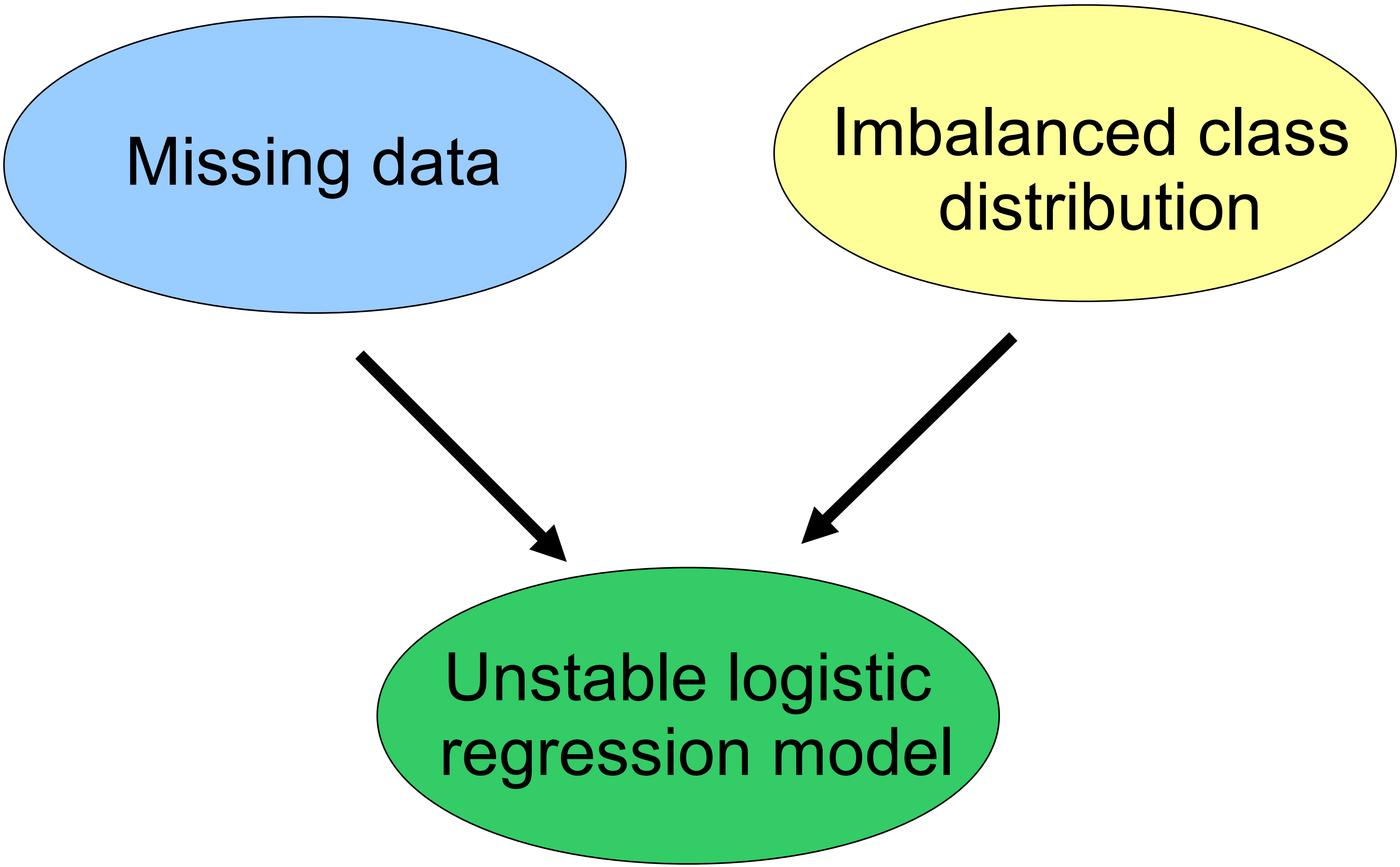
Building a more stable predictive logistic regression model

Anna Elizabeth Campain

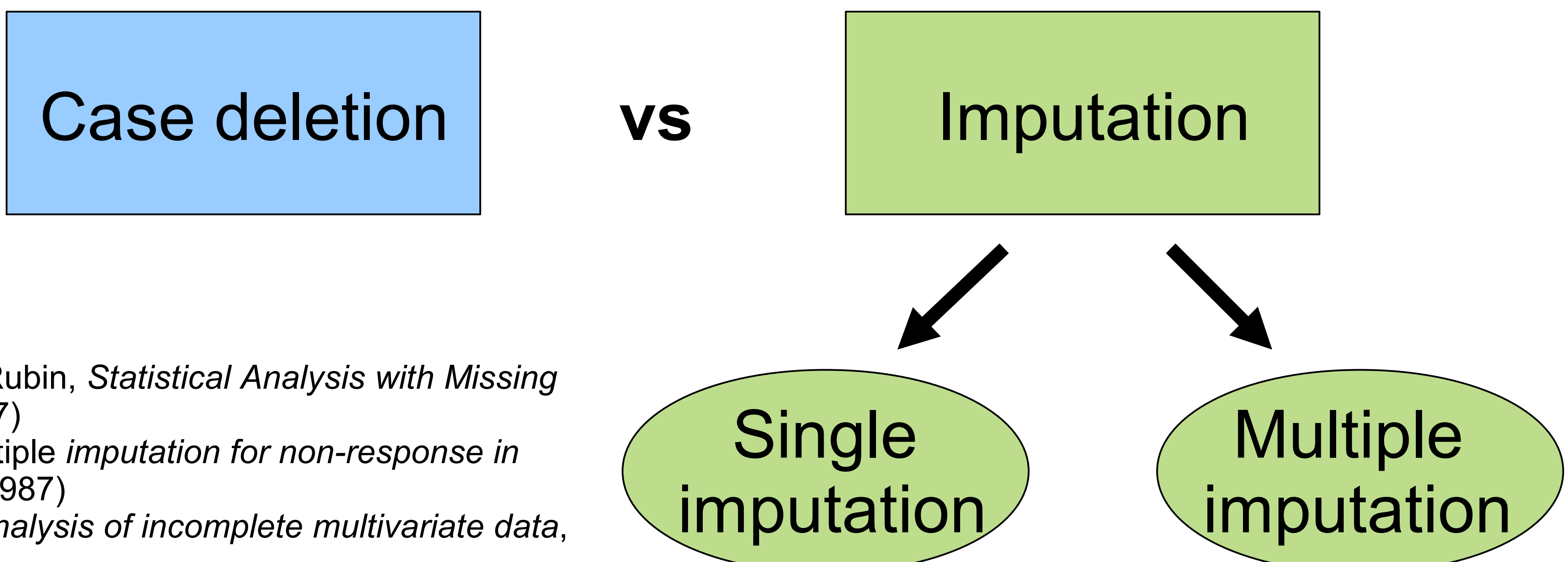
The University of Sydney

Common problems when working with clinical data



Missing data

- Rubin (1987), Little and Rubin (1987), Schafer (1997)
- Consider the missing data structure (MCAR, MAR, MNAR)



Little and Rubin, *Statistical Analysis with Missing Data*, (1987)
Rubin, Multiple *imputation for non-response in surveys*, (1987)
Schafer, *Analysis of incomplete multivariate data*, (1997)

Some imputation methods

Available for R:

Norm, Cat and Mix *(Schafer, 1997)*
AmeliaII *(Honaker et al, 2001)*
MICE *(Buuren and Oudshoorn, 1999)*
Mi *(Gelman et al, 2009)*
Pan *(Schafer, 2000)*

Stand-alone:

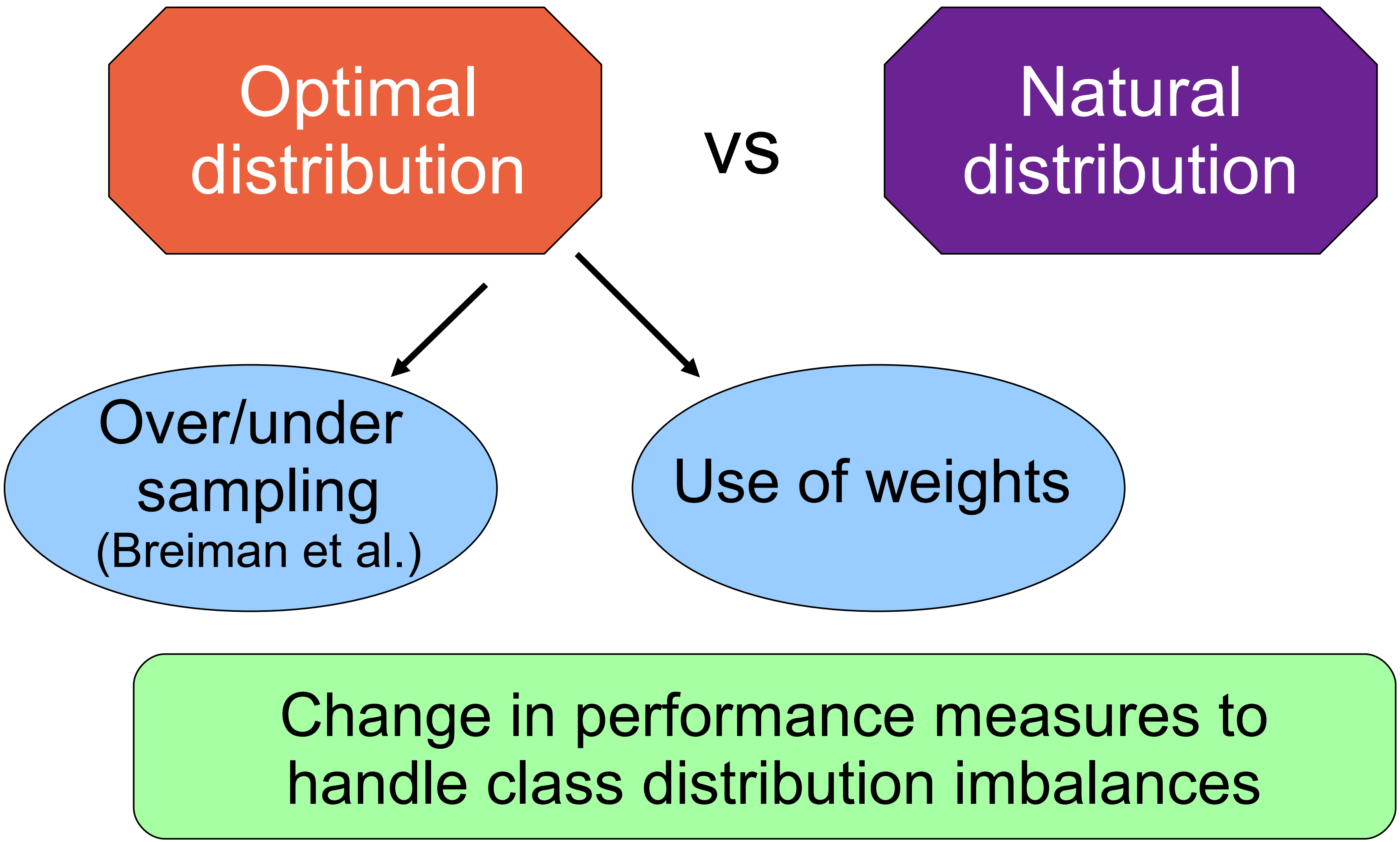
AmeliaII *(Honaker et al, 2001)*
IVEware *(Raghunathan at al. 2001)*

Available for SAS

IVEware (Raghunathan at al. 2001)

R software: http://cran.r-project.org/
AmeliaII:http://gking.harvard.edu/stats.shtml
IVEware: http://www.isr.umich.edu/src/smp/ive/

Imbalanced class distribution



Weiss and Provest, *The effect of class distribution on classifier learning*, (2001)
Breiman, Friedman, Stone and Olshen, *Classification and Regression Trees*, (1984)

Medical/Clinical motivation

- Nepean Early Pregnancy Clinic – Nepean Hospital, Penrith, NSW Australia
- 416 patients, (33 miscarriages)
- Missingness per variable from 0 – 80%

Medical/Clinical motivation

- Nepean Early Pregnancy Clinic – Nepean Hospital, Penrith, NSW Australia
- 416 patients, (33 miscarriages)
- Missingness per variable from 0 – 80%

Aim:

To build a model which aids in the prediction of the first trimester outcome at the initial consultation

Variable missingness

91 Variables

↓

21 Variables

Care was taken to ensure no depletion in 'miscarriage' cases

Remove:

- Redundant/non-informative variables
- Categorical variables with too small sample sizes
- Any variables with missingness greater than 25%

Include: (After expert opinion)

- Subchronic bleed variable (55% missingness)

Existing methods

Case deletion → Exacerbates small sample size issue, leaving only 15%, (miscarriages=7)

Single imputation → Under estimates variability inherent in post-imputation model (Rubin 1987)

Multiple imputation → In this case still produces an unstable model

Unstable models

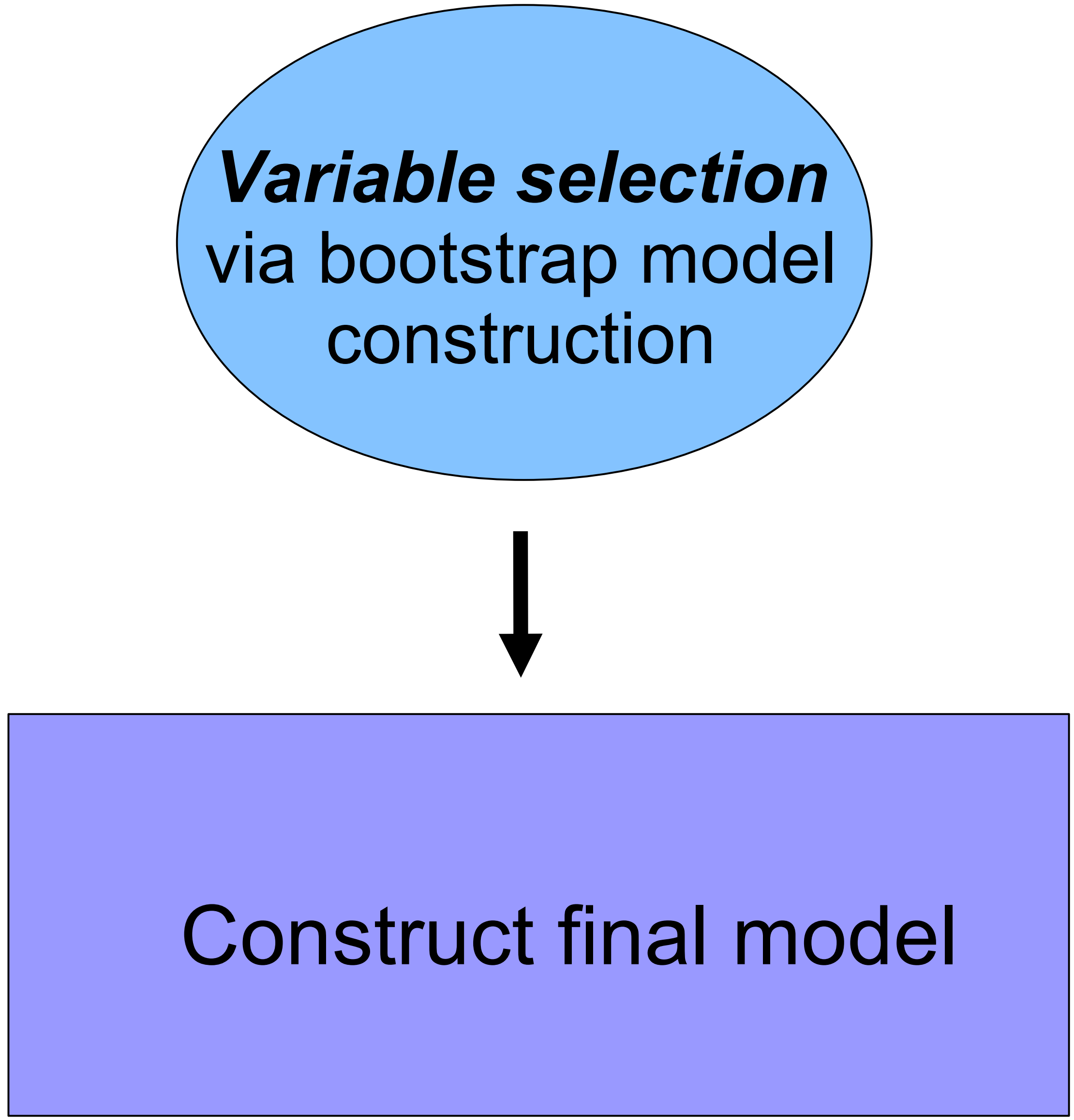
1st Run

$$\ln\left(\frac{\pi_k}{1-\pi_k}\right) = 2.71 \times \text{Clots} - 0.051 \times \text{Foetal heart rate} - 1.31 \times \text{Consistent with menstrual dates}$$

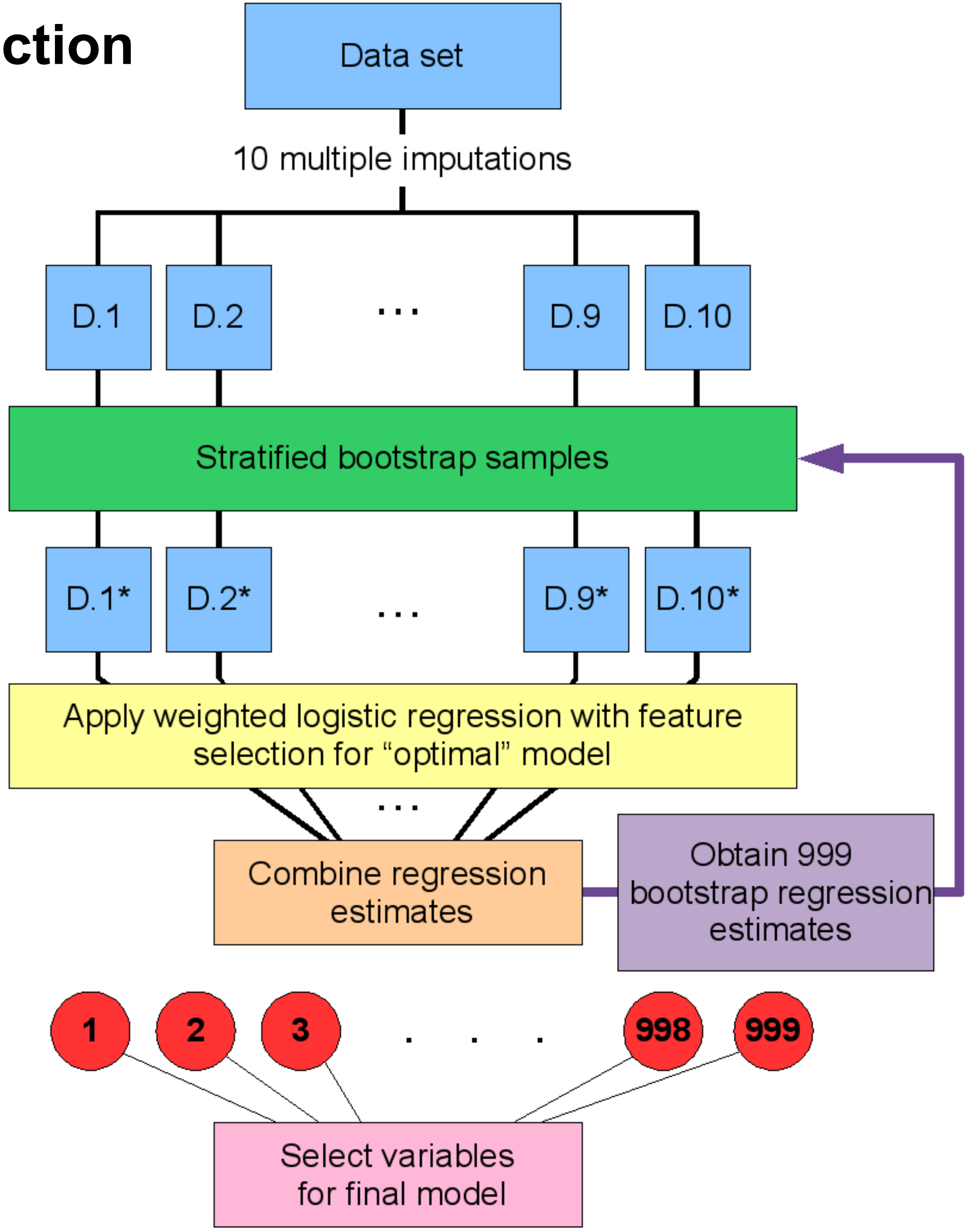
2nd Run

$$\ln\left(\frac{\pi_k}{1-\pi_k}\right) = 2.34 \times \text{Clots} - 0.12 \times \text{GS mean}$$

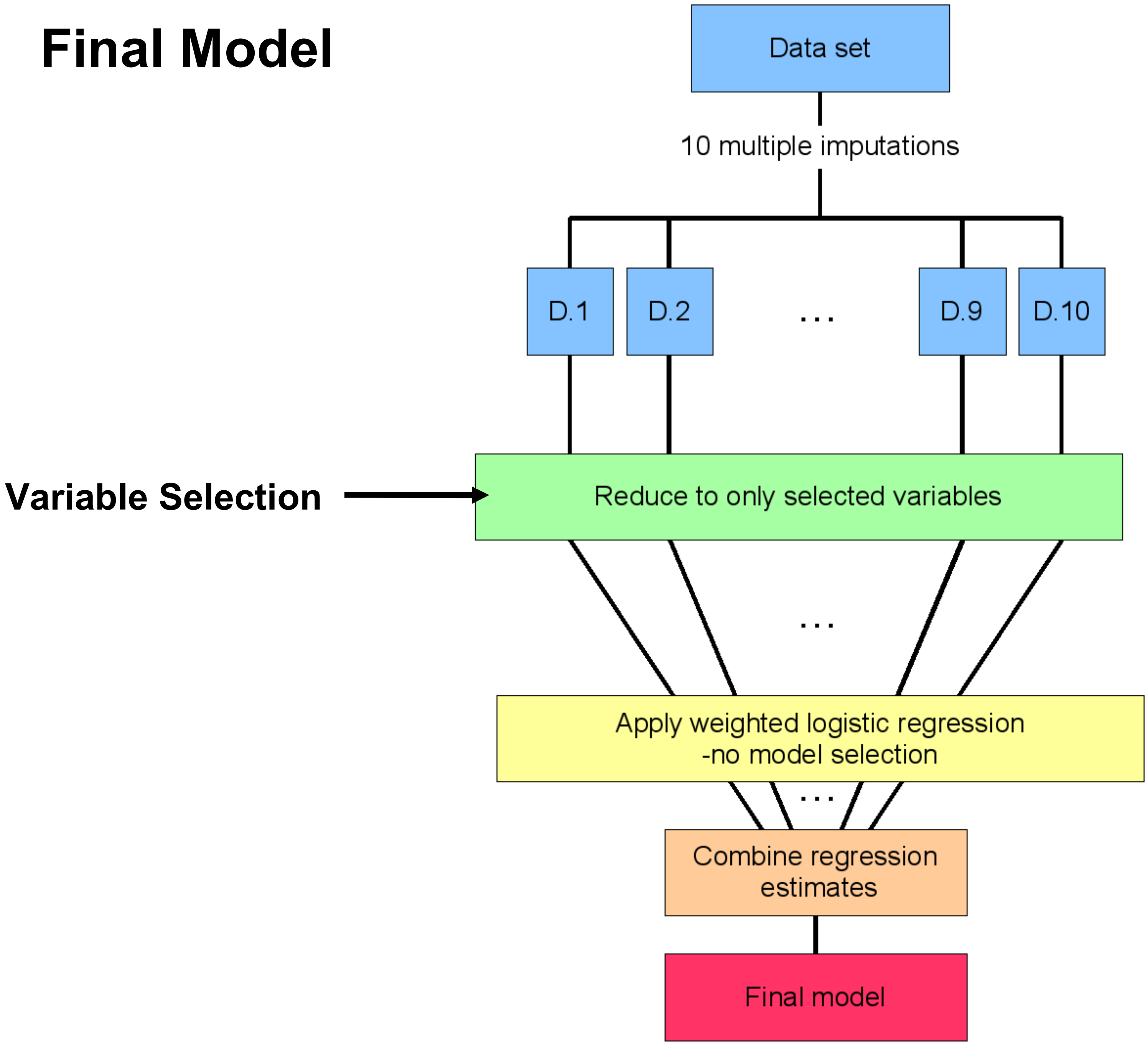
A solution to the 'instability problem'



Variable Selection

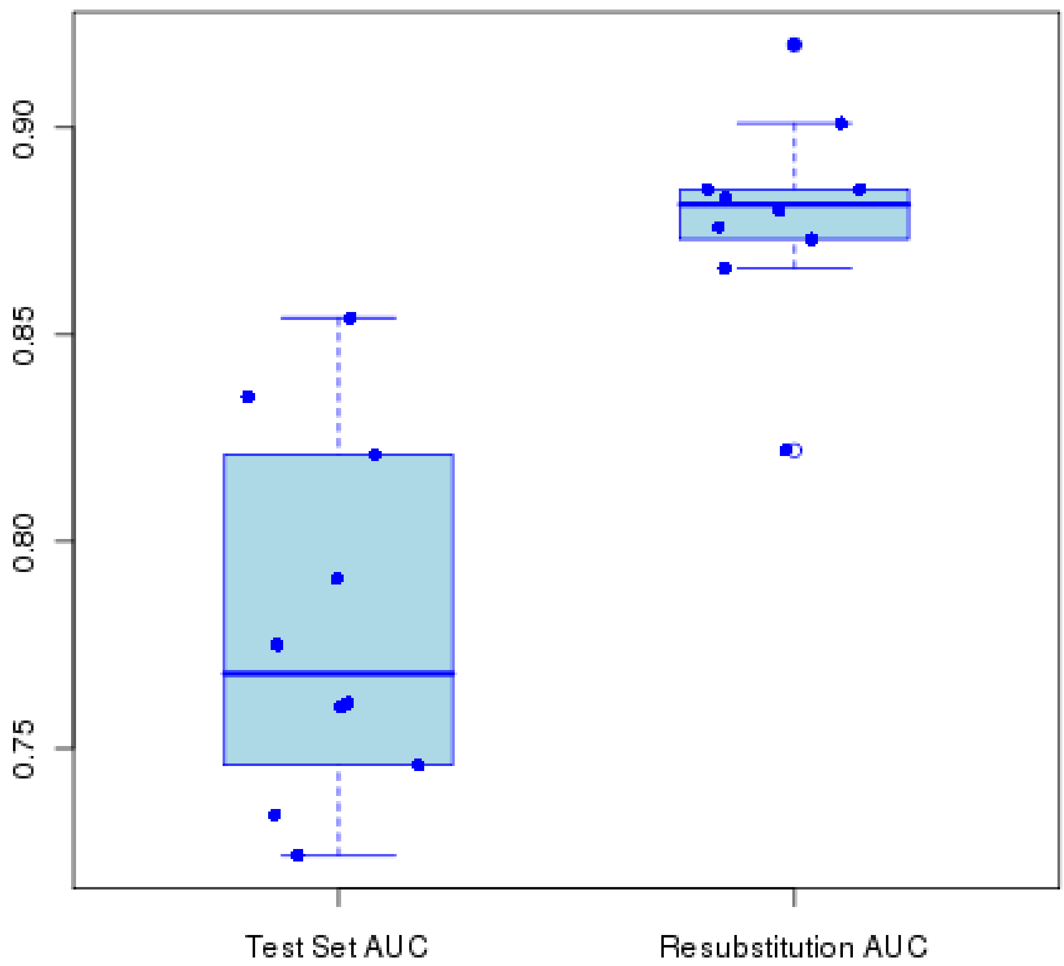


Final Model



Data set
10 multiple imputations
D.1
D.2
. . .
D.9
D.10
Variable Selection
Reduce to only selected variables
. . .
Apply weighted logistic regression
-no model selection
. . .
Combine regression
estimates
Final model

Results



- 10 random test/training set splits.
- Area under the receiver operative characteristic curve was calculated as a predictive measure.

Variable	Odds Ratio
LSCS	0.44
Gestational age days	1.05
Bleeding	1.93
Clots	6.12
USS gestational age days	0.91
Consistent with menstrual dates	0.50
GS mean	0.88
YS mean	1.54

How much missingness is too much missingness?

Contrast

Acuña et al. - "*1-5% is manageable, 5-15% require sophisticated methods... more than 15% may severely impact any kind of interpretation*"

with

Zhang et al. - *Compare results with missingness up to 80%*

Acuna and Rodriguez, *Classification, Clustering and Data Mining Applications* (2004)
Zhang, Qin, Ling and Sheng, *IEEE Transactions in knowledge and data engineering* (2005)

How much missingness is too much missingness?

Contrast

Acuña et al. - "*1-5% is manageable, 5-15% require sophisticated methods... more than 15% may severely impact any kind of interpretation*"

with

Zhang et al. - *Compare results with missingness up to 80%*

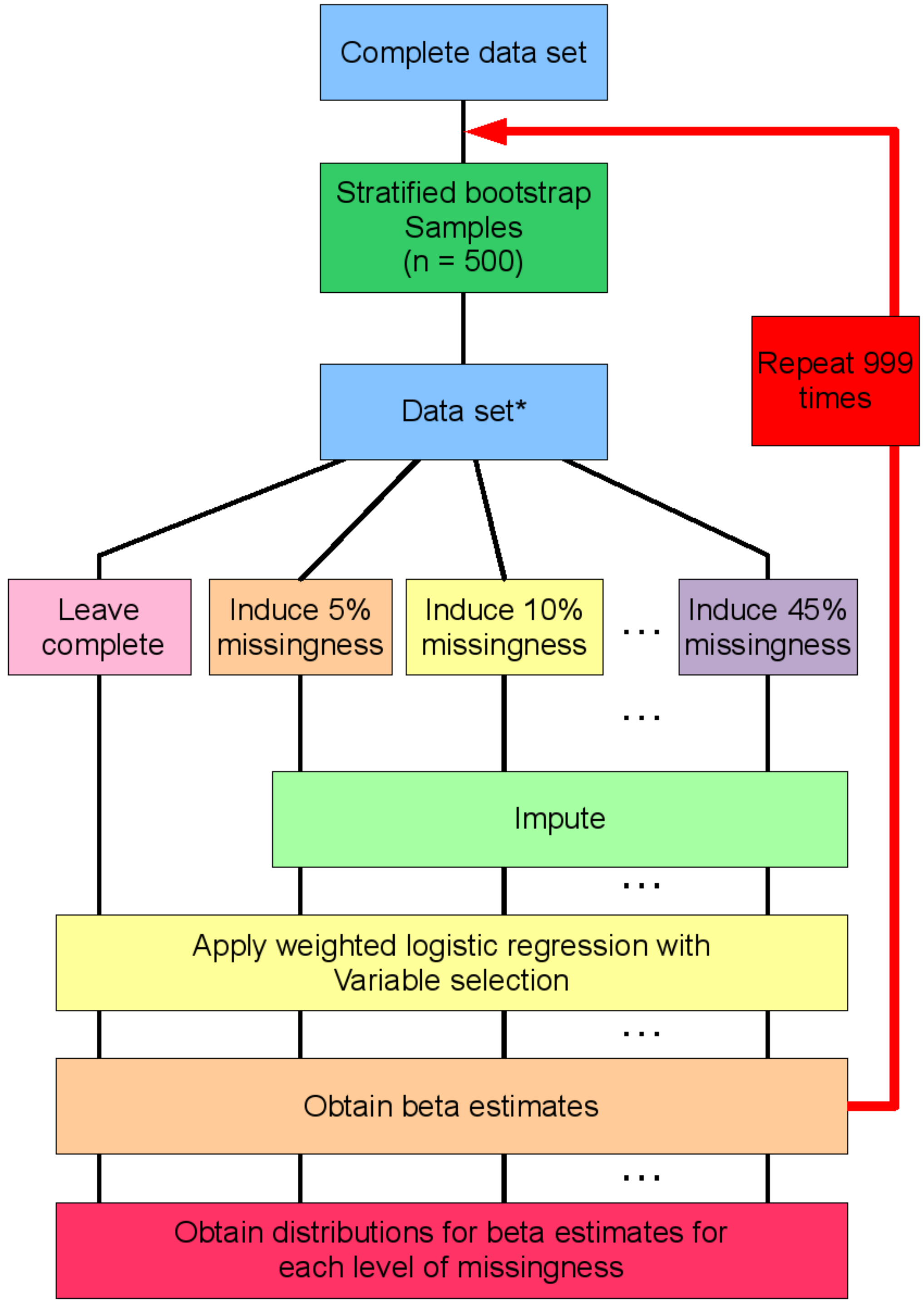
Is there a point where missingness is too great, and imputation is not appropriate?

Acuna and Rodriguez, *Classification, Clustering and Data Mining Applications* (2004)
Zhang, Qin, Ling and Sheng, *IEEE Transactions in knowledge and data engineering* (2005)

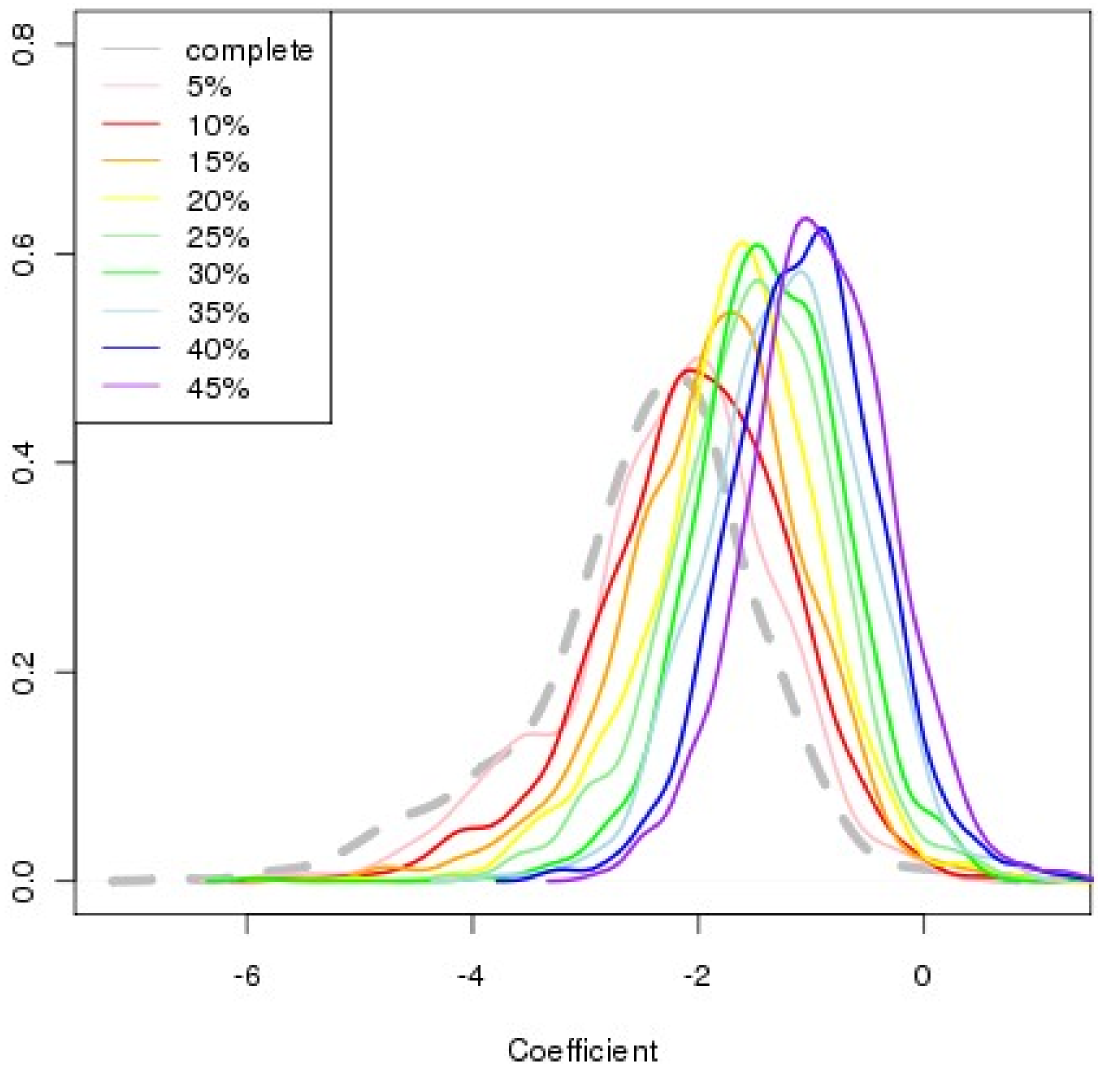
270 samples
(8% miscarriages)

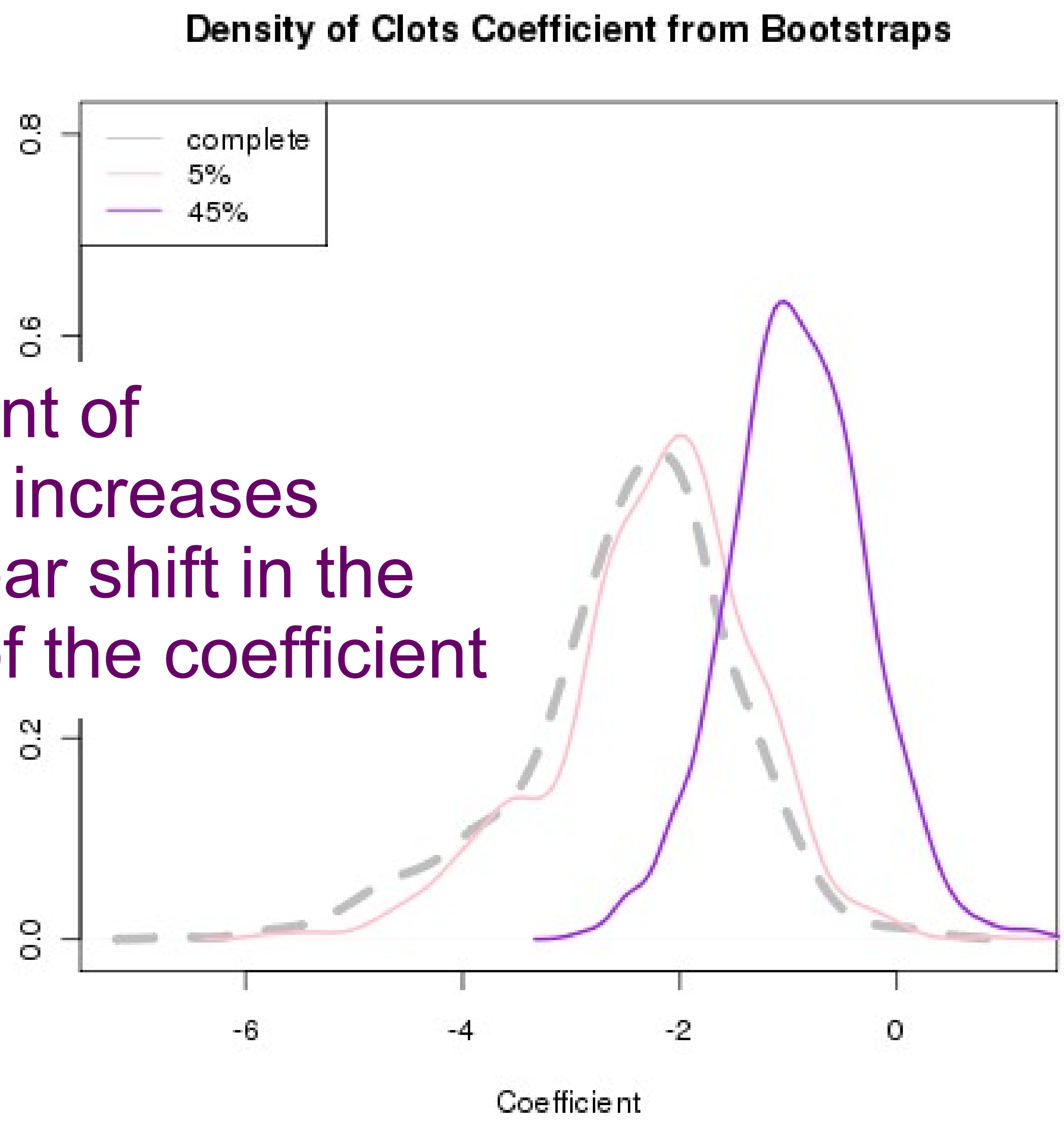
Variables:

- Age
- NVD
- Miscarriages
- Gestational Age
- Bleeding
- Clots
- Smoker
- CRL
- GS Mean
- FHR
- Consistent Dates

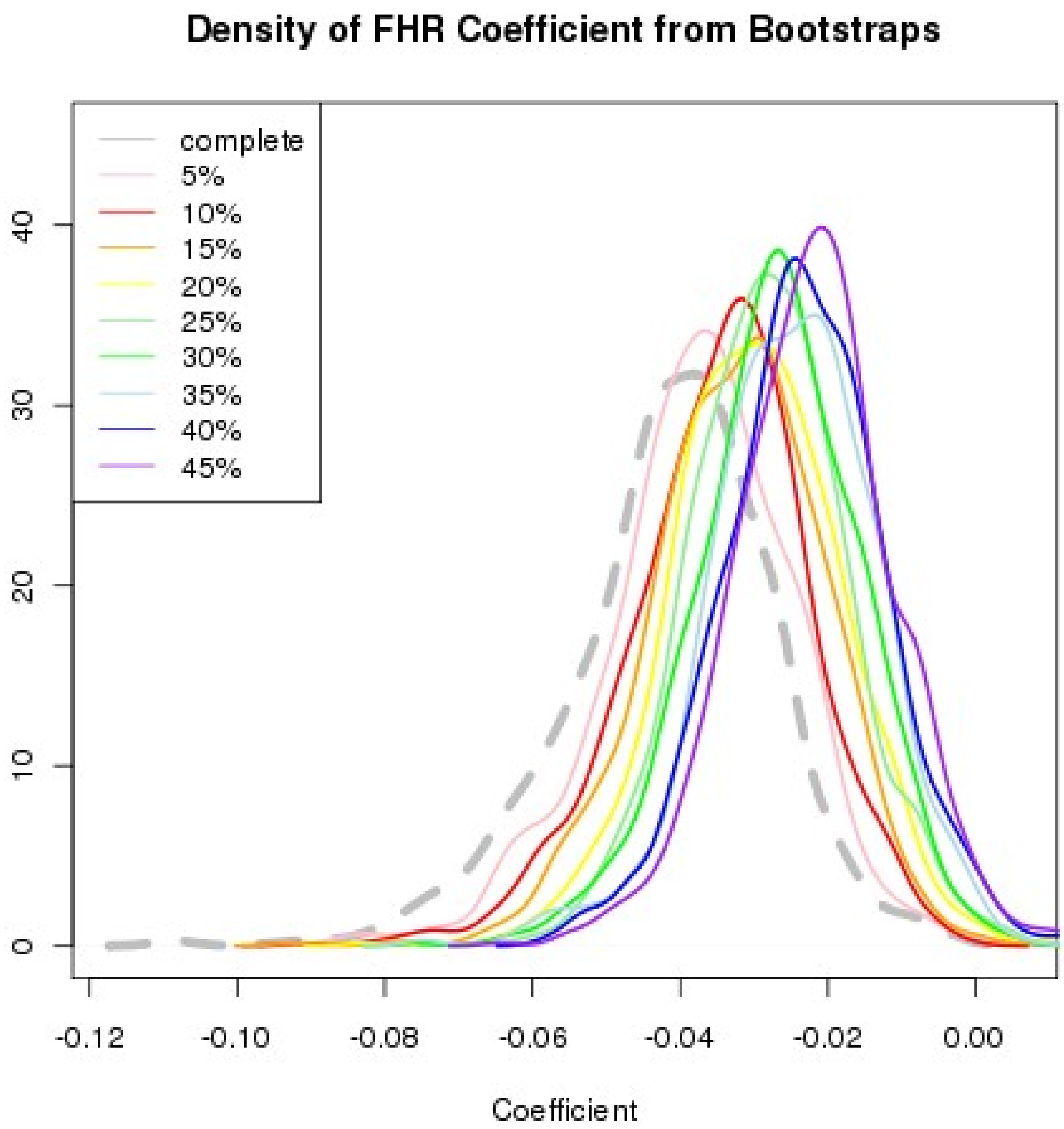


Density of Clots Coefficient from Bootstraps

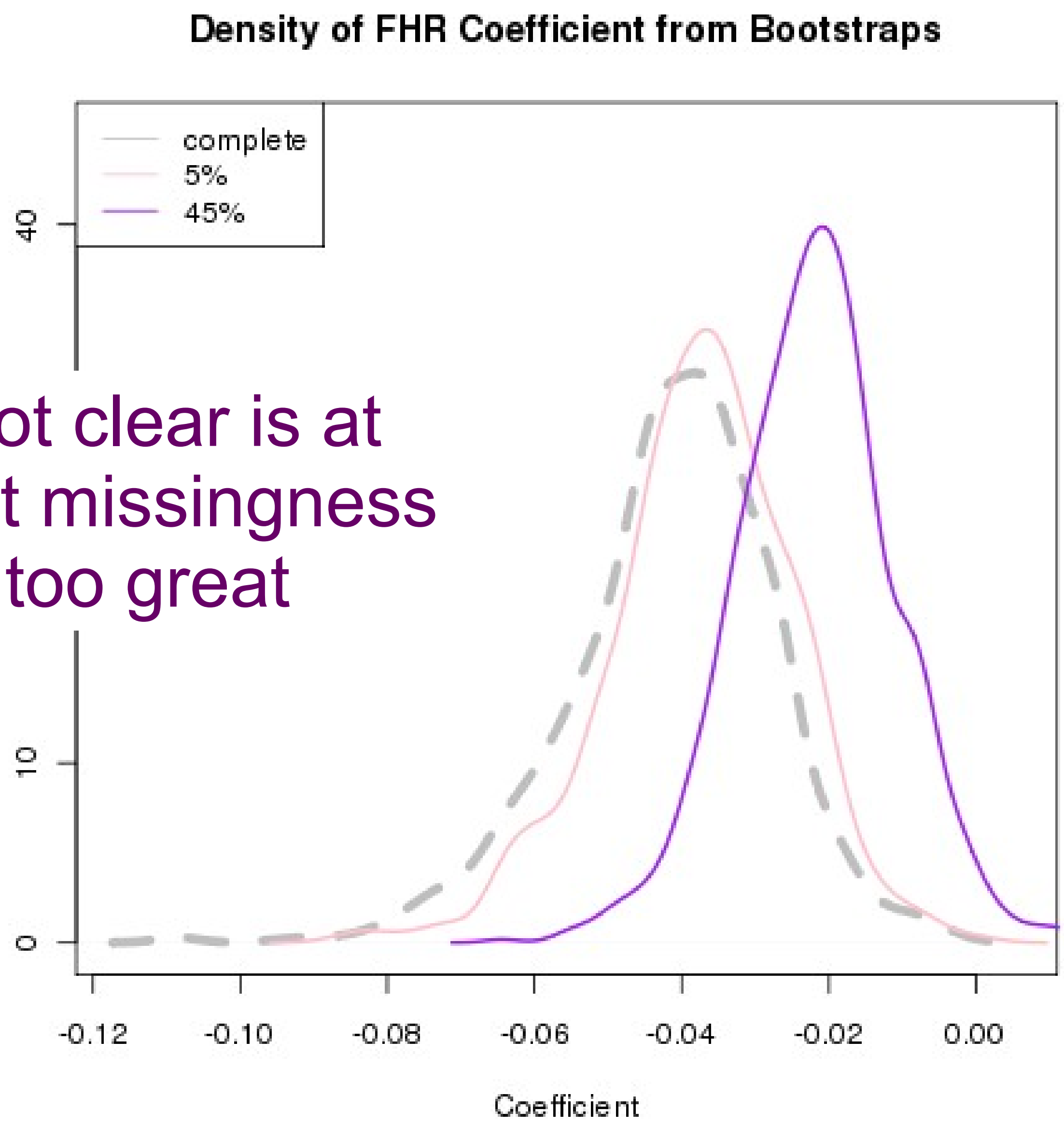




As the amount of missingness increases there is a clear shift in the distribution of the coefficient



Density of FHR Coefficient from Bootstraps
complete
5%
10%
15%
20%
25%
30%
35%
40%
45%
0
10
20
30
40
-0.12
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
Coefficient



What is not clear is at what point missingness becomes too great

Summary

Missingness and uneven class distributions contribute to unstable models – bootstrapping variable selection procedures can aid in overcoming this problem.

Amount of missingness is important to consider

Be considerate of potential problems when considering variables with large amounts of missingness

Special Thanks

- PhD Supervisors:
  - Dr Jean Yang
  - Dr Samuel Müller

The University of Sydney

- Team at Nepean Early Pregnancy Clinic
  - Dr George Condous
  - Dr Jennifer Riemke
  - And others

Funding

APA ARC Biometrics

References

- Acuna and Rodriguez, *Classification, Clustering and Data Mining Applications* in The Treatment of missing values and its effect on the classifier accuracy, page 639-648, 2004.
- Amelia-1.2-12 – R Software, 18th July 2009
- Breiman, Friedman, Stone and Olshen, *Classification and Regression Trees*, 1984.
- Buuren and Oudshoorn, Flexiable multivariate imputation by mice, *Leiden:TNO Preventieen Gezondheid*, TNO/VGZ/PG 99.054, 1999
- Honaker, Joseph and Scheve, and Singh, *Amelia: A program for missing data*, Harvard University, Cambridge, MA, 2001, Software
- King, Honaker, Joseph and Scheve, Analysing incomplete political science data: an alternative algorithm for multiple imputation, *American Political Science Review,* 95(1):49-69, 2001
- Little and Rubin, *Statistical Analysis with Missing Data*, 1987
- Raghunathan, Solenberger and Hoewyk, IVEware: *Imputation and variance estimation software*, University of Michigan, Ann Arbor, MI, 2000, Software
- Rubin, Multiple *imputation for non-response in surveys*, 1987
- Schafer, *Analysis of incomplete multivariate data*, 1997.
- Schafer, *Multiple imputation with PAN*, 2000, Software
- Weiss and Provest, The effect of class distribution on classifier learning: An empirical study, *Technical Report* Department of Computer Science, Rutgers University, 2001.
- Zhang, Qin, Ling and Sheng, Missing is Useful:Missing Values in Cost-Sensitive Decision Trees, *IEEE Transactions in knowledge and data engineering* **17**(12) , 2005.