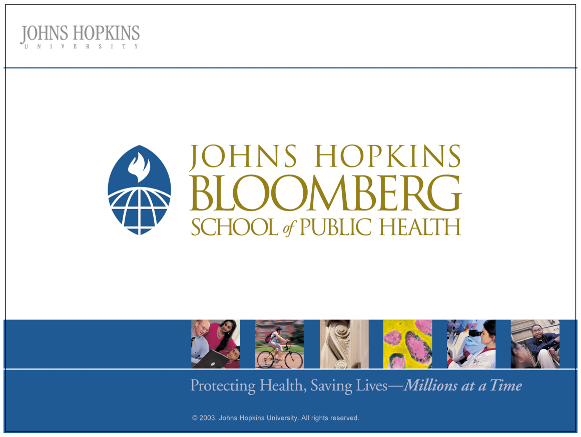
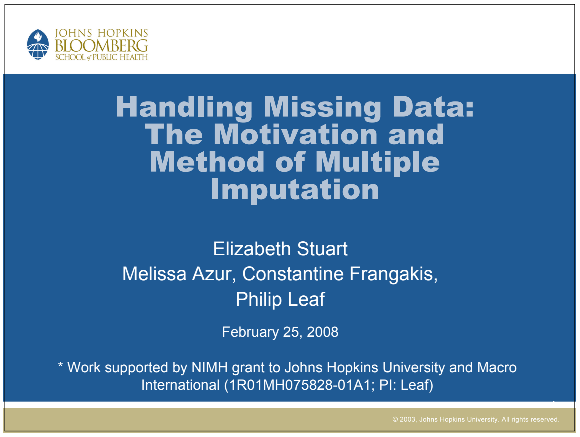


JOHNS HOPKINS UNIVERSITY

JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

Protecting Health, Saving Lives—*Millions at a Time*



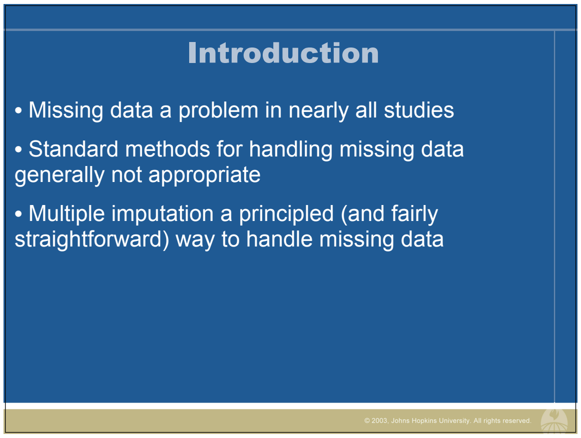
JOHNS HOPKINS BLOOMBERG SCHOOL *of* PUBLIC HEALTH

Handling Missing Data: The Motivation and Method of Multiple Imputation

Elizabeth Stuart
Melissa Azur, Constantine Frangakis,
Philip Leaf

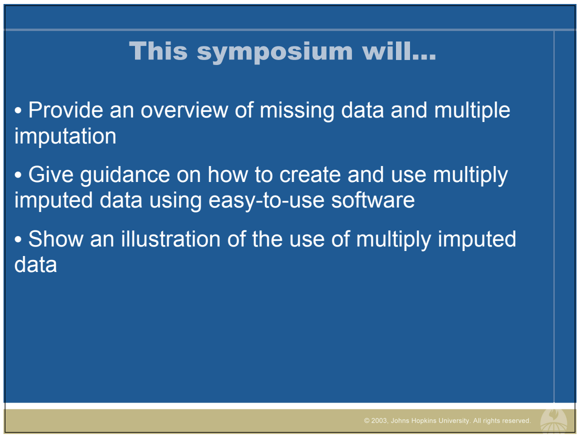
February 25, 2008

* Work supported by NIMH grant to Johns Hopkins University and Macro International (1R01MH075828-01A1; PI: Leaf)



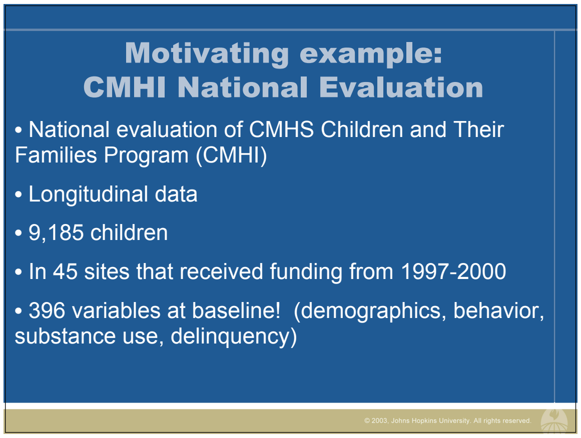
Introduction

- Missing data a problem in nearly all studies
- Standard methods for handling missing data generally not appropriate
- Multiple imputation a principled (and fairly straightforward) way to handle missing data



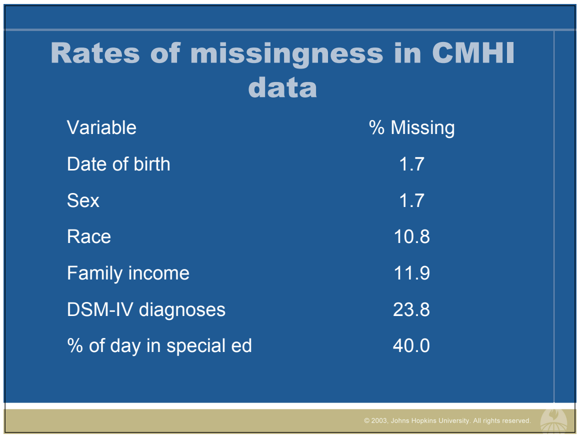
This symposium will…

- Provide an overview of missing data and multiple imputation
- Give guidance on how to create and use multiply imputed data using easy-to-use software
- Show an illustration of the use of multiply imputed data



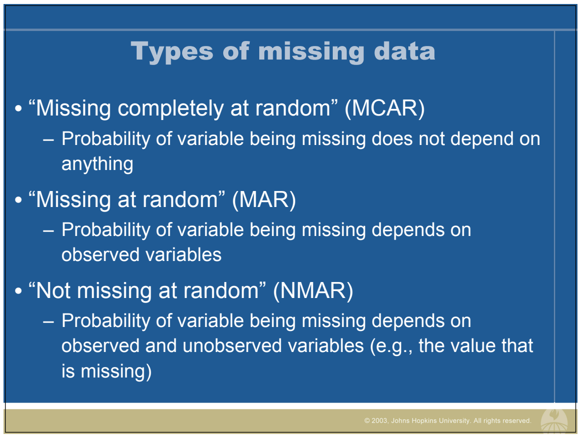
Motivating example: CMHI National Evaluation

- National evaluation of CMHS Children and Their Families Program (CMHI)
- Longitudinal data
- 9,185 children
- In 45 sites that received funding from 1997-2000
- 396 variables at baseline! (demographics, behavior, substance use, delinquency)



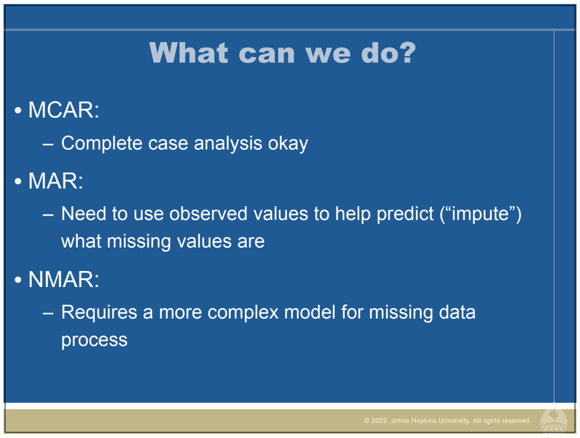
Rates of missingness in CMHI data

Variable	% Missing
Date of birth	1.7
Sex	1.7
Race	10.8
Family income	11.9
DSM-IV diagnoses	23.8
% of day in special ed	40.0



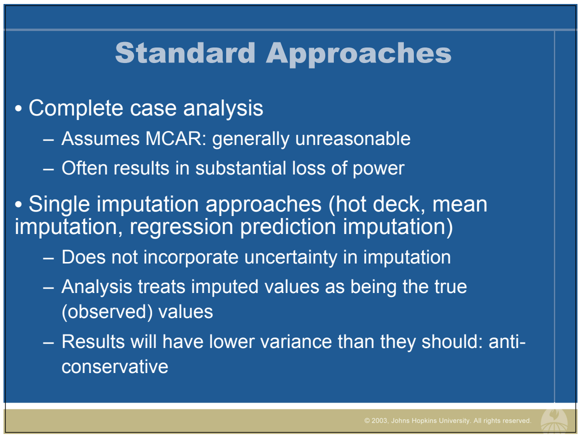
Types of missing data

- "Missing completely at random" (MCAR)
  - Probability of variable being missing does not depend on anything
- "Missing at random" (MAR)
  - Probability of variable being missing depends on observed variables
- "Not missing at random" (NMAR)
  - Probability of variable being missing depends on observed and unobserved variables (e.g., the value that is missing)



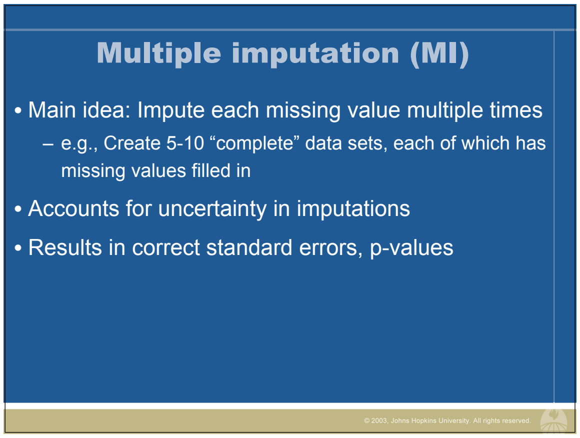
What can we do?

- MCAR:
  - Complete case analysis okay
- MAR:
  - Need to use observed values to help predict ("impute") what missing values are
- NMAR:
  - Requires a more complex model for missing data process



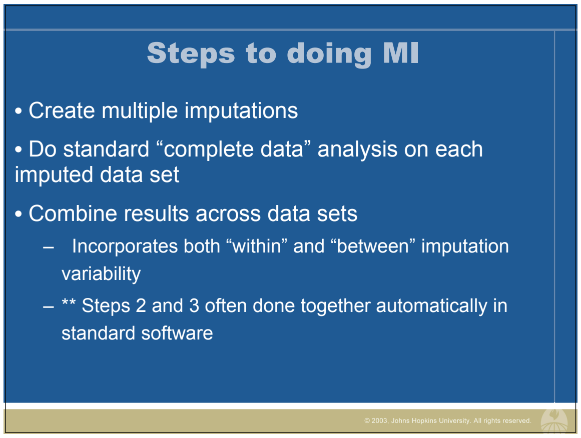
Standard Approaches

- Complete case analysis
  - Assumes MCAR: generally unreasonable
  - Often results in substantial loss of power
- Single imputation approaches (hot deck, mean imputation, regression prediction imputation)
  - Does not incorporate uncertainty in imputation
  - Analysis treats imputed values as being the true (observed) values
  - Results will have lower variance than they should: anti-conservative



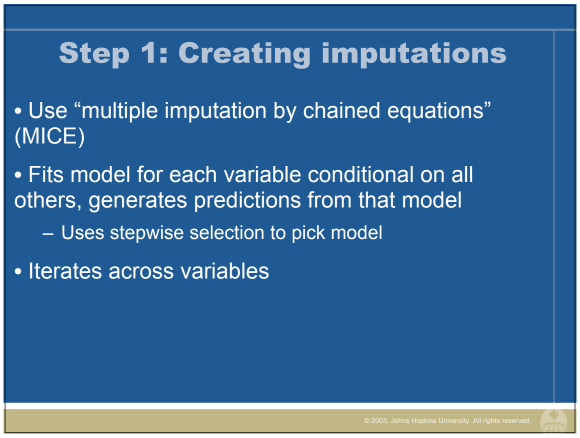
Multiple imputation (MI)

- Main idea: Impute each missing value multiple times
  - e.g., Create 5-10 "complete" data sets, each of which has missing values filled in
- Accounts for uncertainty in imputations
- Results in correct standard errors, p-values



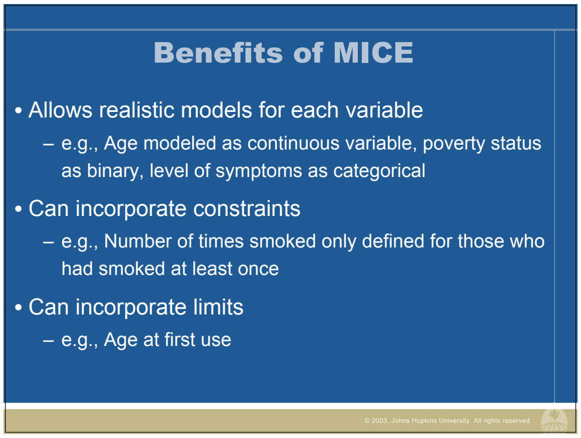
Steps to doing MI

- Create multiple imputations
- Do standard "complete data" analysis on each imputed data set
- Combine results across data sets
  - Incorporates both "within" and "between" imputation variability
  - ** Steps 2 and 3 often done together automatically in standard software



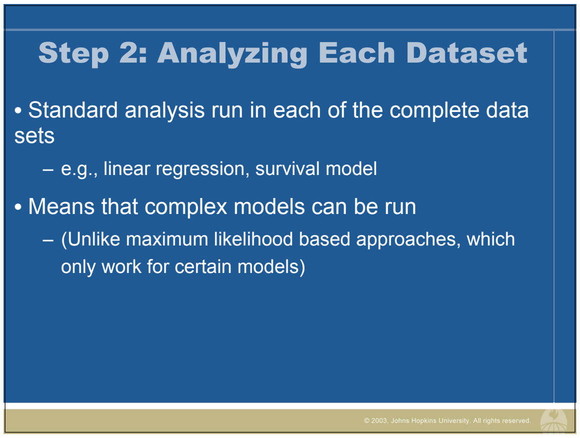
Step 1: Creating imputations

- Use "multiple imputation by chained equations" (MICE)
- Fits model for each variable conditional on all others, generates predictions from that model
  - Uses stepwise selection to pick model
- Iterates across variables



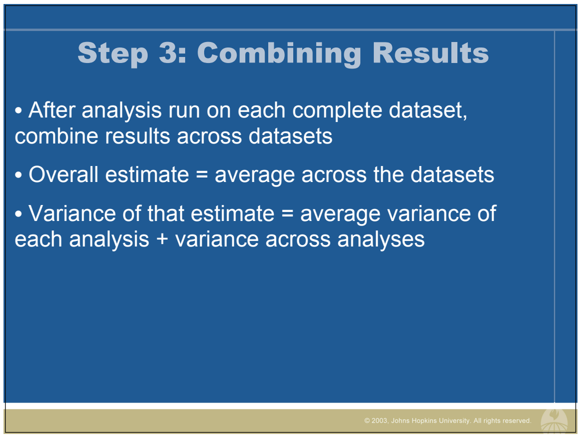
Benefits of MICE

- Allows realistic models for each variable
  - e.g., Age modeled as continuous variable, poverty status as binary, level of symptoms as categorical
- Can incorporate constraints
  - e.g., Number of times smoked only defined for those who had smoked at least once
- Can incorporate limits
  - e.g., Age at first use



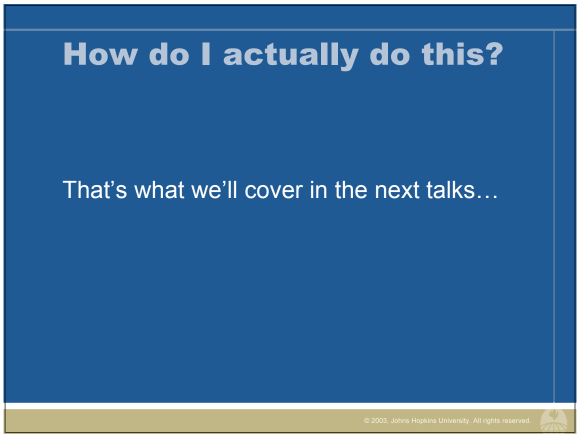
Step 2: Analyzing Each Dataset

- Standard analysis run in each of the complete data sets
  - e.g., linear regression, survival model
- Means that complex models can be run
  - (Unlike maximum likelihood based approaches, which only work for certain models)



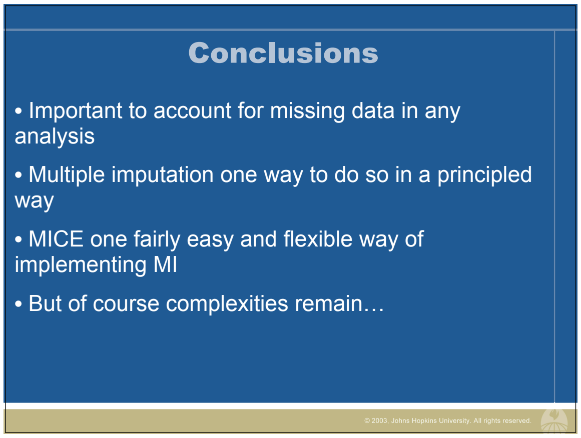
Step 3: Combining Results

- After analysis run on each complete dataset, combine results across datasets
- Overall estimate = average across the datasets
- Variance of that estimate = average variance of each analysis + variance across analyses



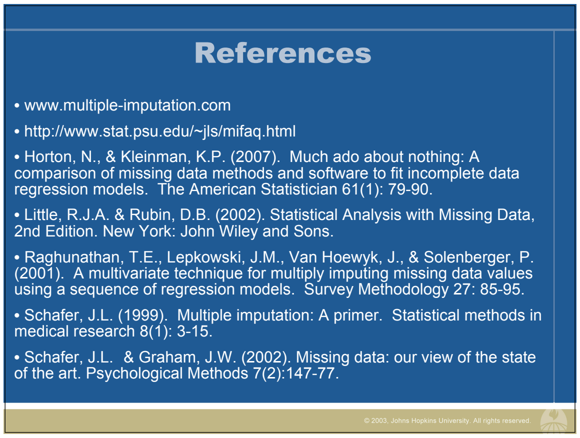
How do I actually do this?

That’s what we’ll cover in the next talks…



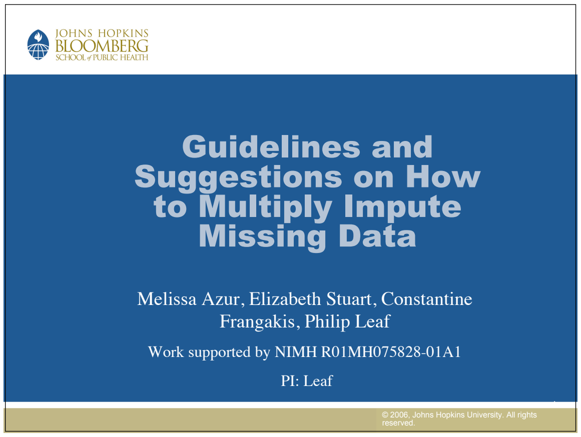
Conclusions

- Important to account for missing data in any analysis
- Multiple imputation one way to do so in a principled way
- MICE one fairly easy and flexible way of implementing MI
- But of course complexities remain…



References

- www.multiple-imputation.com
- http://www.stat.psu.edu/~jls/mifaq.html
- Horton, N., & Kleinman, K.P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician 61(1): 79-90.
- Little, R.J.A. & Rubin, D.B. (2002). Statistical Analysis with Missing Data, 2nd Edition. New York: John Wiley and Sons.
- Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing data values using a sequence of regression models. Survey Methodology 27: 85-95.
- Schafer, J.L. (1999). Multiple imputation: A primer. Statistical methods in medical research 8(1): 3-15.
- Schafer, J.L. & Graham, J.W. (2002). Missing data: our view of the state of the art. Psychological Methods 7(2):147-77.



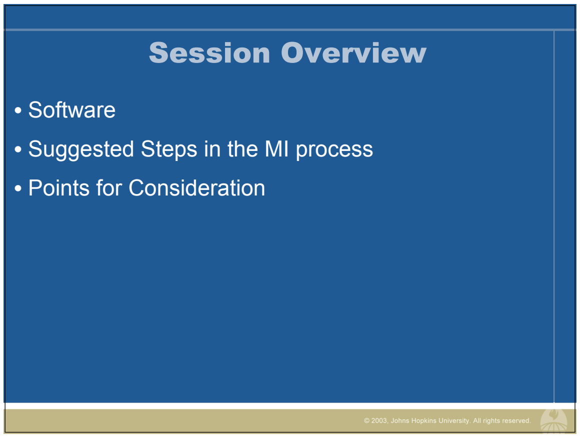
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Guidelines and Suggestions on How to Multiply Impute Missing Data

Melissa Azur, Elizabeth Stuart, Constantine Frangakis, Philip Leaf

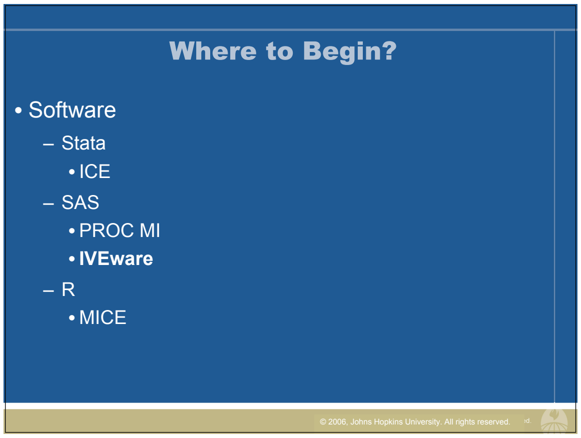
Work supported by NIMH R01MH075828-01A1

PI: Leaf



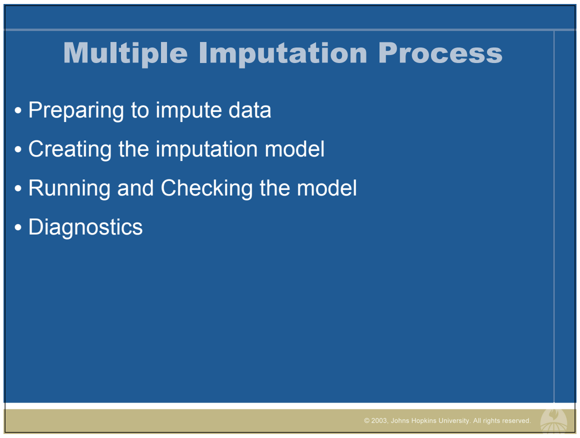
Session Overview

- Software
- Suggested Steps in the MI process
- Points for Consideration



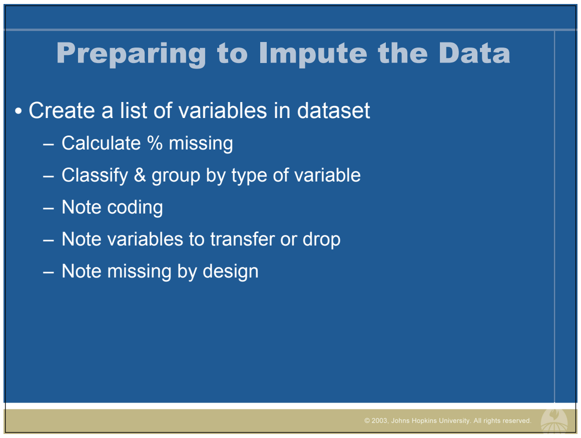
Where to Begin?

- Software
  - Stata
    - ICE
  - SAS
    - PROC MI
    - **IVEware**
  - R
    - MICE



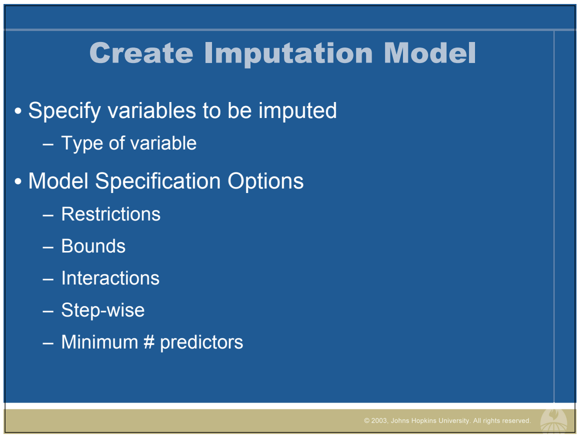
Multiple Imputation Process

- Preparing to impute data
- Creating the imputation model
- Running and Checking the model
- Diagnostics



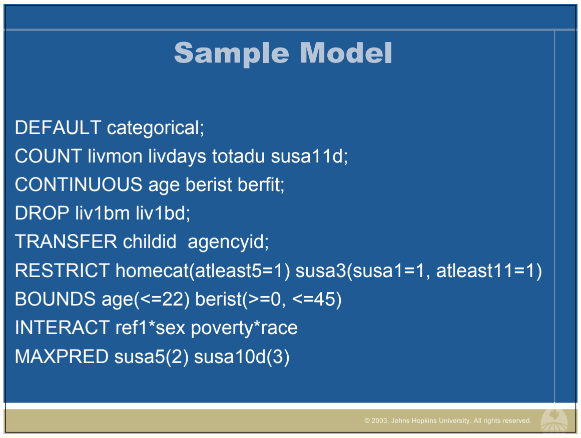
Preparing to Impute the Data

- Create a list of variables in dataset
  - Calculate % missing
  - Classify & group by type of variable
  - Note coding
  - Note variables to transfer or drop
  - Note missing by design



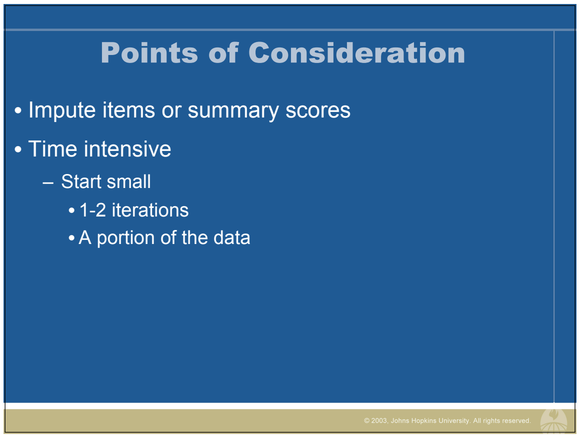
Create Imputation Model

- Specify variables to be imputed
  - Type of variable
- Model Specification Options
  - Restrictions
  - Bounds
  - Interactions
  - Step-wise
  - Minimum # predictors



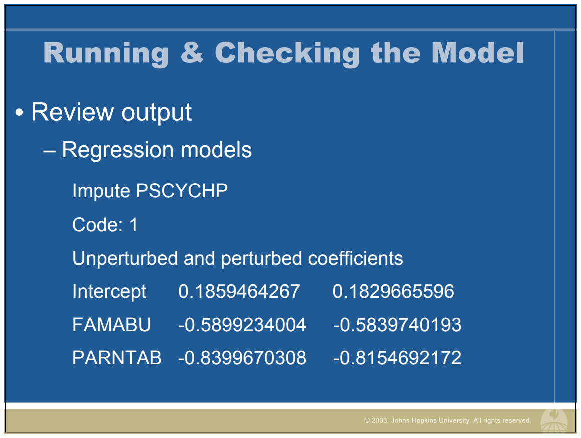
Sample Model

```
DEFAULT categorical;
COUNT livmon livdays totadu susa11d;
CONTINUOUS age berist berfit;
DROP liv1bm liv1bd;
TRANSFER childid  agencyid;
RESTRICT homecat(atleast5=1) susa3(susa1=1, atleast11=1)
BOUNDS age(<=22) berist(>=0, <=45)
INTERACT ref1*sex poverty*race
MAXPRED susa5(2) susa10d(3)
```



Points of Consideration

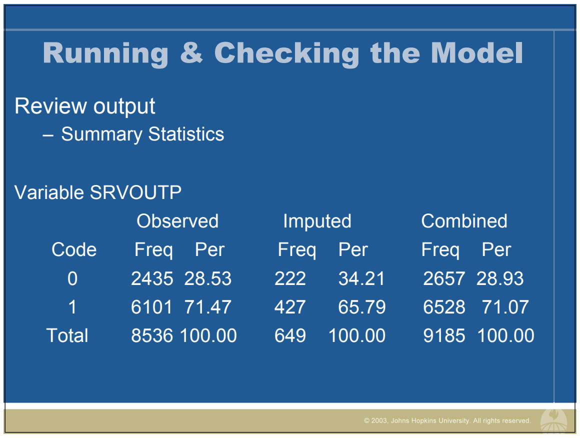
- Impute items or summary scores
- Time intensive
  - Start small
    - 1-2 iterations
    - A portion of the data



Running & Checking the Model

- Review output
  - Regression models

```
Impute PSCYCHP
Code: 1
Unperturbed and perturbed coefficients
Intercept     0.1859464267    0.1829665596
FAMABU       -0.5899234004   -0.5839740193
PARNTAB      -0.8399670308   -0.8154692172
```

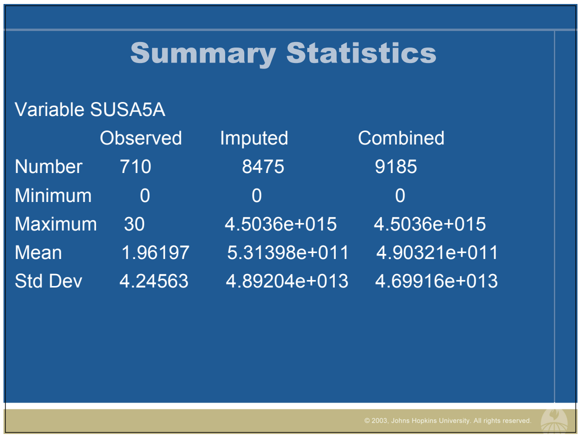


Running & Checking the Model

Review output

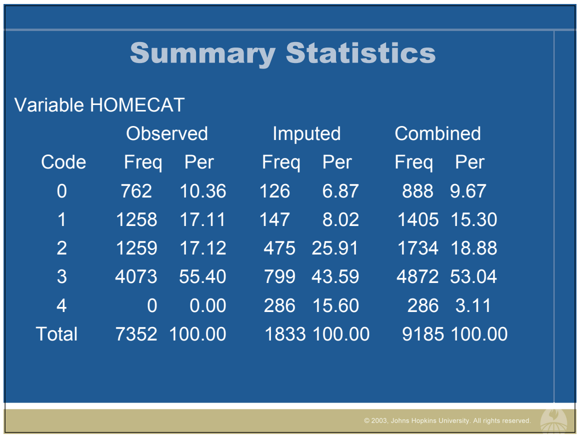
- Summary Statistics

```
Variable SRVOUTP
             Observed        Imputed         Combined
   Code     Freq   Per     Freq   Per      Freq   Per
    0       2435  28.53    222    34.21    2657  28.93
    1       6101  71.47    427    65.79    6528  71.07
   Total    8536 100.00    649   100.00    9185  100.00
```



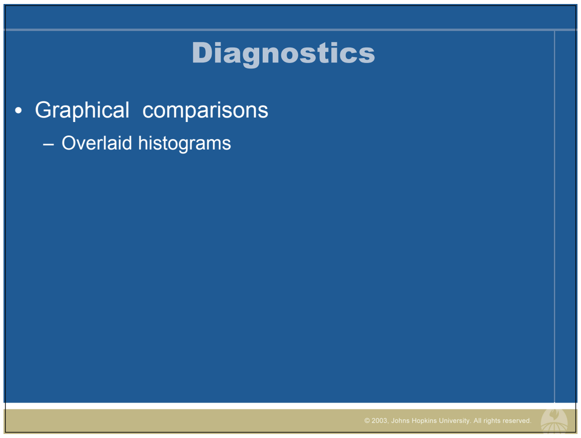
Summary Statistics

```
Variable SUSA5A
            Observed     Imputed         Combined
Number      710          8475            9185
Minimum       0            0               0
Maximum     30           4.5036e+015     4.5036e+015
Mean        1.96197      5.31398e+011    4.90321e+011
Std Dev     4.24563      4.89204e+013    4.69916e+013
```



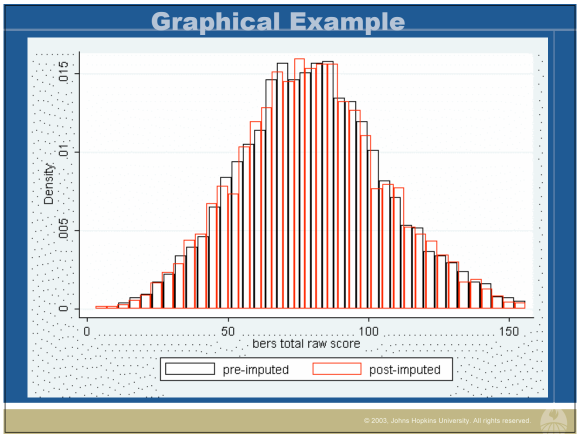
Summary Statistics

```
Variable HOMECAT
            Observed         Imputed          Combined
  Code     Freq    Per      Freq    Per      Freq   Per
   0        762    10.36    126     6.87     888   9.67
   1       1258    17.11    147     8.02     1405  15.30
   2       1259    17.12    475    25.91     1734  18.88
   3       4073    55.40    799    43.59     4872  53.04
   4          0     0.00    286    15.60      286   3.11
  Total    7352  100.00    1833  100.00     9185 100.00
```

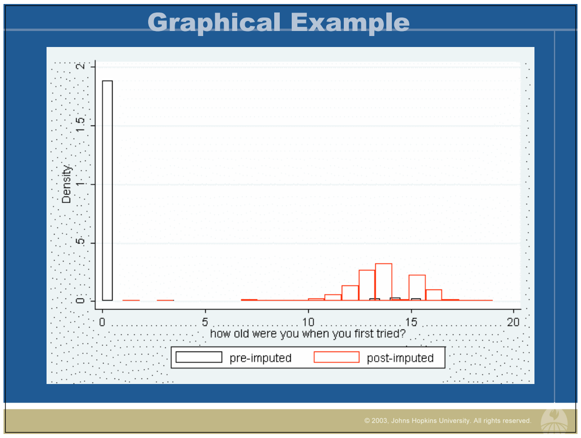


Diagnostics

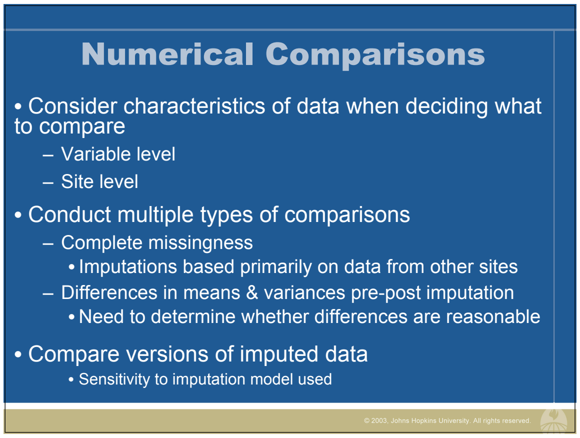
- Graphical comparisons
  - Overlaid histograms



Graphical Example

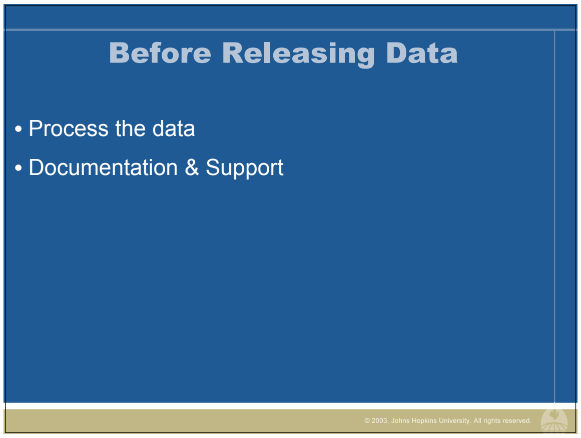


Graphical Example



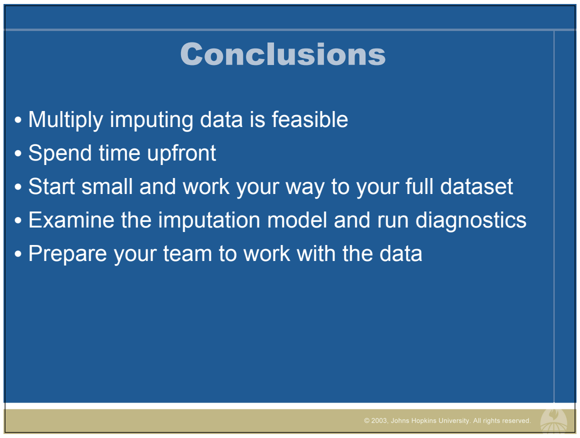
Numerical Comparisons

- Consider characteristics of data when deciding what to compare
  - Variable level
  - Site level
- Conduct multiple types of comparisons
  - Complete missingness
    - Imputations based primarily on data from other sites
  - Differences in means & variances pre-post imputation
    - Need to determine whether differences are reasonable
- Compare versions of imputed data
    - Sensitivity to imputation model used



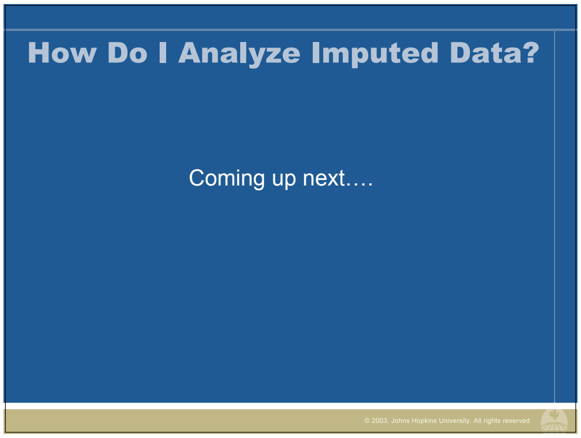
Before Releasing Data

- Process the data
- Documentation & Support



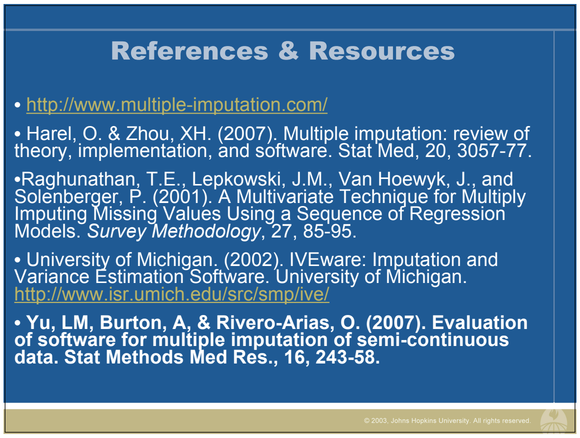
Conclusions

- Multiply imputing data is feasible
- Spend time upfront
- Start small and work your way to your full dataset
- Examine the imputation model and run diagnostics
- Prepare your team to work with the data



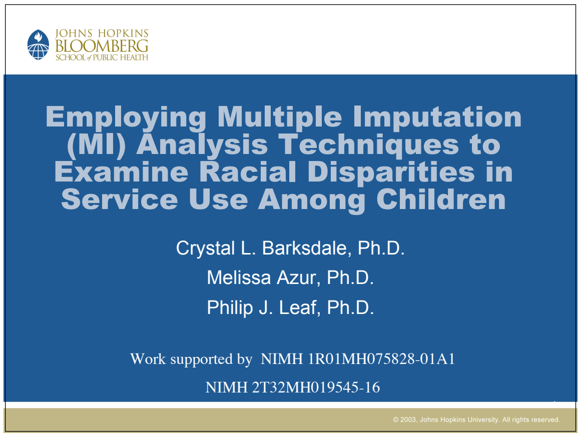
How Do I Analyze Imputed Data?

Coming up next….



References & Resources

- http://www.multiple-imputation.com/
- Harel, O. & Zhou, XH. (2007). Multiple imputation: review of theory, implementation, and software. Stat Med, 20, 3057-77.
- Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, 27, 85-95.
- University of Michigan. (2002). IVEware: Imputation and Variance Estimation Software. University of Michigan. http://www.isr.umich.edu/src/smp/ive/
- **Yu, LM, Burton, A, & Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. Stat Methods Med Res., 16, 243-58.**

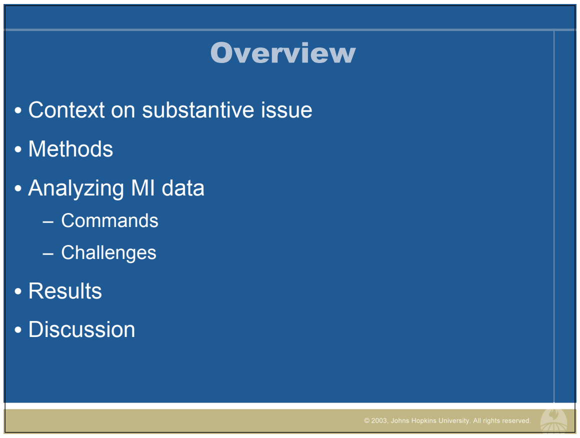


JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Employing Multiple Imputation (MI) Analysis Techniques to Examine Racial Disparities in Service Use Among Children

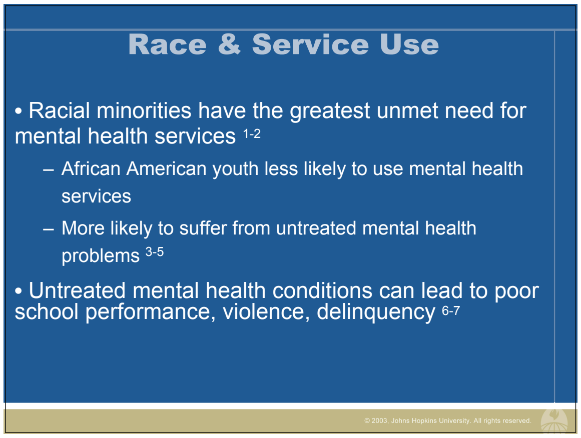
Crystal L. Barksdale, Ph.D.
Melissa Azur, Ph.D.
Philip J. Leaf, Ph.D.

Work supported by NIMH 1R01MH075828-01A1
NIMH 2T32MH019545-16



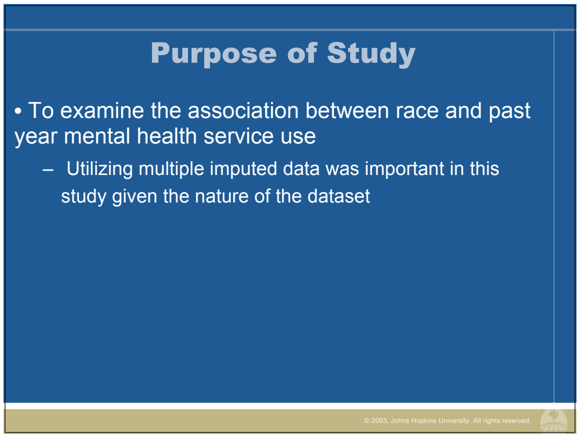
Overview

- Context on substantive issue
- Methods
- Analyzing MI data
  - Commands
  - Challenges
- Results
- Discussion



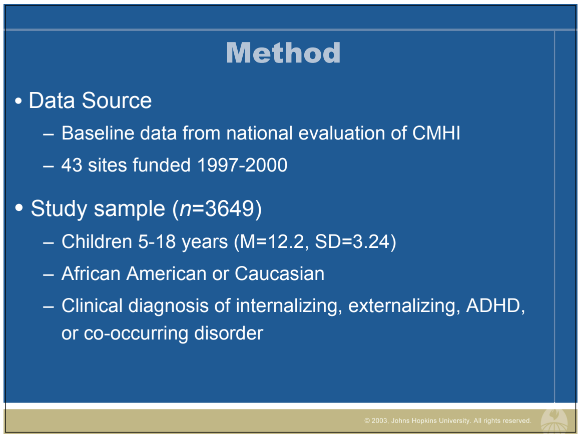
Race & Service Use

- Racial minorities have the greatest unmet need for mental health services [1-2]
  - African American youth less likely to use mental health services
  - More likely to suffer from untreated mental health problems [3-5]
- Untreated mental health conditions can lead to poor school performance, violence, delinquency [6-7]



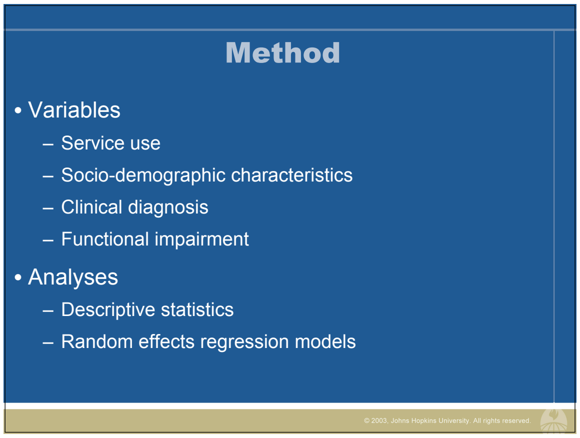
Purpose of Study

- To examine the association between race and past year mental health service use
  - Utilizing multiple imputed data was important in this study given the nature of the dataset



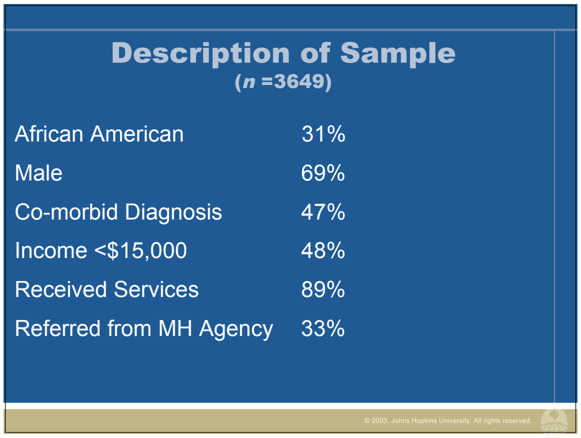
Method

- Data Source
  - Baseline data from national evaluation of CMHI
  - 43 sites funded 1997-2000
- Study sample (*n*=3649)
  - Children 5-18 years (M=12.2, SD=3.24)
  - African American or Caucasian
  - Clinical diagnosis of internalizing, externalizing, ADHD, or co-occurring disorder



Method

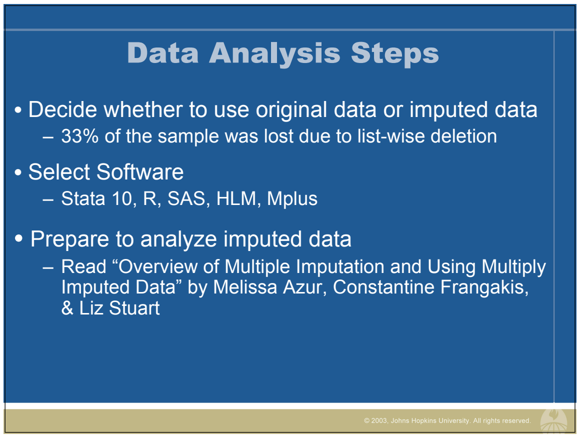
- Variables
  - Service use
  - Socio-demographic characteristics
  - Clinical diagnosis
  - Functional impairment
- Analyses
  - Descriptive statistics
  - Random effects regression models



Description of Sample

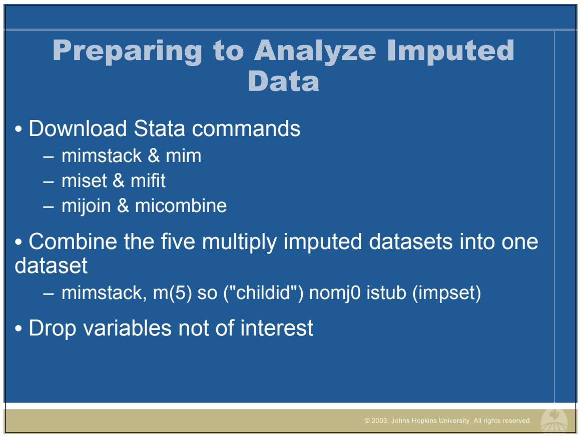
(*n* =3649)

African American	31%
Male	69%
Co-morbid Diagnosis	47%
Income <$15,000	48%
Received Services	89%
Referred from MH Agency	33%



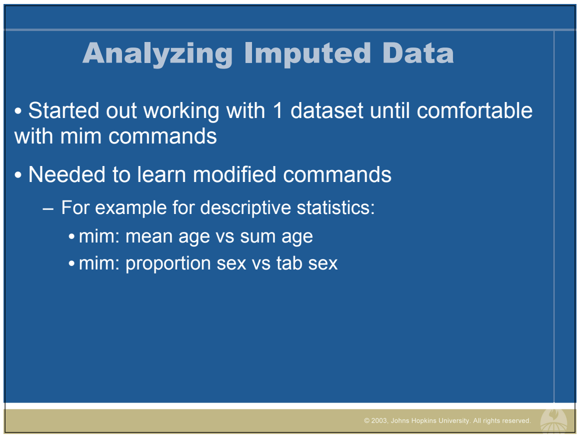
Data Analysis Steps

- Decide whether to use original data or imputed data
  - 33% of the sample was lost due to list-wise deletion
- Select Software
  - Stata 10, R, SAS, HLM, Mplus
- Prepare to analyze imputed data
  - Read “Overview of Multiple Imputation and Using Multiply Imputed Data” by Melissa Azur, Constantine Frangakis, & Liz Stuart



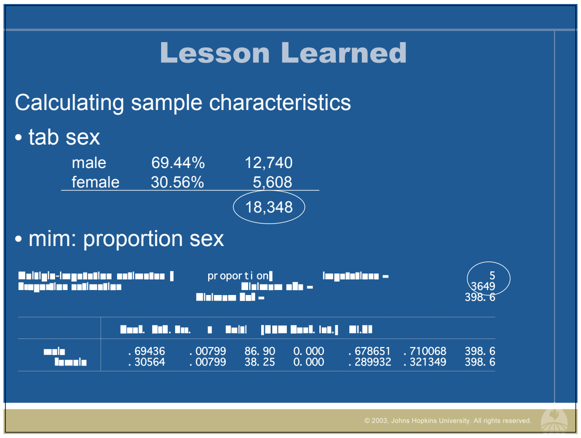
Preparing to Analyze Imputed Data

- Download Stata commands
  - mimstack & mim
  - miset & mifit
  - mijoin & micombine
- Combine the five multiply imputed datasets into one dataset
  - mimstack, m(5) so ("childid") nomj0 istub (impset)
- Drop variables not of interest



Analyzing Imputed Data

- Started out working with 1 dataset until comfortable with mim commands
- Needed to learn modified commands
  - For example for descriptive statistics:
    - mim: mean age vs sum age
    - mim: proportion sex vs tab sex



Lesson Learned

Calculating sample characteristics

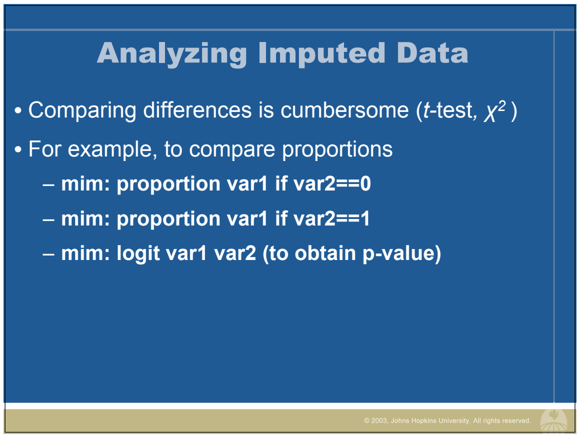
- tab sex

male	69.44%	12,740
female	30.56%	5,608
		18,348

- mim: proportion sex

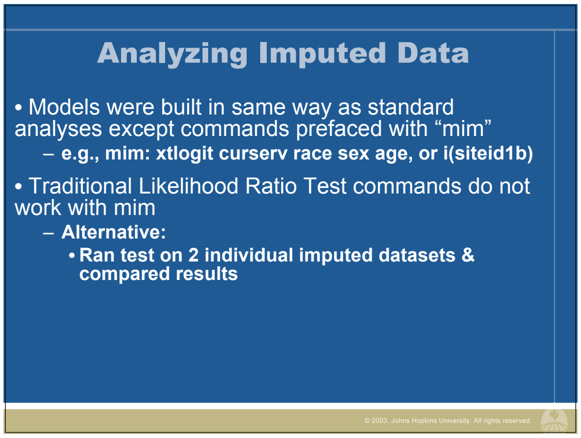
```
[illegible] |       proportion|        [illegible] =          5
[illegible]                   [illegible] =                  3649
                     [illegible] =                           398.6

------------------------------------------------------------------
          |  [illegible]
------------------------------------------------------------------
 [illegible] |  .69436   .00799   86.90   0.000   .678651   .710068   398.6
 [illegible] |  .30564   .00799   38.25   0.000   .289932   .321349   398.6
------------------------------------------------------------------
```



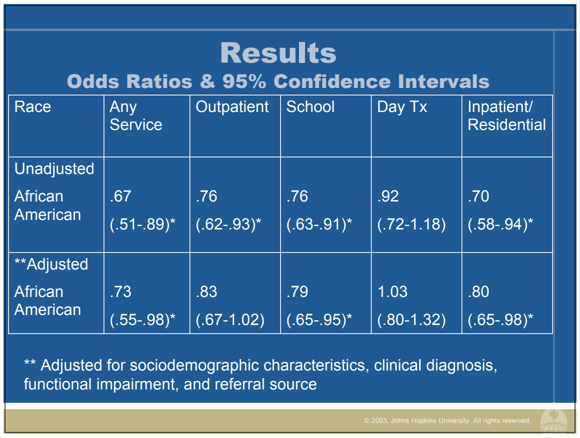
Analyzing Imputed Data

- Comparing differences is cumbersome (*t*-test, χ^2)
- For example, to compare proportions
  - **mim: proportion var1 if var2==0**
  - **mim: proportion var1 if var2==1**
  - **mim: logit var1 var2 (to obtain p-value)**



Analyzing Imputed Data

- Models were built in same way as standard analyses except commands prefaced with “mim”
  - **e.g., mim: xtlogit curserv race sex age, or i(siteid1b)**
- Traditional Likelihood Ratio Test commands do not work with mim
  - **Alternative:**
    - **Ran test on 2 individual imputed datasets & compared results**

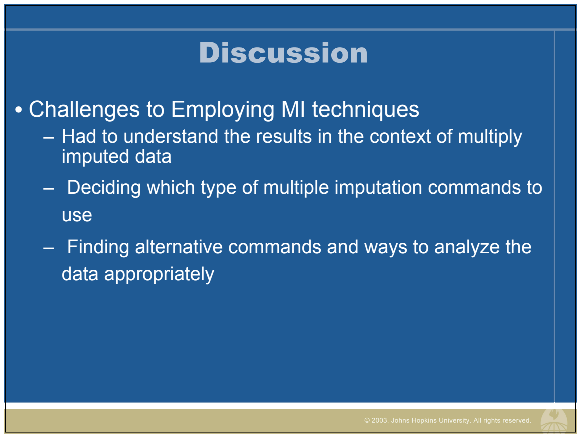


Results

Odds Ratios & 95% Confidence Intervals

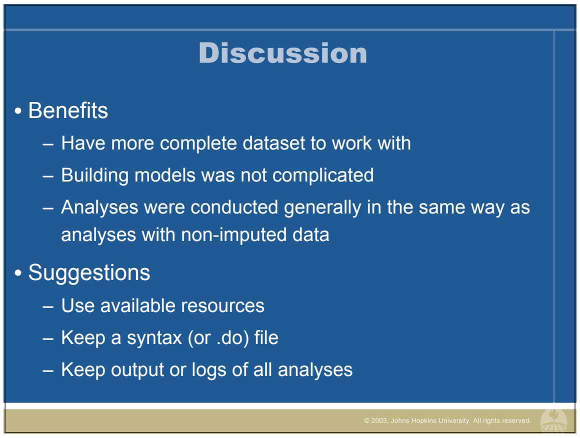
Race	Any Service	Outpatient	School	Day Tx	Inpatient/ Residential
Unadjusted African American	.67 (.51-.89)*	.76 (.62-.93)*	.76 (.63-.91)*	.92 (.72-1.18)	.70 (.58-.94)*
**Adjusted African American	.73 (.55-.98)*	.83 (.67-1.02)	.79 (.65-.95)*	1.03 (.80-1.32)	.80 (.65-.98)*

** Adjusted for sociodemographic characteristics, clinical diagnosis, functional impairment, and referral source



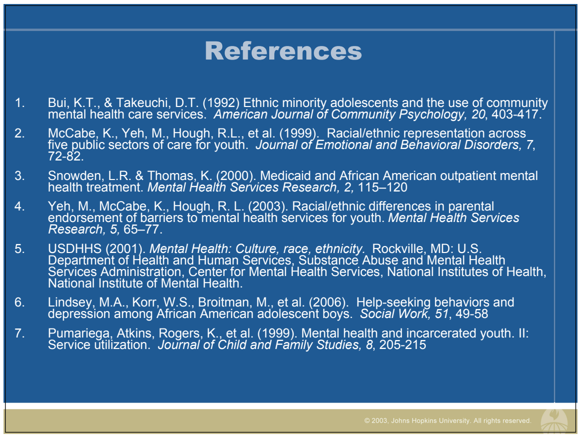
Discussion

- Challenges to Employing MI techniques
  - Had to understand the results in the context of multiply imputed data
  - Deciding which type of multiple imputation commands to use
  - Finding alternative commands and ways to analyze the data appropriately



Discussion

- Benefits
  - Have more complete dataset to work with
  - Building models was not complicated
  - Analyses were conducted generally in the same way as analyses with non-imputed data
- Suggestions
  - Use available resources
  - Keep a syntax (or .do) file
  - Keep output or logs of all analyses

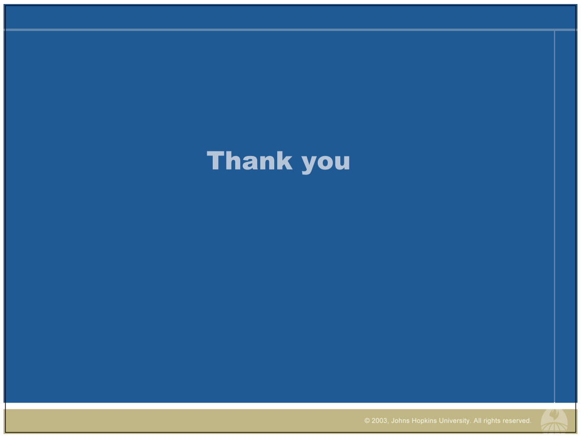


References

1. Bui, K.T., & Takeuchi, D.T. (1992) Ethnic minority adolescents and the use of community mental health care services. *American Journal of Community Psychology, 20*, 403-417.
2. McCabe, K., Yeh, M., Hough, R.L., et al. (1999). Racial/ethnic representation across five public sectors of care for youth. *Journal of Emotional and Behavioral Disorders, 7*, 72-82.
3. Snowden, L.R. & Thomas, K. (2000). Medicaid and African American outpatient mental health treatment. *Mental Health Services Research, 2,* 115–120
4. Yeh, M., McCabe, K., Hough, R. L. (2003). Racial/ethnic differences in parental endorsement of barriers to mental health services for youth. *Mental Health Services Research, 5,* 65–77.
5. USDHHS (2001). *Mental Health: Culture, race, ethnicity.* Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health.
6. Lindsey, M.A., Korr, W.S., Broitman, M., et al. (2006). Help-seeking behaviors and depression among African American adolescent boys. *Social Work, 51*, 49-58
7. Pumariega, Atkins, Rogers, K., et al. (1999). Mental health and incarcerated youth. II: Service utilization. *Journal of Child and Family Studies, 8*, 205-215

Acknowledgements

- Collaborators
  - Macro International
    - Christine Walrath
    - Brigette Manteuffel
    - Bob Stephens
    - Bhuvana Sukumar
    - Lucas Godoy Garraza
  - Johns Hopkins
    - Philip Leaf, PI
    - Constantine Frangakis
  - University of Colorado, Denver
    - Richard Miech



Thank you