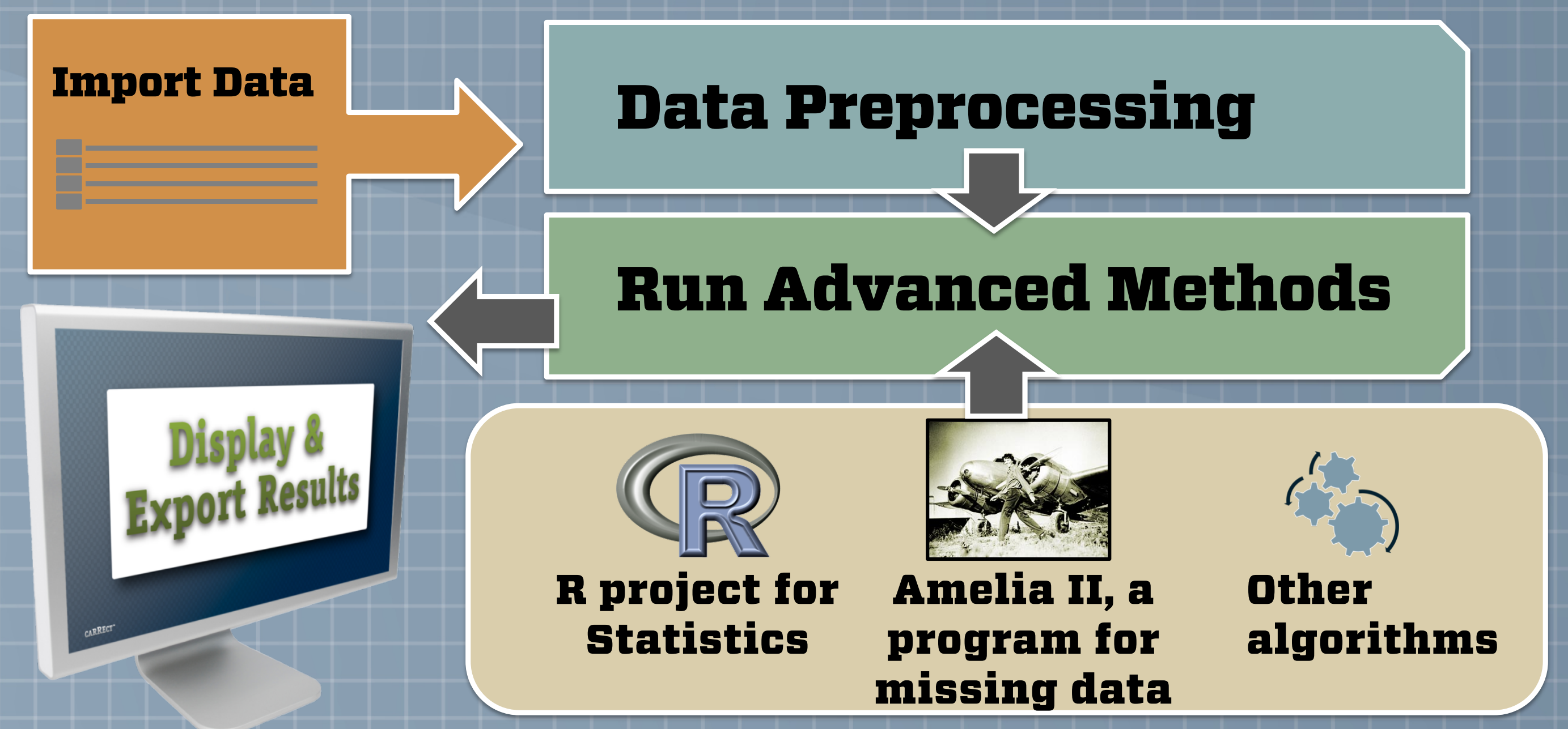
Bias Reduction in Statistical Analyses of Incomplete Data Sets

Jonathan Lansey, Paul Picciano, Ph.D., Ian Yohai, Ph.D., Robert McCormack, Ph.D. - Aptima, Inc.
Fred Grant, Ph.D., Robert Gern, Ph.D - Northrop Grumman Corp.

Introduction

Analyses produced by epidemiologists and public health practitioners are susceptible to bias from a number of sources. It often requires a great deal of expertise to understand and apply the multitude of statistical methods that are avaialble to diagnose and correct for these biases. To address this challenge, Aptima began development of CARRECT, the Collaborative Automation Reliably Remediating Erroneous Conclusion Threats system. When complete, CARRECT will support statistical bias reduction and improved analyses and decision making by engaging the user in a collaborative process in which the technology is transparent to the analyst.

Overview of CARRECT system architecture



Import Data

Data Preprocessing

Run Advanced Methods

Display & Export Results

R project for Statistics

Amelia II, a program for missing data

Other algorithms

Methods

Older approaches to imputing missing data, including mean imputation and single imputation regression methods, have steadily given way to a class of methods known as multiple imputation (MI).

Under MI, the missing values are drawn multiple times, generating m complete datasets along with the estimated parameters of the model (Rubin 1987; King et al. 2001). As a result, MI will lead to valid standard errors and confidence intervals along with unbiased point estimates. CARRECT uses a bootstrapping-based MI algorithm (Honaker and King, 2010) that gives essentially the same answers as the standard Bayesian Markov Chain Monte Carlo (MCMC) or Expectation Maximization (EM) approaches, is significantly faster, and can handle larger datasets.

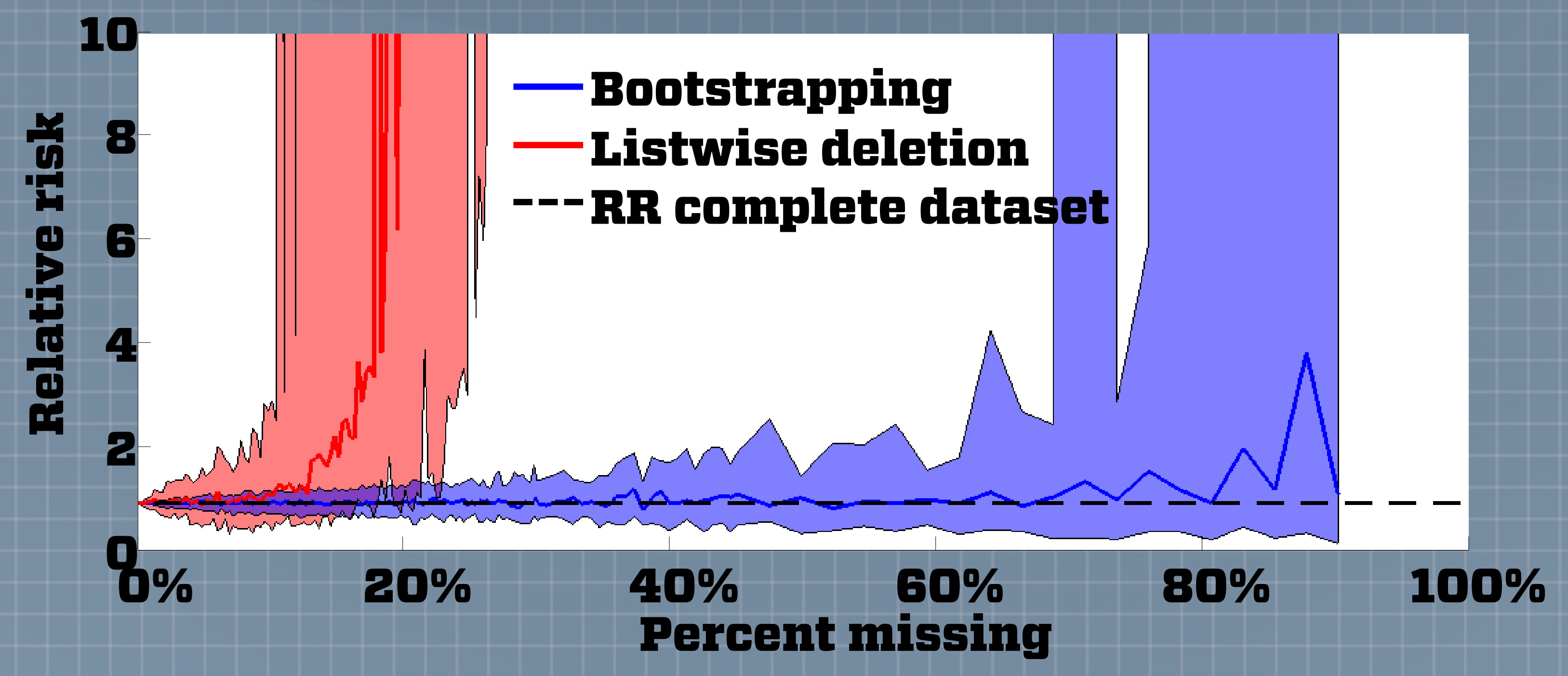
Results

Tests of the bootstrapping MI algorithm were conducted with an epidemiological dataset from the Integrated Health Interview Series (IHIS) producing verifiably unbiased results despite high missingness rates.

Procedure:

- Created "complete" dataset. With 13 variables (selection shown at right).
- Removed increasing amounts of data completely at random
- Performed Listwise Deletion and Bootstrapping Multiple Imputation (MI)
- Calculated Relative Risk (RR) for coronary heart disease, given that the subject drinks.
- Repeated the imputation and RR calculation 50 times for each level of missingness from 1-90%

Name	Description
AGE	Age
WEIGHT	Weight in pounds without clothes or shoes
BMI	Body mass index
HEIGHT	Height
ALC1YR	Ever had 12+ drinks in any one year
WALKFUNX	Times walked for leisure during past 7 days
......	......
CHEARTDIEV	Ever told had coronary heart disease

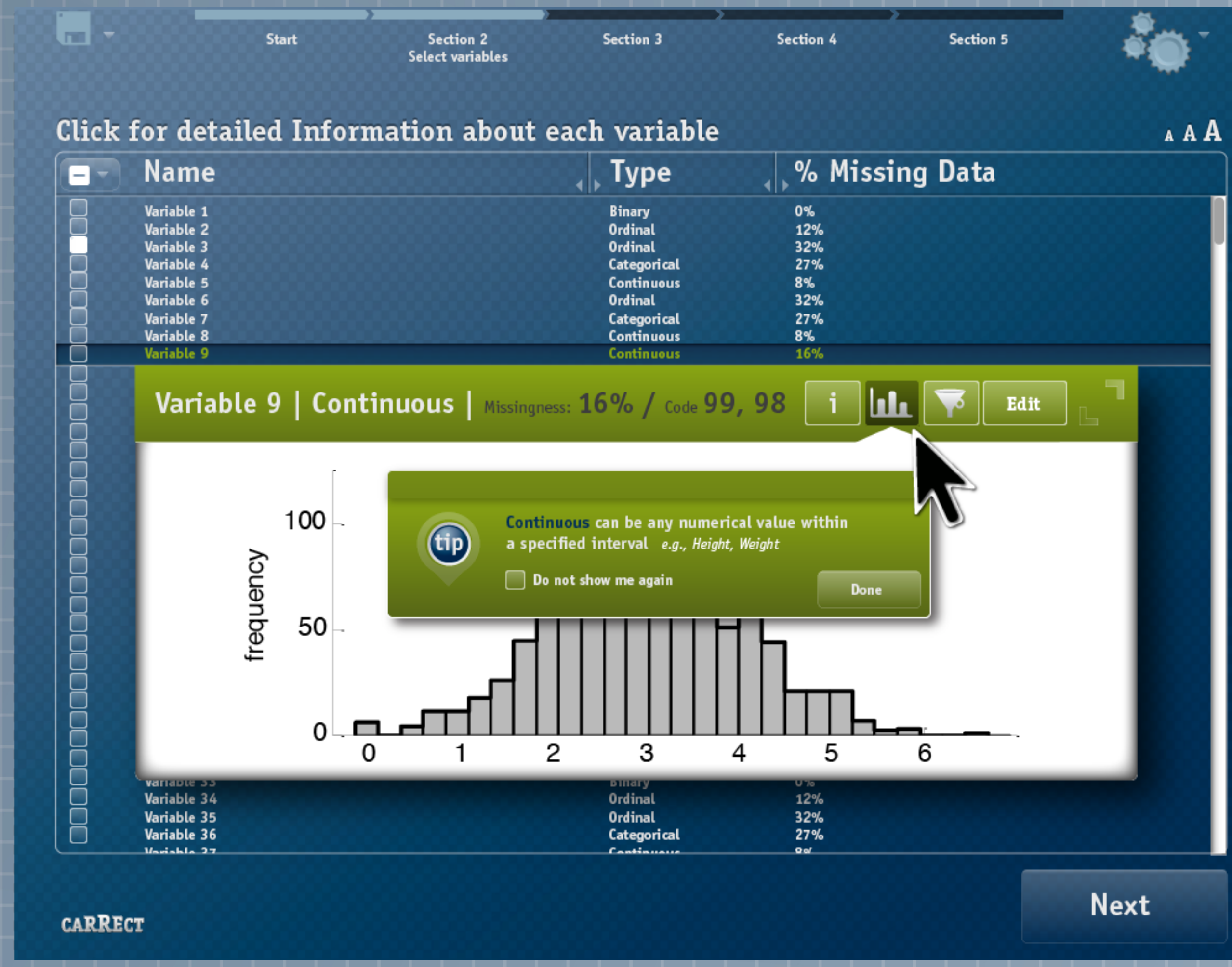


The solid lines show the median RR over the 50 iterations for each missingness rate and the shaded regions show the 95% bounds for those 50 iterations. Bootstrapping MI provides results with greatly reduced bias, and has fewer computational demands than other methods.

User Interface Mockups

An intuitive data wizard guides the user through the analysis processes by analyzing key features of a given dataset. The wizard shows prompts the user to provide additional substantive knowledge to improve the handling of imperfect datasets.

The figure below shows how a user could select a variable type.



Click for detailed Information about each variable

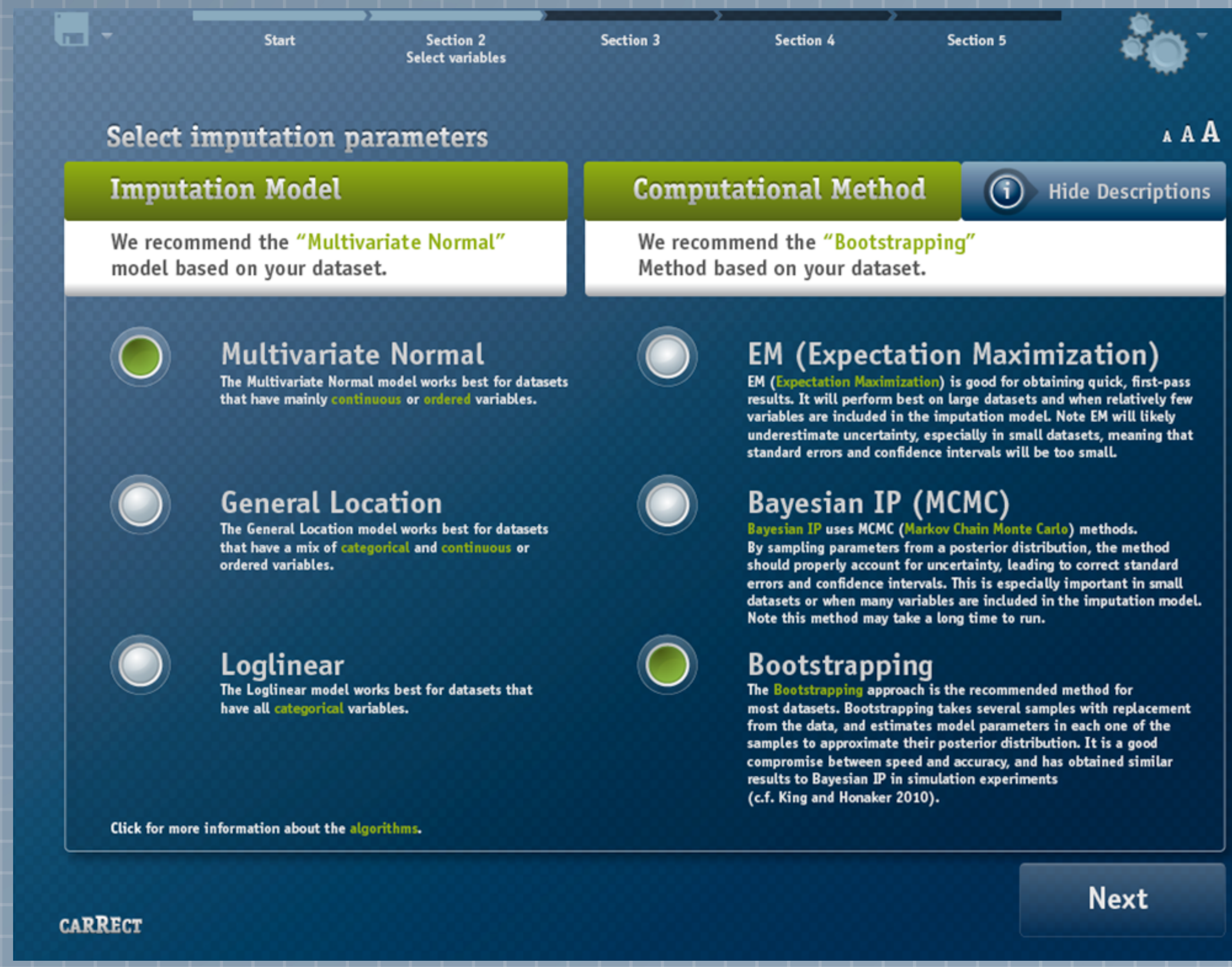
Name	Type	% Missing Data
Variable 1	Binary	0%
Variable 2	Ordinal	12%
Variable 3	Ordinal	32%
Variable 4	Categorical	27%
Variable 5	Continuous	8%
Variable 6	Ordinal	32%
Variable 7	Categorical	27%
Variable 8	Continuous	8%
Variable 9	Continuous	16%
Variable 34	Ordinal	12%
Variable 35	Ordinal	32%
Variable 36	Categorical	27%

Variable 9 | Continuous | Missingness: 16% / Code 99, 98

tip Continuous can be any numerical value within a specified interval e.g., Height, Weight

□ Do not show me again

The figure below shows how a user is helped along with the selection of the most appropriate algorithms and models for the data



Select imputation parameters

Imputation Model

We recommend the "Multivariate Normal" model based on your dataset.

● Multivariate Normal
The Multivariate Normal model works best for datasets that have mainly continuous or ordered variables.

○ General Location
The General Location model works best for datasets that have a mix of categorical and continuous or ordered variables.

○ Loglinear
The Loglinear model works best for datasets that have all categorical variables.

Computational Method

We recommend the "Bootstrapping" Method based on your dataset.

○ EM (Expectation Maximization)
EM (Expectation Maximization) is good for obtaining quick, first-pass results. It will perform best on large datasets and when relatively few variables are included in the imputation model. Note EM will likely underestimate uncertainty, especially in small datasets, meaning that standard errors and confidence intervals will be too small.

○ Bayesian IP (MCMC)
Bayesian IP uses MCMC (Markov Chain Monte Carlo) methods. By sampling parameters from a posterior distribution, the method should properly account for uncertainty, leading to correct standard errors and confidence intervals. This is especially important in small datasets or when many variables are included in the imputation model. Note this method may take a long time to run.

● Bootstrapping
The Bootstrapping approach is the recommended method for most datasets. Bootstrapping takes several samples with replacement from the data, and estimates model parameters in each one of the samples to approximate their posterior distribution. It is a good compromise between speed and accuracy, and has obtained similar results to Bayesian IP in simulation experiments (c.f. King and Honaker 2010).

Click for more information about the algorithms.

Conclusion

Our approach and program were designed to make bias mitigation much more accessible to much more than only the statistical elite. We hope that it will have a wide impact on reducing bias in epidemiological studies and provide more accurate information to policymakers.

Contacts: jlansey@aptima.com, ppicciano@aptima.com, iyohai@aptima.com.

Acknowledgements:

This material is based upon work supported by the Walter Reed Army Institute of Research (WRAIR) under Contract No. W81XWH-11-C-0505. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the WRAIR.