Imputation for National Hospital Care Survey

Iris Shimizu, Diba Khan, Yulei He, Xiang Liu,
Marian Strazzeri, Jin Zhang, Jianmin Xu, Bill Cai

National Center for Health Statistics (NCHS)

Notes:
Marian Strazzeri is currently at the Food and Drug Administration (FDA).

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official positions of the NCHS, Centers for Disease Control and Prevention, or the FDA.

OUTLINE

- Introduction/background
- Methods
- Results
- Discussion
- Summary

Introduction

- National Hospital Care Survey (NHCS)
- Survey objectives: To produce estimates on utilization of health care provided in hospital based settings
- Started recruiting hospitals in 2011
- Replaces National Hospital Discharge Survey

Survey Population

- Non-Federal, non-institutional hospitals
- Discharges from those hospitals and
- Visits made to their emergency rooms and outpatient departments.

Sample design

- Stratified list sample of hospitals

- Stratified by specialty type
    - Acute care
        - Bed size
        - Area type: Inner city, fringe city, small MSA, and non-MSA
    - Children's
    - Psychiatric
    - Long term care and other

Data collected

- Electronic medical records (preferred) or
- Claims data from Uniform Billing-04 (UB-04) form
- For each discharge and ambulatory visit (no sampling within hospitals)

Data collected for discharges include:

- Dates of admission and discharge
- Patient demographics: age, sex
- Diagnoses at admission
- Diagnoses related group (DRG)
- Procedures received during stay
- Discharge status
- Expected sources of payment

Discharge variables targeted in research and percent of times data is missing for them in 2011 data set

- Age - missing in 10.25 % of records
- Sex - missing in 0.01% of records
- Length of Stay - missing in 0.01% of records

The 2011 data set included 764,148 records from 60 hospitals

Research methods

- Compared imputation software
  - IVEWare
  - SUDAAN 11 -

- Compared imputation methods:
  - Hot deck imputation and
  - Model based imputation

- Based comparisons on accuracy of imputed values

- Investigated multiple imputation using best combination of imputation model and software

Software Considered

- IVEWare
  - Imputes continuous and categorical variables
  - Assumes data are missing at random (MAR)
  - Fits a sequence of regression models for values used for imputing
  - Simultaneously imputes several missing variables
    - Imputed values are interdependent
    - Uses correlational structure that exists among the auxiliary variables

Software Considered (continued)

- SUDAAN 11 – Three procedures offered in IMPUTE procedure. All:
  - Assume data are missing at random (MAR).
  - Uses design information, but not automatically

- SUDAAN 11 Weighted Sequential Hot Deck (WSHD)
  - Imputes continuous and categorical variables.
  - Donors are records with data for the targeted variables.
  - Sampling weights can be used in process.
  - Multiple Imputation is supported only in WSHD.

Software Considered (continued)

- SUDAAN 11 Cell Mean
  - Imputes continuous variables
  - Item respondents are records used for weighted cell means used to impute missing values.

- SUDAAN 11 Linear
  - Imputed continuous variables
  - Item respondents are the records used to fit regression models for values used to impute missing values.

Imputation methods

- Developed 5-7 imputation models for each targeted variable

- Variables used in models
  - Design variable (Hospital ID)
    - Included in all models
    - Was the only variable in Model 1 for each targeted variable

  - Auxiliary variables with probable influence on variables targeted

Imputation methods (continued)

- Models and their variables for imputing age
  - Model 1: Hospital ID
  - Model 2: Hospital ID, pediatric status
  - Model 3: Hospital ID, pediatric status, sex
  - Model 4: Hospital ID, pediatric status, major complication condition, other complication condition
  - Model 5: Model 4 variables plus sex

Note: Pediatric status was assigned if 20 percent of patient's diagnostic codes were pediatric specific.

Imputation methods (continued)

- Models and their variables for imputing sex
  - Model 1: Hospital ID
  - Model 2: Hospital ID, major complication condition, other complication condition
  - Model 3: Hospital ID, discharge status
  - Model 4: Hospital ID, length of stay
  - Model 5: Model 2 variables plus discharge status
  - Model 6: Model 2 variables plus length of stay
  - Model 7: Model 2 variables plus discharge status and length of stay.

Imputation methods (continued)

- Models and variables for imputing length of stay
  - Model 1: Hospital ID
  - Model 2: Hospital ID, major complication condition, other complication condition
  - Model 3: Hospital ID, discharge status
  - Model 4: Hospital ID, diagnosis related group (DRG)
  - Model 5: Model 2 variables plus discharge status
  - Model 6: Model 2 variables plus DRG
  - Model 7: Model 2 variables plus discharge status and DRG.

Comparing accuracy of imputed values

Evaluation for each targeted variable

- Created test sample of records
  - Randomly selected complete records.
  - Omitted the reported data for targeted variable

- Determined best combination of model and software
  - Restored test sample to research file
  - Imputed omitted data
  - Compared imputed with omitted values for each record
  - Computed percent of times correctly imputed in test sample
  - Compare accuracy rates across models and software

Comparing accuracy of imputed values: For age groups

Table: Percent of test sample records with correctly imputed values for age groups by software and imputation model: 2011 NHCS

Model	SUDAAN WSHD	SUDAAN Cell Mean	SUDAAN Linear	IVEware
I	16.80%	17.25%	11.40%	13.25%
II	17.25%	17.56%	11.74%	13.43%
III	18.02%	17.76%	12.07%	13.09%
IV	19.11%	19.43%	14.67%	14.76%
V	19.54%	**20.65%**	14.67%	14.73%

Test sample included 103,215 records.

Model V and **SUDAAN's Cell mean** was most accurate combination but Cell mean software does not support multiple imputation. Model V and SUDAAN's WSHD was used instead when investigating multiple imputation for age groups

Comparing accuracy of imputed values: For sex

Table: Percent of test sample records with correctly imputed values for sex by software and imputation model: 2011 NHCS

Model	SUDAAN WSHD	IVEware
I	51.90%	51.42%
II	52.76%	52.13%
III	52.12%	51.82%
IV	52.22%	51.67%
V	52.96%	52.68%
VI	53.18%	52.34%
VII	**53.47%**	52.68%

Test sample included 148,821 records.

Model VII with **SUDAAN's WSHD** had best accuracy for sex.

Comparing accuracy of imputed values: For length of stay

Table: Percent of test sample with correctly imputed values for length of stay by software and imputation model: 2011 NHCS

Model	SUDAAN WSHD	SUDAAN Cell Mean	SUDAAN Linear	IVEware
I	14.67%	8.53%	7.54%	0.64%
II	15.73%	12.20%	10.57%	6.74%
III	15.91%	11.29%	7.83%	6.79%
IV	14.92%	8.98%	7.41%	7.87%
V	17.22%	13.51%	10.92%	6.94%
VI	16.19%	12.99%	10.81%	5.69%
VII	**17.67%**	14.05%	11.12%	5.76%

Test sample included 103,213 records.

Model VII with **SUDAAN's WSHD** had best accuracy for length of stay.

Summary of results from single imputations

- For each targeted variable: the model with the most variables appears most accurate.
- Of tested software, SUDAAN's WSHD and cell mean procedures appeared most accurate.
- Too many variables (or categories) can make SUDAAN's WSHD crash.
- Multiple variables can be imputed in single run of SUDAAN's WSHD and IVEWare.

Research methods – Investigate multiple imputation

Multiple Imputation (MI):

- Accounts for variation caused by imputation of missing data

- Involves
  - Imputing missing data M times
  - Analysis of M completed data sets (after imputation)
  - Combining results of the M completed data sets for inference.

Multiple imputation -Quantities for Inference

M = total number of imputations (completed data sets)

$\hat{Q}_i$ = estimate based on i-th completed data set (i = 1, …, M)

$\bar{Q} = \frac{1}{M}\sum_{i=1}^{M} \hat{Q}_i$ = multiple imputation estimate

Multiple imputation –Quantities for inference (continued)

$\widehat{U}_i$ = estimated variance <u>within</u> i-th completed data set (i=1, ..., M)

$\overline{U} = \frac{1}{M}\sum_{i=1}^{M} \widehat{U}_i$ = <u>within</u> imputation variance for the estimate

$B = \frac{1}{M-1}\sum_{i=1}^{M}\left[\widehat{Q}_i - \bar{Q}\right]^2$ = <u>between</u> imputation variance for estimates

$T = \overline{U} + \left[1 + \frac{1}{M}\right] B$ = the overall multiple imputation variance

Multiple imputation –Quantities for how many imputations

- $r = \left[1 + \frac{1}{M}\right] B / \bar{U}$ = relative increase in variance due to imputation

- $FMI \approx \frac{r+2}{(v_M+3)(r+1)}$ = fraction of missing information

where v_M is adjusted degrees of freedom in multiple imputation variance

Multiple imputation –Quantities for how many imputations (continued)

- Fraction of missing information (continued)
  - $0 < FMI < 1$
  - FMI close to 0 implies small number M of imputations is sufficient
  - FMI close to 1 implies a larger M is necessary
- $RE = \left[1 + \frac{FMI}{M}\right]^{-1}$ = relative efficiency of using finite M imputations.

RE close to "1" denotes M imputations are probably sufficient

Multiple imputations results

Table: Values for the between, within, and total variances, relative increases in variances, relative efficiencies, fractions of missing information due to ten imputations and the design effects for weighted values for average age, proportions male and female, and average LOS: 2011 NHCS

Variable	Between imputation variance	Within imputation variance	Total variance	Relative increase in variance	Relative efficiency	Fraction of missing information	Design effect
Average age	0.00002	0.00098	0.00100	0.01694	0.99833	0.01672	1954.84
Proportion Male	0.00000	0.00000	0.00000	0.00021	0.99998	0.00021	305.02
Proportion Female	0.00000	0.00010	0.00010	0.00137	0.99986	0.00137	305.02
Average Length of stay	0.00000	0.00010	0.00010	0.00137	0.99986	0.00137	1632.64

These results are based on data from 60 hospitals and 10 imputations

Discussion

- Multiple imputations had negligible effect in this study
- Potential reasons for the negligible effect:
  - Percent of records with data missing is small.
  - Within imputation variances dominate total variance
  - Large design effects for the targeted variables (due to single stage sample).
  - Small number (60) of hospitals included in study.

Future research

- Apply research to a larger dataset (more hospitals)
- Investigate more auxiliary variables for correlation with targeted variables to improve imputation accuracy
- Do research on more domains for analysis of multiple imputations.

Thank you!

Contact information: Iris Shimizu at ishimizu@cdc.gov