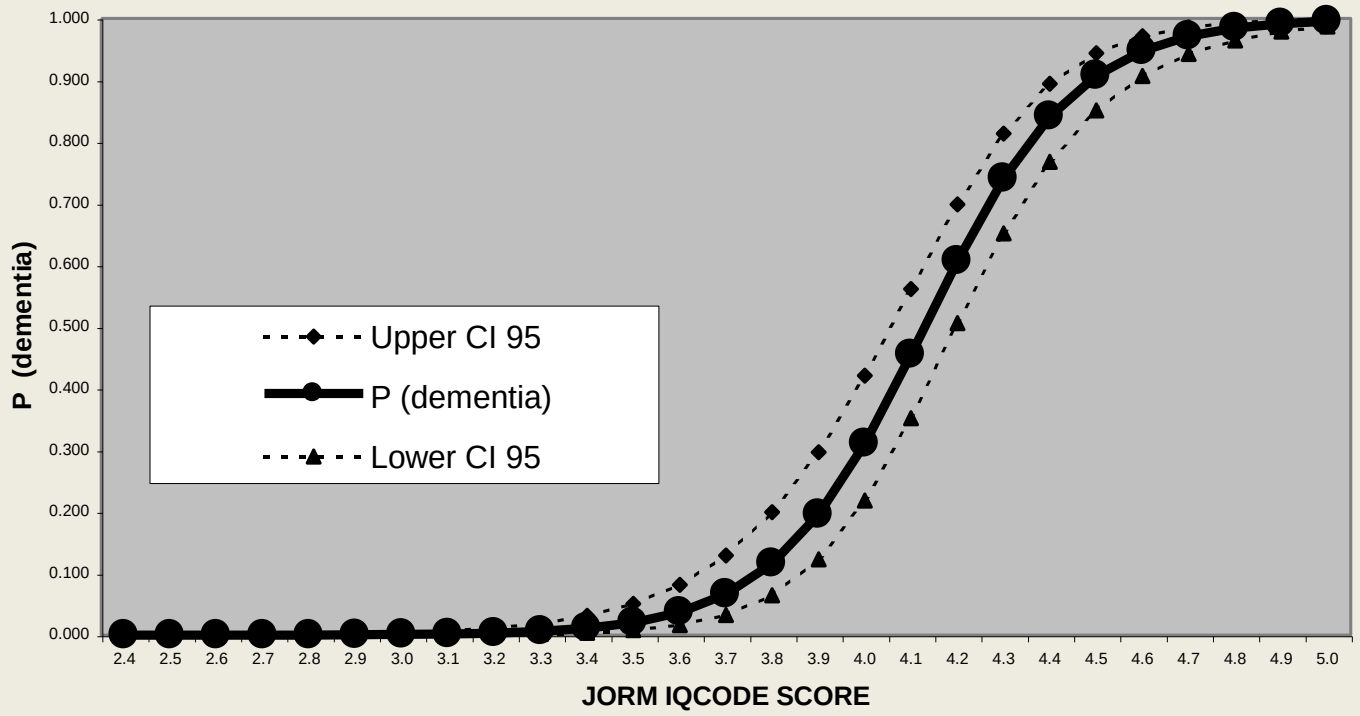
where:

CogScore = TICS 10 for HRS Self-reporters

= JORM IQ Code Score for Proxy Reports

* Source: Veterans Study of Memory and Aging

Figure 1: Predicted Probability of Dementia. Model Estimated from VSMA Data



Aging Demographics and Memory Study (ADAMS) - Direct Estimation

- Langa, K.M., Nlassman, B.L., Wallace, R.B., Herzog, A.R., Heeringa, S.G., Ofstedal, M.B., Burke, J.R., Fisher, G.G., Fultz, N.H., Hurd, M.D., Potter. G.G., Rodgers. W.L., Steffans, D.C., Weir, D.R., Willis, R.J. (2005). "The Aging, Demographics and Memory Study: Study Design and Methods". *Neuroepidemiology, 25,* 181-191.

- Multi-phase design (continued)
  - Step 1: "Screening" and stratified subsampling for follow-up of HRS panel based on a stratification that used an externally estimated model relating probability of dementia to: age, education level , and TICS/JORM cognition test scores (2000 or 2002).
  - Step 2: In-home neurocognitive assessment, medical records collection, followed by consensus diagnostic conference review by expert medical panel to assign diagnosis category: normal, CIND, possible dementia, probable dementia, ALZ
  - Step 3: Two year follow-up to refine probable/possible dementia into CIND and dementia categories. Y→Y*

Aging Demographics and Memory Study (ADAMS) - Direct Estimation

$$\hat{p}_{dementia} = \frac{\sum_i w_i \cdot y_i}{\sum_i w_i}$$

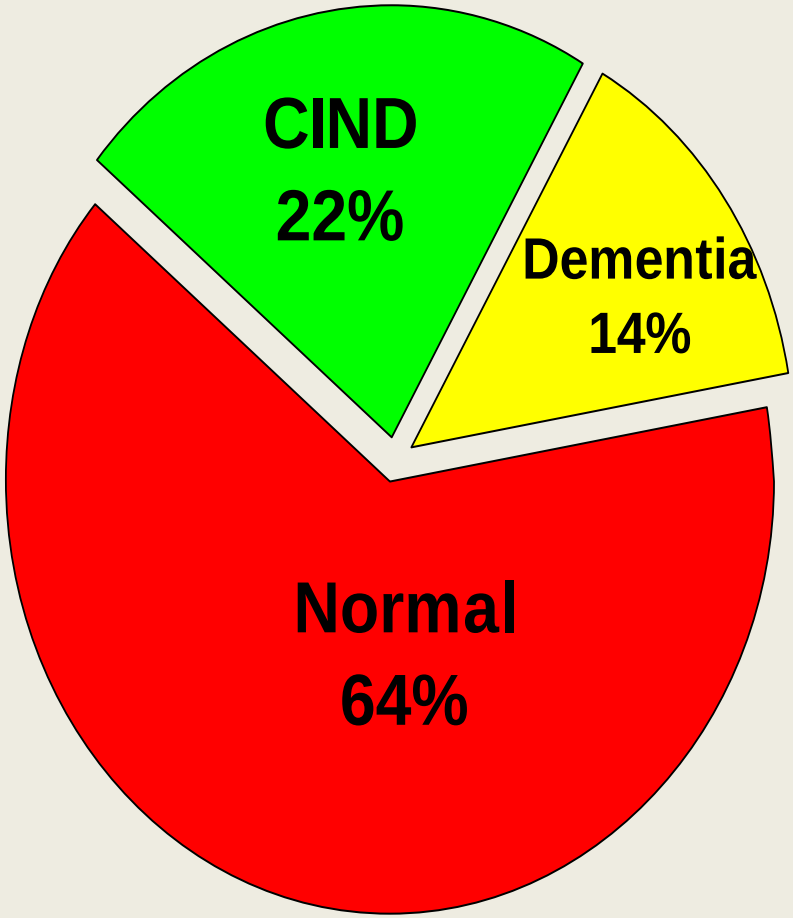
where:

$y_i = 1$ if ADAMS respondent i=1,...,n is classified as dementia, 0 otherwise;

$w_i =$ a case specific (population) weight to reflect sampling probabilities and nonresponse in the HRS panel and the ADAMS subsample.

$se(\hat{p}_{dementia}) \sim$ computed using variance estimation methods appropriate to the ADAMS complex sample.

ADAMS Estimates of 2002 Population Prevalence, Age 71+



US Population, Age 71+:

Dementia	**3.4 million**
CIND	**5.4 million**
Total Pop	**24.3 million**

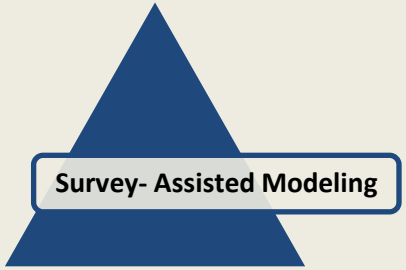
Sources : Plassman B, Langa K, Fisher G et al, 2007, 2008.

ADAMS Direct Estimates of Overall Prevalence of Dementia by Age Categories

Age	Adams Direct
70-79 years	4.95
	(1.27)
80-89 years	24.13
	(2.22)
≥ 90 years	38.18
	(3.79)
Total	13.67
	(1.29)

Percentages and Complex Design Corrected Standard Errors (parentheses) .

Multi-phase Data Collections: Survey-assisted Modeling



- Z, X^0 known for the population or existing probability sample that represents the full population
- Y* is not known. Model of $f(Y^*|Z,X^0)$ is assumed but parameters cannot be estimated from existing data.
- Step 1: Under the assumed model, sample is selected based on known values of Z, X^0. Sample design is optimized to estimate f(Y|Z,X).
- Step 2,3: In depth interview or clinical follow-up for the subsample ascertains Y or Y* and concurrent values of X, Z.
- From the survey data (training set), a "best" predictive model, f(Y|Z,X) is estimated
- The predictive model estimated from the survey is used to predict Y for each element in the existing frame (e.g. population reference, large baseline survey).
- Estimation and inference are based on model-predictions, properly reflecting the uncertainty associated with the modeled values of Y*.

Multi-phase Data Collections: Survey-assisted Modeling

- Predictive modeling approaches assign classification probabilities to all elements in the population frame (or weighted sample):

  Gertrude: "I hear the average American family now has 1.5 automobiles."

  Heathcliffe: "I bet that half a car is tough to drive."

  Red Skelton (ca. 1968)

- Decision is needed to analyze on probability scale or use probabilities to impute discrete classification.

- Inference should reflect prediction (imputation) uncertainty inherent in modeled values.

Estimation and inference: Predicted probability or discrete classification?

- Option 1: Use probability of dx classification directly in analysis

$$\hat{p}_i = \text{ predicted value drawn from } p(Y=1|X, Z_{obs}, \hat{\theta})$$

- Option 2: Impute discrete classification

$$\hat{Y}_i \in (0,1)$$

$$= \text{draw from } B(\hat{p}_i \mid Z, X, \hat{\theta})$$

ADAMS- Survey Assisted Modeling. Estimating the prevalence of dementia in the U.S. household population, age 70+ (2002)*

Statistic	ADAMS Direct Estimate	Predictive Modeling Method Using ADAMS to predict dementia for full HRS.				
		Logistic Regress w/MI	Lasso	Random Forest	Boosting	BART
$\hat{p}_{dementia}$	0.137	0.141	0.156	0.156	0.157	0.155
$se(\hat{p}_{dementia})$	0.013	0.004	0.004	0.004	0.004	0.004

* Covariate data base: HRS 2002. Predictive models fitted based on ADAMS sample data.

HRS: Logistic Regression Model for Overnight Stays in Hospital during the Past Two Years*

	2002	2004
Dementia*	1.30	1.32
	(1.11 - 1.53)	(1.09 - 1.60)
Age	1.03	1.03
	(1.02 - 1.04)	(1.02 -1.04)
White	1.22	1.08
	(1.01 - 1.46)	(0.94 - 1.25)
Female	0.92	0.87
	(0.82 - 1.04)	(0.79 – 0.96)
Odds Ratios, with 95% CI in parentheses		

*Dementia predictor is predicted value from ADAMS dementia logistic prediction model. Multiple imputation of predicted probabilities is used to reflect imputation uncertainty in the model predictions.