A practical introduction to methods for dealing with missing data

Rosie Cornish & Ellie Curnow
University of Bristol

Workshop Objectives

- Recognise the types and patterns of missing data
- Represent a missing data scenario using a causal diagram
- Understand when complete records analysis is likely to be unbiased
- Understand the principles of multiple imputation and when it is likely to be unbiased
- Apply multiple imputation methods to deal with missing data in (relatively) straightforward situations

Workshop Agenda

12:00 – 12:20 Welcome and lunch

12:20 – 13:00 Introduction to missing data

13:00 – 14:00 Computer Practical 1: Exploring missing data and complete records analysis in practice

14:00 – 14:15 Break

14:15 – 15:00 Introduction to multiple imputation

15:00 – 15:30 Computer Practical 2: Demonstration of multiple imputation in practice

15:30 – 16:00 Final summary and discussion

Introduction to missing data

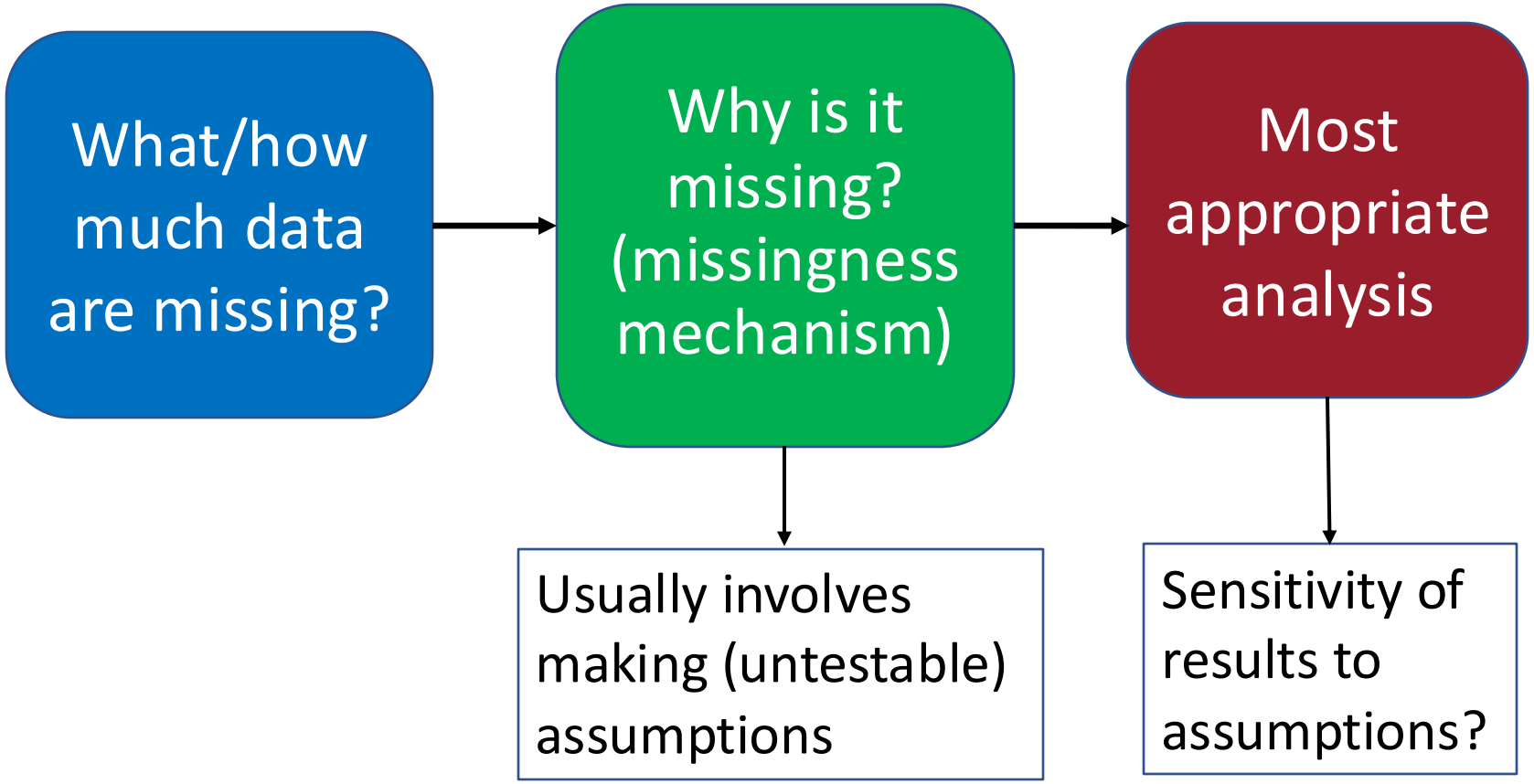
Outline

- Approaches to missing data
- Analyses involving missing data
- Missing data mechanisms
- Introduction to DAGs
- DAGs for missing data

Approaches to missing data

- **Complete case analysis** (complete records analysis, listwise deletion, available case analysis...)
- Maximum likelihood
- Inverse probability (of missingness) weighting
- **Multiple imputation**

Missing data



Missing data mechanisms

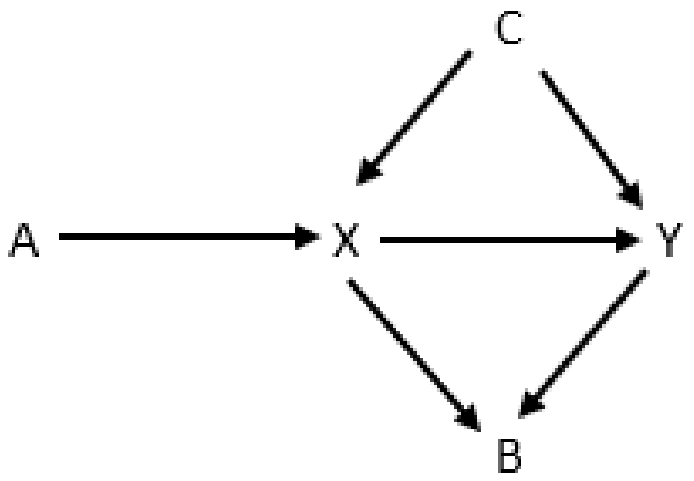
- MCAR – probability of missingness is not related to our data (e.g. batch of questionnaires lost)
- MAR – probability of missingness only depends on observed data (e.g. missing data more likely for older people, but age is fully observed)
- MNAR – probability of missingness is dependent on unobserved data, even after conditioning on observed data (e.g. heavy smokers less likely to provide data on smoking habits)

NOTE: Can rule out MCAR, but never possible to rule out MNAR

A very brief introduction to DAGs

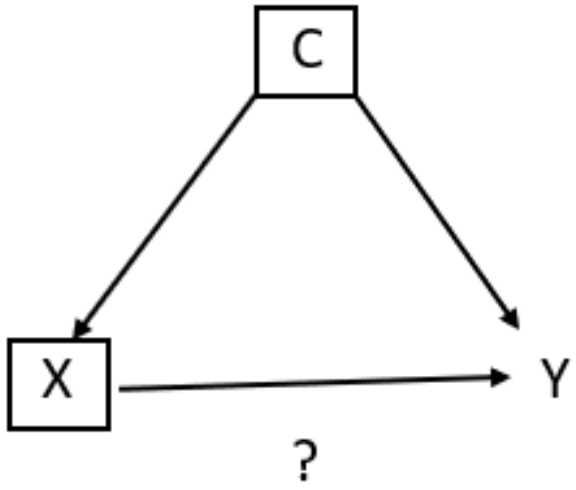
Directed Acyclic Graphs (DAGs)

- Depict our hypotheses/assumptions about (causal) relationships between variables in our analysis



- Arrows between variables that are causally related – direction of arrow indicates direction of the causal relationship (e.g. C is a cause of both X and Y)
- B is a common effect of X and Y (a collider on the path between X and Y – both arrows point into B)
- Paths: sequences of arrows connecting two variables (may be causal or non-causal) – causal if all arrows point in the same direction
- No arrow means no causal effect

Conditioning



We are going to use regression to examine the link between exposure (X) and outcome (Y), adjusting for confounder C

Conditioning on a variable means either:

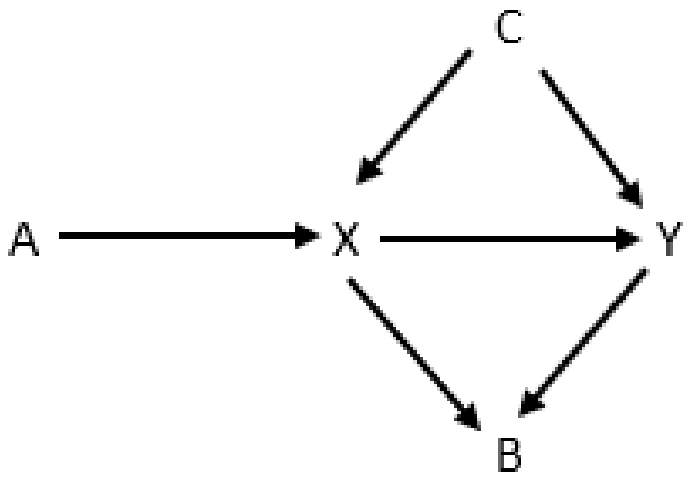
- Sample restriction (e.g. carrying out an analysis only among females; **including only those in your sample who actually provided data**)
- Stratification
- Adjustment/matching (i.e. adjusting for confounders by including them in a regression model)

- In a DAG, we draw a box around the conditioned variables

Open and closed (blocked) paths in DAGs

- Path = sequence of variables
- At each variable, the path is either open or closed/blocked
- Open path -> path that goes through a set of variables where no variable blocks the path
- Two variables are correlated if there is one or more open paths between them
- If two arrows collide at a variable, this blocks the path
- Conditioning on a variable reverses the status:
  - Conditioning on a variable in an open path blocks that path
  - Conditioning on a collider opens that path

Using DAGs

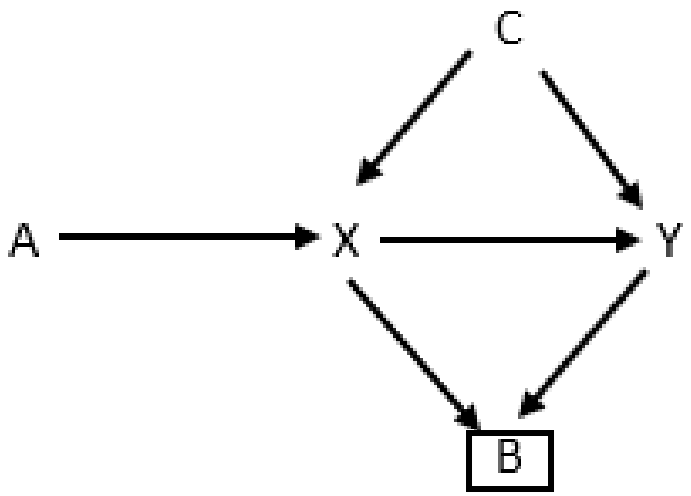


An observed association between X and Y can be because:

- X causes Y
- There is a common cause of X and Y (**confounder** C) that you <u>have not</u> conditioned on

[In DAG speak: conditioning on C blocks/closes the path from X to Y via C]

Using DAGs



An observed association between X and Y can be because:

- X causes Y
- There is a common cause of X and Y (**confounder** C) that you have not conditioned on
- There is a common effect of X and Y (**collider** B) that you have conditioned on

[In DAG speak: conditioning on a collider opens a path from X to Y via B]

DAGs for missing data

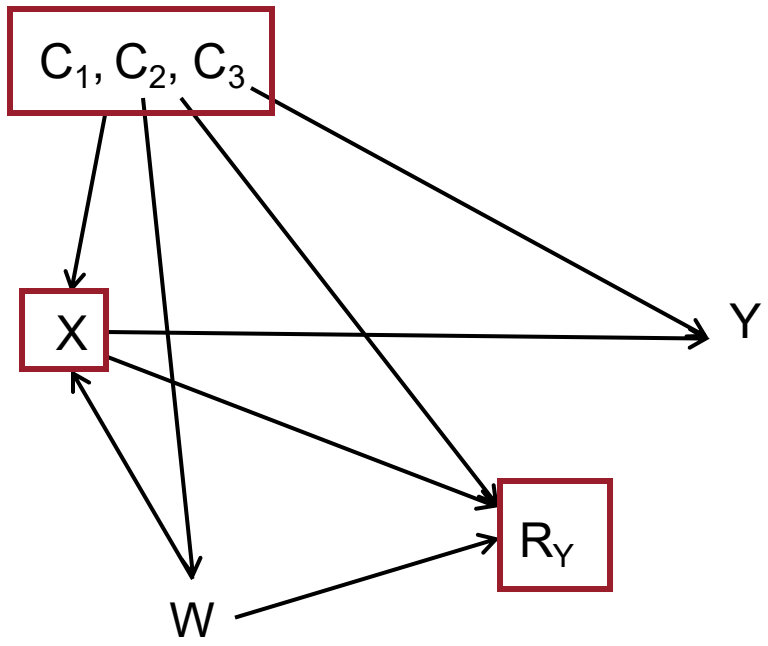
Bias in complete records (case) analysis

- In general, if the outcome Y in your analysis model is unrelated to missingness, conditional on your (analysis model) covariates, then a complete records analysis will be unbiased, although it may be an inefficient use of your data
- IN DAG TERMS: if there is no open pathway between Y and the missingness/response indicator then a complete records analysis will be unbiased

NOTES:

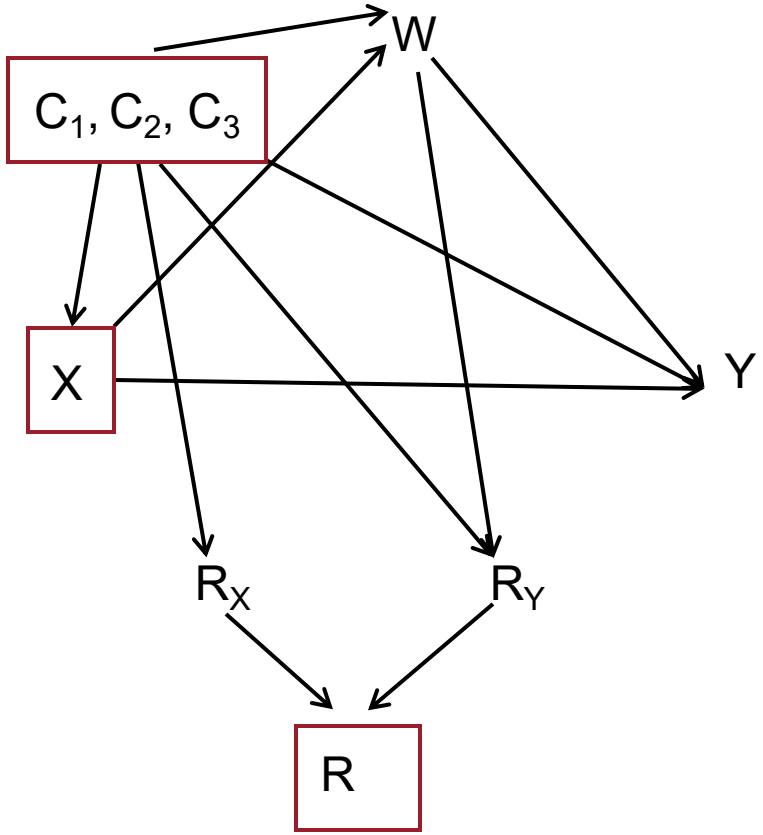
- There are additional situations when complete records analysis will be unbiased, particularly for a binary outcome when the analysis model is logistic regression
- •Here we are talking about bias due to missingness. There may still be bias due to mis-specifying your model or excluding important confounders

Will CRA result in bias? – one missing variable



- Analysis model is regression of Y on X, adjusting for confounders C_1, C_2 and C_3
- We have an additional variable W
- After conditioning on confounders and R, are there any open pathways between the outcome Y and the missingness indicator?
- If the answer is NO (as here), then complete records analysis* (CRA) will give an unbiased estimate of the X-Y relationship
- Regardless of which variable(s) are incomplete, the crucial question always involves Y

Will CRA result in bias? – more than one missing variable



- After conditioning on confounders and R, are there any open pathways between the outcome Y and the missingness indicator?
- Here, the answer is yes (through W), so CRA will give a biased estimate of the X-Y relationship
- Regardless of which variable(s) are incomplete, the crucial question always involves Y

Computer Practical 1: Exploring missing data and complete records analysis in practice

Background – TARMOS framework

Journal of Clinical Epidemiology 134 (2021) 79–88

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework

Katherine J. Lee[a,b,*], Kate M. Tilling[c], Rosie P. Cornish[c], Roderick J.A. Little[d], Melanie L. Bell[e], Els Goetghebeur[f], Joseph W. Hogan[g], James R. Carpenter[h], on behalf of the STRATOS initiative

[a]*Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Melbourne, Australia*
[b]*Department of Paediatrics, University of Melbourne, Melbourne, Australia*
[c]*MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK*
[d]*Department of Statistics, University of Michigan, MI, USA*
[e]*Department of Epidemiology and Biostatistics, University of Arizona, AZ, USA*
[f]*Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium*
[g]*Department of Biostatistics, Brown University, RI, USA*
[h]*MRC Clinical Trials Unit, London School of Hygiene and Tropical Medicine, London, UK*

Accepted 13 January 2021; Published online 2 February 2021

Background – TARMOS framework

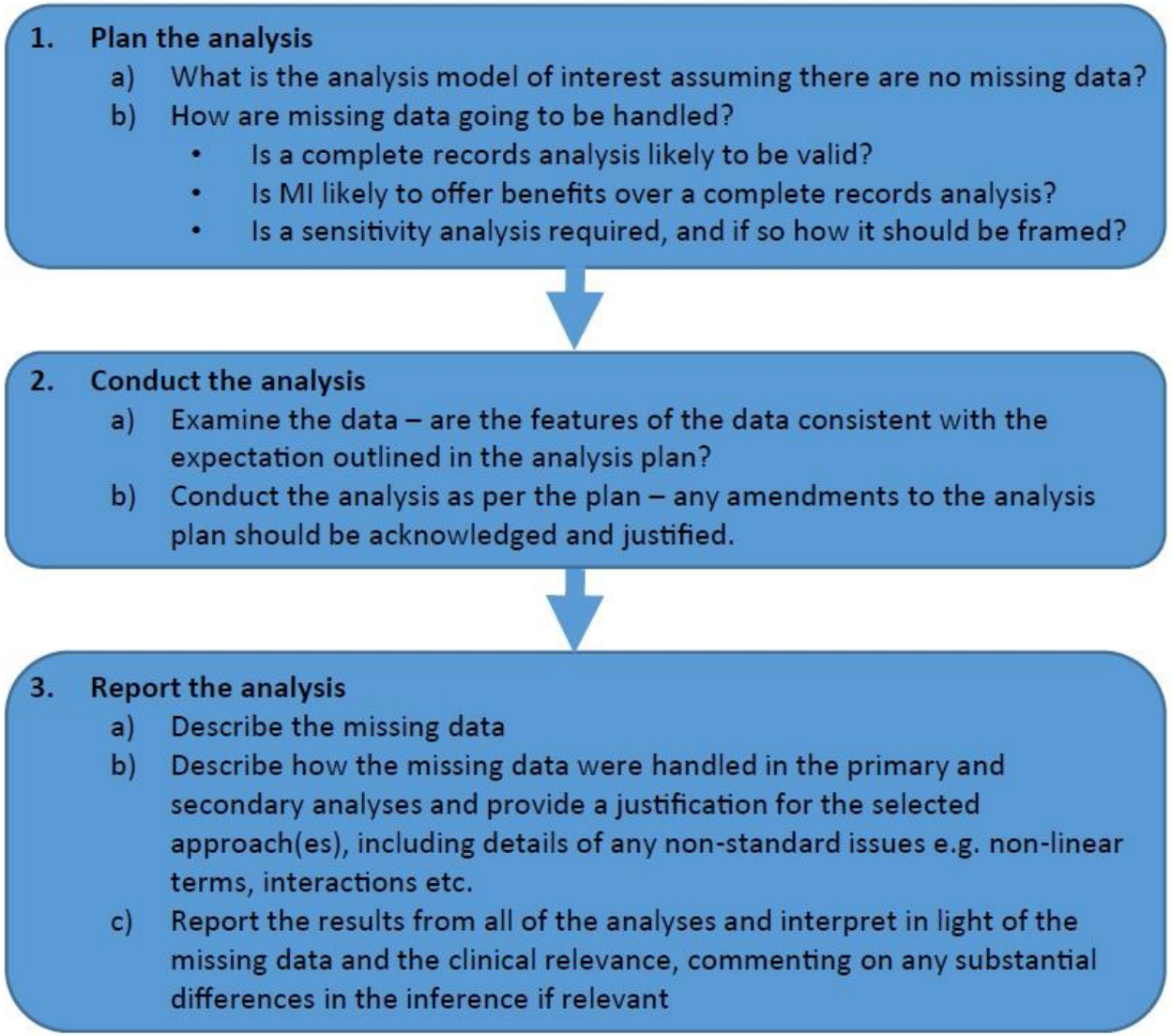


Fig. 1. The framework.

Background – TARMOS framework

1. **Plan the analysis**
   a) What is the analysis model of interest assuming there are no missing data?
   b) How are missing data going to be handled?
      - Is a complete records analysis likely to be valid?
      - Is MI likely to offer benefits over a complete records analysis?
      - Is a sensitivity analysis required, and if so how it should be framed?

Background – TARMOS framework

2. Conduct the analysis

a) Examine the data – are the features of the data consistent with the expectation outlined in the analysis plan?

b) Conduct the analysis as per the plan – any amendments to the analysis plan should be acknowledged and justified.

Background – TARMOS framework

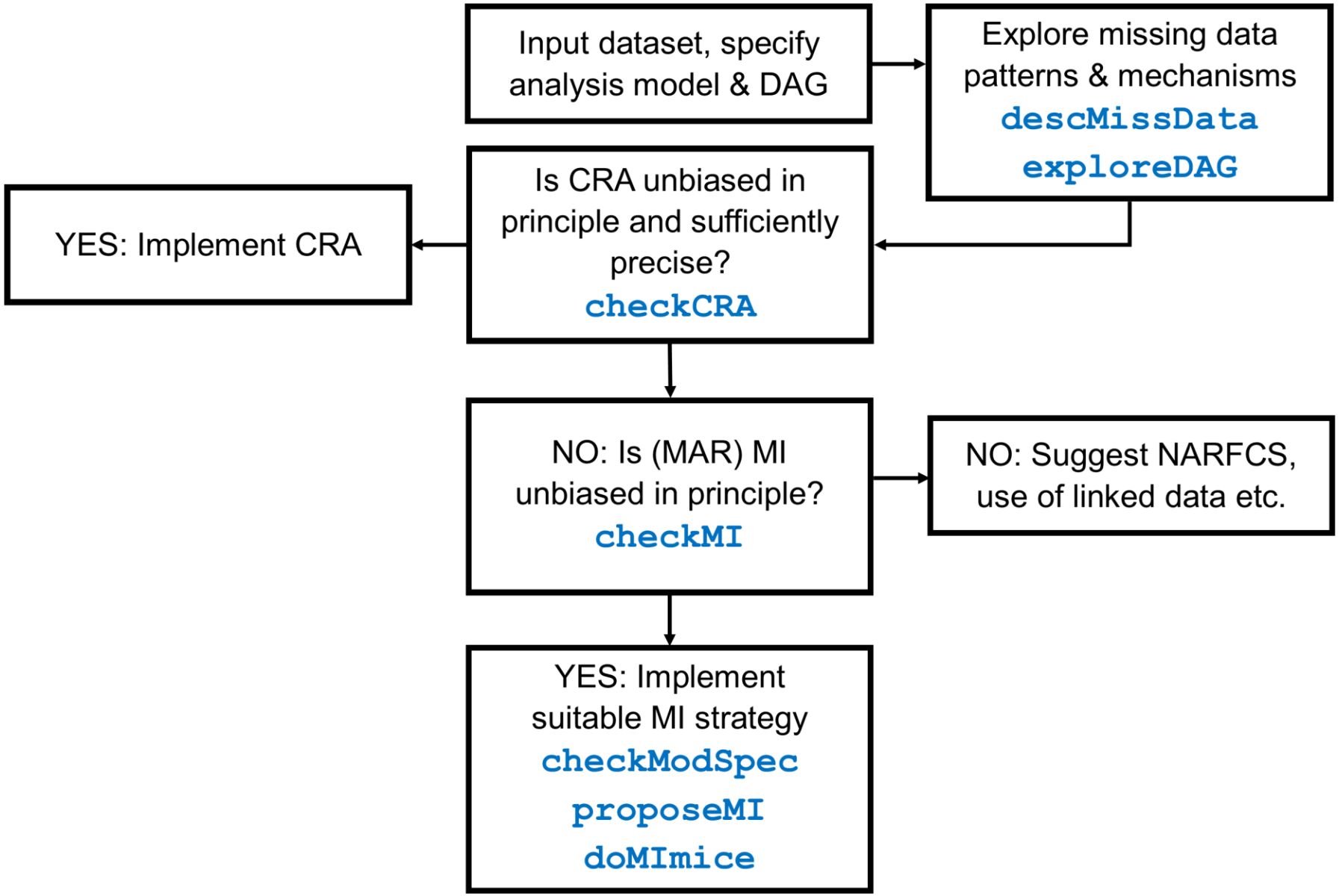
3. Report the analysis

a) Describe the missing data

b) Describe how the missing data were handled in the primary and secondary analyses and provide a justification for the selected approach(es), including details of any non-standard issues e.g. non-linear terms, interactions etc.

c) Report the results from all of the analyses and interpret in light of the missing data and the clinical relevance, commenting on any substantial differences in the inference if relevant

midoc – decision making system



Worked example

- We will use the ‘asrds’ dataset that is included in the ‘midoc’ R package (https://elliecurnow.github.io/midoc/)
- This is a simulated dataset of male adolescents, based on features of real data
- We will investigate the association between family income per annum when the child is age 5y (on the log scale) and child's Adapted Self-Report Delinquency Scale (ASRDS) value at age 17 years (using a transformed, scaled version of this variable)
- For simplicity, we only consider one confounder of this relationship, mother’s education level
- We also have information about the child’s ASRDS value at age 14 years

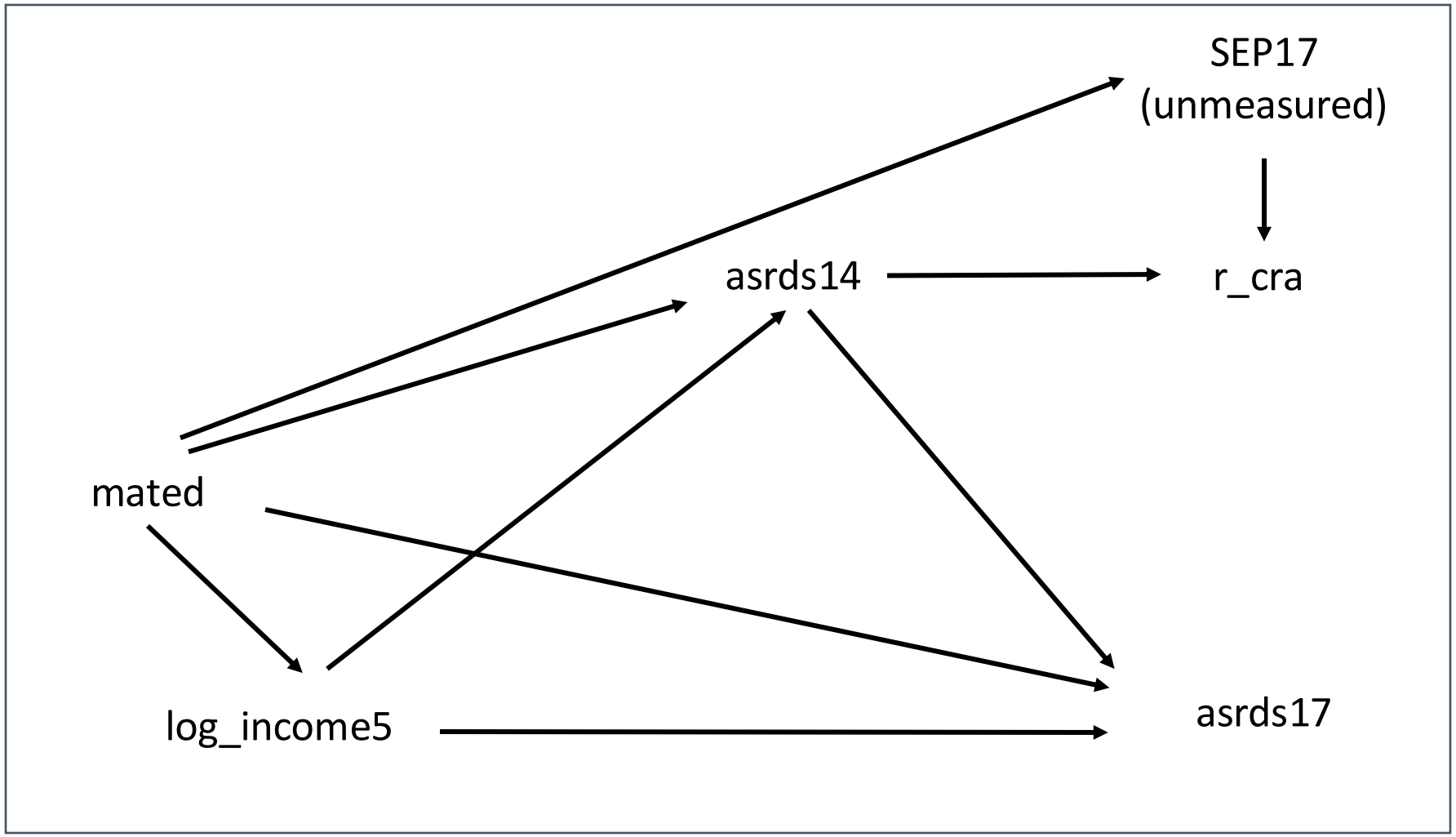
Worked example

- A description of all the variables in the dataset is shown below

Variable	Description
asrds17	Continuous: child's Adapted Self-Report Delinquency Scale (ASRDS) value at age 17y, as a percentage of the maximum possible value of the square-root of ASRDS on the original scale
log_income5	Continuous: logarithm of the family income per annum, in UK pounds, when the child is aged 5y; we believe this is a cause of asrds17
mated	Binary: mother's educational level (post-16y qualification or not) at time of pregnancy; we believe this is a cause of both asrds17 and log_income5
asrds14	Continuous: child's ASRDS value at age 14y, as a percentage of the maximum possible value of the square-root of ASRDS on the original scale; we believe this is a cause of asrds17, and is caused by both log_income5 and mated

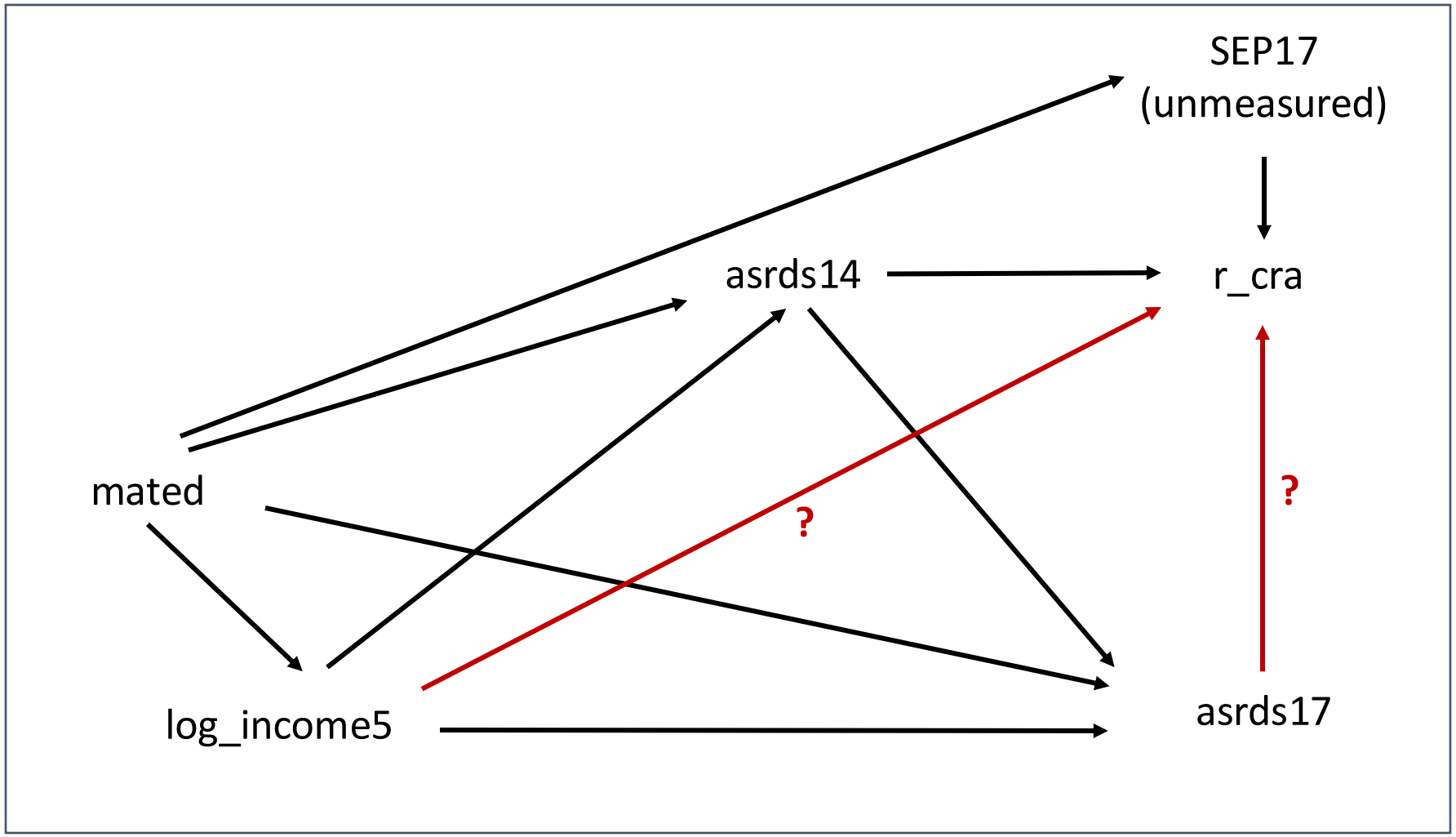
Computer Practical 1: Summary

Worked example – drawing a missingness DAG



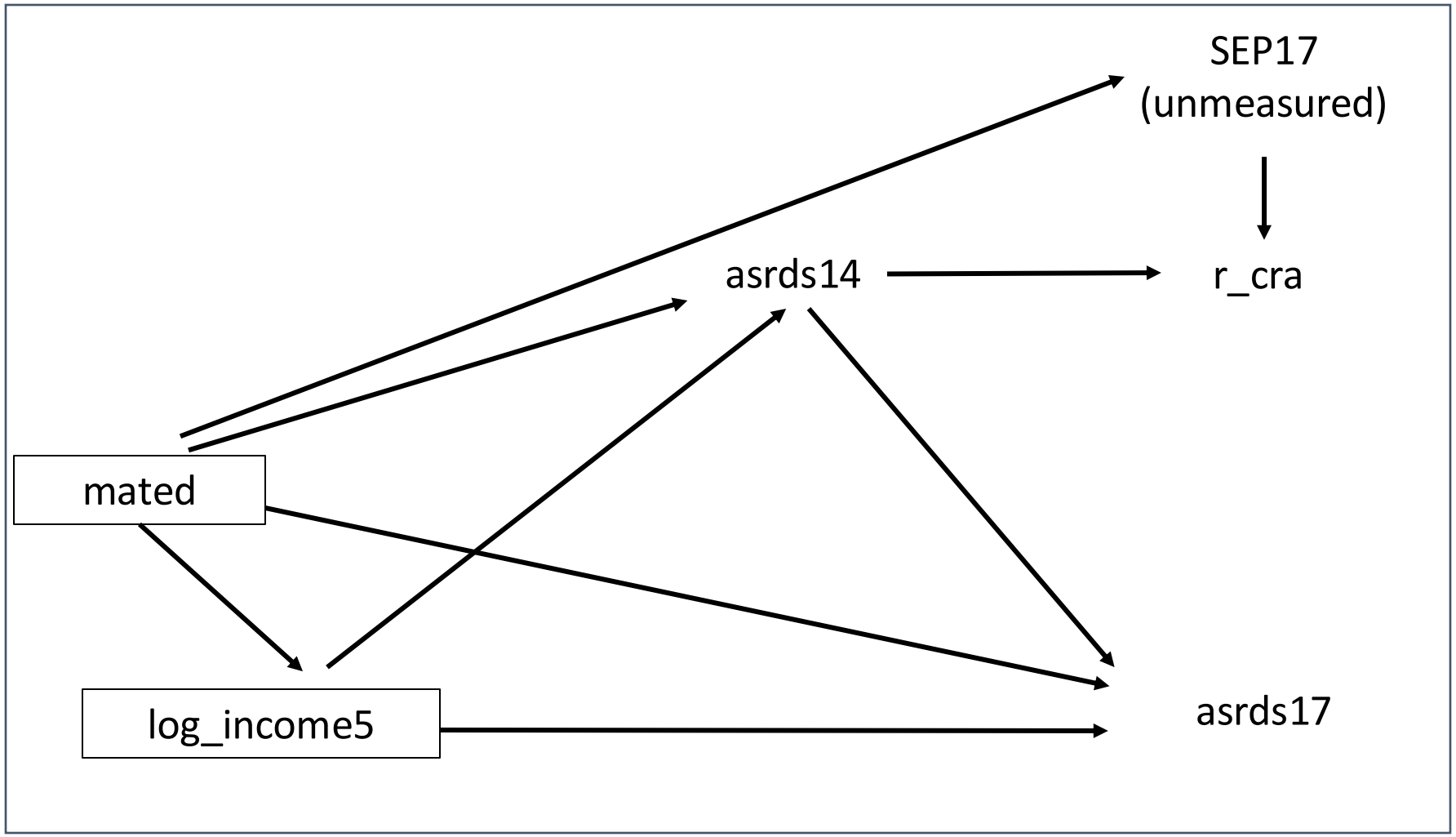
SEP = socio-economic position; suffixes refer to the age of the child at the time the variable was measured

Worked example – explore the missingness DAG



- We cannot use our observed data to determine whether asrds17 is related to missingness - we would need the missing values of asrds17 to explore this

Worked example – check whether CRA is valid in principle



- CRA is not valid for our analysis model because there is an open path between asrds17 (our outcome) and r_cra (missingness indicator) via asrds14
- Including asrds14 in our analysis will change our interpretation

Worked example – drawing a missingness DAG

Recap

- Examining the underlying causal relationships between the variables, and the causes of missing data, will help us choose the best analysis approach
- Necessary even when our question of interest is not causal *e.g.* descriptive or association study

Tips for constructing missingness DAGs

- Missingness DAGs can be constructed using prior knowledge about the study setting, data collection process, *etc.* as well as data exploration
- Can be helpful to start with a simplified DAG e.g. by using 'Z' in place of the set of confounders or covariates
- First specify the DAG for the analysis model, as it would be if there were no missing data
- Next add missingness indicator(s) to your DAG. If you have multiple variables with missing data, you may want to start by including just the complete records indicator in your DAG

Break

Introduction to multiple imputation

Outline

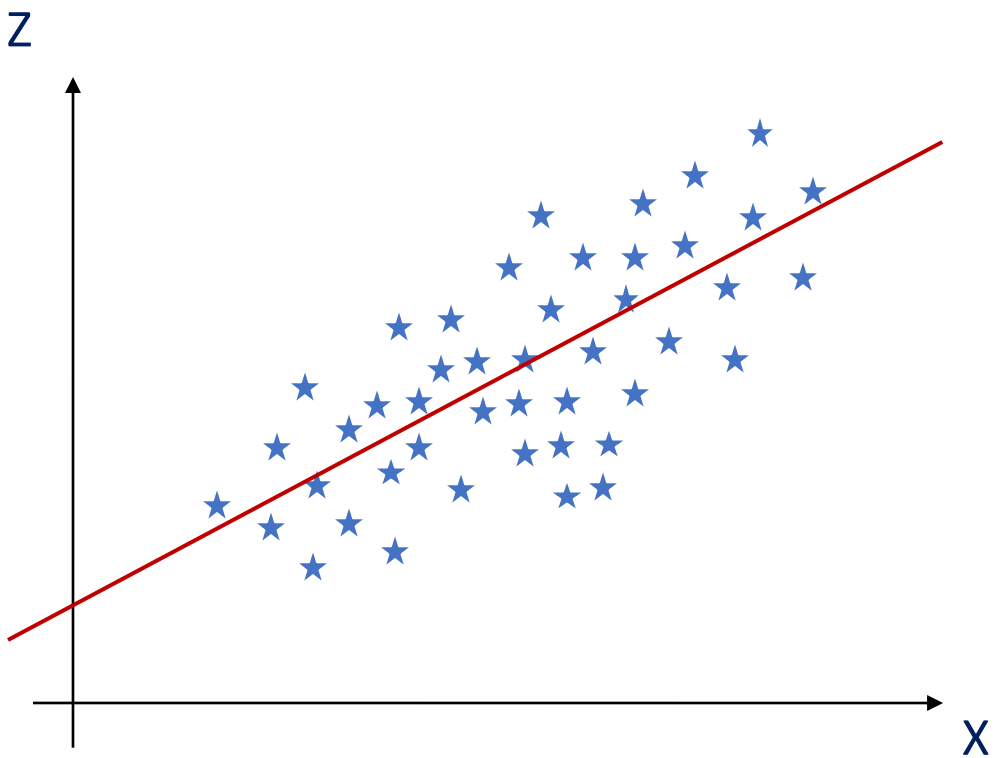
- Explain what multiple imputation (MI) by chained equations (MICE) is and how regression models are used to fill in missing values
- Using a DAG to decide what to include in the imputation model
- Describe what an auxiliary variable is and explain why they are important in MI

Three stages to multiple imputation (MI)

1. Create m (≥ 2) imputed datasets
   - We build regression models and use these to predict the missing values
2. Analyse each imputed (complete) dataset
   - Using standard methods, we estimate our model on each imputed dataset
3. Combine the results in an appropriate way

- We'll start with the simple scenario where on one variable (Z) suffers from missing data

Use of regression to fill in missing data for Z

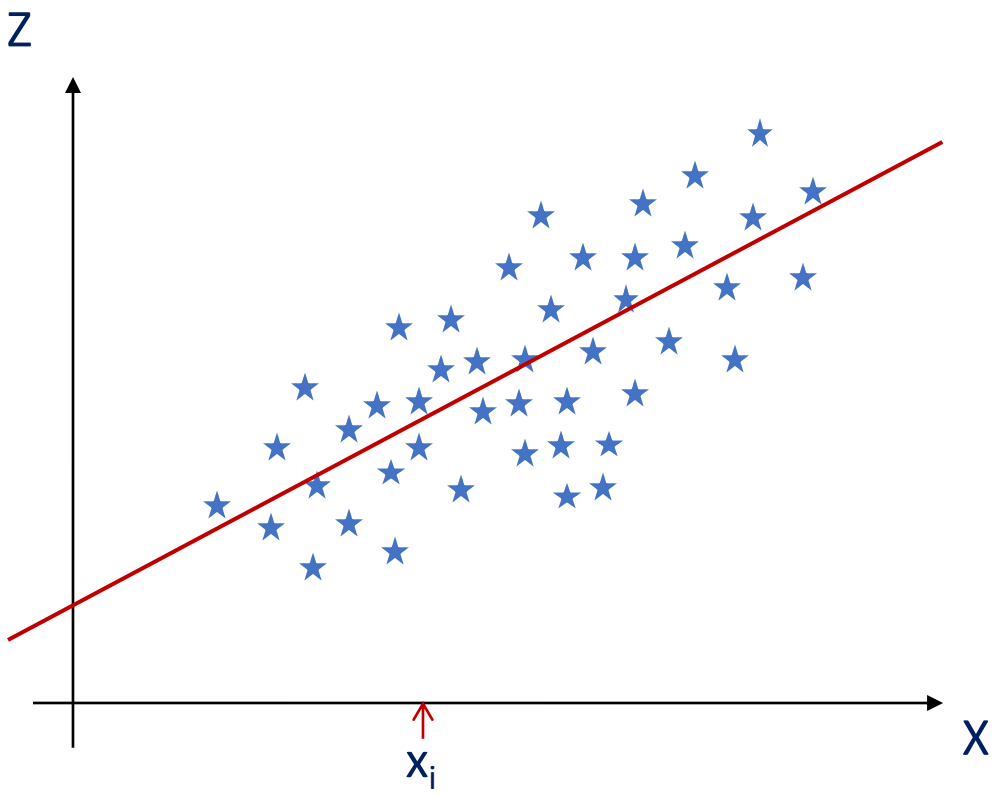


- X is complete
- Z is incomplete
  - z_{obs} are observed
  - z_{mis} are not observed

- Fit a regression model for Z using X as the only predictor
- This gives us the line $z_j = \hat{\alpha} + \hat{\beta} x_j$

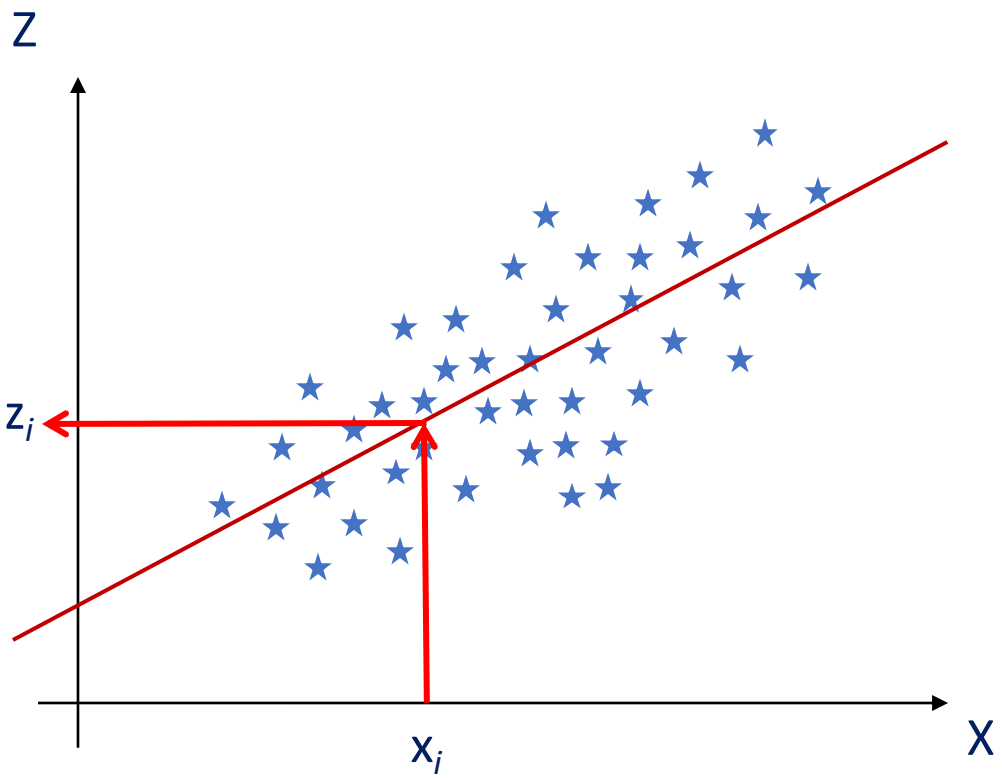
  for any participant *j* in the sample with observed Z (i.e. z_{obs})

Use of regression to fill in missing data for Z



- Say we have a person *i* with a recorded value for X (call this x_i) but a missing value for Z

Use of regression to fill in missing data for Z



- We can use the regression line

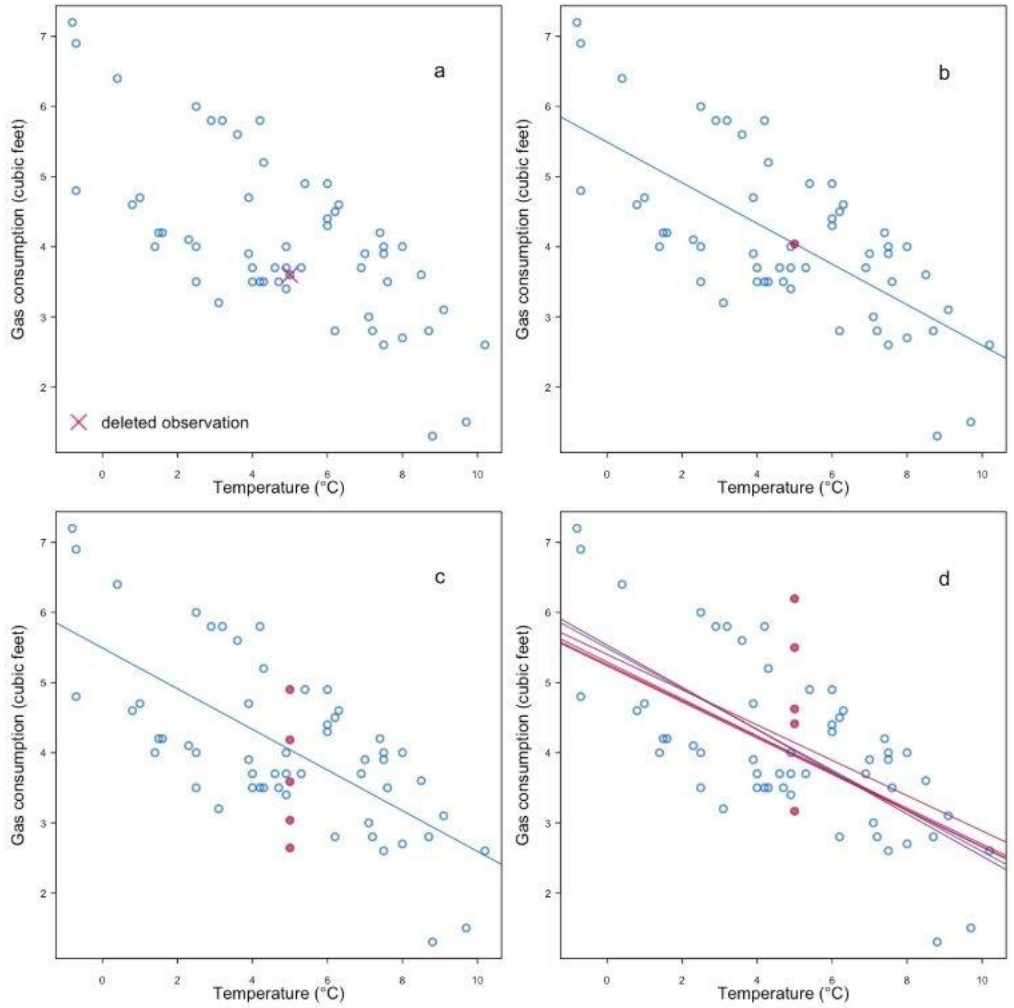
$$z_j = (\hat{\alpha} + \hat{\beta} x_j)$$

to predict what the value of Z might have been, *had we observed it*

$$z_i = (\hat{\alpha} + \hat{\beta} x_i)$$

- Here we are assuming that we can use the sample with observed Z (i.e. z_{obs}) and the *observed* relationship between Z and X, to inform us about people who are missing Z

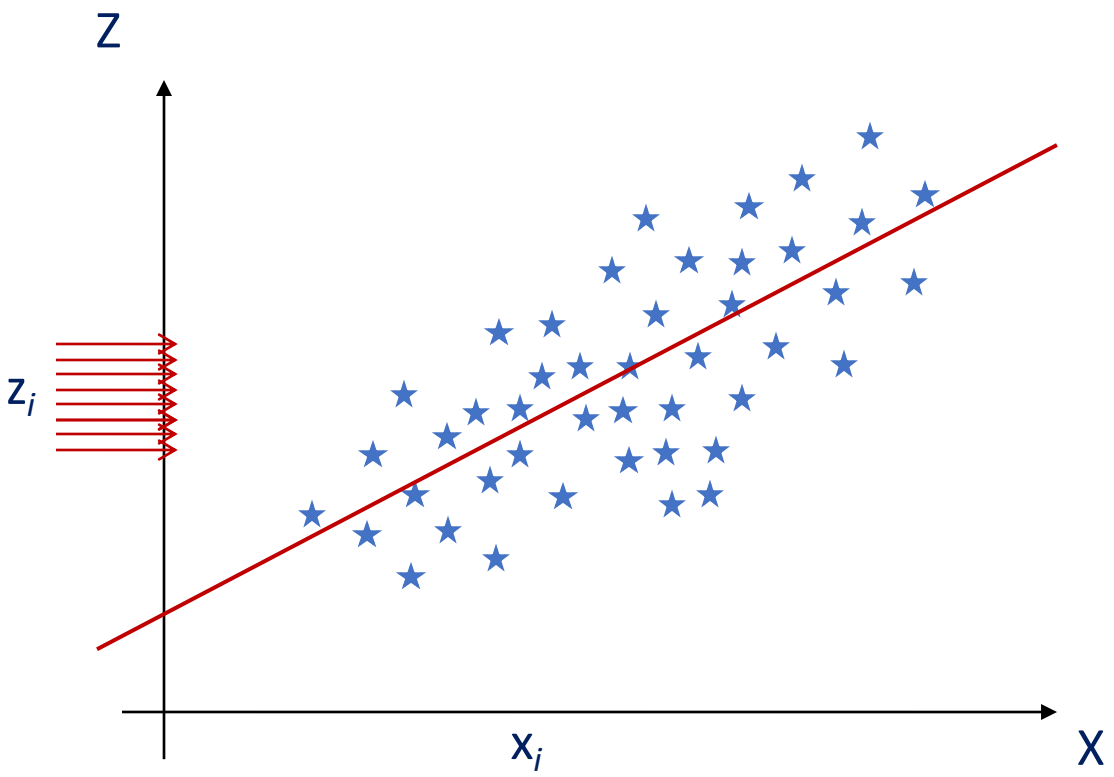
Use of regression to fill in missing data for Z



- We don't actually know the TRUE value of Z for person *i*
- We want the imputed values to reflect this uncertainty
- To do this we "perturb" the parameter values from the fitted regression line by taking draws from their posterior distributions

Images: van Buuren:
Flexible imputation
of missing data

Use of regression to fill in missing data for Z

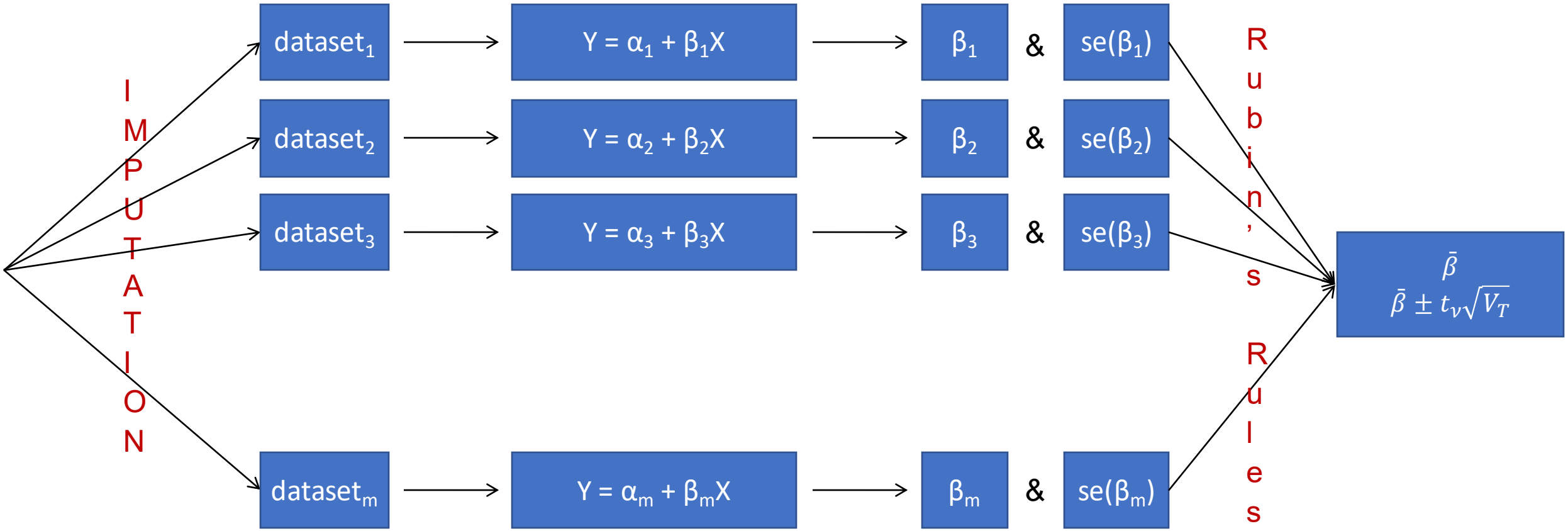


- We end up with lots of different estimates of what z_i might be
- These imputed value of z_i will vary across the imputed datasets that we create

Resulting data shows variation in imputed values

			Ten imputed versions of Z									
ID	X	Zobs	$Zimp_1$	$Zimp_2$	$Zimp_3$	$Zimp_4$	$Zimp_5$	$Zimp_6$	$Zimp_7$	$Zimp_8$	$Zimp_9$	$Zimp_{10}$
1	150	89	89	89	89	89	89	89	89	89	89	89
2	378	104	104	104	104	104	104	104	104	104	104	104
3	366	99	99	99	99	99	99	99	99	99	99	99
4	234	74	74	74	74	74	74	74	74	74	74	74
5	342	103	103	103	103	103	103	103	103	103	103	103
6	307		90	92	97	80	90	101	88	96	89	61
7	318	79	79	79	79	79	79	79	79	79	79	79
8	396	91	91	91	91	91	91	91	91	91	91	91
9	448	111	89	70	88	98	120	116	120	111	130	106
10	191		85	86	89	90	89	92	62	81	92	65
11	296	106	106	106	106	106	106	106	106	106	106	106
12	404		101	81	87	105	118	124	104	109	95	119
13	523		96	93	95	111	118	108	126	107	118	124
14	427	99	99	99	99	99	99	99	99	99	99	99
15	249		92	81	96	97	89	96	80	82	88	91

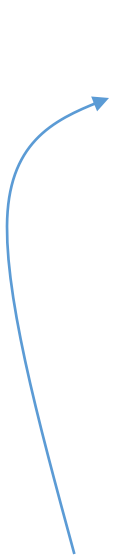
MI steps


IMPUTATION
dataset$_1$
dataset$_2$
dataset$_3$
dataset$_m$
$Y = \alpha_1 + \beta_1 X$
$Y = \alpha_2 + \beta_2 X$
$Y = \alpha_3 + \beta_3 X$
$Y = \alpha_m + \beta_m X$
β_1
β_2
β_3
β_m
&
se(β_1)
se(β_2)
se(β_3)
se(β_m)
Rubin's Rules
$\bar{\beta}$
$\bar{\beta} \pm t_\nu \sqrt{V_T}$

Multiple Imputation by Chained Equations (MICE)

MICE procedure

1. Crudely fill in all the missing values for all the Z-variables



2. Select the Z variable that had the least amount of missing data, replace the crudely filled in values with missing data again, and estimate a regression model for Z_{obs} using any X-variables and/or other crudely filled-in Z's
3. Fill-in the values Z_{mis} by making predictions from the Step-2 regression model, perturbing the model's parameter estimates to account for uncertainty
4. Go back to Step-2 for the next Z variable that contains missing values, etc.

5. Repeat steps 2-4 for several cycles and at the end save the imputed dataset
6. Repeat this whole process m times. The imputed values for Z_{mis} will vary across these m datasets

MICE: simple illustrative example

Analysis: regress % bodyfat (outcome) on age and weight (predictors)

bodyfat	age	weight
14.2	28	151.25
5.0	47	127.50
12.0	61	143.00
.	47	197.00
14.0	50	161.00
.	26	223.00
24.6	36	226.75
22.6	43	170.75
11.3	55	.
28.3	72	186.75

Missing data for bodyfat (2 obs) and weight (1 obs)

MICE: simple illustrative example

bodyfat	age	weight
14.2	28	151.25
5.0	47	127.50
12.0	61	143.00
16.5	47	197.00
14.0	50	161.00
16.5	26	223.00
24.6	36	226.75
22.6	43	170.75
11.3	55	**176.33**
28.3	72	186.75

1. Fill in all missing values in the dataset by some crude method (e.g. randomly sampling from the observed values)

- Here we've replaced missing values with the sample mean of the observed data

MICE: simple illustrative example

bodyfat	age	weight
14.2	28	151.25
5.0	47	127.50
12.0	61	143.00
16.5	47	197.00
14.0	50	161.00
16.5	26	223.00
24.6	36	226.75
22.6	43	170.75
11.3	55	.
28.3	72	186.75

2. Select the Z variable that had the least amount of missing data and estimate a regression model for Z_{obs} using any X-variables and/or other crudely filled-in Z's

- We throw away the crudely imputed value for weight since this was the variable with the least missing data
- We then regress weight on bodyfat and age using the remaining nine observations

Weight = 162.7 – 1.08*age + 3.67*bodyfat

MICE: simple illustrative example

bodyfat	age	weight
14.2	28	151.25
5.0	47	127.50
12.0	61	143.00
16.5	47	197.00
14.0	50	161.00
16.5	26	223.00
24.6	36	226.75
22.6	43	170.75
11.3	55	**146.25**
28.3	72	186.75

3. Fill-in the values Z_{mis} by making predictions from the Step-2 regression model, perturbing the model's parameter estimates to account for uncertainty

- Prediction would be

```
predwt = 162.7 – 1.08*(55) + 3.67*(11.3)
```

- However, we draw from the parameters' posterior distributions - including the variance of the residuals σ^2 - (rather than using 162.7, 1.08 and 3.67) to account for uncertainty

MICE: simple illustrative example

bodyfat	age	weight
14.2	28	151.25
5.0	47	127.50
12.0	61	143.00
.	47	197.00
14.0	50	161.00
.	26	223.00
24.6	36	226.75
22.6	43	170.75
11.3	55	**146.25**
28.3	72	186.75

4. Go back to Step 2 for the next Z variable that contains missing values, etc.

- The 9th observation now contains its freshly imputed variable.
- Bodyfat was the other incomplete variable so we have now removed the crudely imputed values so we're ready to insert new predicted values

MICE: simple illustrative example

bodyfat	age	weight
14.2	28	151.25
5.0	47	127.50
12.0	61	143.00
22.8	47	197.00
14.0	50	161.00
28.1	26	223.00
24.6	36	226.75
22.6	43	170.75
11.3	55	**146.25**
28.3	72	186.75

4. Go back to Step-2 for the next Z variable that contains missing values, etc.

- Regress bodyfat on age and weight:

bodyfat = -27.12 + 0.22*weight + 0.14*age

- Perturb the parameters from this regression model (including the variance of the residuals σ^2) and use the perturbed parameters to predict new values for the two missing observations.

MICE: simple illustrative example

5. Repeat steps 2-4 for several cycles and at the end save the imputed dataset

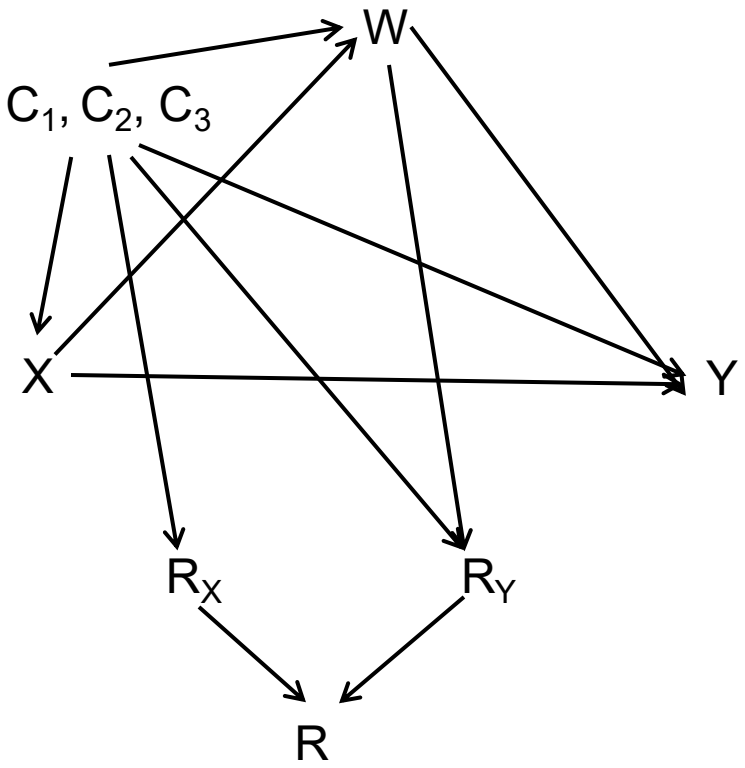
- i.e. make weight missing again, regress weight on bodyfat and age, draw from posterior distribution, fill in missing weight, etc.) and cycle through all variables again.
- After a few “cycles” the predicted values will settle down, changing by gradually smaller amounts. Save current values (imputed dataset 1)

6. Start whole process again, repeat until you have m imputed datasets…

Imputer’s vs analyst’s model

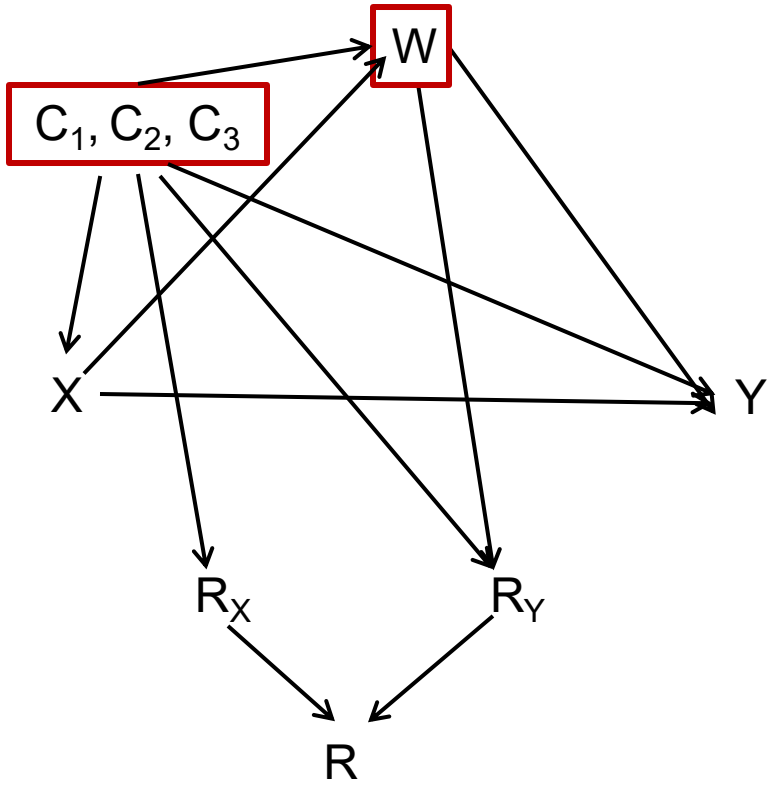
- Imputation model should be compatible with the substantive model
  - Contain ALL variables required in the analysis model
  - Preserves associations or relationships among variables, including any non-linear relationships or interactions
  - Reflect other aspects of the data e.g. if numeric data are skewed
- Converse is not necessary
  - You can include extra variables in the imputation model that are not in the analysis model (key advantage over CRA) -> auxiliary variables

Using a DAG to inform multiple imputation



- For CRA to be unbiased we require no open path from Y to R
- When we impute missing data using chained equations (MICE) we build a prediction model for each incomplete variable
- Each incomplete variable is an outcome in the imputation step so we must consider whether there is an open path from it to R when we condition on the other data

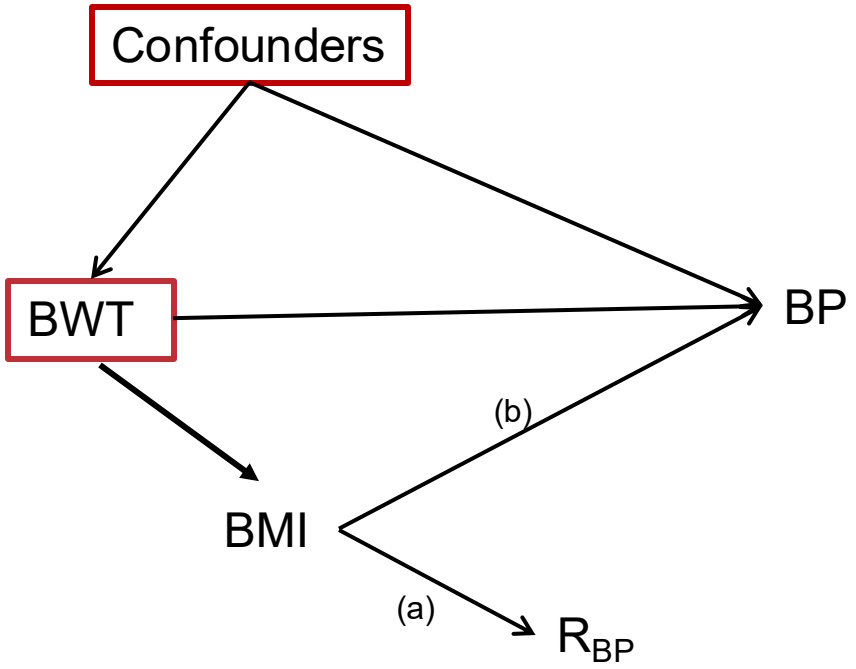
Using a DAG to inform multiple imputation



- For CRA to be unbiased we require no open path from Y to R
- When we impute missing data using chained equations (MICE) we build a prediction model for each incomplete variable
- Each incomplete variable is an outcome in the imputation step so we must consider whether there is an open path from it to R when we condition on the other data
- In this example, we must condition on C and W to block the paths from X and Y to R
- Provided our imputation models contain these variables (and we specify their relationships correctly) we can use imputation to achieve an unbiased estimate of the effect of X on Y
- It helps to have available good quality auxiliary data

Example: Regression of adult blood pressure on birthweight and confounders (age, sex, family history of CHD, education level)

- Missing blood pressure more likely with higher BMI (a)
- BMI associated with BP (b)
- Could include BMI as covariate to make complete records analyses unbiased – BUT BMI is on the causal pathway – so we don't want to include it
- Impute blood pressure based on all other variables in the model and BMI (to block open pathway from BP to its missingness)

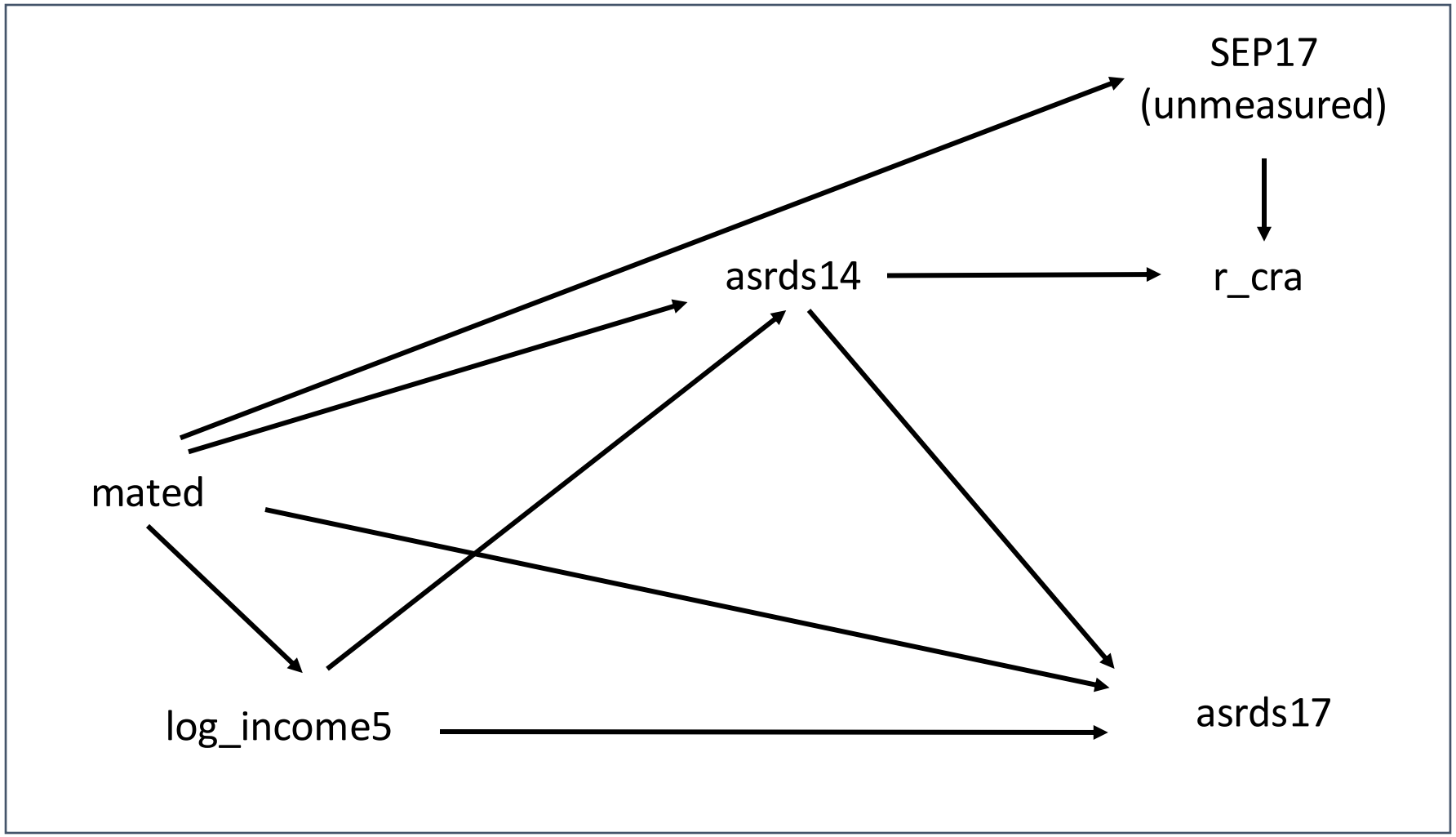


Summary – checking whether CRA is valid in principle

- In general, CRA will be valid (unbiased) if our outcome is independent of missingness, given the covariates in our analysis model
- If CRA is not valid, we may be able to obtain an unbiased estimate using multiple imputation (MI)
  - Our dataset includes an additional variable, Adapted Self-Report Delinquency Scale score at age 14, which is likely to be predictive of our missing data
  - We can use this as an **auxiliary variable** in our imputation model
- Note that even when CRA is valid, we may be able to obtain a more precise estimate using MI

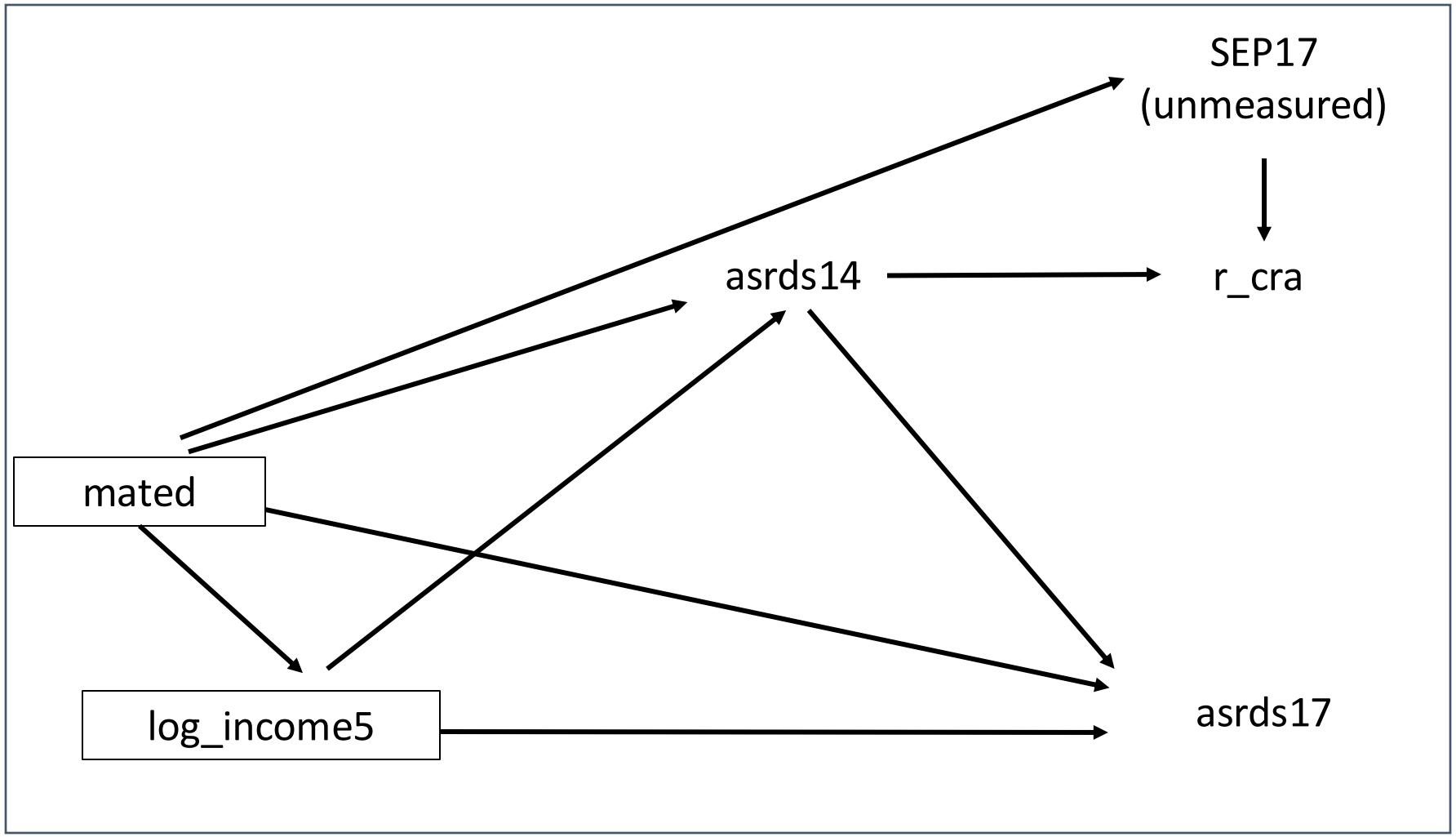
Computer Practical 2: MI in practice

Worked example – checking whether MI is valid in principle



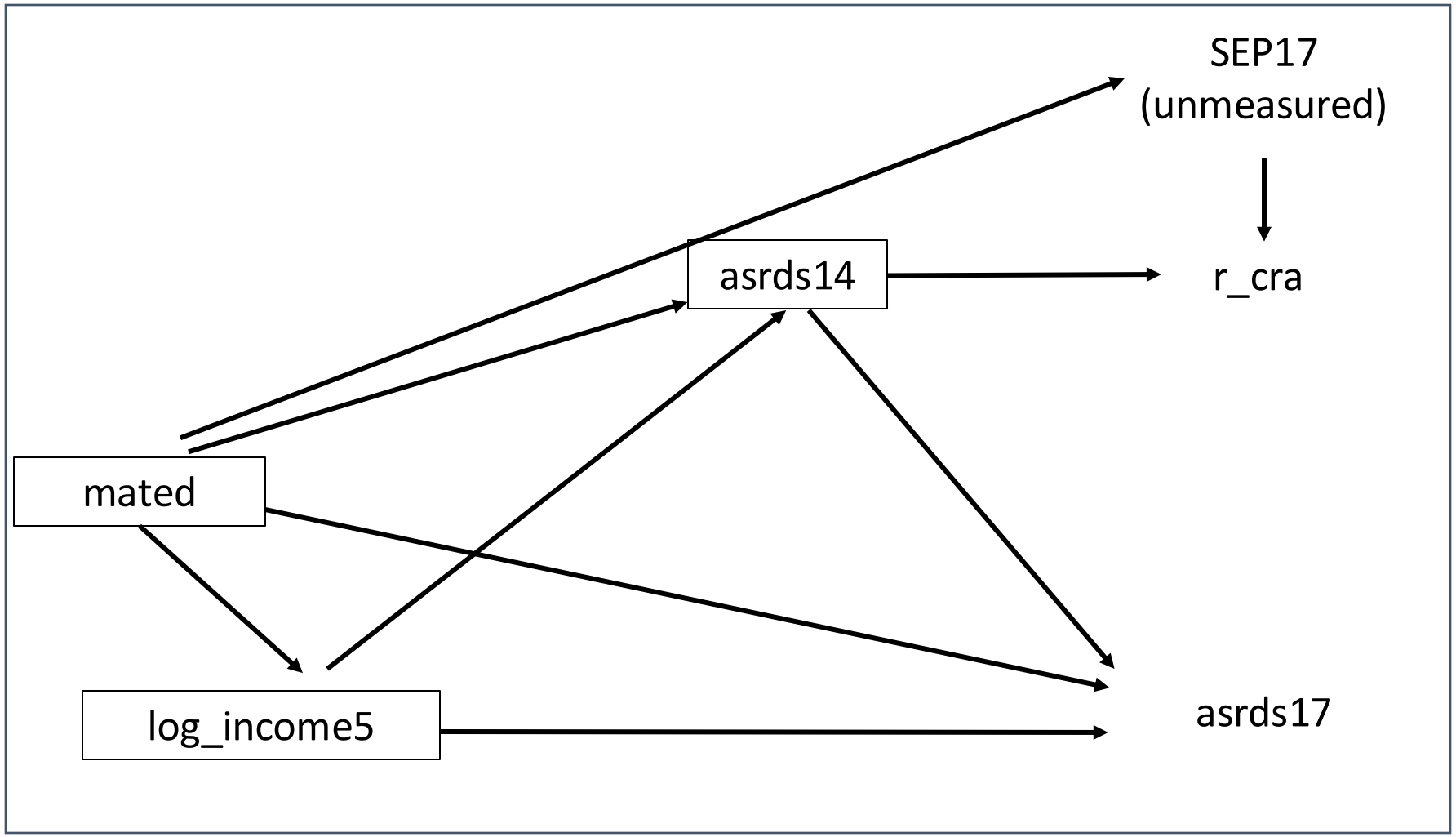
- We will apply MI, imputing missing values of asrds17
- Which variables should we include in the imputation model for asrds17?

Worked example – checking whether MI is valid in principle



- We must include all other analysis model variables in the imputation model for asrds17 (so that our imputation and analysis models are compatible)

Worked example – checking whether MI is valid in principle



- We must include all other analysis model variables in the imputation model for asrds17 (so that our imputation and analysis models are compatible)
- We must also include asrds14 as an **auxiliary variable** (to block the path between asrds17 and r_cra)

Summary – checking whether MI is valid in principle

- In general, MI will be valid (unbiased) if **all** partially observed variables are independent of missingness, given the (observed data on) predictors in the imputation model
- Imputation models should always include all analysis model variables, in the same form as in the analysis model *e.g.* including interactions
- Auxiliary variables need to be chosen carefully
  - Choose variables that are most predictive of the missing data
  - Avoid variables that predict missingness but not the missing data
  - Avoid colliders

Worked example – explore the specification of the imputation model

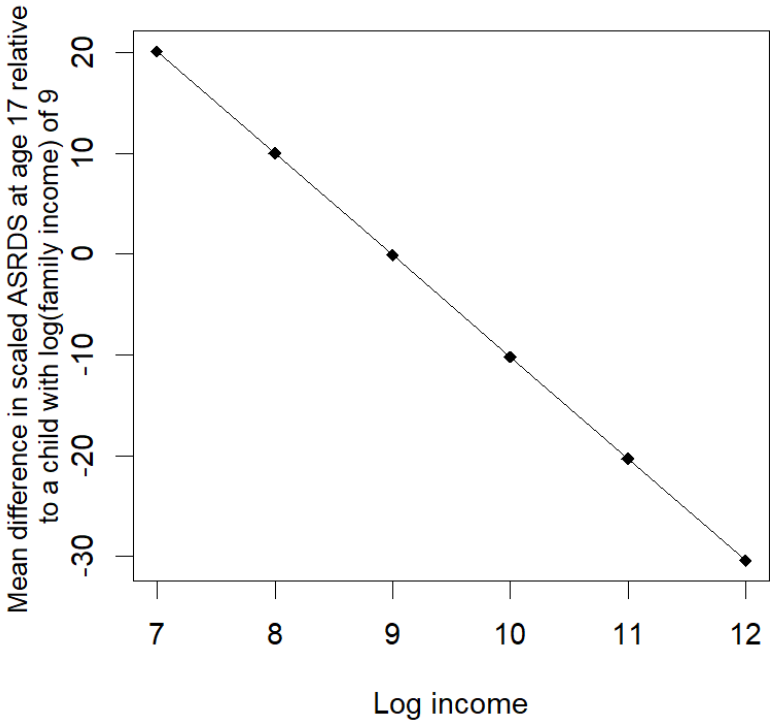
- We have explored whether CRA and MI are valid in principle using our missingness DAG
- We have not made any assumptions about the form of our variables, or their relationships with each other
- For MI to give unbiased estimates, imputation models must be both compatible with the analysis model and correctly specified:
  - Containing all the variables required for the analysis model
  - Including all relationships implied by the analysis model *e.g.* interactions
  - **Specifying the form of all relationships correctly**

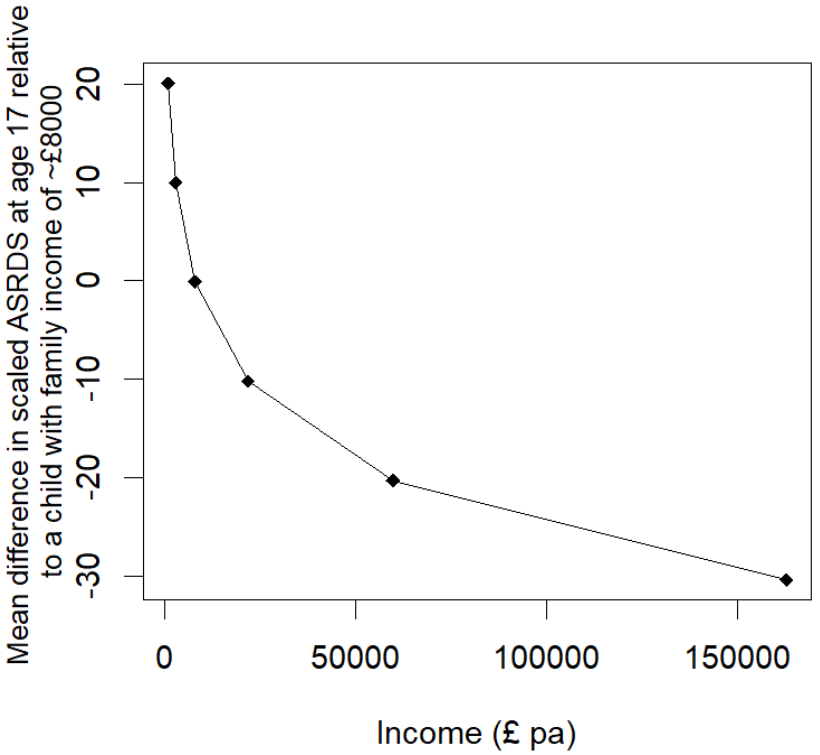
Worked example – perform MI using the proposed imputation model

- We have explored both the validity of MI in principle, using our missingness DAG, and the specification of our imputation model, based on our observed data
- We will now use the **midoc** functions **proposeMI** and **doMImice** to specify and apply the best options when using the **mice** R package (https://doi.org/10.18637/jss.v045.i03)
- What is the MI estimate of the association between family income per annum when the child is age 5y (on the log scale) and child's ASRDS value at age 17y?

Worked example – interpretation of the MI estimate

- Our MI estimate of the association between family income per annum when the child is age 5y (on the log scale) and child's ASRDS value at age 17y is: -10.0 (95% CI: -11.1, -9.0) – this is illustrated in the left-hand plot below
- We can transform family income back to its original scale for easier interpretation – as shown in the right-hand plot below





Code to produce the plots

```
# Income on the log scale:
x<-c(7:12)
plot(x=x, y=c(-10*(x-9)), type='o', pch=18,  xlab="Log
income", ylab="")
title(ylab="Mean difference in scaled ASRDS at age 17
     relative \nto a child with log(family income) of 9",
     mgp=c(2,1,0),cex.lab=0.9)

# Income on the original scale:
plot(x=exp(x), y=c(-10*(x-9)), type='o', pch=18,
     xlab="Income (£ pa)", ylab="")
title(ylab="Mean difference in scaled ASRDS at age 17
     relative \nto a child with family income of ~£8000",
     mgp=c(2,1,0),cex.lab=0.9)
```

Worked example – compare MI and CRA estimates

- We have found that CRA is not valid in principle
- In practice, how different is the (biased) CRA estimate from the (unbiased) MI estimate?
  - The MI estimate is: -10.0 (95% CI: -11.1, -9.0)
  - The CRA estimate is: -9.3 (95% CI: -10.3, -8.2)
- The CRA estimate is attenuated (closer to the null)
- The CRA estimate is also (very slightly) less precise than the MI estimate (larger standard error)
  - In general, MI does not recover more information than CRA when only the outcome is partially observed (although we gain some precision by including auxiliary variables that are predictive of the missing data)
  - We may gain greater precision from MI when covariates are partially observed

Summary - when is MI unbiased?

- In general, MI relies on the missing at random (MAR) assumption:

  Given the observed data, the probability that data are missing is independent of the true values of the partially observed variable(s)

- MI estimates are generally biased if data are missing not at random (MNAR) *i.e.* if the probability that data are missing is related to the (unobserved) values of the partially observed variable(s) even after conditioning on the observed data

What do we do when data may be MNAR?

Where we suspect data are MNAR, we can:

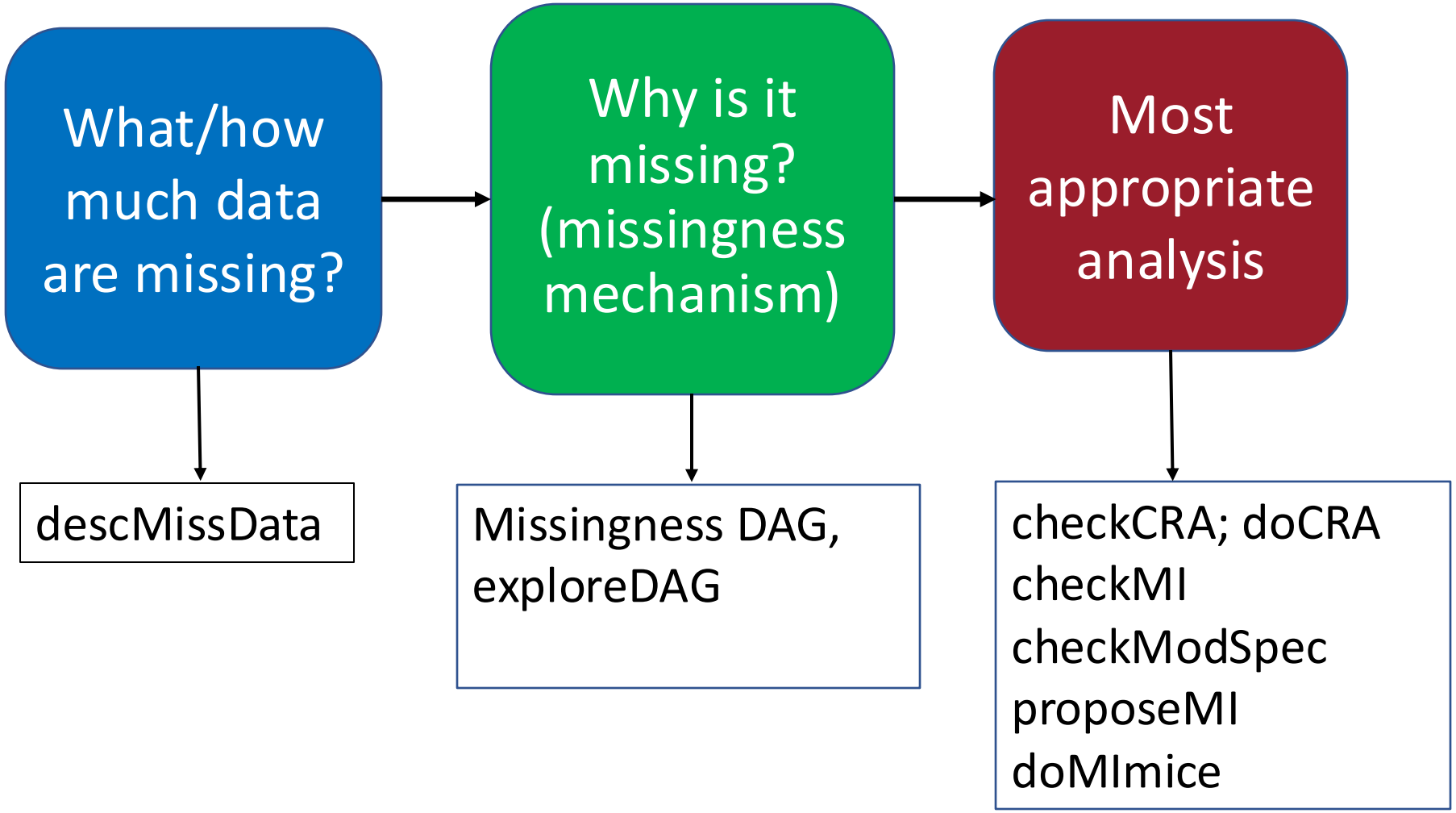
- Explore the sensitivity of MI results to departures from the MAR assumption
- Use auxiliary variables (variables not in the analysis model but which may contain information about the missing data)
  - ➢ We aim to reduce the bias due to data MNAR
  - ➢ See Cornish *et al* (2017) and Curnow *et al* (2024) for further details

Cornish R, Macleod J, Carpenter J, Tilling K. 2017. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: A simulation study. *Emerging Themes in Epidemiology.*

Curnow E, Cornish RP, Heron JE, Carpenter JR, Tilling K. 2024. Multiple imputation using auxiliary imputation variables that only predict missingness can increase bias due to data missing not at random. *BMC Med Res Meth.*

Final Summary and Discussion

Summary


What/how much data are missing?
Why is it missing? (missingness mechanism)
Most appropriate analysis
descMissData
Missingness DAG, exploreDAG
checkCRA; doCRA
checkMI
checkModSpec
proposeMI
doMImice

Feedback

We are very grateful for your feedback on **midoc**

Please email any comments to

Ellie Curnow: elinor.curnow@bristol.ac.uk

Rosie Cornish: rosie.cornish@bristol.ac.uk

White Rose Social Sciences DTP

Economic and Social Research Council

WRDTP Feedback Form

If QR not available, use this link: https://forms.gle/WcnsQgNQ2eH3yjcP9

University of Bradford · University of Hull · University of Leeds · Manchester Metropolitan University · Sheffield Hallam University Knowledge Applied · University of Sheffield · University of York

Thank you