



Statistics Can Lie But Can Also  
Correct for Lies:  
Reducing Response Bias in  
NLAAS via Bayesian Imputation

Jingchen Liu

Columbia University



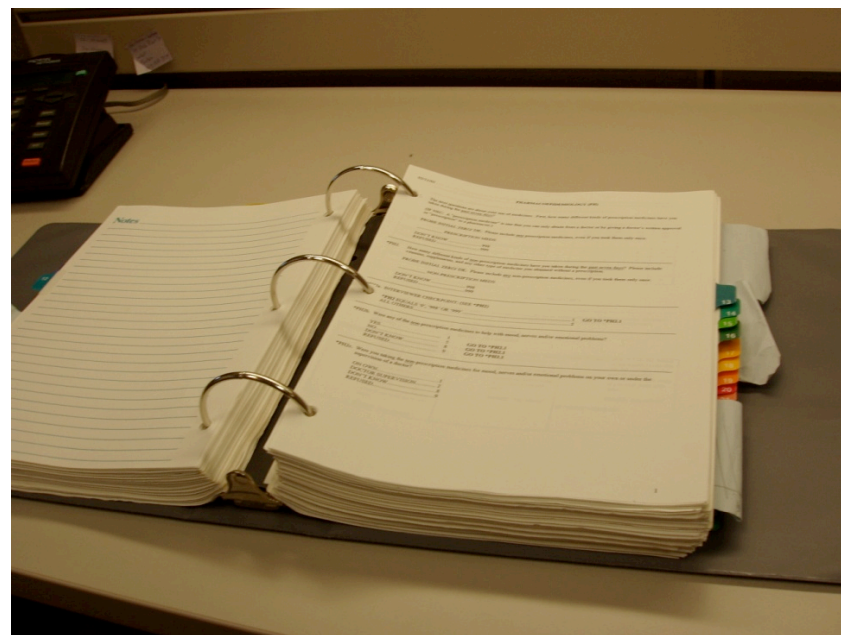
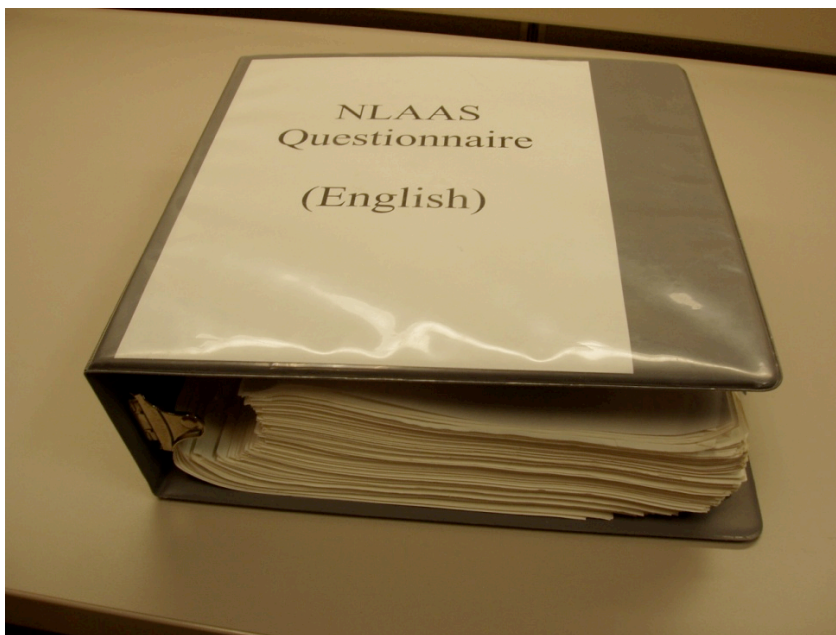
# Overview of NLAAS

## National Latino and Asian American Study (NLAAS)

- NLAAS, conducted in 2002-2003, and made public **July 2007**, is a national psychiatric epidemiologic study conducted to measure psychiatric disorders and mental health service usage in a nationally representative household sample of Asians and Latinos.
- There are more than **5000** variables (and the number is still growing!)
- Total sample size is **4864**: 2554 Latinos + 2095 Asians + 215 Whites

<http://www.multiculturalmentalhealth.org/nlaas.asp>

# A HUGE Questionnaire!






# Traditional Questionnaire

Duan et.al. (2007)

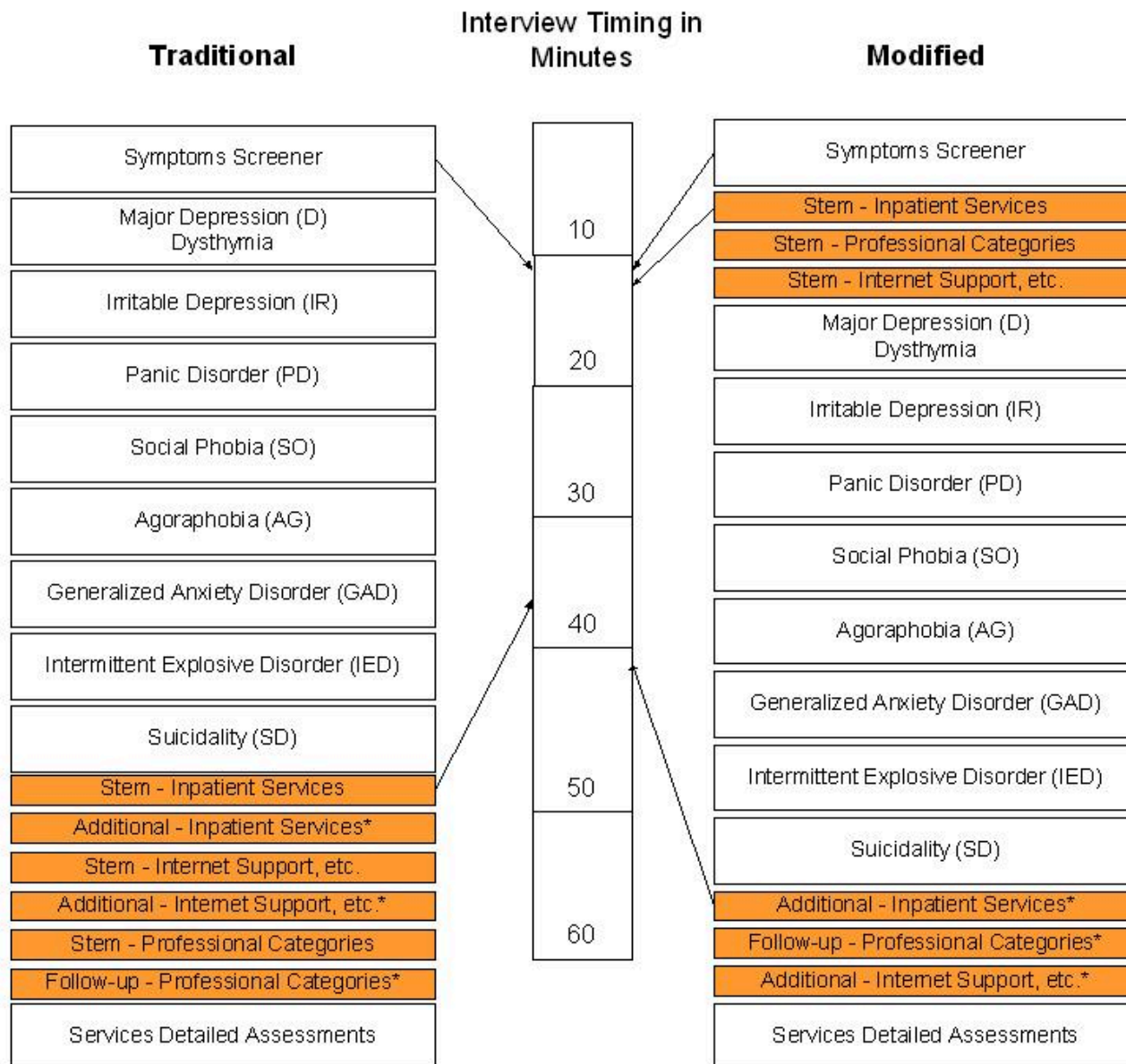
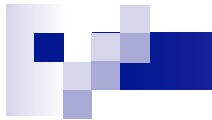
- At any time in your life did you ever see a psychiatrist about problems with your emotions, nerves or your mental health or your use of alcohol?  
(Yes/No)
  - How old were you at the time of this admission?
  - How much time did you stay in the hospital during this admission?
  - Was the treatment helpful?
  - ...
  
- Another stem question...
  - ...
  - ...
  - ...

**75% people have this design**

- 
- Which of the following types of professionals did you ever see about problems with your emotions or nerves or your use of alcohol or drugs?

|                          |    |   |    |
|--------------------------|----|---|----|
| <input type="checkbox"/> | A. | PSYCHIATRIST  | 1  |
| <input type="checkbox"/> | B. | GENERAL PRACTITIONER OR FAMILY DOCTOR                         | 2  |
| <input type="checkbox"/> | C. | ANY OTHER MEDICAL DOCTOR                                      | 3  |
| <input type="checkbox"/> | D. | PSYCHOLOGIST  | 4  |
| <input type="checkbox"/> | E. | SOCIAL WORKER   | 5  |
| <input type="checkbox"/> | F. | COUNSELOR   | 6  |
| <input type="checkbox"/> | G. | ANY OTHER MENTAL HEALTH PROFESSIONAL                          | 7  |
| <input type="checkbox"/> | H. | A NURSE, OCCUPATIONAL THERAPIST, OR OTHER HEALTH PROFESSIONAL | 8  |
| <input type="checkbox"/> | I. | A RELIGIOUS OR SPIRITUAL ADVISOR                              | 9  |
| <input type="checkbox"/> | J. | ANY OTHER HEALER  | 10 |
| <input type="checkbox"/> | K. | DON'T KNOW  | 11 |
| <input type="checkbox"/> | L. | REFUSED   | 12 |

**25% people have this design**



Duan et.al. (2007)



## Compare the Old and New Design – Service Use

|                                   | New Design | Old Design |
|-----------------------------------|------------|------------|
| 1. Psychiatrist                   | 14.9%      | 10.4%      |
| 2. General Practitioner           | 17.6%      | 13.1%      |
| 3. Other Medical Doctor           | 9.2%       | 3.8%       |
| 4. Psychologist                   | 13.4%      | 9.7%       |
| 5. Social worker                  | 7.6%       | 3.4%       |
| 6. Counselor                      | 13.2%      | 8.7%       |
| 7. Other Mental Health Prof       | 5.3%       | 3.2%       |
| 8. Nurse, Occupational Therapist, | 4.0%       | 2.0%       |
| 9. Religious/Spiritual Advisor    | 15.3%      | 5.9%       |
| 10. Other Healer                  | 5.9%       | 1.9%       |
| 11. Hot Line                      | 2.3%       | 1.2%       |
| 12. Internet Group or Chat Room   | 2.9%       | 1.1%       |
| 13. Self Help Service             | 5.9%       | 4.1%       |



## Compare the Old and New Design – Other Variables

|                                 | New Design | Old Design |
|---------------------------------|------------|------------|
| Major Depression                | 0.14       | 0.15       |
| Any Affective Disorder          | 0.14       | 0.15       |
| Any Disorder                    | 0.15       | 0.16       |
| Any Affective Disorder 12 month | 0.08       | 0.10       |
| Number of disorders             | 0.60       | 0.63       |
| k10 distress                    | 13.6       | 13.4       |
| Gender                          | 0.49       | 0.50       |
| Age                             | 38.9       | 40.7       |
| Social Status                   | 5.7        | 5.8        |
| Immigration Status              | 0.53       | 0.54       |

Duan et.al. (2007)



# Correcting/Reducing Response Biases via Multiple Imputation

- Predicting the service use of the old design group, had they received new design.
- Create multiply imputed public data sets.
- Negative responses in the traditional design group



# Grand Challenge

- Goal: Correcting/reducing the underreporting of service use in the old design by using the data from the new design
- A Grand Challenge: Imputation  $\neq$  Randomization

DESIRE: Imputed rates for old design group match the observed rates from new design group for any potential subpopulation of interest.

REALITY: Can only include a handful number of covariates due to identifiability, computational, and time constraints.



# Basic Assumptions

- New design:
  - Responded truthfully
  
- Traditional design
  - Respond truthfully, if the respondent reported positively
  - May not respond truthfully, if the respondent reported negatively



# Basic Model Setup

## ■ Notations

- I: Group indicator: 0 for new design, and 1 for old design.
- y: self-reported service use: 0 for no service, 1 for having service;
- $\xi_s$ : true service use: 0 for no service, 1 for having service;
- $\xi_l$ : lying behavior the old design: 0 for lying and 1 for telling the truth.

$$y = \begin{cases} \xi_s \cdot \xi_l, & I = 1 \\ \xi_s, & I = 0 \end{cases}$$

- Of interest is the distribution of  $\xi_s \mid y=0, I=1$ .



# Multivariate Probit Model

- We have 13 lifetime service use variables:  $(\xi_s^1, \dots, \xi_s^{13})^\top$
- Associated with them, there are 13 lying indicators:  $(\xi_l^1, \dots, \xi_l^{13})^\top$



# Continuation Ratio Model

- Let  $\xi_s^j = \eta^j \cdot \zeta^j$ .

$\eta^j$  and  $\zeta^j$  follows multivariate probit model, that is,

$$\eta^j = I(z_\eta^j > 0), \quad \zeta^j = I(z_\zeta^j > 0)$$

- $z_\eta = \{z_\eta^j\}$  and  $z_\zeta = \{z_\zeta^j\}$  are independent and

$$z_\eta \sim N(\beta_\eta, \Sigma_\eta),$$

$$z_\zeta \sim N(\beta_\zeta X + W, \Sigma_\zeta),$$

- Cox (1972, JRSSB) proportional hazard model  
Heagerty and Zeger (2000, Biometrics) multivariate continuation ratio model.



# Covariates

- Ideally, we should put in as many variables as possible.
- Nonidentifiability, computation, time constraint.
- Criteria: important for prediction, potential analysis
- Selections are “negotiated” with psychologists



# Covariates

- **Categorical variables:** marital status, insurance status, working status, region in the country, ethnicity, immigration status, gender, psychiatric disorder diagnostics.
- **Continuous variables:** logarithm of annual income, total number of psychiatric disorders, social status, age, k10 distress (psychiatric disorder related variable).



# Prior Distribution Specification

- $\beta = (\beta^>_1, \dots, \beta^>_{26})^>, \beta_i \sim \text{Normal}(\mu_0, \Sigma_0)$ .

- It is desirable to choose  $\mu_0$  and  $\Sigma_0$  such that

$$\beta_i x \sim N(0, I) \Leftrightarrow P(\xi_s = 1) \sim U(0, 1)$$

$\mu_0 = 0$  and  $\Sigma_0 = n(X^> X)^{-1}/p$  as an approximation.

- $\Sigma_{11}$  » **correlation matrix** obtained from Inverted-Wishart
- $\alpha$  »  $N(0, I)$



# Simulation of Posterior Distribution

Markov chain Monte Carlo: Gibbs sampling, Gelman et al. (2003)

$P(z, \beta, \Sigma_{11}, \rho \mid y)$

- $P(z \mid \beta, \Sigma_{11}, \rho, y)$  is a truncated multivariate normal.
- $P(\beta \mid z, \Sigma_{11}, \rho)$ , multivariate normal.
- $P(\Sigma_{11} \mid \beta, z)$ , Corr-Inv-Wish( $S, d$ ), where  $S = (z - X\beta)^{\top}(z - X\beta)$ .
- Draw  $\rho$ , using a data augmentation via

$$Z = \rho N(0, 1) + \sqrt{1 - \rho^2} N(0, 1).$$

- Convergence monitor: Gelman and Rubin R statistic

# Lifetime Service Use Results

|                                 | Puerto Rican |      |      | Cuban |      |      | Mexican |      |      | Other Latino |      |      |
|---------------------------------|--------------|------|------|-------|------|------|---------|------|------|--------------|------|------|
|                                 | New          | Imp  | Old  | New   | Imp  | Old  | New     | Imp  | Old  | New          | Imp  | Old  |
| <b>Specialist (1,4,7,11)</b>    | 34.2         | 32.3 | 25.5 | 26.7  | 22.8 | 18.0 | 14.9    | 14.8 | 10.2 | 20.8         | 19.4 | 13.9 |
| 1. Psychiatrist                 | 28.8         | 22.9 | 16.0 | 19.7  | 16.6 | 12.9 | 11.0    | 9.4  | 6.5  | 9.4          | 11.0 | 8.0  |
| 4. Psychologist                 | 20.6         | 18.6 | 14.0 | 16.2  | 13.3 | 9.7  | 10.4    | 8.4  | 5.6  | 13.5         | 12.7 | 8.7  |
| 7. Other M. H. Prof.            | 10.7         | 7.1  | 5.5  | 4.7   | 3.6  | 2.3  | 5.0     | 4.0  | 2.7  | 4.5          | 4.6  | 3.0  |
| 11. Hot Line                    | 1.7          | 2.3  | 1.5  | 2.3   | 0.9  | 0.2  | 0.9     | 1.9  | 1.3  | 3.0          | 2.3  | 1.3  |
| <b>Generalist (2,3,8)</b>       | 32.5         | 30.0 | 25.4 | 21.4  | 23.9 | 19.6 | 18.3    | 16.8 | 12.3 | 16.8         | 15.4 | 10.3 |
| 2. General Practitioner         | 28.5         | 27.1 | 23.9 | 18.5  | 19.7 | 16.7 | 16.6    | 14.1 | 10.3 | 12.7         | 11.3 | 8.4  |
| 3. Other Med. Doctors           | 15.5         | 12.7 | 7.5  | 10.3  | 9.7  | 4.9  | 7.0     | 5.1  | 1.9  | 10.2         | 8.4  | 4.3  |
| 8. Other Professionals          | 13.4         | 5.6  | 3.5  | 4.0   | 4.2  | 3.2  | 3.0     | 2.7  | 1.7  | 3.3          | 2.4  | 1.2  |
| <b>Human Services (5,6,9)</b>   | 38.5         | 30.5 | 22.7 | 17.4  | 13.6 | 8.1  | 20.1    | 16.6 | 10.2 | 24.9         | 21.1 | 13.3 |
| 5. Social Worker                | 17.9         | 14.0 | 10.1 | 5.1   | 3.3  | 2.6  | 7.4     | 4.9  | 2.6  | 5.5          | 6.0  | 3.8  |
| 6. Counselor                    | 29.6         | 17.9 | 12.9 | 11.0  | 5.6  | 3.5  | 11.2    | 9.2  | 7.1  | 13.7         | 12.6 | 9.5  |
| 9. Religious Advisor            | 16.3         | 17.2 | 10.4 | 11.8  | 10.6 | 5.2  | 14.0    | 10.7 | 5.2  | 16.5         | 12.3 | 5.0  |
| <b>Alt. Services (10,12,13)</b> | 24.9         | 15.3 | 8.9  | 10.2  | 9.0  | 5.3  | 9.1     | 7.8  | 3.7  | 9.4          | 12.8 | 6.1  |
| 10. Other Services              | 13.0         | 8.4  | 1.6  | 6.6   | 5.7  | 2.7  | 4.0     | 3.2  | 1.1  | 4.0          | 6.8  | 2.3  |
| 12. Internet                    | 1.8          | 4.4  | 2.5  | 3.2   | 2.1  | 0.6  | 1.4     | 1.6  | 0.1  | 3.7          | 3.3  | 0.6  |
| 13. Self Service                | 13.0         | 7.4  | 6.3  | 2.3   | 3.7  | 3.2  | 5.4     | 4.6  | 3.2  | 4.8          | 5.8  | 4.5  |
| <b>Formal Services (1-8)</b>    | 46.3         | 47.2 | 40.7 | 33.9  | 35.5 | 29.9 | 25.0    | 26.9 | 20.8 | 29.5         | 31.6 | 25.0 |
| <b>Any Services (1-10)</b>      | 50.0         | 50.4 | 42.4 | 37.6  | 37.6 | 30.6 | 30.0    | 30.1 | 21.9 | 35.3         | 35.9 | 26.2 |
| <b>Any Services (1-13)</b>      | 50.9         | 51.4 | 42.9 | 38.4  | 38.3 | 31.0 | 32.5    | 31.5 | 22.7 | 36.5         | 37.4 | 26.8 |



# Model Checking

- From the imputer's view: “data mining” tools for discovering problematic subpopulations (almost the same as discovering “bad genes”) over the 5000 variables.
- Sources of mismatching between the two groups
  - Imputation model
  - Randomization has large variation in small strata
  - Finite imputation: 10 samples from the posterior distribution
- What is the right criteria?



# Model Checking

- From the potential analysts' view: if the data set is suitable for their particular analyses
- In general, very hard problem  
Incomplete or no information about the imputation model
- For our particular dataset, there is one assumption one can check
- Diagnostic tools – graphical analysis



# Summary

- Underreporting to service use questions in NLAAS
- Correct the bias in the old design according to data collected by the new design
- Multivariate probit model – continuation ratio (CR) model
- Challenges:
  - Potential subpopulations
  - Higher order interactions
  - Criteria to evaluate the imputation quality



# Continuation Ratio Model

- Two extreme cases

- $\beta_\eta \rightarrow \infty$ , then  $\eta \rightarrow 1$  and  $\xi_s = \zeta$

$$z_\zeta \sim N(\beta_\zeta X + W, \Sigma_\zeta), \quad z_\eta \sim N(\beta_\eta, \Sigma_\eta).$$

- No restriction on  $\beta_\eta$  – non-identifiable.

- Tradeoff between identifiability and flexibility

- $\Sigma_\eta$  contributes more flexibility

- $\beta_\eta \sim N(2, 0.01)$