ASSESSING CORRESPONDENCE BETWEEN TWO DATA SOURCES ACROSS CATEGORICAL COVARIATES WITH MISSING DATA: APPLICATION TO ELECTRONIC HEALTH RECORDS

2018 American Statistical Association
Conference on Statistical Practice
Presenter: Emile Latour

KNIGHT CANCER *Institute*

Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data.

-Allison (2002)

Introduction

- Missing data complicates any analysis and needs to be addressed

- Ignoring the missing data or editing leads to problems
  - Inefficiency – loss of information leading to loss of power,
  - Systematic difference – leading to biased results, and
  - Unreliable results

- Values that are not available and that *would be meaningful* for analysis had they been observed

Introduction

- Electronic health records (EHRs) typically are not collected specifically for the purpose of answering research questions

- Susceptible to issues with missing data which may lead to
  - Bias results
  - Invalid conclusions

- Due to increased use and the increased volume of data, EHR will continue to be a vital source of patient information

Introduction

- Motivating example
  - 2014 study by Heintzman, et al. (the “original study”)
  - Assessed agreement of electronic health records (EHR) with Medicaid claims data for documentation of 11 preventive care procedures using kappa statistics
  - Aimed to validate the EHR data in order to ensure quality and accuracy
  - Population of continuously insured adult Medicaid recipients being served by a network of Oregon community health centers (CHCs) during 2011

Measuring agreement

- Compare two data sources
  - EHR
  - Medicaid claims
- Examine the level of agreement
  - How well does the documentation of preventive services *agree* between EHR and Medicaid claims
  - Medicaid claims billing data used as “gold standard”
- Independent single rating for each subject
    - Yes, received screening/procedure
    - No, did not receive screening/procedure

Measuring agreement

- The 2×2 table

		Claims data		
		Yes	**No**	**Total**
EHR data	**Yes**	a	b	m_1
	No	c	d	m_0
	Total	n_1	n_0	n

Research question

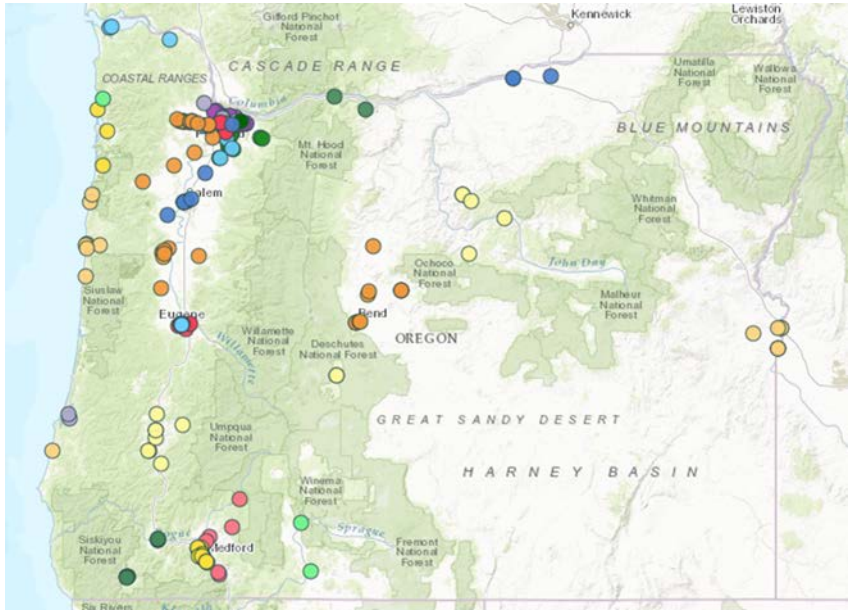
- The original study assessed agreement between the two data sources for all eligible patients

- Does agreement differ when patients are stratified by race and by federal poverty level?
  - Important to ensure the same quality of care is provided across demographic groups
  - Seek to assess any systematic differences among patients to ensure the same quality of care for all

Missing data

- Missing data on Race and Federal Poverty Level (FPL)
  - Opportunity to apply missing data methods
  - Multiple imputation
    - Multivariate imputation by chained equations, MICE (Stef Van Buuren)

Study data



- OCHIN provided all data
  - Nonprofit community health information network of over 300 community health centers (CHCs) in 13 states
- EHR data
  - Extracted for the year 2011 from data storage for 43 Oregon CHCs
- Medicaid claims data
  - Oregon's Medicaid program
  - Data for 2011 obtained 18 months after end of year to account for lag in processing

Table 1

	Patients appearing in both EHR and claims (N=13,101)	
	No.	%
Gender		
Female	8,600	65.6
Male	4,501	34.4
Race		
Asian/Pacific Islander	772	5.9
American Indian/Alaskan native	180	1.4
Black	1,409	10.8
White	9,720	74.2
Multiple Races	143	1.1
Unknown	877	6.7
Race, ethnicity		
Hispanic	1,186	9.1
Non-Hispanic, white	8,943	68.3
Non-Hispanic, other	2,441	18.6
Unknown	531	4.1
Primary Language		
English	10,927	83.4
Spanish	589	4.5
Other	1,585	12.1
Federal poverty level		
≤138% FPL	10,153	77.5
≥138% FPL	234	1.8
Missing/Unknown	2,714	20.7
Age in years (as of January 1, 2011)		
19–34	4,632	35.4
35–50	5,033	38.4
51–64	3,436	26.2
Mean (SD)	40.6 (12.3)	

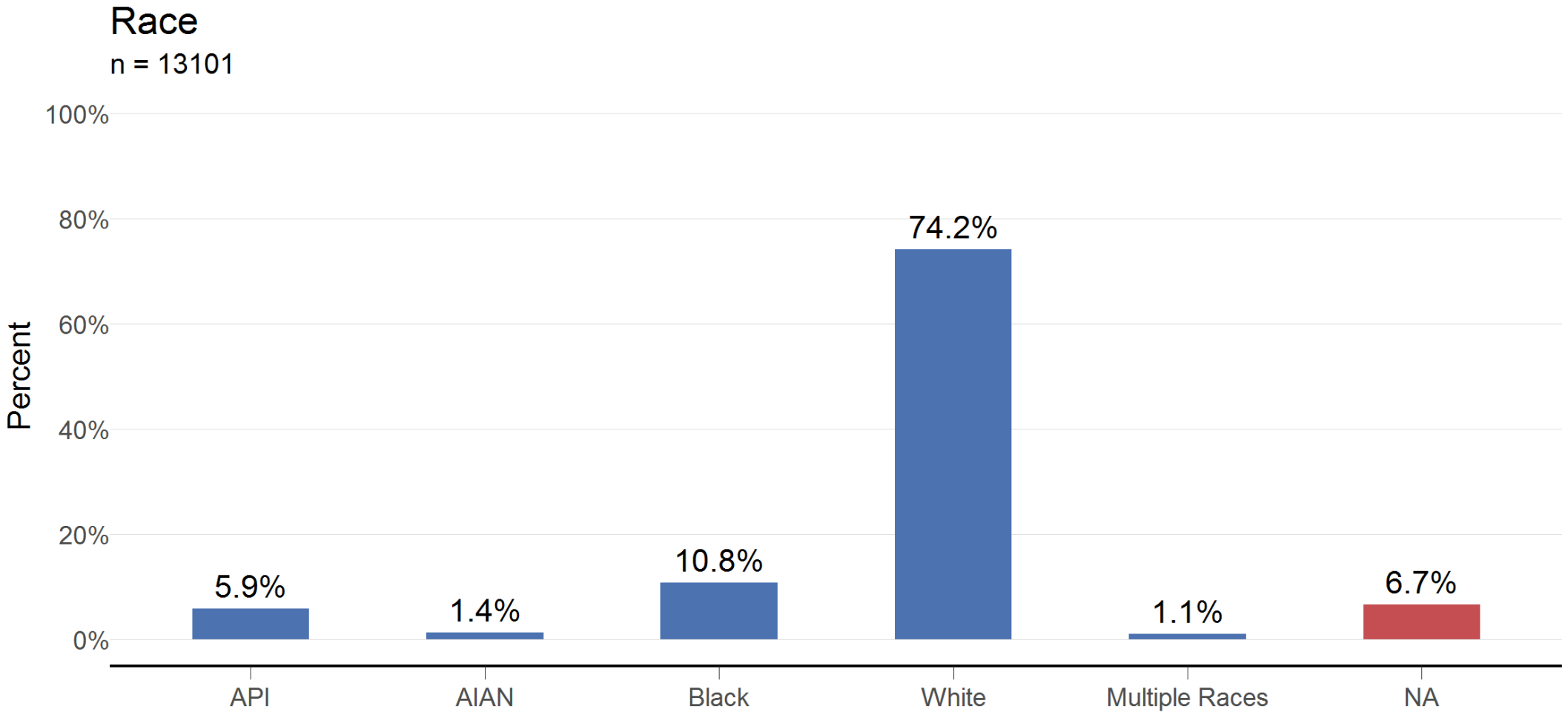
Variables

- Demographics
- Eligible for screening (Y/N)
- Screened in Medicaid data (Y/N)
- Screened in EHR data (Y/N)
- Primary department code

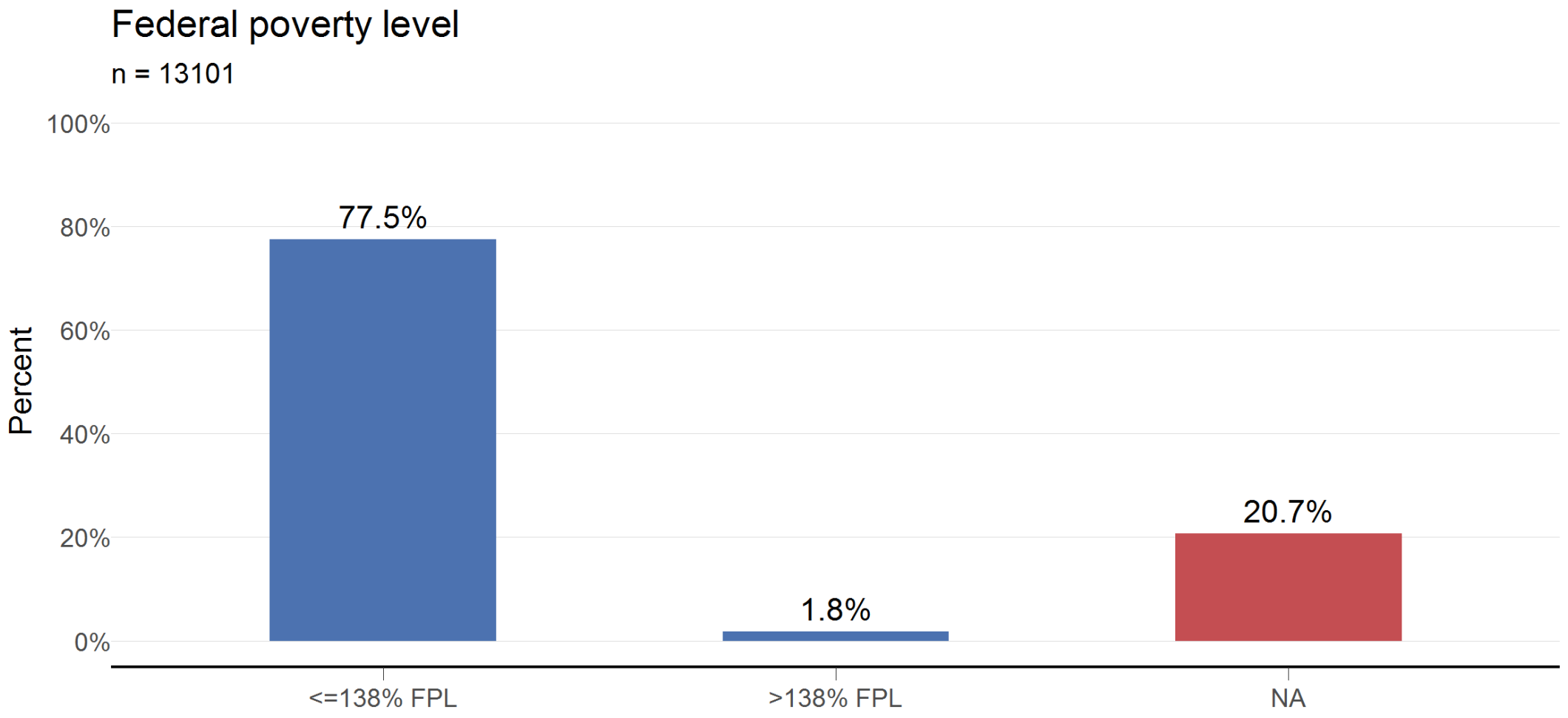
3 procedures

Procedure	n	Kappa	Likelihood done in clinic
Cholesterol screening	High (12817)	High (0.80)	High
Chlamydia screening	Low (523)	Medium (0.52)	Medium
Colonoscopy	Medium (3761)	Low (0.26)	Low

Missing data



Missing data

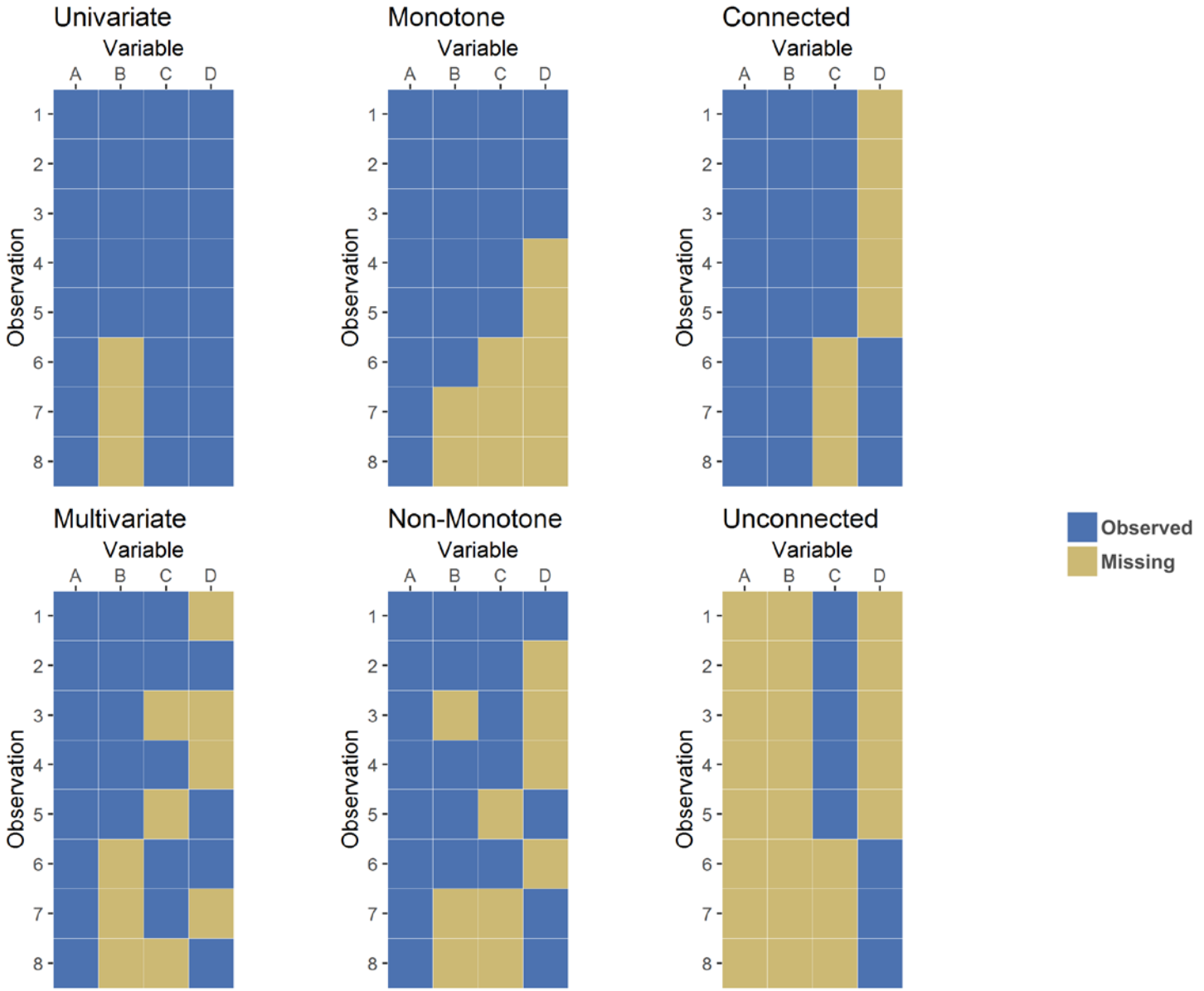


Pattern and mechanism

- Working with missing data, need to consider
  - Pattern – which values are observed and which are missing
  - Mechanism – relationship between missingness and values of the variables in the data
- Pattern and mechanism dictate which missing data methods are appropriate

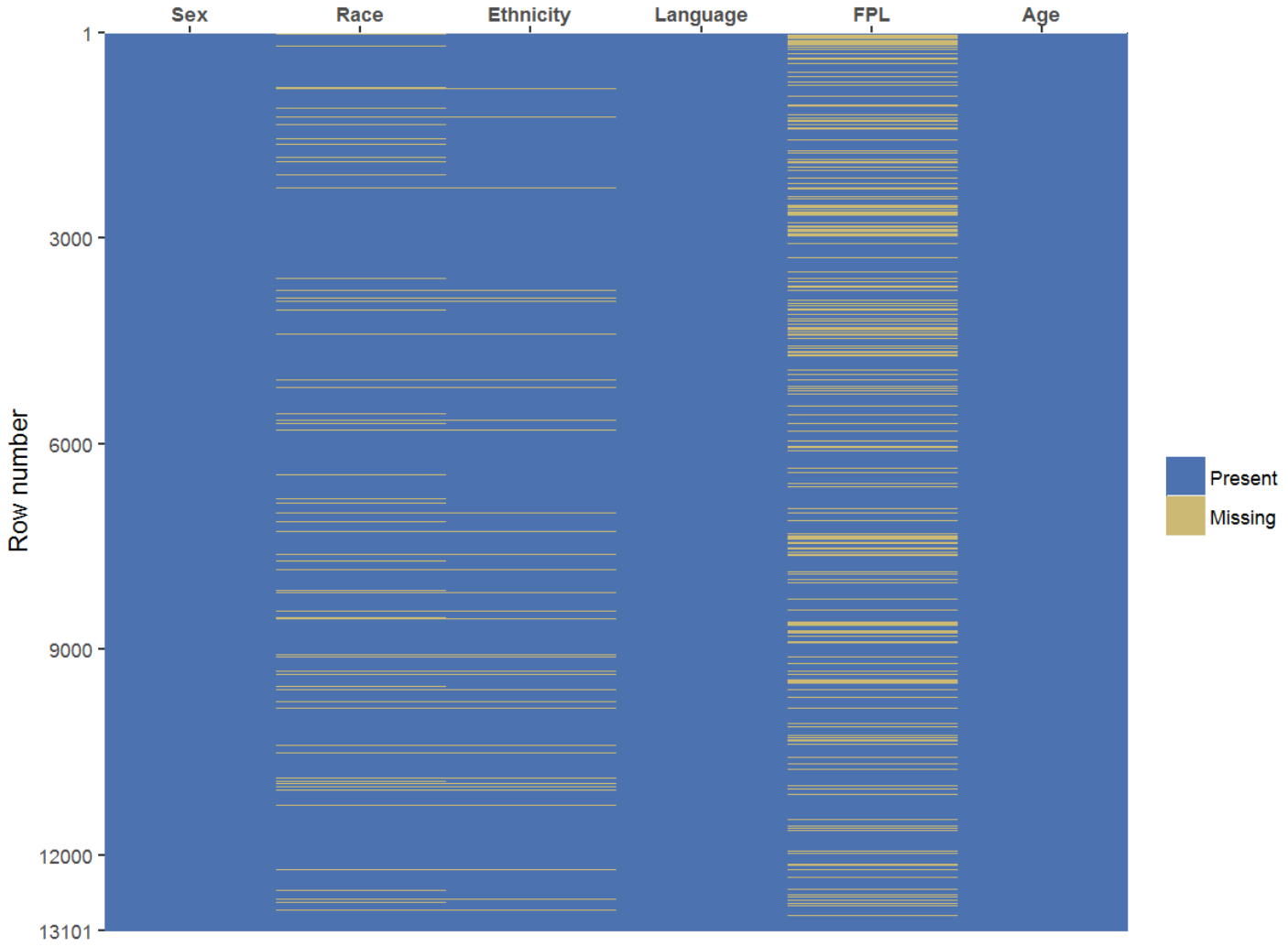
Pattern

- Examples of



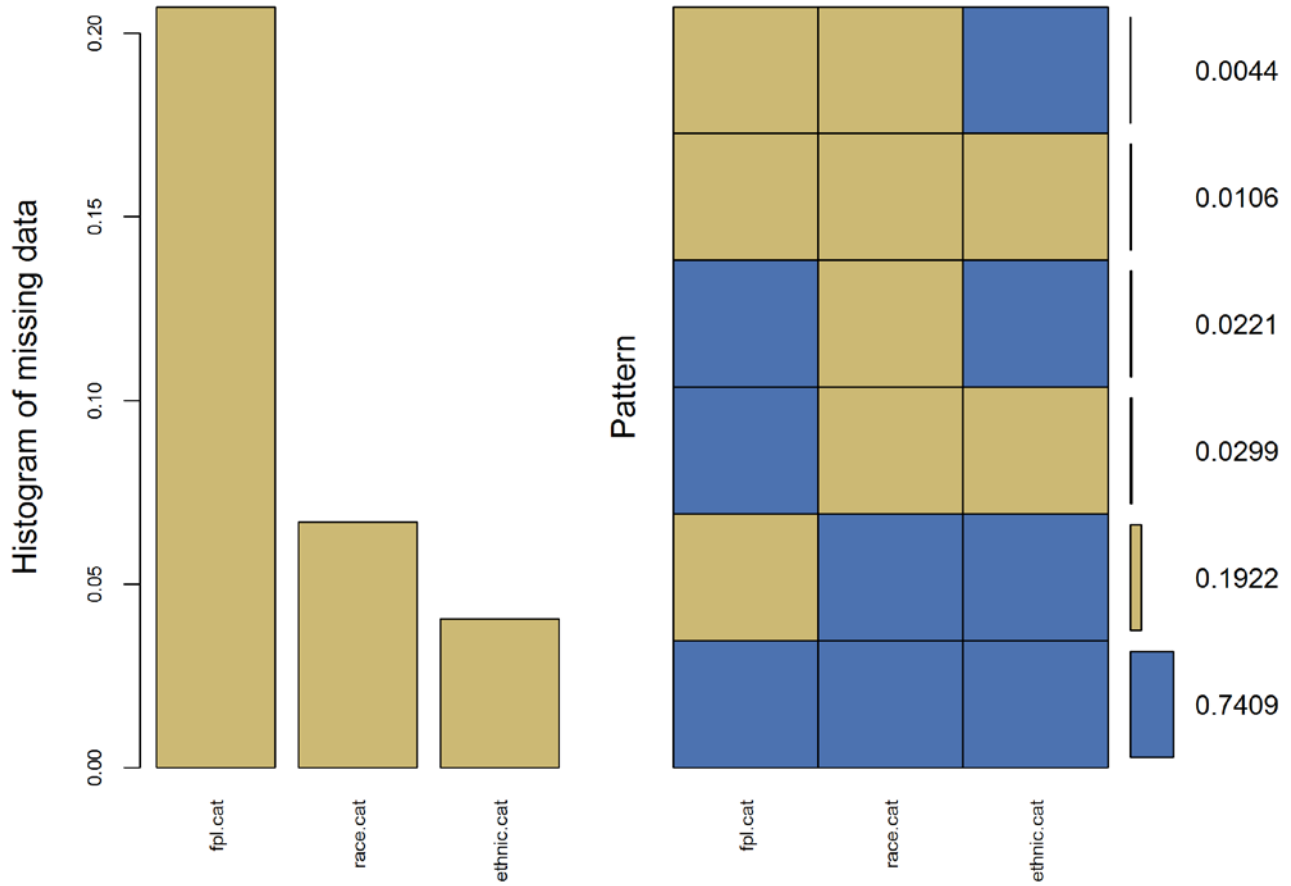
Adapted from Van Buuren (2012)

Examine the missing data



Adapted from the `vis_miss()` plot in the `naniar` and `visdat` packages in R by Nicholas Tierney

Examine the missing data



Aggregation plot made with VIM package in R by Matthias Templ, et al

Examine the missing data

	All patients		Cholesterol		Chlamydia		Colonoscopy	
	n	%	n	%	n	%	n	%
Total observations	13101		12817		523		3761	
Complete	9706	74%	9506	74%	374	72%	2828	75%
Incomplete	3395	26%	3311	26%	149	28%	933	25%
Missing variable								
Any	4122	31%	4014	31%	162	31%	1111	30%
Ethnicity	531	4%	513	4%	8	2%	127	3%
Race	877	7%	841	7%	46	9%	178	5%
FPL	2714	21%	2660	21%	108	21%	806	21%
No. missing columns								
1	2807	21%	2745	21%	136	26%	796	21%
2	449	3%	429	3%	13	2%	96	3%
3	139	1%	137	1%	0	0%	41	1%

Mechanism

- Examines the “reason” for missing values
- Tries to determine whether variables that are missing are related to the underlying values of the variables
- Rubin (1976) classified missing data mechanisms into three categories

Mechanism

- Missing completely at random (MCAR)
  - Missing value does not depend on the observed data or the missing data
  - Missing data are a random subset
  - *Example*: flip a coin before deciding to answer FPL question

- Missing at random (MAR)
  - Missing value may depend on observed data, but it does not depend on the missing data
  - Possible that missingness can be predicted from other available data
  - *Example*: men are more likely to not answer FPL, but it does not depend on FPL

Mechanism

- Missing not at random (MNAR)
  - Not MCAR and not MAR
  - Probability that a missing value is associated with the missing variable itself and with other variables
  - *Example*: people with low FPL are less likely to answer questions about their FPL

Mechanism

- Ignorability – the missing data mechanism is said to be ignorable if
  - The data are MAR (or MCAR), and
  - The parameters that govern the missing data process are not related to the parameters to be estimated
- This implies that the distribution of the data is the same for the response and non-response groups
- Don’t have to include information about the missing data

Ways to handle missing data

- Common approaches
  - Complete case analysis
  - Single imputation (e.g. mean substitution)
  - Missing indicator approach (e.g. add categorical variable missing or not)

- Can be bad
  - Assumes missingness is MCAR
  - Results in many dropped cases
    - Decreased power
  - May lead to biased results

Ways to handle missing data

- Some more appropriate ways
  - Likelihood methods (e.g. full information maximum likelihood)
  - Obtaining alternative sources of information
  - Weighting methods (e.g. inverse probability weighting)
  - Multiple imputation

Multiple imputation

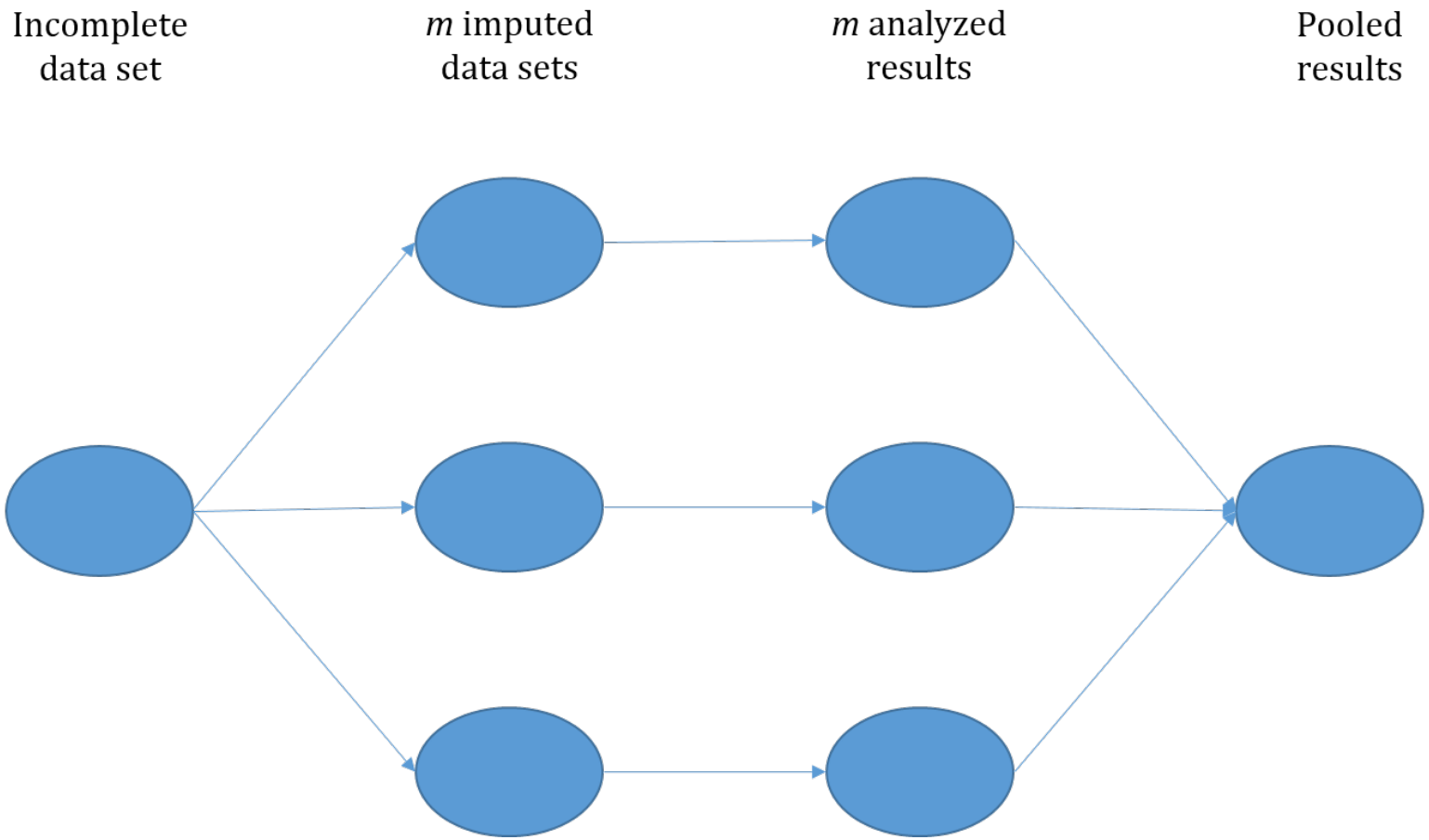
The goal of multiple imputation is to obtain statistically valid inferences from incomplete data.

-Van Buuren (2012)

Multiple imputation

- A statistical method for analyzing data sets with missing values
- Basic idea is to substitute a reasonable guess (imputation) for each missing value, multiple times
- Creates multiple “complete” data sets
- Analyze each data set separately and then combine (or pool) the results

Multiple imputation



Adapted from Van Buuren (2012)

Multiple imputation

- Joint model of all variables (traditional approach)
  - Multivariate normal distribution
  - Fit using observed cases
  - Used to predict the missing values
  - Sometimes use multivariate normal with categorical variables

Multivariate imputation by chained equations

- MICE
  - Sometimes called “fully conditional specification” or “sequential regression multiple imputation”
  - Specify multivariate imputation model on a variable-by-variable basis
  - Iteratively fits a model and imputes each variable
  - Model depends on the type of variable (binary, categorical, ordinal, continuous)
  - Raghunathan et al. (2001), Van Buuren et al. (2006)

MICE

- Flexible approach to multiple imputation
- Unnecessary to assume that the variables share a common distribution
- Can more easily work with large data sets with complex data structures
- Models more accurately reflect the distribution of each variable

MICE

- Theoretically weaker than joint modelling
- Incompatibility of conditionals
  - No joint distribution exists for the specification of conditional distributions
- In simulation and in practice, the method seems to be robust when the conditions are not met

MICE algorithm

- MICE algorithm developed by Stef Van Buuren
- Implemented in his R software package `mice 2.3`
- Other MICE software packages exist and vary somewhat in their exact implementation, but the general strategy is the same

Some notation

- Let Y represent a $n \times p$ matrix with data for n observations (rows) and p variables (columns)

- Define a response matrix, R, as a $n \times p$ matrix of $0 - 1$ values which indicate missingness

$$r_{ij} = \begin{cases} 1, & if\ y_{ij}\ is\ observed \\ 0, & if\ y_{ij}\ is\ missing \end{cases}$$

- The complete data values Y are made in two parts Y^{obs} and Y^{mis}

MICE algorithm

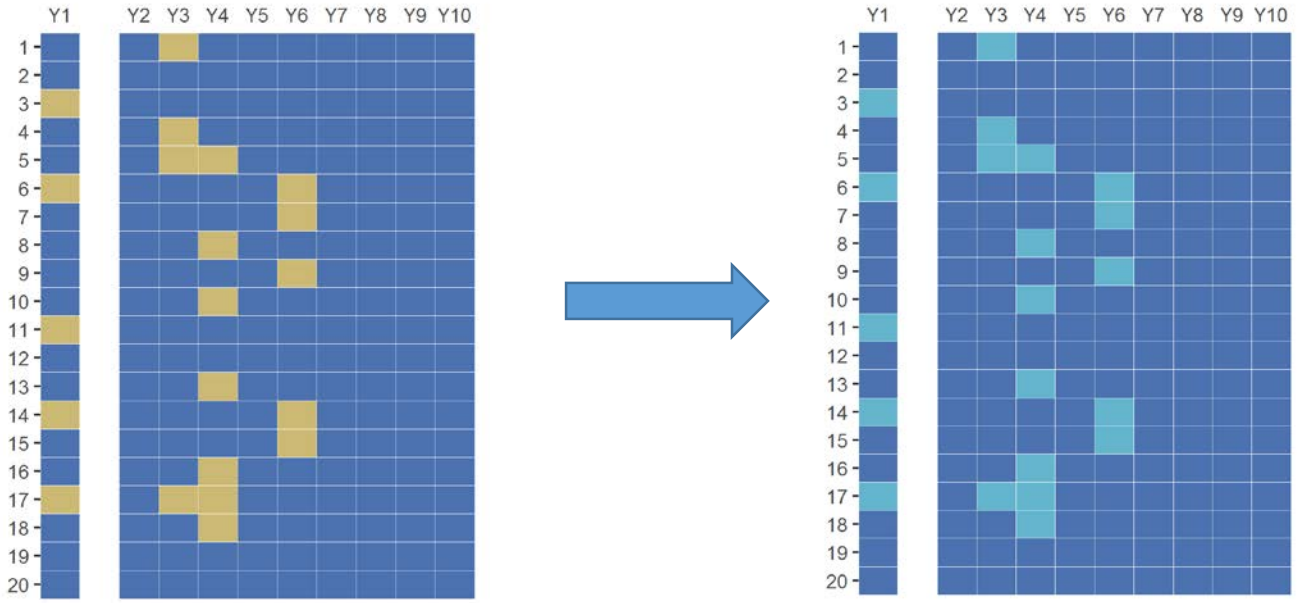
- Step 1
  - The analyst/researcher decides on an imputation model $P\left(Y_j^{mis} \middle| Y_j^{obs}, Y_{\_j}, R\right)$ for each variable Y_j with $j = 1, \ldots, p$.

  - Default choices for some variables type in `mice`

Method	Variable type
Predictive mean matching	Numeric
Logistic regression	Binary
Multinomial logit model	Nominal
Ordered logit model	Ordinal

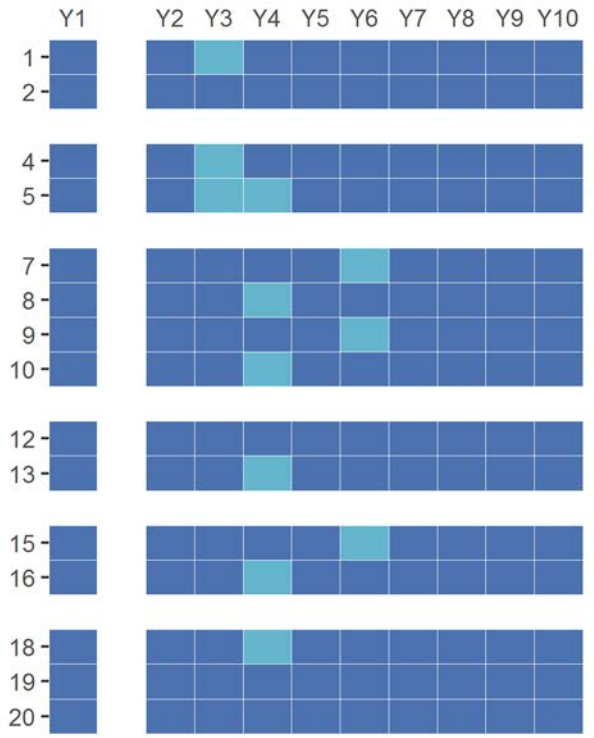
MICE algorithm

- Step 2
  - For each j, missing values are filled in with starting imputations Y_j^0 by a simple imputation using the observed values Y_j^{obs} (e.g. random sampling of observed value with replacement or mean substitution).
  - This initialization is repeated for $t = 1, \ldots, T$ and for $j = 1, \ldots, p$.



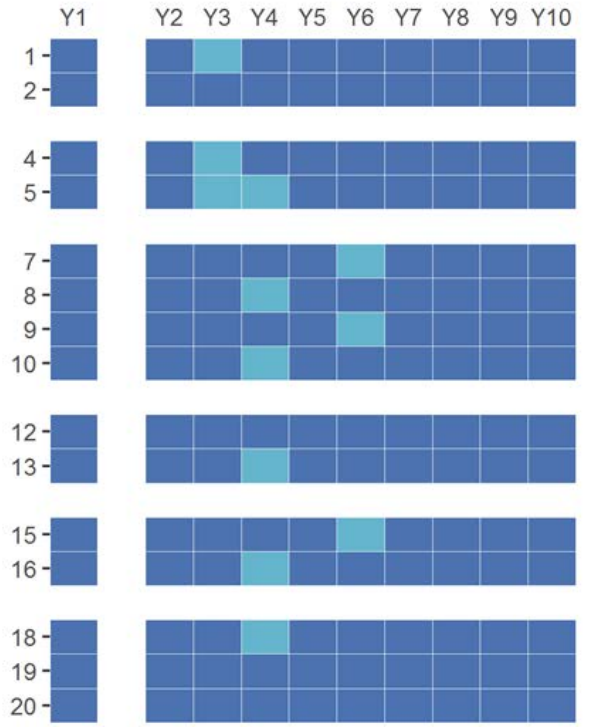
MICE algorithm

- Step 3
  - The values for the variable to be imputed, Y_j, are set back to missing.
  - The currently complete data except Y_j is defined $Y_{-j}^t = \left(Y_1^t, \ldots, Y_{j-1}^t, Y_{j+1}^{t-1}, \ldots, Y_p^{t-1}\right)$.



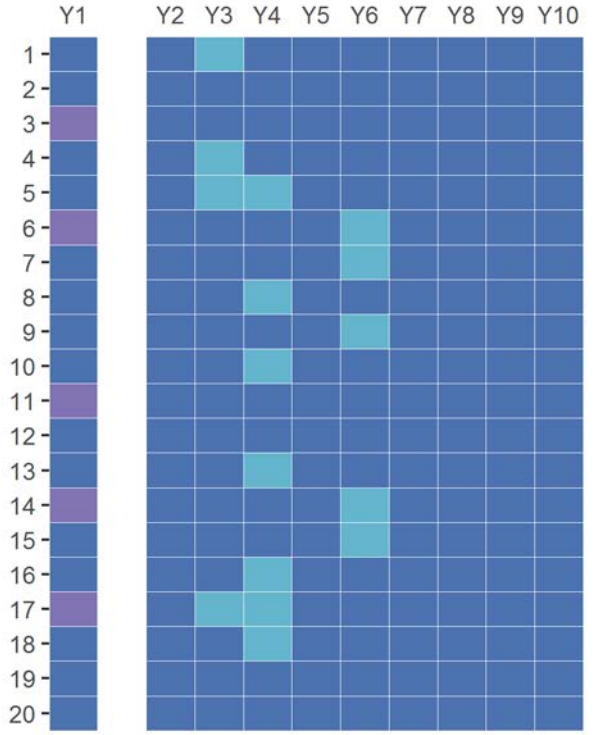
MICE algorithm

- Step 4
  - Draw $\phi_j^t \sim P\left(\phi_j^t \middle| Y_j^{obs}, Y_{-j}^t\right)$ where ϕ_j represents the unknown parameters of the imputation model.
  - The observed values Y_j^{obs} are regressed on the other variables in the imputation model Y_{-j}^t in order to obtain estimates of the regression model parameters ϕ_j^t.



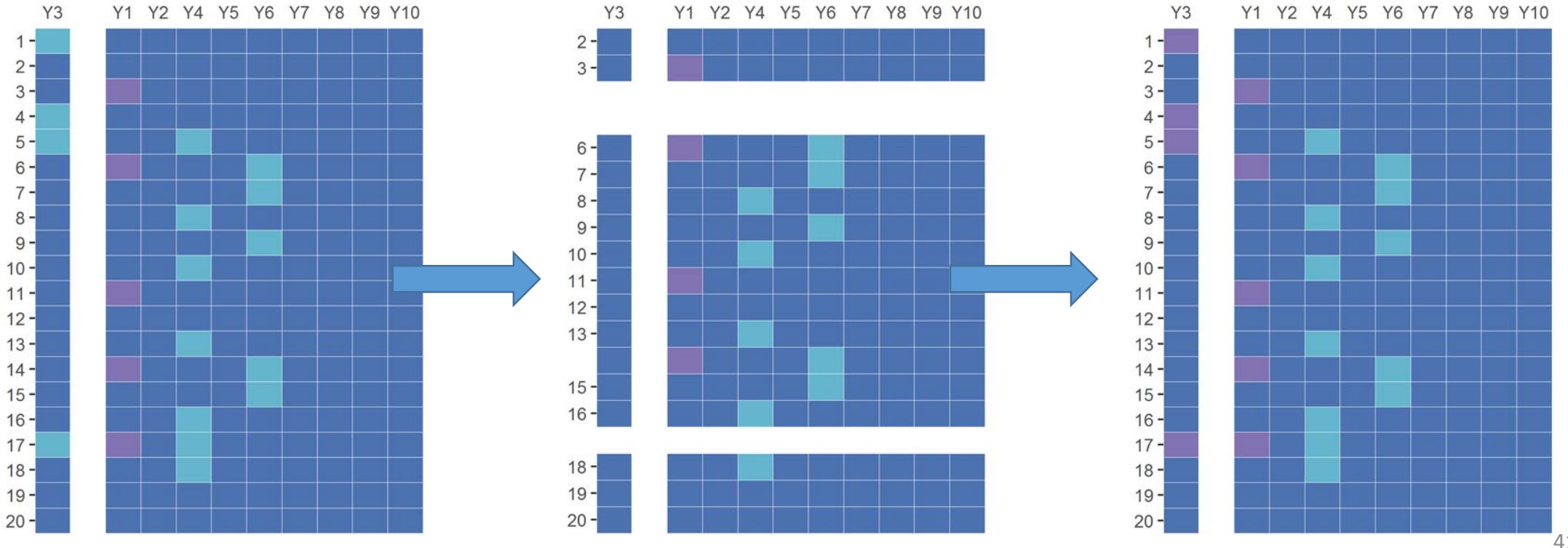
MICE algorithm

- Step 5
  - Draw imputations $Y_j^t \sim P\left(Y_j^{mis} \middle| Y_j^{obs}, Y_{-j}^t, \phi_j^t\right)$, the corresponding posterior predictive distribution of Y_j^t.
  - So missing Y_j^{mis} are replaced with predictions (imputations) from the regression model that was fit in step 4.



MICE algorithm

- Step 6
  - End and repeat (3–5) for the next variable, j.
  - End and repeat for the next iteration, t.



MICE algorithm

- Steps 1 through 6 are repeated to create m imputed data sets.
  - After one cycle, all of the missing values have been replaced with predictions from regression models that reflect relationships observed in the data
  - The researcher decides how many cycles to perform so that the results have converged or stabilized (generally 10 to 20) which will produce one imputed data set.
  - This whole process is repeated *m* times to produce *m* imputed data sets.

Specify the imputation model

- Decisions and set up prior to running the MICE algorithm
- Van Buuren describes this as the most challenging step
- The model should
  - Account for the process that created the missing data,
  - Preserve the relations in the data, and
  - Preserve the uncertainty about these relations.
- To help, he provides 7 ordered considerations to take

Specify the imputation model

- Step 1
  - Decide if the missing at random (MAR) assumption is reasonable
- Step 2
  - Decide on the form of the imputation model

Variable	Variable Type	Imputation model
Ethnicity	Categorical factor with >2 levels	Multinomial logit regression
Race	Categorical factor with >2 levels	Multinomial logit regression
FPL	Categorical factor with 2 levels	Logistic regression

Specify the imputation model

- Step 3
  - Decide the set of predictors to include in the imputation model
- Step 4
  - Decide whether to impute variables that are function of other (incomplete) variables.
- Step 5
  - Decide the order in which variables should be imputed

Specify the imputation model

- Step 6
  - Decide the number of iterations
    - 10 to 20 are recommended; we chose 20

- Step 7
  - Decide m, the number of multiply imputed data sets.
    - Rule of thumb from more recent authors
      - The number of imputations should be similar to the percentage of cases that are incomplete (at least 5)
      - Our data set had 26% incomplete cases, so we chose m = 30

Variable selection

- The general advice cited is to include as many variables as possible
  - Tends to make MAR assumption more reasonable
  - Reasonable to include all variables for small to medium data sets (20 to 30 variables)
  - For large data sets, the advice is to select a subset of 15 to 25
    - Avoid multicollinearity issues
    - Avoid computational issues

Additional advice for variable selection

- Include variables
  - In the model of scientific interest
  - Related to occurrence of missing data (i.e. related to non-response)
  - Where distributions differ between response and non-response
  - Correlated with the target variable

Additional advice for variable selection

- Remove variables
  - With too many missing values
  - If missing on same cases as target variable

Specify the imputation model

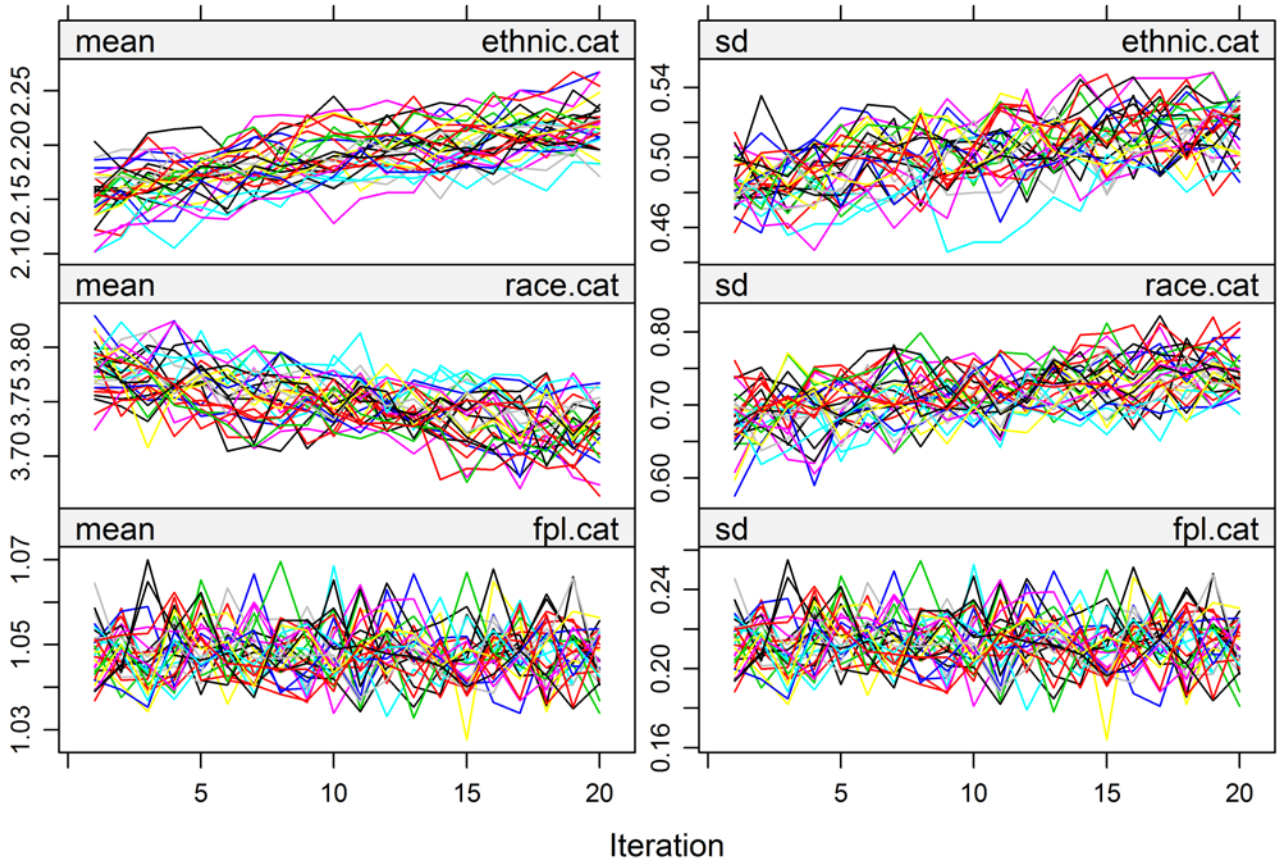
- Final imputation model
  - 3 planned scenarios
    - Full with 37 variables
    - Reduced with 21
    - QPM (quick predictor matrix chosen by software) with 2 to 3
  - Impute from highest percent missing (FPL, 21%) to lowest (Ethnicity, 4%)
  - 20 iterations
  - m = 30 imputed data sets

Assessing convergence

- Models must be reviewed for convergence
- Advice is to plot parameters against the iteration number
- `mice` makes this easy
- Streams should be (1) well mixed, and (2) show no signs of trend

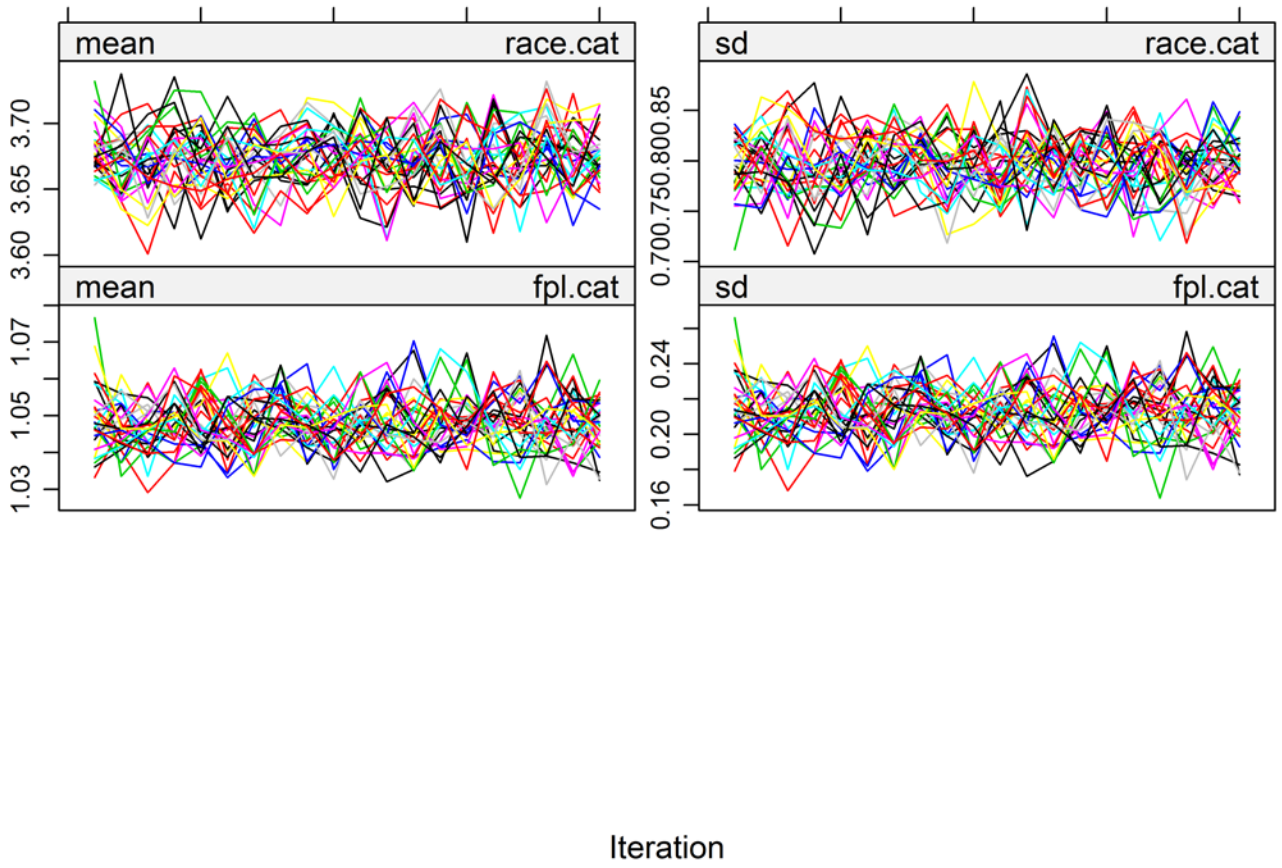
Assessing convergence

- Full scenario
  - 3 imputed variables



Revisit the model

- Full scenario
  - 2 imputed variables



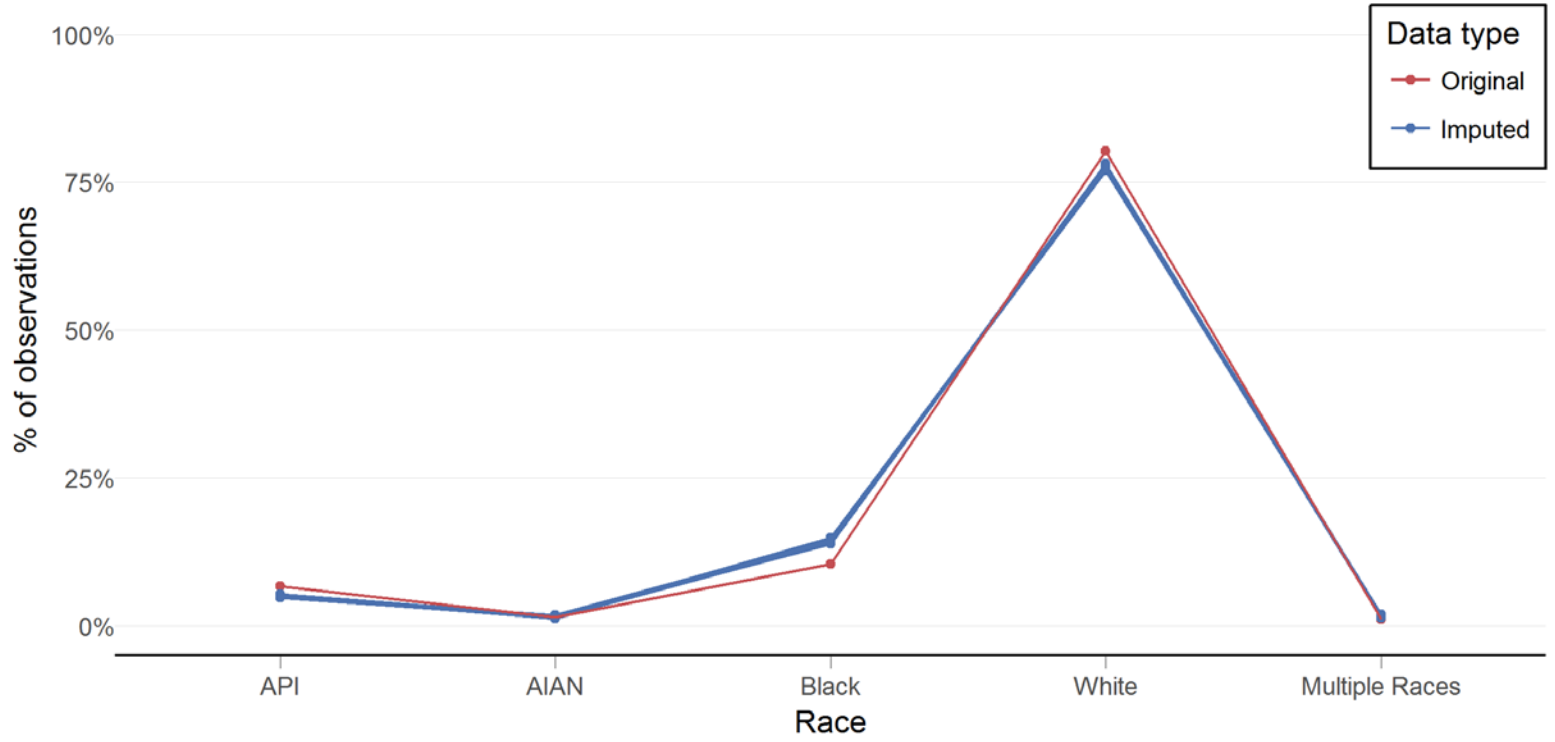
Assessing the imputations

- Numerical diagnostics exist
  - Most examples use continuous variables and regression models.

- Data visualization is also recommended to assure that imputed data is
  - Close to the original data,
  - Plausible
    - Within an appropriate range
  - Make common sense
    - e.g. pregnant fathers

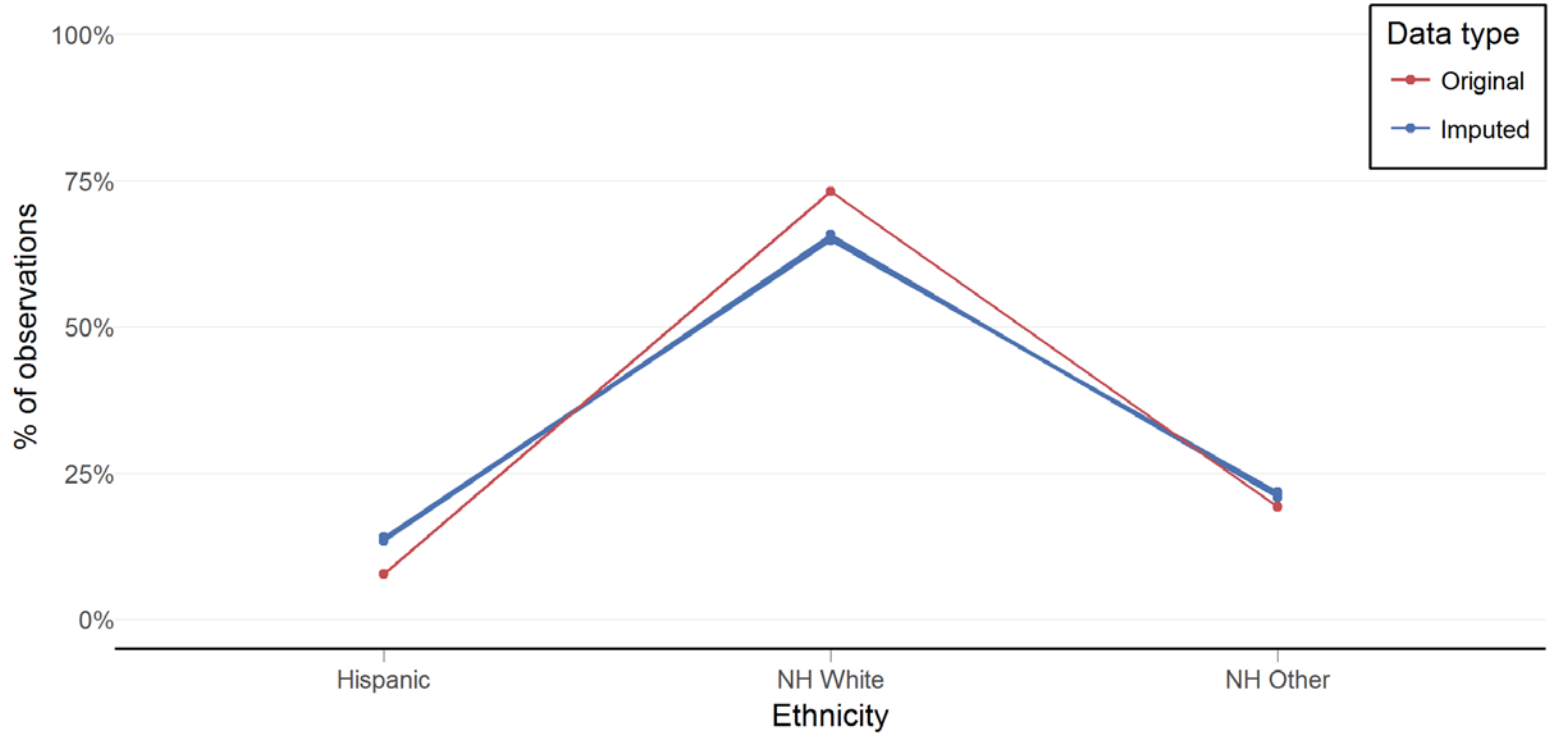
Assessing the imputations

- Race
  - Full Scenario



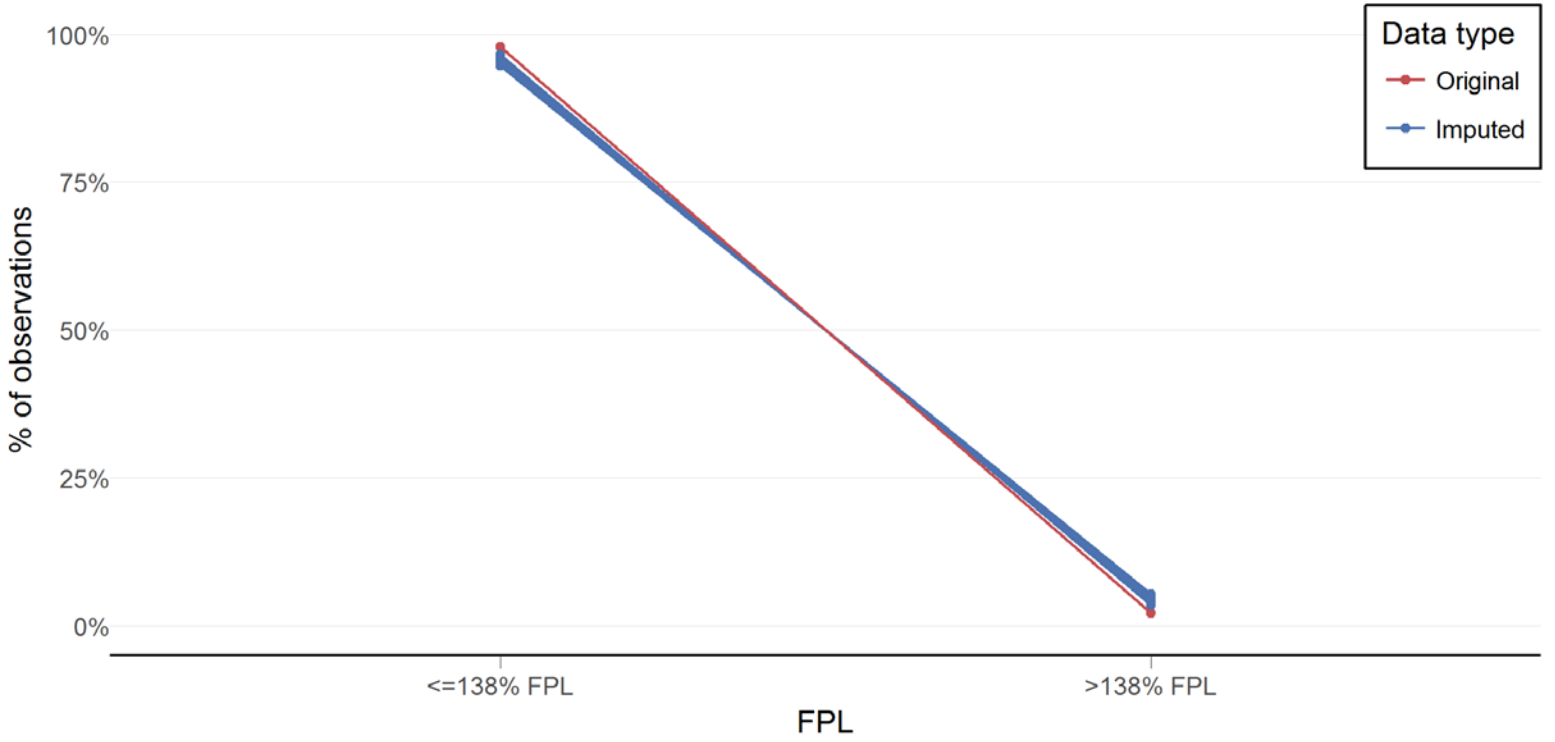
Assessing the imputations

- Ethnicity
  - Full Scenario



Assessing the imputations

- FPL
  - Full Scenario



Analyzing m imputed data sets

- Software packages have helpful tools for analyzing the m data sets
- Tend to be designed to neatly handle models
- We had to do it "by hand"
  - Obtain point estimates and variances for each of the m data sets
  - 11 procedures × 22 strata = 242 2×2 tables
  - 23 statistics, 30 data sets, and 6 scenarios
  - Approx. 1 million estimates and variances

Pooling the results

- Rubin's rules (1987)
  - Provide the method to pool *m* parameter estimates, $\hat{Q}_1, \ldots, \hat{Q}_m$ into a single estimate $\bar{Q}$ and to estimate its variance-covariance matrix
  - Accounts for both within- and between- imputation variance
  - Software is available to help, but designed for models
  - `mice` has some functionality to help "by hand"

Pooling the results

- $\hat{Q}_l$ is the complete-data estimate of the scalar quantity of interest (e.g. regression coefficient, kappa statistic) from the lth imputed data set

- The overall estimate is the average of the estimates from the m complete data sets

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_l$$

Pooling the results

- The combined within-imputation variance $\bar{U}$ is equal the average of the complete data variances

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m} U_l$$

- The between-imputation variability reflects the uncertainty due to missing information, i.e. the variance between (among) the *m* complete data estimates

$$B = \frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{Q}_l - \bar{Q}\right)^2$$

Pooling the results

- The total variance of $\bar{Q}$ is given by

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

- An additional term B/m is included to reflect the additional variance since $\bar{Q}$ is estimated for finite m

Pooling the results

- For multi-parameter inference, approaches are available such as Wald Test, likelihood ratio test, and χ^2-test

- For single parameter or scalar inference, like kappa statistics and others examined in this thesis, Wald-type significance tests and confidence intervals can be calculated in the usual way

Pooling the results

- Since the total variance of T is not known, $\bar{Q}$ follows a t-distribution rather than normal

- So univariate tests are based on the approximation

$$\frac{Q - \bar{Q}}{\sqrt{T}} \sim t_v$$

where t_v is Student's t-distribution with v degrees of freedom

- For degrees of freedom see Van Buuren (2012), Schafer (1997), or Barnard and Rubin (1999)

Pooling the results

- The $100(1-\alpha)\%$ confidence interval for Q is calculated as

$$\bar{Q} \pm t_{v,1-\alpha/2}\sqrt{T}$$

Pooling the results

- Quantities with non-normal distributions may require transformation

Can be combined without transformation	May require sensible transformation before combination	Cannot be combined
· mean · proportion · regression coefficient · linear predictor · C-index · area under the ROC curve	· odds ratio · hazard ratio · baseline hazard · survival probability · standard deviation · correlation · proportion of variance explained · skewness · kurtosis	· *p* -value · likelihood ratio test statistic · model chi-squared statistic · goodness-of-fit test statistic

Choose a scenario

- Select one imputation scenario to represent the results after imputation

- Selected **Full Scenario with Race, Ethnicity, and FPL**
  - Results seemed robust to suspect convergence
  - Similar results across scenarios suggest that a small number of variables dominate the imputation model
  - Advice from literature is including more variables makes MAR a more reasonable assumption

Reporting results

- Suggested guidelines exist for reporting statistical methodology

- Van Buuren (2012) curated well-chosen sources in to a unified list
  - Amount of missing data
  - Reasons for missing
  - Method to handle missing data
  - Software used
  - Number of imputed data sets
  - Complete case analysis

Reporting results

- 3 of 41 variables in the data set had missing values
- Of 13,101 observations, 3,395 (26%) were incomplete
- Reason Race and ethnicity are missing is not known. High rate of missing for FPL is due to a data discrepancy that was left unresolved intentionally for this analysis

Reporting results

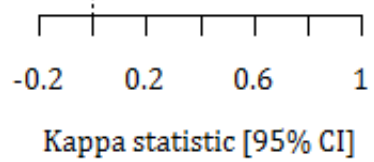
- Multiple imputation (Rubin 1987) was used to create and analyze 30 multiply imputed data sets

- Imputed under fully conditional specification (Van Buuren et al. 2006)

- Calculations were done in R version 3.4.0 (2017-04-21) using the default settings of the `mice 2.3` package (Van Buuren et al. 2011), as these were deemed appropriate for our study setting

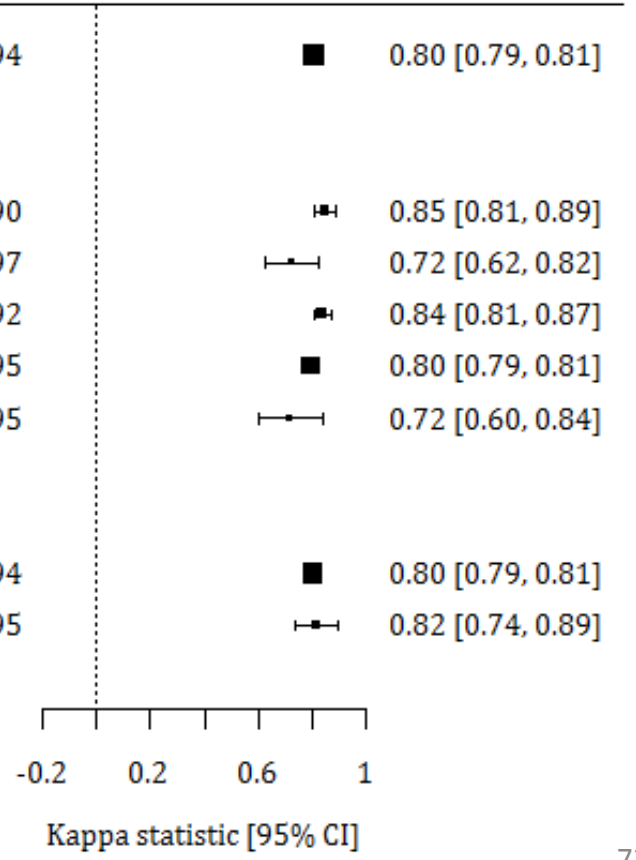
Cholesterol screening, pre-imputation

	Total eligible patients	EHR (+) p1, %	Claims (+) p2, %	EHR (+), Claims (+) p1+p2, %	Sensitivity	Specificity	Kappa statistic [95% CI]
All eligible patients	12817	39.5	42.1	45.5	0.86	0.94	0.80 [0.79, 0.81]
Race							
Asian/Pacific Islander	756	50.1	47.5	52.5	0.95	0.90	0.85 [0.81, 0.89]
American Indian/Alaskan native	175	33.7	44.0	45.7	0.73	0.97	0.71 [0.61, 0.82]
Black	1388	41.8	40.4	44.9	0.92	0.92	0.84 [0.81, 0.87]
White	9518	38.4	42.1	45.1	0.84	0.95	0.80 [0.79, 0.81]
Multiple Races	139	28.8	36.0	38.9	0.72	0.96	0.71 [0.58, 0.83]
Unknown	841	41.4	41.6	46.2	0.88	0.92	0.80 [0.76, 0.84]
Federal poverty level							
<=138% FPL	9930	41.5	43.5	46.9	0.88	0.94	0.82 [0.81, 0.83]
>138% FPL	227	35.2	35.2	38.8	0.90	0.95	0.85 [0.77, 0.92]
Missing/Unknown	2660	32.2	37.8	40.9	0.77	0.95	0.74 [0.71, 0.77]

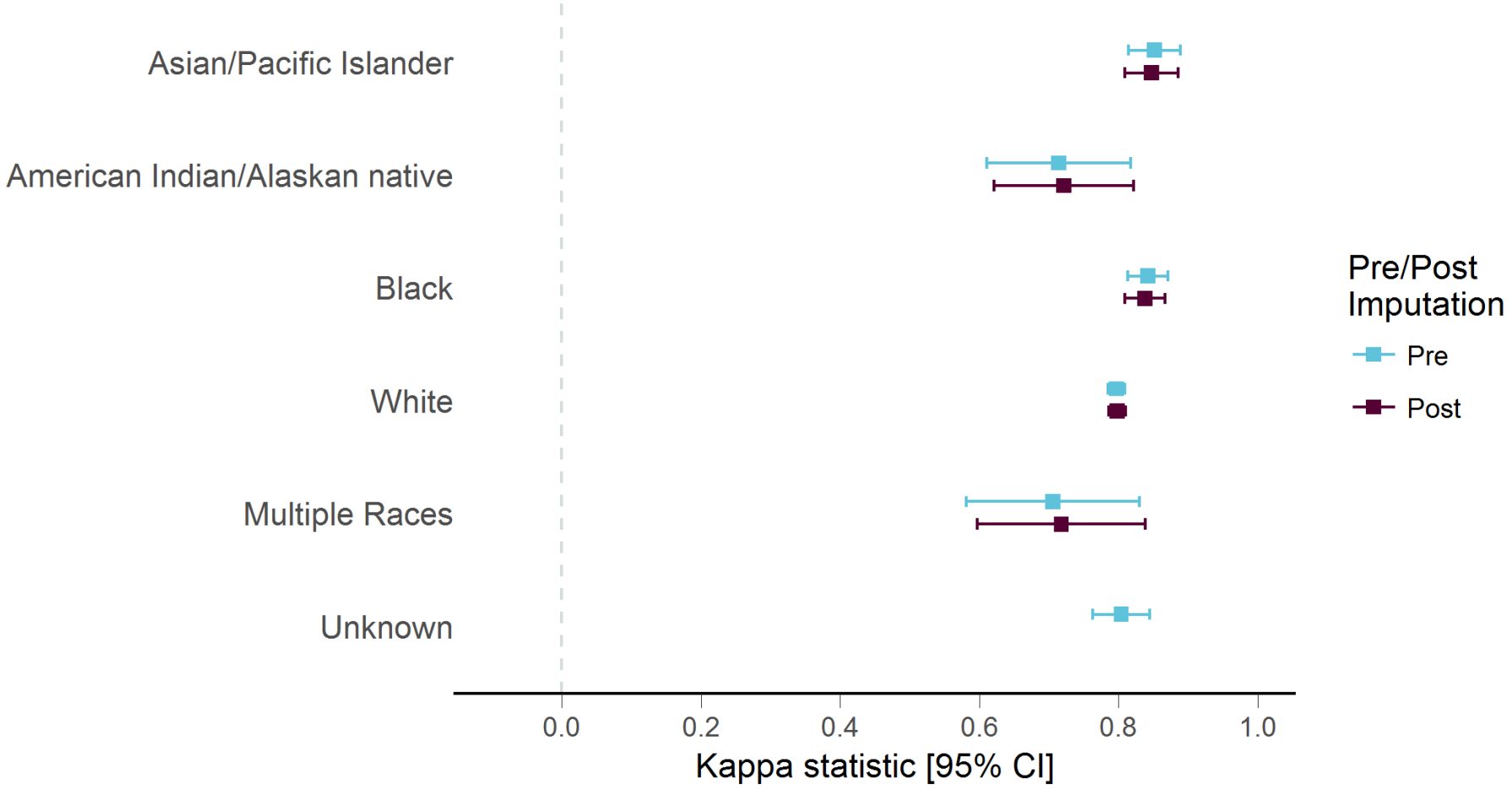


Cholesterol screening, post-imputation

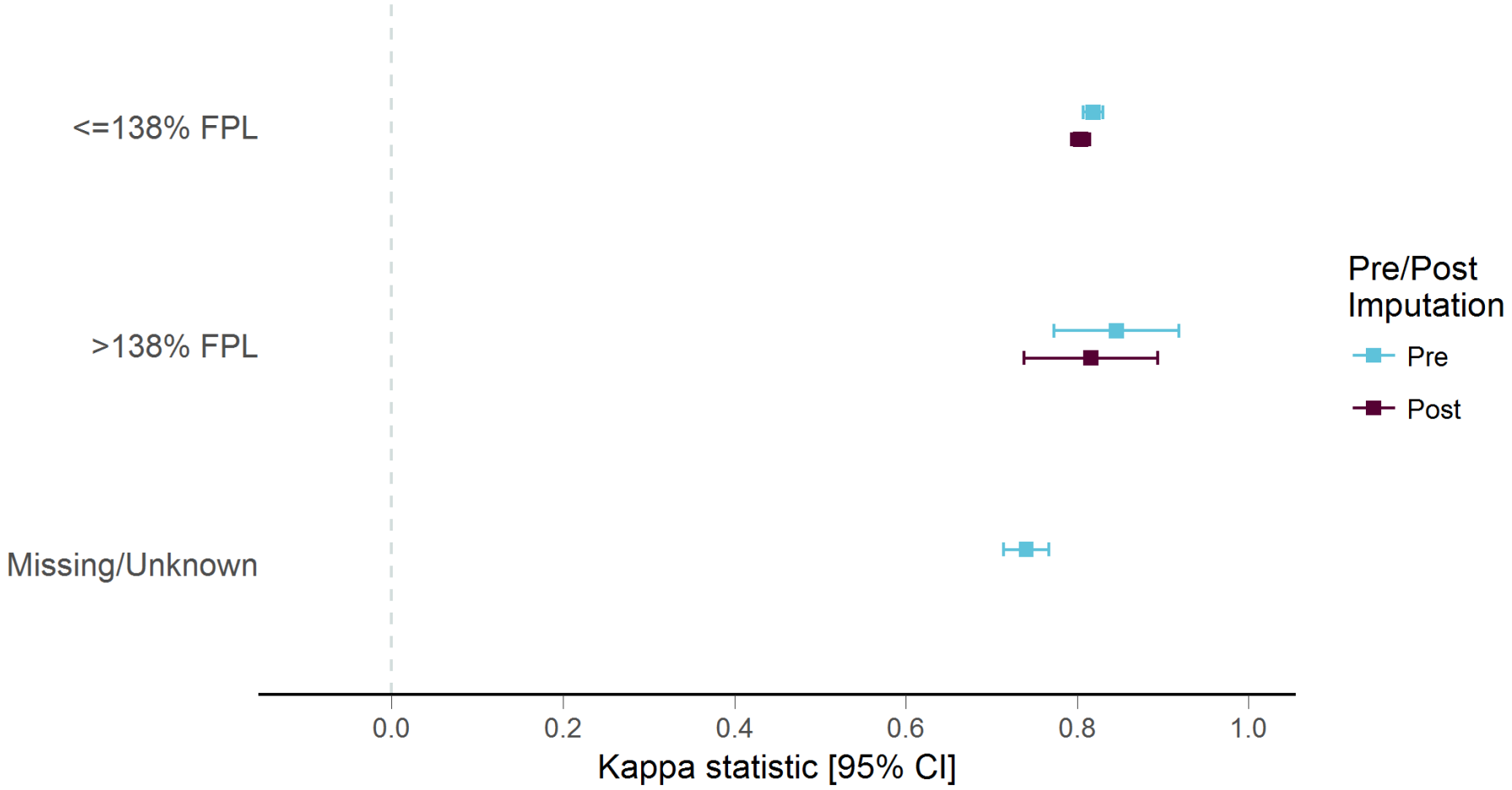
	Total eligible patients	EHR (+) p1, %	Claims (+) p2, %	EHR (+), Claims (+) p1+p2, %	Sensitivity	Specificity	Kappa statistic [95% CI]
All eligible patients	12817	39.5	42.1	45.5	0.86	0.94	0.80 [0.79, 0.81]
Race							
Asian/Pacific Islander	795	49.9	47.5	52.5	0.95	0.90	0.85 [0.81, 0.89]
American Indian/Alaskan native	190	33.9	43.8	45.5	0.73	0.97	0.72 [0.62, 0.82]
Black	1483	41.9	40.6	45.1	0.92	0.92	0.84 [0.81, 0.87]
White	10191	38.6	42.0	45.1	0.84	0.95	0.80 [0.79, 0.81]
Multiple Races	155	29.6	35.8	38.9	0.74	0.95	0.72 [0.60, 0.84]
Federal poverty level							
<=138% FPL	12464	39.7	42.4	45.8	0.86	0.94	0.80 [0.79, 0.81]
>138% FPL	352	32.7	34.0	37.4	0.86	0.95	0.82 [0.74, 0.89]



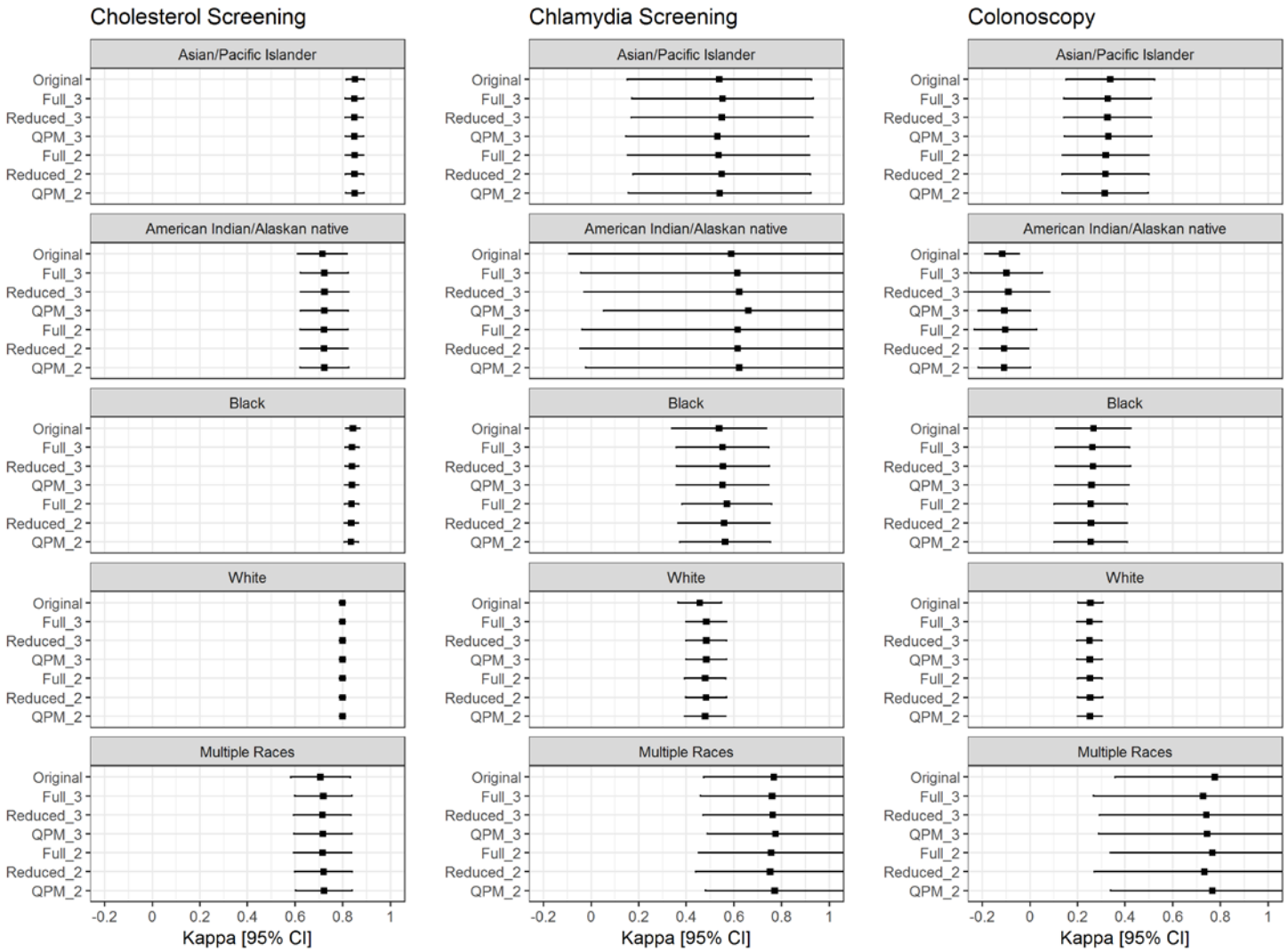
Cholesterol screening



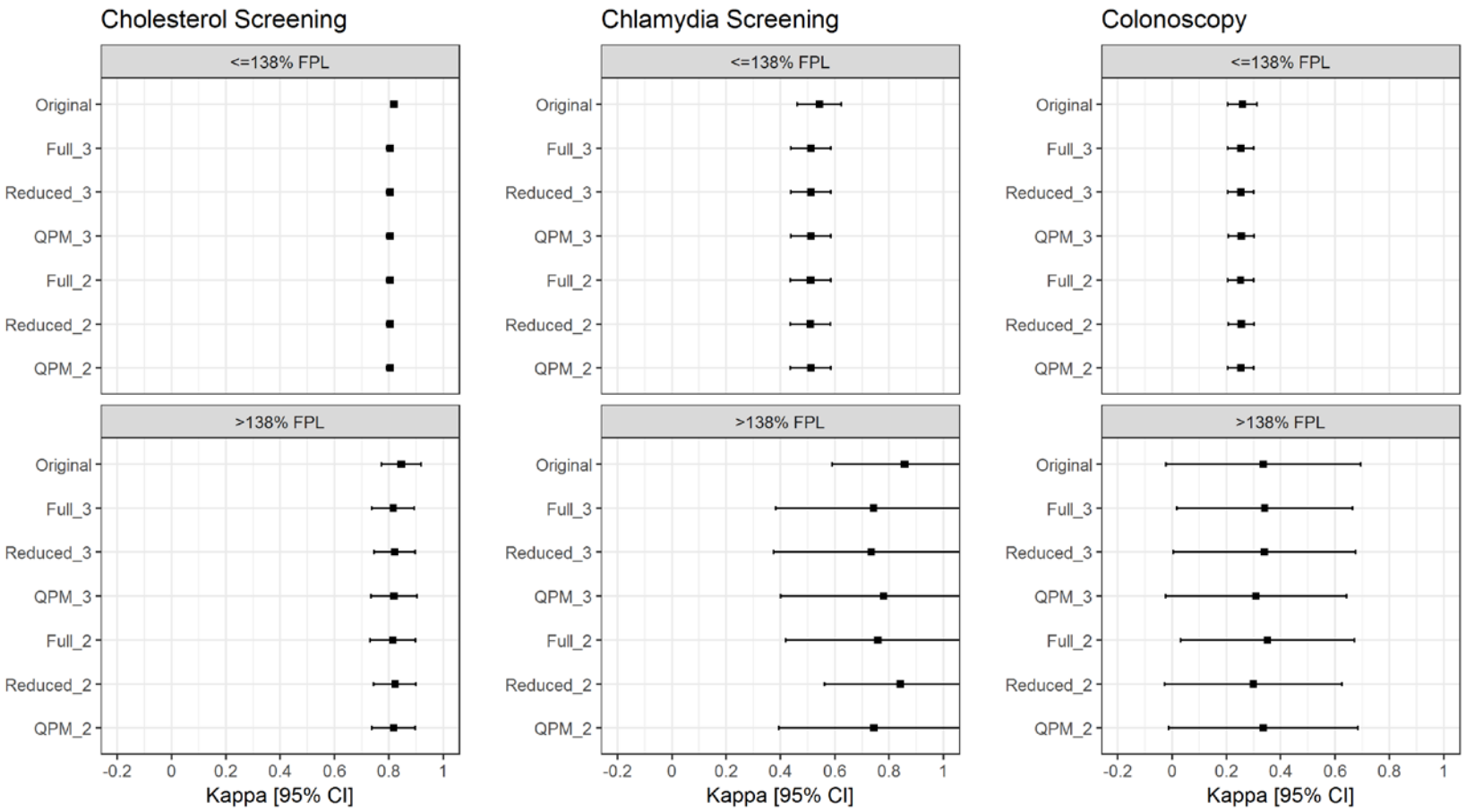
Cholesterol screening



Compare results – Race



Compare results – FPL



Discussion

- Similarity between post-imputation results and pre-imputation results
  - Not always the case
- Multiple imputation with non-standard statistics like kappa are not fully developed
  - Rubin's rules
- Further diagnostics are available to
  - Assess MAR assumptions
  - Assess distributions of the imputed data

Conclusions

- By our example, we've shown MICE to be a viable method to address missing data in EHR
  - Flexible approach
  - Adaptable to many different types of variables
  - Suitable for large data sets

Conclusions

- MICE appears to be an approach worth further investigation and application to obtain statistically valid kappa statistics with incomplete EHR data

Further reading and resources

- Software:
  - R
    - `mice` package by Stef Van Buuren
    - `visdat` and `naniar` packages by Nicholas Tierney
  - SAS
    - PROC MI: FCS statement
    - IVEware callable software
  - Stata
    - mi impute chained

Further reading and resources

- References
  - J. Heintzman, S. Bailey, M. Hoopes, T. Le, R. Gold, J. P. O'Malley, S. Cowburn, M. Marino, A. Krist and J. E. DeVoe, "Agreement of Medicaid claims and electronic health records for assessing preventive care quality among adults," J Am Med Inform Assoc, vol. 24, p. 720–724, 2014
  - www.multiple-imputation.com, Stef Van Buuren's website
  - S. van Buuren, Flexible Imputation of Missing Data, Boca Raton, FL: CRC Press, 2012
  - I. R. White, P. Royston and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine,* vol. 30, p. 377–399, 2011
  - M. J. Azur, E. A. Stuart, C. Frangakis and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?," International Journal of Methods in Psychiatric Research, vol. 20, no. 1, p. 40–49, 2011.
  - R. J. Little and D. B. Rubin, Statistical Analysis with Missing Data, Second Edition ed., Hoboken, NJ: Jonh Wiley & Sons, Inc., 2002
  - J. L. Schafer, "Multiple Imputation: a primer," Statistical Methods in Medical Research, vol. 8, pp. 3-15, 1999

Sample code on GitHub

emilelatour/csp-2018

Acknowledgments

- OHSU Knight Cancer Institute Biostatistics Shared Resource
- John Heintzman, MD, MPH at Department of Family Medicine, OHSU
- Megan Hoopes, Jennifer DeVoe, and rest of team at OCHIN, Inc.
- Miguel Marino, PhD at Department of Family Medicine, OHSU


+Dunken k Bliths

Questions

latour@ohsu.edu