Missing Values in Epidemiology:

A Summary of Literatures Discussing Industry Practices

Yian Wang

November 23, 2020

1. Missing Data and Multiple Imputation in Clinical Epidemiological Research

Published Date: March 15, 2017
Authors: Alma B Pedersen, Ellen M Mikkelsen, Deirdre Cronin-Fenton, Nickolaj R Kristensen, Tra My Pham, Lars Pedersen, Irene Petersen
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5358992/
Cited by: 284

1.1 Overview

- Some methods of handling missing data: complete-case analyses, missing indicator method, single value imputation, sensitivity analyses incorporating worst-case and best-case scenarios"
- Under MCAR assumption, some of these methods provide unbiased (but less precise) estimates
- However, in clinical epidemiological research, missing data are rarely missing completely at random (MCAR) so the results are usually biased and either have too large standard errors (lack of precision) or have too small standard errors (overestimation)
- Suggests multiple imputation as an alternative as it accounts for uncertainty associated with missing data. Implemented under MAR assumption (can also handle MNAR) and gives unbiased estimates based on the available data.
- "In order to increase the transparency and understanding of the research results, [. . .] the use of extended STROBE guidelines for reporting of multiple imputation analyses" is recommended.

1.2 Other Notes

None

2. Multiple Imputation of Missing Values was not Necessary Before Performing a Longitudinal Mixed-Model Analysis

Published Date: September 2013
Authors: Jos Twiskab, Michiel de Boer, Wieke de Vente, Martijn Heymans
URL: https://www.sciencedirect.com/science/article/pii/S0895435613001236?via%3Dihub
Cited by: 218

2.1 Overview

- By comparing a mixed-model analysis with and without the use of multiple imputations (MI), this paper shows that MI should/need not be used on missing values prior to longitudinal data analysis (eg. mixed-model analysis).
- It states that these longitudinal studies consist of the most missing values, and that when missing data depends on unobserved data, the missing data is not at random (MNAR).
- They find that when the number of imputations is low, using MI can be extremely unstable.

2.2 Other Notes

None

3. The Handling of Missing Data in Molecular Epidemiology Studies

Published Date: August 2011
Authors: Manisha Desai, Jessica Kubo, Denise Esserman, Mary Beth Terry
URL: https://cebp.aacrjournals.org/content/20/8/1571.full
Cited by: 21

3.1 Overview

- To understand how missing values were addressed in the field and how prevalent the missing data is, this article investigates numerous epidemiological studies.
  - It was found that "missing data methods are underutilized in molecular epidemiology studies"
- Common method is the complete-case (CC) analysis: excluding subjects with any missing values.
- The papers that used a missing data method (13%) used a combination of single imputation and missing data indicators
- Single Imputation Advantages:
  - only one set of imputations generated, standard complete data methods can be applied, and the ability to integrate the investigator's knowledge
- Single Imputation Disadvantages:
  - the one input value does not take into consideration sampling variability of the value itself or the variability corresponding to multiple models
- Lists MI as a potential practical solution
- MI Advantages:
  - replies on "assumption of missingness that is more flexible than that of CC and closer to what is expected from a typical molecular epidemiology study"
  - can accommodate missing data on one or more variables at the same time
  - methods are already available on SAS, SPSS, STATA, and R
  - auxiliary variables can be incorporated - estimation is better.
- MI Disadvantages:
  - requires "careful specification of imputation model" - more effort to the user
  - "varying performance of imputation methods"
  - "varying answers for each application of MI"
  - "need for specialized software"

3.2 Other Notes

Only cited by 21 - is this sufficient?

4. Unpredictable Bias when using the Missing Indicator Method or Complete Case Analysis for Missing Confounder Values: An Empirical Example

Published Date: July 2010
Authors: Mirjam J.Knol, Kristel J.M.Janssen, A. Rogier T.Donders, Antoine C.G.Egberts, E. Rob Heerdink, Diederick E.Grobbee, Karel G.M.Moons, Mirjam I.Geerlings
URL: https://www.sciencedirect.com/science/article/pii/S0895435610000181?via%3Dihub
Cited by: 129

4.1 Overview

- This paper looks at 2 methods of handling missing data: missing indicator method (MIM), complete case analysis (CC), and compares the degree and direction of bias in estimating the effect to that of multiple imputation (MI).
- When missing values are completely random, MIM overestimated the odds ratio - CC and MI gave unbiased results
- When missing values depended on observed values, MIM and CC over- or underestimated
- "Bias increased with increasing percentage of missing values"
- They found that MIM gave an unpredictable direction of bias and explained why MIM should not be used to handle missing confounder data, and stated that when missing values are completely random, CC can be applied but its use loses statistical power

4.2 Other Notes

None

5. Multiple Imputation for Missing Laboratory Data: An Example from Infectious Disease Epidemiology

Published Date: December 2009
Authors: Zuber D Mulla, Byungtae Seo, Ramaswami Kalamegham, Bahij S Nuwayhid
URL: https://www.sciencedirect.com/science/article/pii/S1047279709002853?via%3Dihub
Cited by: 14

5.1 Overview

- This paper provides an application of multiple imputation (MI) to a dataset of patients hospitalized for invasive group A streptococcal infections to non-statisticians to show that it is a valid method to address missing values.
- It is common practice in clinical epidemiology to perform a complete-subject analysis, ie. just delete anything that has missing values
- Imputation involves using observed data to predict and fill in missing values - MI replaces a missing value with a group of possible values

- “Other novel missing data methods such as likelihood based approaches, weighting methods, and MI using an iterative hot-deck with distance-based donor selection should be explored by epidemiologists”

5.2 Other Notes

- From 2009, slightly dated but I still think it provides good information
- Only cited by 14 - is this sufficient?

6. [Additional] Automating Data Analysis Methods in Epidemiology

Published Date: January 2019
Authors: George Choueiry, Pascale Salameh
URL: https://www.researchgate.net/publication/330015778_AUTOMATING_DATA_ANALYSIS_METHODS_IN_EPIDEMIOLOGY
Cited by: 2

6.1 Overview

- Medical research often contains statistical errors that could be prevented with appropriate data analysis
- Using machine learning to predict normality of a distribution, they created a package (White) to automate data analysis and out-perform Shapiro-Wilk test.
- Used R package missForest to use random forests to impute missing values
- Their package additionally allows users to replace numerical outliers with missing values and the software suggests options to handle categorical outliers

6.2 Other Notes

- Not heavily cited
- Not super relevant in content, I just thought this was an interesting idea/tangential concept

Summary and Additional Notes:

- In practice, a number of common practices exist, including complete-case (CC) analysis, which consists of just deleting the entries with missing values, which leads to errors in analysis.
- Other common methods include missing indicator method (MIM) and single value imputation.
- The seemingly best and most common method presented was the missing indicator method (MI) method (or other likelihood based methods); however, according to Paper #2, it is shown that it is not required on missing values prior to longitudinal data analysis.
- Additionally, Paper #6 uses random forest R package to impute missing values.