Multiple Imputation for Multivariate Incomplete Data: A Practical Guide

Yulei He
Department of Health Care Policy
Harvard Medical School
he@hcp.med.harvard.edu

Acknowledgement: Sharon-Lise Normand, Alan Zaslavsky, Haiden Huskamp
Grant Support: U01-CA93344 from the National Cancer Institute

Outline

- Part I: Introduction
  - Motivation
  - Analysis goals
  - Missing data mechanisms
  - Ad-hoc missing data methods
  - Principled methods
    - Nonresponse weighting
    - Likelihood-based methods

Outline (ctd.)

- Part II: Multiple Imputation
  - Concepts
  - Analysis procedures
  - Imputation models
  - Software
  - Practical guidance
  - Example
  - Comments
  - References

Motivation

- Missing data a common problem
- Maybe due to design or happenstance
- Inappropriately dealing with missing data may lead to bias and inefficiency
- Particularly salient for health outcomes and services research
  - More opportunity for missingness in larger studies

Illustrative Study: Hospice Discussion

Huskamp et al. “Discussions With Physicians About Hospice Among Patients With Metastatic Lung Cancer” Arch Intern Med. 2009;169(10):954-962.

- Objective: identify cancer patient characteristics associated with early discussion about hospice
- Data: subset from CanCORS database
- Outcome: early discussion of hospice (within 4 months of diagnosis) with a provider
- Independent variables: patient’s clinical and demographic variables
- For illustrative purpose in this talk

Table 1: Variables in the "Hospice Discussion" study

Variable	Label and Classification	Missingness freq. (%)
Hospice discussion	1=hospice discussed, 0=no	3.56
Income	1= <20k, 2= 20-40k, 3=40-60k, 4=>60k	19.48
Gender	0=male, 1=female	0
Race	1=white, 2=black, 3=hispanic, 4=asian, 5=other	0.16
English	1=yes, 0=no	0.65
Education	1= less than high school, 2=high school/some college, 3=college degree or more	2.34
Marital status	1= married/live with partner, 2=widowed, 3=divorced/separated, 4=never married	2.79
Myocardial infarction	1=heart attack, 0=no	12.37
Congestive heart failure	1=heart failure, 0=no	13.06

Variable	Label and Classification	Missingness freq. (%)
Stroke	1=stroke, 0=no	12.53
Lung disease	1=lung disease, 0=no	12.93
Diabetes	1=diabetes, 0=no	12.21
Depression	1=depression, 0=no	12.37
Chemotherapy	1=chemo, 0=no	1.78
Insurance	1=medicare, 2=medicaid, 3=private, 4=other	8.61
Hospice	1=hospice used, 0=no	3.23
Age group	1=21-55 yrs, 2=56-60, 3=61-65, 4=66-70, 5=71-75, 6=76=80, 7=81+	0.04
Deceased	1=deceased within 1 yr of dx, 0=no	0
Cancer stage	1=stage IV, 0=stage III	0
PDCR site/code	10=CRN, 20=NCCC, 30=UAB, 40=UCLA, 50=Iowa, 70=VA	0

Table 2: Missing data matrix

Subject	Myocardial Infarction	Heart Failure	Stroke	Age	. . .
1	Yes	No	No	No	56-60
2	Yes	No	No	No	56-60
3	No	Yes	?	Yes	76-80
4	?	Yes	No	Yes	?
5	?	No	?	Yes	?

Note: ? Indicates missing data.

What Do We Do with Missing Data?

Make statistically valid inferences about population parameters from an incomplete dataset

- Not to estimate, predict, or recover missing values themselves
- Good to understand reasons for/causes for missingness
- Good to make reasonable assumptions based on data and substantive knowledge
- Sensitivity analyses are helpful

What Is a Missing Value?

- There is a well-defined "true" value underlying the missing-value code
  - In the hospice study, a subject refuses to answer the income question
  - In a pre-election poll, subject won't say which candidate he/she supports
  - In a clinical trial, a patient drops out of a study because treatment does not work
- Types of nonresponse
  - Unit nonresponse: no data are collected from the study patients (e.g., not at home, refused to participate, etc.)
  - Item nonresponse: partial data collected for the patient, but some items missing (e.g., refused to answer income question but respond to age question)

What Is a Missing Value? (ctd.)

Sometimes may be more appropriate to consider a nonresponse to be a qualitatively different category

- E.g. “If you had to make a choice now, would you prefer treatment that extends life as much as possible, even if it means having more pain and discomfort, or would you want treatment that focuses on relieving pain and discomfort as much as possible, even if it means not living so long?”

  1=Extend life as much as possible; 2=Relieve pain or discomfort as much as possible; 3=Don’t know; 4=Refused

Missing Data Mechanisms

- Missingness: defined as binary random variables that can be characterized by statistical models

Table 3: Missingness matrix

Subject	Myocardial Infarction	Heart Failure	Stroke	Diabetes	Age	...
1	0	0	0	0	0	
2	0	0	0	0	0	
3	0	0	1	0	0	
4	1	0	0	0	1	
5	1	0	1	0	1	

Note: "1" for missing response and "0" for observed response.

Missing Data Mechanisms: Main Categories

- Missing completely at random (MCAR): the probability of missingness does not depend on the data values, missing or observed
  - A bunch of data files are lost
- Missing at random (MAR): the probability of missingness depends only on the observed data (ignorability assumption)
  - Older patients might be more likely to have missing "insurance" than younger patients
  - "Insurance" is MAR if patient age for all patients is available
- Not missing at random (NMAR): the probability of missingness is related to unobserved values (non-ignorable)
  - "Income" is NMAR if higher-income group is less likely to respond

Missing Data Mechanisms: More Comments

- MCAR rarely holds in real scenarios
- Can differentiate MCAR and MAR if meaningful differences exist between those with and without missing data for some variables
- Under MAR, do not have to model the missingness
- Most of the practical applications of multiple imputation assume MAR
- Cannot differentiate MAR and NMAR based only on the information from the observed data
- Under NMAR, need to model both missingness and data and analyses are harder
- The limitation of NMAR can be mitigated by including more variables in model for missing data and bringing it closer to MAR
  - E.g, whites and persons with college degrees tend to have higher-than-average incomes, so controlling for race and education predictors may somewhat correct for the hypothetical higher rate of nonresponse among higher-income people

Ad-hoc Missing Data Methods

Complete-case analysis (CC): using only subjects who have all variables observed

- Default option for an incomplete dataset in many statistical software packages
- If data are not MCAR, CC analysis results can be biased
- If many variables are included in a model, CC analysis will result in dropping many subjects and lead to the reduction of statistical power

Ad-hoc Missing Data Methods (ctd.)

Imputation: “fill-in” missing values, keep the full sample size and facilitate easy completed-data analysis

- Unprincipled single imputation approaches: mean imputation, last values carried forward, missing data indicator method, etc.
  - Do not reflect missing data uncertainty: overstate sample size and produce underestimated standard errors
  - Impractical in multivariate settings with arbitrary missingness

Principled Missing Data Methods: Nonresponse Weighting

- Modify the complete-case dataset so that it becomes representative of the full sample
- Observed data are a sample selected from completed data
- The weights are the inverse of the predicted probabilities of response
- Might be more suitable for unit nonresponse
- Becomes less tractable for multivariate item missingness with arbitrary pattern
- Extreme estimates of weights can lead to erratic variance estimates

Principled Missing Data Methods: Likelihood-based Methods

- Maximize the observed-data likelihood to obtain parameter estimates
- Observed-data likelihood function: complete-data likelihood averaged over unknown missing data
  - Y_i: complete data on subject $i = 1, 2, \ldots, n$;
  - $Y_{i,obs}$: observed components, $Y_{i,mis}$: missing components, $Y_i = (Y_{i,obs}, Y_{i,mis})$
  - Complete-data model: $f(Y_i|\theta) = f(Y_{i,obs}, Y_{i,mis}|\theta)$
  - θ is the unknown parameter to be estimated (e.g., regression coefficients)
  - Observed-data likelihood:
    $L_{obs}(\theta|Y_{obs}) = \prod_{i=1}^{n} L(\theta|Y_{i,obs}) \propto \prod_{i=1}^{n} \int f(Y_{i,obs}, Y_{i,mis}|\theta) dY_{i,mis}$
- Theories and applications have been extensively developed
- Special computational techniques are often needed
- Not widely implemented in software and packages

Multiple Imputation Concepts

- $L_{com}(\theta|Y_{obs}, Y_{mis})$ is the complete-data likelihood if there are no missing data
- Observed-data likelihood:

$$\begin{aligned} L_{obs}(\theta|Y_{obs}) &= \int L_{com}(\theta|Y_{obs}, Y_{mis}) Pr(Y_{mis}|Y_{obs}) dY_{mis} \\ &\approx \frac{1}{M} \sum_{m=1}^{M} L_{com}(\theta|Y_{obs}, Y_{mis}^{m}) \end{aligned}$$

- M imputations, $(Y_{mis}^{m}, m = 1, \ldots, M)$, are independent draws from the posterior predictive distribution, $Pr(Y_{mis}|Y_{obs})$
- Multiple imputation analysis that combines the likelihood-based analysis from each completed dataset is approximately equivalent to the analysis based on the observed-data likelihood
- Missing data uncertainty is embedded in the M imputations.

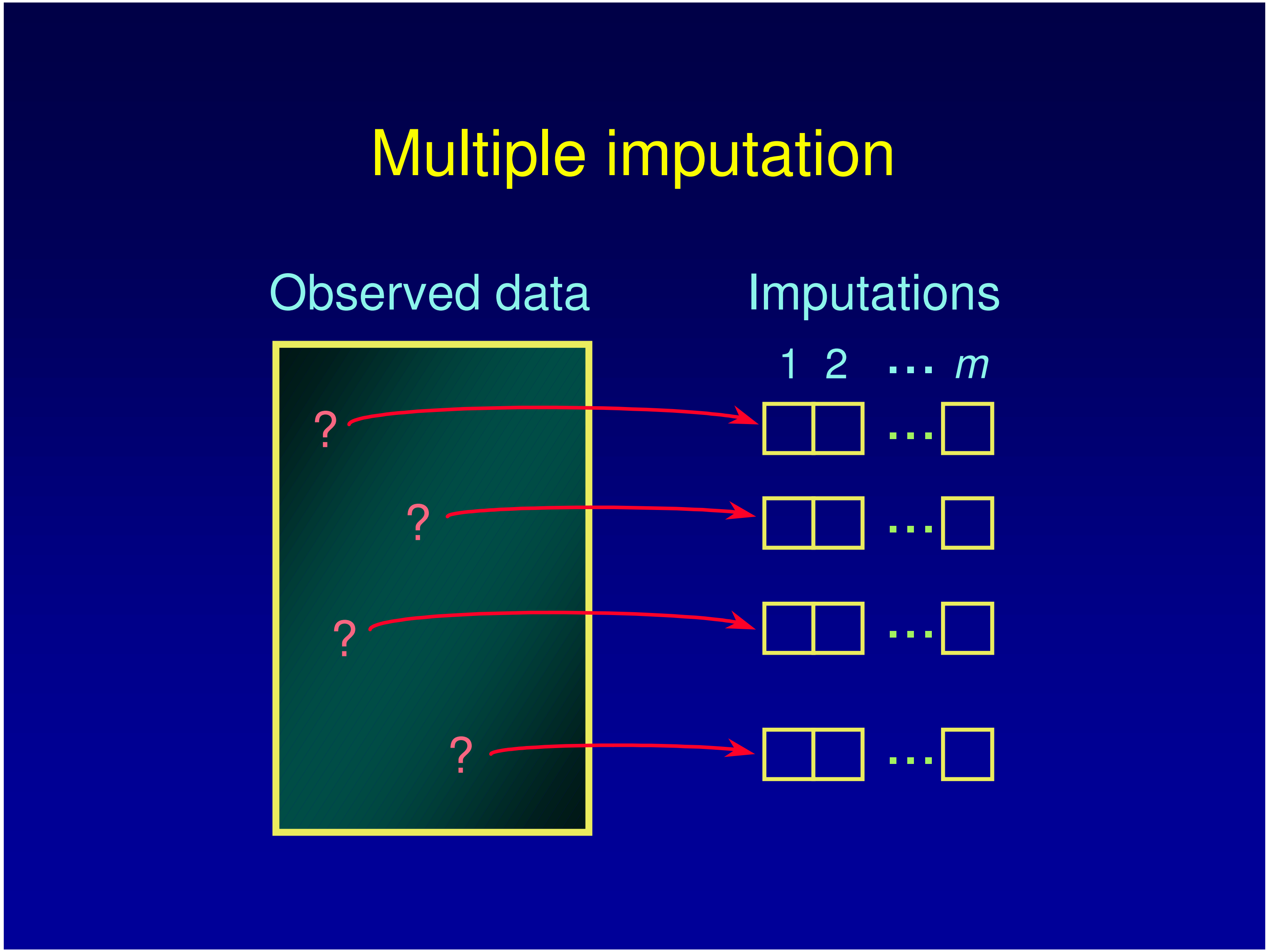


Figure 1: The scheme of multiple imputation, where ? indicates missing data

Multiple Imputation Procedure

- Impute Y_{mis} from a complete-data model independently multiple (say M) times
- For each completed data set, obtain the inference for θ: point estimate $\hat{\theta}_m$ and variance (standard error) estimate $V\hat{a}r(\hat{\theta}_m)$, $(m = 1, \ldots, M)$ using typical complete-data procedures
- Combine M sets of point estimates and standard errors to obtain a single point estimate, standard error, and the associated confidence interval or significance test (p-value)

Multiple Imputation Combining Rules

- Point estimate is the average $\hat{\theta}_{MI} = \frac{\sum_{m=1}^{M} \hat{\theta}_m}{M}$
- The variance of the multiple imputation estimate is $\hat{Var}(\hat{\theta}_{MI}) = U_{MI} + (1 + \frac{1}{M})B_{MI}$
  - $U_{MI} = \frac{\sum_{m=1}^{M} \hat{Var}(\hat{\theta}_m)}{M}$ (within-imputation variance)
  - $B_{MI} = \frac{\sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta}_{MI})^2}{M-1}$ (between-imputation variance)
- To obtain test statistics and confidence intervals $\frac{\theta - \hat{\theta}_{MI}}{\sqrt{\hat{Var}(\hat{\theta}_{MI})}} \dot{\sim} t_\nu$
  - $r = (1 + \frac{1}{M})B_{MI}/U_{MI}$ (relative increase in variance due to nonresponse), $\nu = (M-1)(1 + r^{-1})^2$

An Example for Combining Estimates

Imputation 1		Imputation 2		Imputation 3		Imputation 4		Imputation 5	
Y	X	Y	X	Y	X	Y	X	Y	X
3	1	3	1	3	1	3	1	3	1
4	4	4	4	4	4	4	4	4	4
3.6	3	4.3	3	4.0	3	2.7	3	2.9	3
2	2	2	2	2	2	2	2	2	2
8	7	8	7	8	7	8	7	8	7
...	...	...	...	...	...	...	...	...	...

- Completed-data analysis: regress Y on X
- The slopes of Y on X from 5 imputed datasets are $\{1.5, 1.4, 1.6, 1.0, 1.3\}$
- The variances of slopes are $\{.21, .24, .30, .27, .25\}$

An Example for Combining Estimates (ctd.)

- The combined point estimate is $\frac{1.5+1.4+1.6+1.0+1.3}{5} = 1.36$
- The within-imputation variance, U_{MI}, is $\frac{.21+.24+.30+.27+.25}{5} = .244$
- The between-imputation variance, B_{MI} is $\frac{(1.5-1.36)^2+(1.4-1.36)^2+(1.6-1.36)^2+(1.0-1.36)^2+(1.3-1.36)^2}{5-1} = .053$
- The total variance is $.244 + (1 + 1/5) \times .053 = .3076$, and the standard error is $\sqrt{.3076} = .55$

An Example for Combining Estimates (ctd.)

- $r = (1 + 1/5) \times .053/.244 = .261$, and $\nu = (5 - 1) \times (1 + 1/.261)^2 = 93.6$
- The critical value for the t-distribution with $\nu = 93.6$ and .025 upper tail probability is 1.986
- The 95% confidence interval for the slope is $(1.36 - 1.986 \times .55, 1.36 + 1.986 \times .55) = (.27, 2.45)$, and the p-value for testing if the slope is 0 (two-sided) is .015.
- Automatic calculations are enabled in multiple imputation packages

Imputation Models for Univariate Missing Variable

Set up a regression model between the missing variable (as the outcome) and complete variables (as the predictors)

- Linear normal regression model for continuous data (or with transformation)
- Generalized linear models (logistic/probit/Poisson) for categorical data
- Semiparametric or nonparametric extensions regression models for robustness

Imputation Models for Multivariate Incomplete Variables

Joint modeling approach

- Assume a joint model for the incomplete variables
  - Multivariate normal model for continuous variables
  - Multinomial/loglinear model for categorical variables
  - General location model for a mixture of continuous and categorical variables
  - Multilevel model for repeated measurements
- Maybe difficult to implement if
  - Many variables are involved
  - Have different types: continuous, binary, ordinal, counts, etc.
  - Exhibit restrictions and boundaries

Imputation Models for Multivariate Incomplete Variables (ctd.)

Sequential regression multiple imputation strategy (SRMI)

- Impute each variable through a regression model using other variables as predictors
  - In a dataset $\{Y_1, \ldots, Y_p\}$, impute Y_1 from $P(Y_1|Y_2, \ldots, Y_p)$, Y_2 from $P(Y_2|Y_1, Y_3, \ldots, Y_p)$,..., Y_p from $P(Y_p|Y_1, \ldots, Y_{p-1})$
- Possible regression models in SRMI include:
  - linear normal model for continuous outcomes
  - logistic model for categorical outcomes
  - Poisson model for count data
  - Two-part model for the variable of mixed type
- One imputation cycle goes through all incomplete variables sequentially and such cycles are repeated
- Relatively easy to incorporate complex data features compared to the joint modeling approach

Multiple Imputation Software

- SAS:
  - PROC MI for imputation
  - PROC MIANALYZE combines multiple estimates
- S-plus
  - libraries: impGauss, impLogin, and impCgm
- R
  - libraries: norm, cat, mix, pan, mi and Hmisc
- IVEware: Imputation and Variance Estimation software for SRMI, callable by SAS (http://www.isr.umich.edu/src/smp/ive).
- MICE: Multiple Imputation by Chained Equations, SRMI library available in both S-plus and R
- ICE: SRMI library available in STATA.

Some Practical Guidance for Multiple Imputation

- Understand the analytic objective of the study
- Initial data processing and identify missing data
- Identify working variables in imputation.
  - Include at least all variables involved in the planned analysis.
  - When imputing missing predictors, the outcome variables should be included in imputation
  - Other variables that have strong correlation with incomplete variables might be included
  - The more variables involved, the more predictive power the model has.
  - However, including more variables increases the complexity of model fitting

Some Practical Guidance for Multiple Imputation (ctd.)

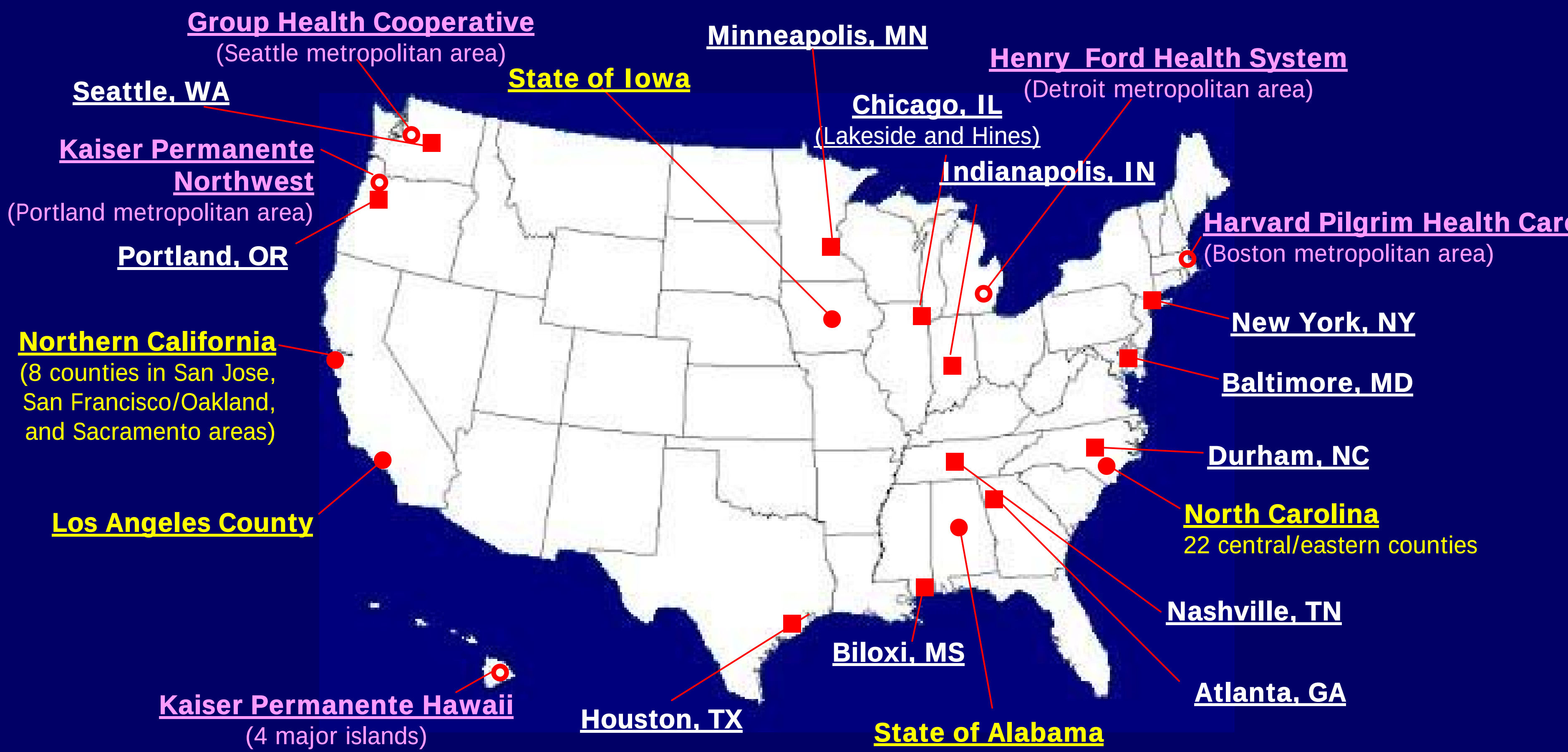
- Construct the imputation model: seeking a balance between sophistication and feasibility of models
  - For most empirical analyses, we recommend using models based on developed imputation routines provided by available software
- Carry out imputation diagnostics and sensitivity analysis.
- Post-imputation data processing
- Combine completed-data estimates from multiple datasets and report the results.
  - Analysis results can include a sensitivity analysis from several candidate models, and accompany with results from other missing data methods
- Store multiply imputed datasets and observed data together
- Document the imputation procedures

Hospice Care Analysis: Study Design

CanCORS (Cancer Care Outcome Research and Surveillance Consortium)

- Objective: study the patterns of cancer care using observational data
- Multi-site study: 5 geographical sites and 2 provider collections
  - Enroll 10,000+ newly diagnosed colorectal and lung cancer patients
- Collect information from various sources
  - Patient/surrogate surveys
    - Multiple forms: baseline(full, brief, surrogate live, and surrogate death) and follow-up (survivor follow-up and decedent follow-up)
    - Multi-wave: baseline (4 months post-dx) and follow-up (12 months post-dx)
  - Medical records abstraction (up to 12 months post-dx)
  - Provider survey

CanCORS Sites



● Patients from population-based cohorts in geographic areas

○ Patients from integrated health-care delivery systems

■ Patients at Veterans Health Administration hospitals

Hospice Care Analysis: Data Feature

- The analytic cohort includes 20 variables from both the baseline survey and medical records data
- The cohort had 2474 patients with stage IIIB or IV lung cancer
- The main analysis is a logistic regression for "early discussion of hospice use" with predictors including patient's clinical and sociodemographic characteristics
- The missingness rates of variables range from .04% to 19.48%
- Complete-case analysis deletes around 30% of the subjects
- Missingness is not likely to be completely at random
  - Missingness of income and insurance are significantly related to other variables

Missing Data Methods

- Complete-case analysis
- Missing indicator method
  - Treat missing data as a separate category of the variable in regression analysis
- Nonresponse weighting and likelihood-based method are difficult to implement

Multiple Imputation Strategies

- Multivariate normal model: code all categorical variables as binary dummies, impute them as continuous, and round the imputations
- General location model
  - Treat race, marital status, and insurance as nominal variables and model them using a loglinear model
  - Treat other variables (binary or ordinal) as continuous variables and model them using a multivariate normal model conditional on the nominal variables
- Sequential regression imputation
  - Model each variable given others using a logistic regression
  - Retain the categorical feature of all variables
- Applying three imputation models as a sensitivity analysis

Table 4: Logistic Model Estimates

	CC			Missing Data Indicator			SRMI		
Predictor	EST	SE	P-value	EST	SE	P-value	EST	SE	P-value
Race									
White	Ref.								
Black	.03	.21	.90	.03	.17	.86	-.01	.16	.97
Hispanic	-.51	.31	.10	-.68	.24	.00	-.68	.25	.01
Asian	.24	.29	.42	.20	.25	.41	.19	.24	.43
Other	.46	.27	.08	.46	.23	.04	.43	.22	.06
Marital status									
Married/Living with partner	Ref.								
Widowed	-.24	.18	.19	-.10	.15	.50	-.06	.15	.68
Divorced/Separated	.24	.17	.17	.29	.14	.05	.30	.14	.03
Never Married	.66	.32	.04	.57	.25	.03	.60	.25	.02
Missing				.09	.42	.83			
Age group									
55 or less	Ref.								
56-60	-.10	.24	.67	.07	.19	.71	.04	.19	.82
61-65	.23	.25	.35	.10	.20	.64	.09	.20	.65
66-70	.18	.26	.50	.17	.21	.43	.12	.21	.58
71-75	.20	.28	.48	.13	.22	.55	.08	.22	.72
76-80	.25	.29	.39	.25	.24	.29	.19	.24	.45
81+	.50	.31	.10	.61	.25	.01	.51	.25	.04
Missing				-9.25	276.5	.97			
Myocardial infarction									
No	Ref.								
Yes	-.41	.17	.02	-.24	.15	.11	-.24	.14	.09
Missing				-.62	.56	.27			

Comparison among Methods

- Multiple imputation analysis reduced standard error of the regression estimates compared to CC
- Missing data indicator method produces similar results to multiple imputation analysis
- Multiple imputation analysis identified more significant predictors than CC
- Different imputation models produce similar results (not shown)

Substantive Results

- Early discussion of hospice is more likely for
  - patients with stage IV lung cancer
  - patients aged 81+ years than those 55 and under
  - Whites compared to Hispanics
  - divorced/seprated/never married patients compared to married ones;
  - patients who did not receive chemotherapy
  - patients who have depression
  - patients who have other types of insurance than those who have Medicare;
  - patients who died within 1 year of diagnosis
  - patients at CRN compared to those at UAB or VA
- Refer to Huskamp et al. (2009) for formal analysis results

Multivariate Normal Imputation Using SAS PROC MI

```
proc mi data=cohort_one out=cohort_one_nmmi_ori nimpute=10 seed=197789;
mcmc chain=multiple;
var mddishsp gender race2 race3 race4 race5 marital2 marital3 marital4 agegroup2
agegroup3 agegroup4 agegroup5 agegroup6 agegroup7 english_ edu2 edu3 ... ;
run;
```

- "data=" contains the incomplete data to be imputed
- "seed" initializes the random number generator
- "out=" is the output data that will contain 10 concatenated imputed datasets, which has a variable called "_Imputation_" referring to the number of imputed data
- "mcmc" refers to the Monte Carlo Markov Chain algorithm for imputation
- "var" statement identifies the variables to be included in the imputation, and all variables must be numeric

General Location Model Imputation Using "mix" in R

```
s_cohort=prelim.mix(cohort_one_mat, 4)
for (simu in 1:10) {
rngseed(simu);
theta_initial=ecm.mix(s=s_cohort, margins=cohort_margins, design=cohort_design_mat,
showits=TRUE);
theta_final=dabipf.mix(s=s_cohort, margins=cohort_margins, design=cohort_design_mat,
start=theta_initial, steps=200, showits=TRUE);
cohort_impute=imp.mix(s=s_cohort,theta=theta_final,x=cohort_one_mat); }
```

- s_cohort is the summary list of the incomplete data cohort_one_mat
- cohort_design_mat and cohort_margins are the design matrices for the models for the continuous and categorical variables, respectively
- theta_initial is the initial value for the imputation algorithm
- cohort_imptue is the imputed dataset for each imputation

SRMI Code Using IVEware

```
%impute(name=mysetup, setup=new, dir=/home/he/imputation_diagnostics);
datain work.cohort_one;
dataout work.cohort_one_semi_ori all;
* continuous ;
categorical inc20_60 education agegroup mddishsp gender pdcr_combined ... ;
transfer case;
iterations 20;
multiples 10;
seed 19770809;
print coef;
run;
```

- "transfer" includes variables not to be imputed
- "print coef" prints out the regression coefficients in the imputation
- The imputed data cohort_one_semi_ori contains a variable _mult_ as the indicator of the number of imputations

Combining Estimates Using PROC MIANALYZE

```
proc logistic data=cohort_one_semi_recode;
model mddishsp= gender race2 race3 race4 race5 marital2 marital3 marital4 agegroup2
agegroup3 agegroup4 agegroup5 agegroup6 agegroup7 ... /covb;
by _imputation_;
ods output ParameterEstimates=semi_parms CovB=semi_covb;
run;
proc mianalyze parms=semi_parms covb(effectvar=stacking)=semi_covb;
modeleffects Intercept gender race2 race3 race4 race5 marital2 marital3 marital4
agegroup2 agegroup3 agegroup4 agegroup5 agegroup6 agegroup7 ...;
run;
```

- First run logistic regression for each imputed dataset and then combine
- The logistic procedure goes through the 10 imputed datasets through "by _imputation_"
- Point estimates and standard errors are in datasets "semi_parms" and "semi_covb"
- Can combine results obtained from proc corr,reg,genmod,mixed

Additional Comments

- Other analytic problems can be approached from the perspective of missing data
  - Causal inferences, measurement error, confidential use of public database
- Simple yet ad-hoc methods are generally invalid
- Principled methods tend to yield more valid results based on plausible assumptions
  - Not necessarily more significant results
- Multiple imputation: an effective approach to regression analysis with multiple missing variables
  - A single researcher analyzing a particular incomplete dataset for a unique goal
  - Multiple researchers using different portions of a database for various aims

Major References for Multiple Imputation

Introductory

- "Multiple imputation in health-care databases: an overview and some applications", Rubin DB and Schenker N. (1991), *Statistics in Medicine*, 10, 585-598.
- "Multiple imputation: a primer", Schafer JL. (1999), *Statistical Methods in Medical Research*, 8, 3-15
- "Application of multiple imputation in medical studies: from AIDS to NHANES", Barnard J and Meng XL. (1999), *Statistical Methods in Medical Research*, 8, 17-36.
- "Using multiple imputation to incorporate cases with missing items in a mental health study", Belin et al. (2000), *Health Services and Outcomes Research Methodology*, 1, 7-22.
- "Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models", Horton NJ and Kleiman KP. (2007) *American Statistician*, 61, 79-90.
- "Multiple imputation: review of theory, implementation, and software", Harel O and Zhou XH (2007), *Statistics in Medicine*, 26, 3057-3077.

Major References for Multiple Imputation (ctd.)

Technical

- “Multiple Imputation for Nonresponse in Surveys” Rubin DB. (1987) Wiley
- “Analysis of Incomplete Multivariate Data” Schafer JL. (1997) Chapman and Hall
- “Statistical Analysis with Missing Data”, Little RJ and Rubin DB (2002) Wiley