Missing Data

ehsan.karim@ubc.ca

Oct 13, 2020

SPPH 504/007

Ref

SWC

[HTML] Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls

JAC Sterne, IR White, JB Carlin, M Spratt, P Royston... - Bmj, 2009 - bmj.com

Missing data are unavoidable in **epidemiological** and **clinical research** but their **potential** to the validity of **research** results has often been overlooked in the **medical** literature.1 This is par because statistical methods that can tackle **problems** arising from **missing data** have, until recently, not been readily accessible to **medical researchers**. However, **multiple imputation**—a relatively flexible, general purpose approach to dealing with **missing data**—is now available in standard statistical software,2 3 4 5 making it possible to handle **missing data**

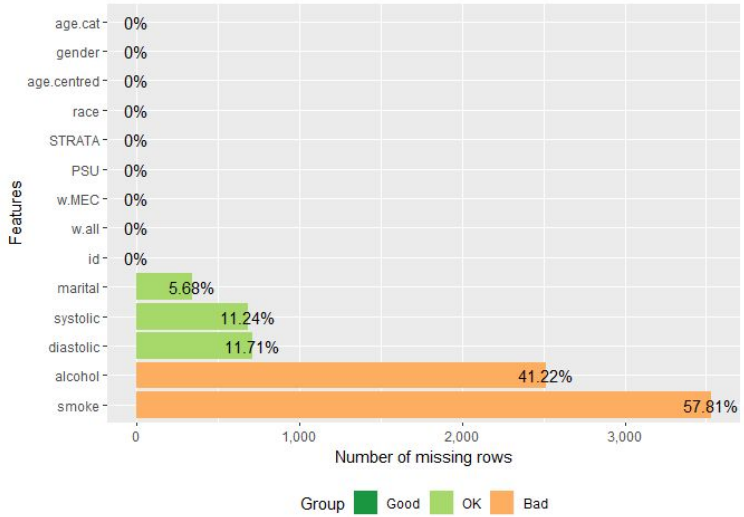
☆ 99 Cited by 3332 Related articles All 16 versions Web of Science: 1992 Import into

Content

1. Missing data / consequences
2. Types of missingness
3. Addressing missing data problem
    a. Ad-hoc approaches
    b. Imputation
        i. Single imputation
        ii. Multiple imputation
4. Special cases
    a. Variable selection in the presence of missing data
    b. Outcome missing
    c. Design-based analysis

Missing data

- Common in clinical and epidemiologic research
- Complete case analysis is very common
  - Without proper consideration of the implication of this choice
- Referees often ask for missing data analysis/sensitivity
- As a result:
  - Authors often do not explicitly state whether they had missing values
  - Often use suboptimal missing data analysis methods



Consequences of missing data

- Bias
- Incorrect SE/precision
- A substantial loss in power

Types of missing data

- Type 1: Missing Completely at Random (MCAR)
  - Random phenomenon
    - Issues with data entry
  - Missing values vs. observed values
    - no systematic difference
  - Complete case analysis is valid
    - but may have low power
  - This is rarely the case.

Types of missing data

- Type 2: Missing At Random (MAR)
  - Missingness can be explained by observed
  - Missing values vs. observed values
    - We know how they differ
  - This assumption is cited more often to justify a missing data analysis.
  - Complete case analysis is NOT valid
  - How plausible is the assumption?
    - In practice, including large # of predictor in the imputation model may be helpful

Types of missing data

- Type 3: Not Missing At Random (NMAR)
  - missingness may be explained by "unobserved"
  - Missing values vs. observed values
    - We know that they differ, we don't know how
  - Under this case, including large number of predictors in the imputation model will not help.

Testing missingness pattern

- No easy way / formal statistical test to identify which types of missingness we are dealing with.
  - Subject area understanding necessary
  - Analyst needs to have a good understanding of the data collection process
- Some diagnostics are available
  - MCAR test
  - "it is not possible to distinguish between MAR and MNAR using observed data."

Addressing missing data problem

(a) Wrong/ad-hoc ways

(b) Correct ways

Ad-hoc approaches

- Ignore and not think about it/blindly go with software
- Replacing missing values with the mean of the observed values
- using a missing category indicator
  - Add missingness indicator in the regression
- replacing missing values with the last measured value
  - Useful for longitudinal data (LOCF)

None of these are valid approaches statistically.

- Complete case is only valid for MCAR.

Imputation

- General rule of thumb: > 5% missing
- Filling in missing data to produce a complete data set
  - Single imputation
  - Multiple imputation

Single imputation

- Imputed values can be taken from other subjects
  - Within the same sample at random or
  - matched on key variables (**hot-deck imputation**) (same data) or
  - **cold-deck imputation** (from external data)

Single imputation

- More generalized methods
  - Simply impute the "mean"
  - Use regression prediction
    - Use observed values to fit a regression, and then used predicted values to impure
    - Also known as conditional mean imputation
    - Y-hat = intercept + slope*predictor
  - 'stochastic regression imputation'
    - Y-hat = intercept + slope*predictor + error term
  - Predictive mean matching (pmm)
    - A type of hot-deck, but uses regression to obtain match
    - (similar to propensity score matching; imputes the best match)
    - Generally associated with good properties

Single imputation

- No variation
  - Analysis treats observed vs imputed values the same way
  - uncertainty in missing data is not represented in the imputed data.
  - Results overly precise! P-values could be more significant!
  - Correlation may increase!

What is the goal of the missing data analysis?

Obtain best prediction of the missing values

Obtain accurate estimation of the treatment effect

Understanding the missingness pattern

Imputing the missing value with a realistic value

Using more data

Satisfy reviewer and get the paper published

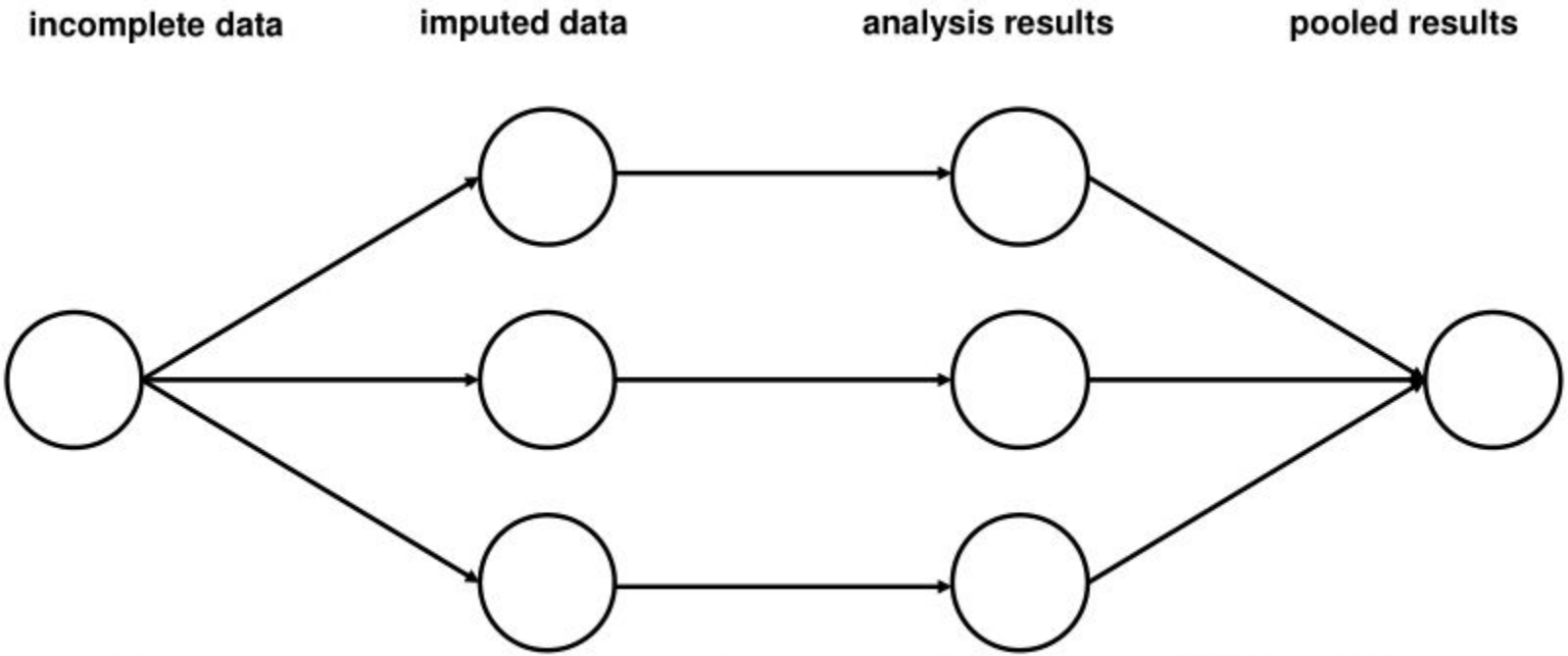
Multiple imputation (MI)

- [s0] construct a imputation model to predict the missing
  - fit this model to the observed data
  - missing data are sampled from the predictive distribution p() of the fitted model
- [s1] Create m (5-20) copies of the dataset (40?)
  - impute the missing values with from p()
  - to generate m complete-case datasets
  - induces variation
- [s2] Perform the same analysis on all of the m datasets.
  - get individual estimates
- [s3] pool/average results to get single estimate & SE

Lumley 2010 ch9

Multiple imputation (MI)

MI process



Multiple imputation (MI) advantages vs. not

- Uncertainty associated with imputations are taken into account
- Analysis is complicated and time consuming

Lumley 2010 ch9

Multiple imputation (MI): Predictors

- MI analyses will avoid bias
  - only if enough variables predictive of missing values are included
- If predictor associations are not incorporated in MI process:
  - such relationships will not be present in the imputed data
  - estimated association will be falsely weaken
- Outcome may carry information about the missing values
  - such information should be utilized. Generally outcomes are considered as predictors in the imputation model (surprising?).
- subject-area knowledge useful to build imputation model

Multiple imputation (MI): Predictors

- Auxiliary variables
  - Predictors of the missing values, which are not used elsewhere in the analysis
  - Research shows not much harm by adding these otherwise noise variables (with respect to the relationship of interest)
  - Adding more variables could make MAR assumption more plausible
  - Makes sense if these variables have not much to do with the relationship of interest
    - Survey features?

Multiple imputation (MI): Predictors

- Imputation model
  - Could include higher order term
  - Might make sense to model exposed vs unexposed group separately
    - if you have a reason to believe the missingness in these two groups are different

Diagnostics and Convergence

- Diagnostics
  - Focus on the variables that seem difficult
  - Compare before after imputation
  - Compare imputations from different imputed datasets
- Convergence

Multiple imputation (MI): % missing

- % of missing information: compares
  - the precision of an estimate to
  - the precision that would be available in a study of the same size without missing/non-response.
- similar to DEff

What do we do if the outcome is missing?

We impute outcome values as usual

We only work with outcome values that are observed

We build a separate imputation model for the outcome

Special cases

van Buuren section 5.4.1

a) Variable selection in the presence of missing

- If we perform stepwise in s2, we may come up with different models
  - Then it is not obvious how to pool them? [Hint: no s3 part necessary]
  - Which variables should be used finally?
- Three methods are proposed
  - Majority: (similar to bootstrap approach)
    - finally select those variables that gets selected more than $\frac{1}{2}$ of the times
    - Note that, this approach is completely independent of bootstrap approach!
  - Stack
    - Stack all imputed data into one big data and do variable selection there
  - Wald-test based stepwise
    - Preferred!

b) What if the Outcome is missing?

- multiple imputation, then deletion (MID)
  - When imputing, use the imputed outcome to impute other variables
  - When imputation complete, delete the imputed outcomes
  - Operates under the assumption that imputed outcomes have nothing to add in the regression; but just adds noise to the analysis
    - To keep those imputed outcomes, one may need to justify whether they really add anything to the analysis
    - Subject to some controversy in recent years

4. Regression with missing Ys: An improved strategy for analyzing multiply imputed data

PT Von Hippel - Sociological Methodology, 2007 - journals.sagepub.com

When fitting a generalized linear model—such as linear regression, logistic regression, or hierarchical linear modeling—analysts often wonder how to handle missing values of the dependent variable Y. If missing values have been filled in using multiple imputation, the usual advice is to use the imputed Y values in analysis. We show, however, that using imputed Ys can add needless noise to the estimates. Better estimates can usually be obtained using a modified strategy that we call multiple imputation, then deletion (MID) …

☆ 99 Cited by 986 Related articles All 10 versions Import into BibTeX

Lumley 2010 ch9

c) Survey data?

- MI steps can be naturally extended to survey data analysis
- Software available
- Will see in the labs/exercises

[BOOK] **Complex surveys**: a guide to analysis using **R**

T **Lumley** - 2011 - books.google.com

A complete guide to carrying out complex survey analysis using R As survey analysis continues to serve as a core component of sociological research, researchers are increasingly relying upon data gathered from complex surveys to carry out traditional …

☆ 99 Cited by 575 Related articles All 7 versions Import into BibTeX

Common problems in the literature

SWC page 5

- Use of suboptimal approaches
  - Not doing anything
  - Complete case
- Not reporting
  - missing data %s properly (extent for each variable)
  - what types of missingness going on (nature)
  - What type of imputation was done (single, multiple)
  - Justification of how it was dealt (plausibility of MAR)
    - imputation model / predictors; whether auxiliary/interactions were used
  - how many imputed datasets use
  - What software/version was used

Additional ref

Longitudinal

Missing data analysis: Making it work in the real world

JW Graham - Annual review of psychology, 2009 - annualreviews.org

This review presents a practical summary of the missing data literature, including a sketch of missing data theory and descriptions of normal-model multiple imputation (MI) and maximum likelihood methods. Practical missing data analysis issues are discussed, most notably the inclusion of auxiliary variables for improving power and reducing bias. Solutions are given for missing data challenges such as handling longitudinal, categorical, and clustered data with normal-model MI; including interactions in the missing data model; and ...

☆ 99 Cited by 4221 Related articles All 13 versions Web of Science: 2377 Import i

MI tutorial

Multiple imputation using chained equations: issues and guidance for practice

IR White, P Royston, AM Wood - Statistics in medicine, 2011 - Wiley Online Library

Multiple imputation by chained equations is a flexible and practical approach to handling missing data. We describe the principles of the method and show how to impute categorical and quantitative variables, including skewed variables. We give guidance on how to specify the imputation model and how many imputations are needed. We describe the practical analysis of multiply imputed data, including model building and model checking. We stress the limitations of the method and discuss the possible pitfalls. We illustrate the ideas using a ...

☆ 99 Cited by 3461 Related articles All 6 versions Web of Science: 2327 Import into BibTeX

Thanks!

ehsan.karim@ubc.ca
www.ehsankarim.com