Missing data and imputation

Assoc. Prof. Cameron Hurst
cameron.hurst@cdu.edu.au

Faculty of Health,
Charles Darwin University

20th October, 2023

Who am I?

- Academic biostatistician (and epidemiologist) for over 20 years
- Have considerable history with Thailand. I lived here for almost 6 years and was Head of Biostatistics Centre of Excellence, Chulalongkorn's Faculty of Medicine.
- LOVE THAILAND. In fact, we (family and I) plan to move back in the next 6 months or so.
- Research encompasses the full range of health science: Biomedical $\rightarrow$ Clinical $\rightarrow$ Population-based
  - Coming up on 200 published articles
- Main interest has been in epidemiology of chronic diseases (esp. diabetes), studying clinical outcome trajectory and pyschometric/clinimetric scale development
- Last few years, focused more on reproductive epidemiology ('Mums & Bubs')
- Most recently in the context of Australia's indigenous people who have had very poor maternal and perinatal outcomes

Introduction to missing data

- Missing data will ALWAYS have some impact on statistical analysis and therefore our conclusions
- Impact of missingness depends on the extent and **mechansim** of missing data
- Conventional approach (complete case analysis) is perhaps one of the worst approaches
- The aim of this session is to discuss the implications of missing data and consider methods for dealing with this problem

What we will cover...

1. Introduction
2. Missing data
   - Mechanism of missing data
   - Strategies for dealing with missing data
3. Intro to imputation methods
   - Basic (naive) imputation methods
   - Model-based single imputation
   - Multiple imputation (MI)
4. Imputation for longitudinal data

Scope of this session

Main focus will be on imputation for

- Both continuous and categorical data
- Both outcomes AND covariates
- Both cross-sectional (single time point) and longitudinal (or otherwise clustered) designs

Mechanism of Missing Data

The mechansim of missing data is represented by the process that generated it.

In practice data can be missing for many reasons including, *lost data, CRFs incorrectly filled out, transcription error, patient refusal, patient too ill to participate, drop-out, faulty measurement devices, etc...*

But regardless of the reasons, the mechanisms of missing data fall into one of three main categories:

1. **MCAR**: Missing Completely at Random
2. **MAR**: Missing at Random
3. **MNAR**: Missing Not at Random

MCAR: Missing Completely At Random

- When the data are missing comletely at random, we can think of our sample as a RANDOM sample of the full dataset
- In other words, missingness does not depend on any other data (either present or missing)
- For this reason, **MCAR reduces the precision of our estimates** (i.e. an effective reduction in sample size $\rightarrow$ Wider CIs)
- But **MCAR should NOT result in biased estimates**

Hint: MCAR

Think of MCAR as getting a complete dataset and RANDOMLY removing elements from this dataset. The only real damage is a reduced sample size $\Rightarrow$ reduced precision (wider 95%CIs)

MAR: Missing At Random

- Data are said to be MAR if, given the observed data, **failure to observe (missingness) does NOT depend on the VALUE of variable (after accounting for other variables)**.
- In other words, the missing data is related to some other factors in the data, but not to the specific variable that has the missing value
- This implies that our missing data may depend on something we have observed elsewhere in the dataset, and therefore we can predict based on the relationships in non-missing values

Note: MAR

For MAR, **missingness** (of a variable) is associated with **other** variables in the dataset

MNAR: Missing Not At Random

- Data are said to be MNAR, if failure to observe data (missingness) depends on what **value** would have been observed or other missing values in the dataset
- One common example of MNAR is in longitudinal studies where missingness is a result of Loss-To-Follow-Up, and this drop-out depends on study variables (too sick, young etc)
  - EG the likelihood of obtaining a coping score in a Quality of life (QoL) study depends on the coping score that would have been observed i.e. $\downarrow$ coping score $\Rightarrow \uparrow$ P(drop-out)

Note: Ignorable and Non-ignorable mechansims of missingness

- MNAR is often called **non-ignorable** because this mechanism CANNOT be ignored when we make inferences about the parameters. We must MODEL the nature of missingness
- But we can ignore MCAR and MAR missingness; **ignorable**

How to find what mechanism is driving the missing data

- Little's test for MCAR can be used to examine whether data are missing completely at random.
- In terms of MAR vs MNAR we can go back to fundamentals:
  1. For each variable create a missingness dummy variable
  2. Check for associations between all variables in the dataset with each missingness variable (best to use standard logistic regression) - Suggests (at least) MAR
  3. Check for association among the MISSINGNESS variables (e.g. χ^2 or Logreg) - **Associations among missingness variables $\Rightarrow$ MNAR**

Why is it important to investigate mechanism?

Because what mechanism drives missingness of the data governs how we should best deal with the missingness. If we can at all. Technically, MNAR is a non-solvable problem. Our only recourse is to very naive (and conservative) imputation methods

Example: Singleton births

Study of Australian indigenous mothers and babies.

	GestationalAgeBirth	Birthweight	SmokedInPregnancy	NA.SmokedInPregnancy
1	40	3590	1	0
2	35	2050	*NA*	1
3	39	3660	0	0
4	39	4855	0	0
5	38	2950	0	0
6	38	3400	0	0
7	0	3650	1	0
8	39	7000	0	0
9	40	4710	0	0
10	39	3620	*NA*	1

We create a dummy variable indicating whether maternal smoking during pregnancy is missing (1) or not (0)

Example: Singleton births (MCAR or MAR/MNAR)

Now using (standard) binary logistic regression let's see if missingness (smoked in pregnancy) is associated with birthweight and gestational age at birth.

Variable.names	OR	L95	U95	P.val
Gestation Age at Birth	0.967	0.936	1	0.04791
Birthweight	0.999	0.999	1	0.00108

- Logistic regression $\rightarrow$ GA at birth and BW both associated with Missingness of SmokedInPregnancy ($\Rightarrow$ MAR or MNAR)
- Incidently, missingness (smoking) is associated with LOWER birthweight and GA at birth.
- **Makes sense:** Low birth weight and premature birth associated with smoking (non-response probably due to 'shame' or 'guilt')

Example: Singleton births (MAR or MNAR)

Now Can we rule out MNAR?

To examine this we see if the **missingness** of gestational age (at birth) (NA.GA) and the **missingness** of Birthweight (NA.BW) are associated with the **missingness** of smoking (NA.Smoke), I will just use a basic χ^2 test of independence

- NA.Smoking vs NA.GA $\chi^2 = 263.65$, $p < 0.0001$
- NA.Smoking vs NA.BW $\chi^2 = 259.65$, $p < 0.0001$
- and now NA.GA with NA.BW $\chi^2 = 1160.7$, $p < 0.0001$

All these results suggest a strong pattern in our missing values.

Looks like our data are MNAR

Note there are lots of tools available in most imputation libraries to graphically examine patterns in missing values (not enough time today)

Options for analysis: Complete case analysis

What are our options if we have missing values?

Complete case analysis (a.k.a. Listwise deletion) is where we exclude any observation which has **ANY** missing value (outcome or covariate).

- Two main problems with this approach:
  1. **loss of precision** (wider 95% CIs) and lower power ($\Downarrow$N $\Rightarrow$ $\Downarrow$power
  2. If values aren't missing at random, the sample may no longer represent the target population $\Rightarrow$ **biased estimates**
- Complete case analysis ('ignore the problem') is the most common 'strategy' for dealing with missing values

Note: Complete case analysis

1. *Complete case analysis* is the default for all statistical software
2. The bigger our model, the more we reduce our dataset (also an issue in propensity score analysis)

Options for analysis: Available case analysis

Available case analysis (a.k.a. pairwise deletion) is where a value is missing in a specific bivariate analysis. Consider:

Observation	X	Y	Z
A	10	5	*NA*
B	20	*NA*	100
C	*NA*	9	150
D	25	7	125

Now let's consider a bivariate correlation analysis

- Corr(X, Y) would exclude obserations B and C
- Corr(X, Z) would exclude A and C
- Corr(Y, Z) would exlcude A and B

Pitfall:

Available case analysis is an special case of complete case analysis at the 'bivariate level'. Avoid this method if conducting both a bivariate and multivariable analysis. WHY???

Options for analysis: Imputation

Imputation involves **replacing missing values with values that 'should be there'** (at least in theory).

In the next section we will cover a number of methods for replacing missing values (IMPUTATION): some naive, and some relatively sophisticated.

But before we do, we should note that the mechanism of the missing data (MCAR, MAR or MNAR) should govern the choice of imputation approach

Note: Which imputation approach to use...

The choice of imputation method should depend on the mechanism of the missing data

Something to think about

I have already mentioned that missing data can (possibly) impact the analysis in two ways:

1. Inefficiency, or **Loss of precision** (e.g. Wider Confidence Intervals)
2. **Biased estimates** (e.g βs higher [or lower] than they should be). That is we over[under]estimate our effect size.

Consider the following problem

Scenario

- We determine that the missing values in our data are MCAR
- We perform a multivariable logistic regression on our analysis using a **Complete case analysis**
- **What impact will this have on our odds ratios (based on the two possibilities given above)????**

Imputation

We will discuss:

- Naive methods: Mean and median imputation (substution)
- More advanced methods:
  - Single Imputation (SI): Regression-based and Maximum likelihood methods
  - Multiple imputation (MI)

In the end, I will only go over Multiple Imputation in any detail. This is because Naive methods are trivial (obvious), and Single Imputation methods have largely been superseded by Multiple Imputation methods. In fact you can see multiple imputation as just running multiple realizations of a single given imputation model.

Naive methods: Substitution-Mean or Median replacement

The simplest approach for imputation is just to replace each missing value by the mean or median.

Q1. Why would I choose the median over the mean????

Before we go any further, I would like you to think about the problems with a 'substitution approach'.

Remember: Imputation can cause two possible problems:

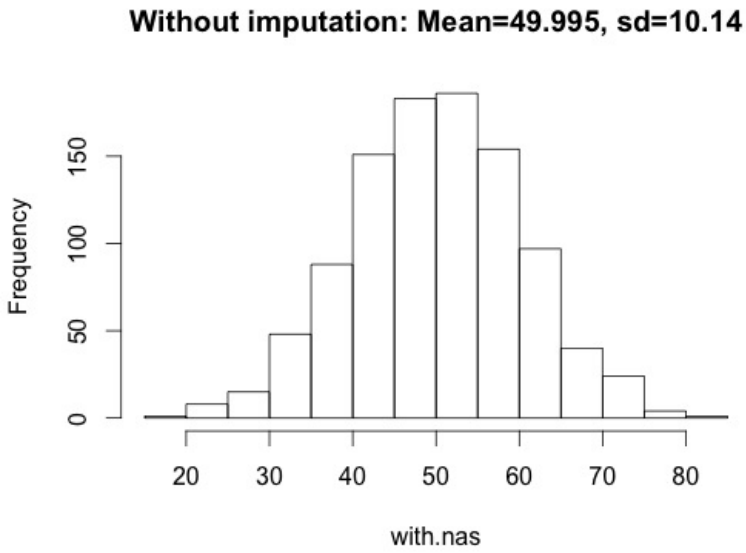
1. **Changes in precision: More or less noise then expected**
2. **Bias: Changes in model parameters ($\Delta\beta$s or ΔORs etc)**

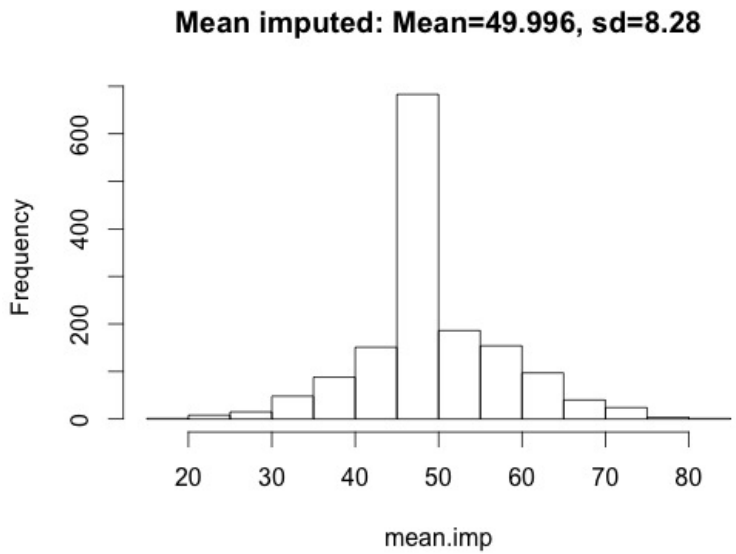
Q2: WHAT MIGHT MEAN/MEDIAN IMPUTATION DO??

Substitution for categorical variable

For categorical variables the mode (most common category) is often used for substitution

Problems with substitution (Mean/median imputation)





Here we have a sample from a normal distribution (mean=50, sd=10) where n=1000. Note that we are assuming that we have 1000 non missing values and 500 missing values....i.e. n SHOULD BE 1500)

500 missing values now replaced with the mea
$\Rightarrow$ 500 extra observations of 50 and we now have n=1500.**Q:** What effect has substitution on the mean (bias??), and the standard deviation ('noise')?

MORE problems with substitution: Associations

The previous example was only in terms of a single variable. What about if we use substituion for several variables. We need to think about what effect this might have on our **associations** (correlations, βs, ORs etc)

REM: Association represents how variables move together (co-vary). What effect will substitution have on strength of associations?

Substitution is likely to weaken the correlation (association) between variables. That is, bias $\beta \to 0$

Upshot: Mean and median imputation (substition)

Substitution generally a poor imputation approach:

- For a given variable: Mean will remain the same, but sd $\downarrow$
- Pairwise associations will generally decrease (ρ, $\beta \to 0$)

Only use subsitution if MNAR and small % of missing values

Random sampling imputation

Another naive approach used for imputation is to randomly sample a value from the empirical or theoretical distribution of the non-missing values.

- Empirical distributions work well for discrete data. If we note that 71% of our sample are younger and 29% older, the we can just take a random value from this empirical distribution of existing observations (we should have about a 71% chance of getting a younger patient, and 29% for - an older patient). We could also use a bootstrap sample
- A theoretical distribution might be a normal distribution which has a mean and standard deviation based on our sample (sampling empirical distributions much more problematic for continuous data **WHY?**)

SI: Model-based Single Imputation

The idea of model based imputation is that we take advantage of the assocations between variables in the dataset (regardless of whether they are outcomes or covariates) and predict missing values based on these associations. Two main types of models available:

- Regression-based methods
- Maximum likelihood methods

We could argue that both of these methods are really regression-based, as they take advantage of associations that exist in the datasets and use a 'regression' model to predict missing values. The only real difference is the methods used to obtain estimates. As single imputation isn't our main focus (and they are embedded in MI methods anyway), I will just briefly cover regression imputation

(Straight) Regression imputation

- Regression imputation replaces NAs with conditional means
- That is, $X_{a,i} = \beta_0 + \beta_b X_{b,i} + \beta_c X_{c,i} + ...$
- The problem with this approach is as this model is deterministic, (the same value of X_b and X_c would give you the same X_a everytime) $\Rightarrow$ leads to an underestimation in the noise for a variable
- **Upshot:** (Straight) regression imputation can lead to reduced standard deviation in an imputed variable (same problem as mean substitution)

Pitfall: Regression imputation

(Straight) regression imputation reduces imputed variables' standard deviations, in turn, reducing our standard errors. This, along with **reinforcing of associations through our imputation model**, can lead to type 1 errors (via narrower 95% CIs)

Adding noise: Stochastic regression imputation

- The problems with (deterministic) regression imputation led to the development of the stochastic regession imputation model
- In this model, noise (random error) is added to the estimates based on either an empirical or theoretical probablility distribution

That is,

$$E(X_{a,i}|X_{b,i}, X_{c,i}, ...) = \beta_0 + \beta_b X_{b,i} + \beta_c X_{c,i} + ... + Noise$$

where (for example)

$E(X_{a,i}|X_{b,i}, X_{c,i}, ...)$ is the conditional mean of X_a; and
Noise $\sim N(0, \sigma_a)$ or $t(0, S_a)$

Noise in regression imputation

Noise can be added using either theroetical distributions (Normal or T-distribution), or via random sampling (e.g. bootstrap samples)

Advantages of Single Imputation

Some people believe single imputation has distinct advantages over multiple imputation, such as:

Advantages of single imputation

Advantages of SI include:

- Single imputation provides a single (complete) dataset
- The imputation model and the analysis model can be separated (i.e. our imputation and statistical modeling done in two different steps). This is good for advanced models that haven't yet been implemented in our MI routines
- Single imputation requires fewer 'uncertain' decisions. E.g
  - How many imputations (datasets) is enough?
  - How many iterations
  - Prior distributions
- Single imputation hass less runtime (MI often takes hours)

Disadvantages of Single Imputation

There are also (MAJOR) problems with single imputation as well:

ML imputation limitations

Disadvantages of SI include:

- All of the conclusions of our study rely on a single imputed dataset
- In particular, a lot of variability (uncertainty) in our estimates is not captured in our single imputed vlaues (MIs will capture this uncertainty)
- Suspicion: In this day and age of MI methods being readily avauiable, reviewers are going to wonder why you chose MI
- There is an advantage to separating (de-coupling) imputation and final models (e.g. easier model building etc), but also a disadvantage (messier)

Multiple Imputation (MI)

Multiple imputation involves the generation of several 'realizations' of complete datasets. That is, one dataset containing missing values will result in multiple imputed ('complete') datasets. MI involves three basic steps"

1. Introduce random variation into the imputation process and generate several datasets, each with (potentially) different imputed values
2. Perform the desired analysis on each dataset (and examine **sensitivity**)
3. Combine the results into a single set of parameter estimates, standard errors (and CIs), test statistics and p-values.

SI $\rightarrow$ MI

We can see Step 1 as just conducting a set (chain) of 'single imputation' datasets.

Multiple Imputation (MI) properties

If MI is valid (assumptions met, and MI conducted properly), the resulting estimators are:

- Consistent $\Rightarrow$ approximately unbiased
- Asymptotically efficient (almost) $\Rightarrow$ minimal sampling variability
- Asymptotically normal $\Rightarrow$ can use normal distribution (table) for p-value and CIs

Multiple Imputation using Chained Equations (MICE)

There are a number of methods available that can be used for multiple imputation. Although most MI methods fall into one of two main groups:

1. **Joint Modelling:** An MI approach based on the joint posterior distribution of incomplete variables. Generally employs a multivariate normal distribution so not that useful if we are considering categorical variables; and
2. Fully conditional specification (FCS), also frequently know as **Multiple Imputation using Chained Equations** (MICE). This approach imputes missing values using univariate conditional distributions for each incomplete variable given all the others, cycling iteratively through the univariate imputation models

For reasons which should be obvious, we will focus on MICE (FCS).

How does MICE really work

Very basic idea:

1. Generate multiple imputed datasets (where variable are mutually predicted using each other)
2. To these predicted values add a little noise (as we did with our regression imputation)
3. Calculate average estimates of β (and its standard error) across the multiple imputed dataset.

Calculating pooled estimates in MICE

The idea behind calculating our pooled estimates is based on a similar idea to Meta-analysis

Pooling estimates in MI: Rubin's pooling formula

Meta-analysis

$$\theta = \frac{1}{k}\sum_{i=1}^{k}\theta_i$$

$$S_\theta = \sqrt{\frac{1}{m}\sum_{i=1}^{k}S_{\theta,i^2}}$$

where

$$m = \sum_{i=1}^{k} m_i$$

and m is the **sum** of the sample sizes from k 'studies'

MICE

$$\theta = \frac{1}{k}\sum_{i=1}^{k}\theta_i$$

$$S_\theta = \sqrt{\frac{1}{\bar{m}}\sum_{i=1}^{k}S_{\theta,i^2}}$$

where

$$\bar{m} = \frac{1}{k}\sum_{i=1}^{k} m_i$$

and m is the **average** of the sample sizes from k imputations

Only true difference is that meta-analysis takes advantage of an 'accumulating' sample size.

A case study: Acute Kidney Injury in South-East Asia

I am going to use a subset of a dataset from a multinational study of AKI led by Professor Nattachai Srisawat at Chula.

Nephrol Dial Transplant (2020) 35: 1729–1738
doi: 10.1093/ndt/gfz087
Advance Access publication 10 May 2019

ORIGINAL ARTICLE

The epidemiology and characteristics of acute kidney injury in the Southeast Asia intensive care unit: a prospective multicentre study

Nattachai Srisawat[1,2], Win Kulvichit[1,2], Noppathorn Mahamitra[1], Cameron Hurst[3], Kearkiat Praditpornsilpa[1], Nuttha Lumlertgul[1], Anan Chuasuwan[4], Konlawij Trongtrakul[5], Adis Tasnarong[6], Ratapum Champunot[7], Rangsun Bhurayanontachai[8], Manasnun Kongwibulwut[9], Pornlert Chatkaew[9], Petchdee Oranrigsupak[10], Theerapon Sukmark[11], Thanachai Panaput[12], Natthapon Laohacharoenyot[13], Karjbundid Surasit[14], Thathsalang Keobounma[15], Kamol Khositrangsikun[16], Ummarit Suwattanasilpa[17], Pattharawin Pattharanitima[6], Poramin Santithisadeekorn[18], Anocha Wanitchanont[19], Sadudee Peerapornrattana[1,2], Passisd Loaveeravat[1], Asada Leelahavanichkul[20], Khajohn Tiranathanagul[1], Stephen J. Kerr[21], Kriang Tungsanga[1], Somchai Eiam-Ong[1], Visith Sitprija[1,22] and John A. Kellum[2]; SEA-AKI study group

I really enjoyed this study as it involved an ordinal longitudinal outcome (progression) and a very nasty model (Proportional odds ordinal logistic mixed-effect regression) - More later

Missing data in longitudinal studies

Before we dig deeper into MICE, I also want to discuss missing data in longitudinal studies (as this will be my MI case study). Missing data in longitudinal studies is generally based on one of two processes:

1. **Non-monotonic** (intermittent) processes:
   This happens when study information is not available for a patient at a given time-point, but the patient then returns to the study.
2. **Monotonic** (drop-out) processes:
   This happens when a patient is lost-to-follow-up. That is, they do not return to the study

Both of these processes can occur simultaneously (and for different reasons)

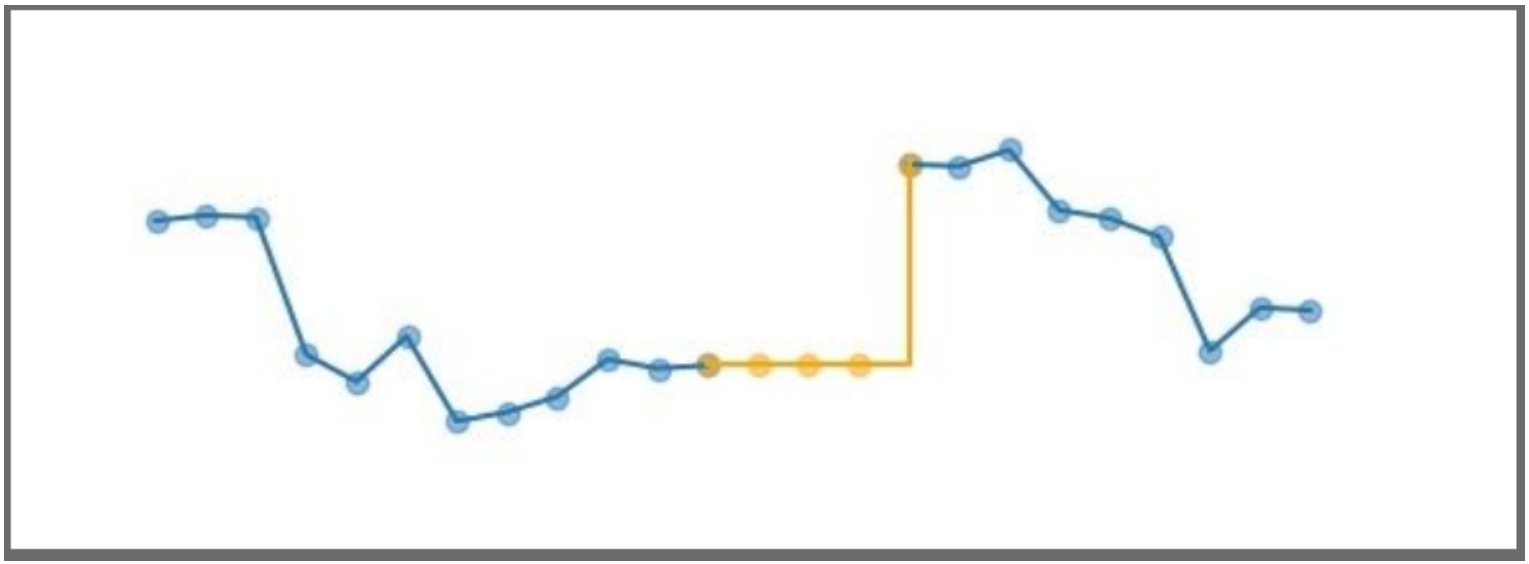
Naive imputation in longitudinal data

What NOT to do!

Perhaps one of the worst approaches to imputing longitidual data is the **Last-Observation-Carried-Forward (LOCF)** or the **Next-Observation-Carried-Back (NOCB)** approaches.
These approaches are still commonly used, and you don't have to think very hard about how they are likely to both:

- Lead to biased estimates
- Artificially reduce noise

LOCF in an intermittent missing values



As you can see, Last-Observation-Carried-Forward is likely to be a very poor imputation method.

MICE methods

Today, I will consider a Acute Kidney Injury data which is a longitudinal design. However, it should be noted:

- Both JM and MICE methods can be used for both cross-sectional and longitudinal (or repeated-measure) data
- Indeed, the two situations only really differ for our **longitudinal outcome** and **time-varying covariates**
- In other words, **baseline covariates** in a longitudinal study represent exactly the **same situation as a standard cross-sectional problem** (single time point study)

Longitudinal or 'cross-sectional' imputation

When we have datasets that contain BOTH **time-varying variables** AND **baseline covariates**, then we use **Longitudinal imputation**, for the former, and standard **'Cross-sectional' imputation** for the latter, and we can do this **simultaneously**.

MICE for repeated measures data

I will TALK about two main approaches for repeated-measures data using MICE:

1. Simplest: Widen our longitudinal data so measurements at different times (outcome and time-varying covariates) just become distinct variables in a 'cross-sectional' dataset. This can be implemented in any (decent) stats package (Stata, SAS and R). **PROBLEM: Only works for balanced designs****
2. MICE-GLMM:
   - Repeated measurements of time-dependent variables are imputed using hierarchical models (i.e. Mixed models)
   - Assumes a conditional GLMM for incomplete binary and categorical variables (and LMM for normal variables)
   - A constant residual error variance is assumed for all individuals
   - As far as I know, only implemented in R (using the `mice` and other package) at this time (Don't know about Stata/SAS)

In our case study we will use the MICE-GLMM approach

Multiple imputation of longitudinal (or clustered) data

So to clarify, our analytical approach will be:

1. Conduct a complete case analysis using a standard Logistic mixed effect regression model (unlike marginal models, like GEE, conditional models like GLMMs are pretty forgiving about missing values and unbalanced data)
2. Specify hierarchical models that capture the structure of our correlated data and the conduct an MICE (more difficult, but probably better practice) *For baseline covariates we'll use standard MICE models*
3. Then we will perform a sensitivity analysis running our two models:
    1. Complete case data
    2. Hierarchical structured MICE approach

I will provide my R syntax to anyone who wants it.

Our dataset

To make life easier we will only consider a single time-varying variable (the outcome, severe AKI). Regardless of whether a time-varying variable is an outcome, or a covariate, the approach is the same.

- **Incomplete:**
  - TIME-VARYING: **Outcome: Severe AKI** (class 3 and 4)
  - BASELINE COVARIATES: **Sex, Age, BMI, Community-acquired AKI, HTN, T2D, CAD and CKD**
- **Complete:**
  - **SiteID (Random effect), PatientID (Random effect)**
  - **Record_day (Fixed, Within-subject effect)**

Missing value patterns

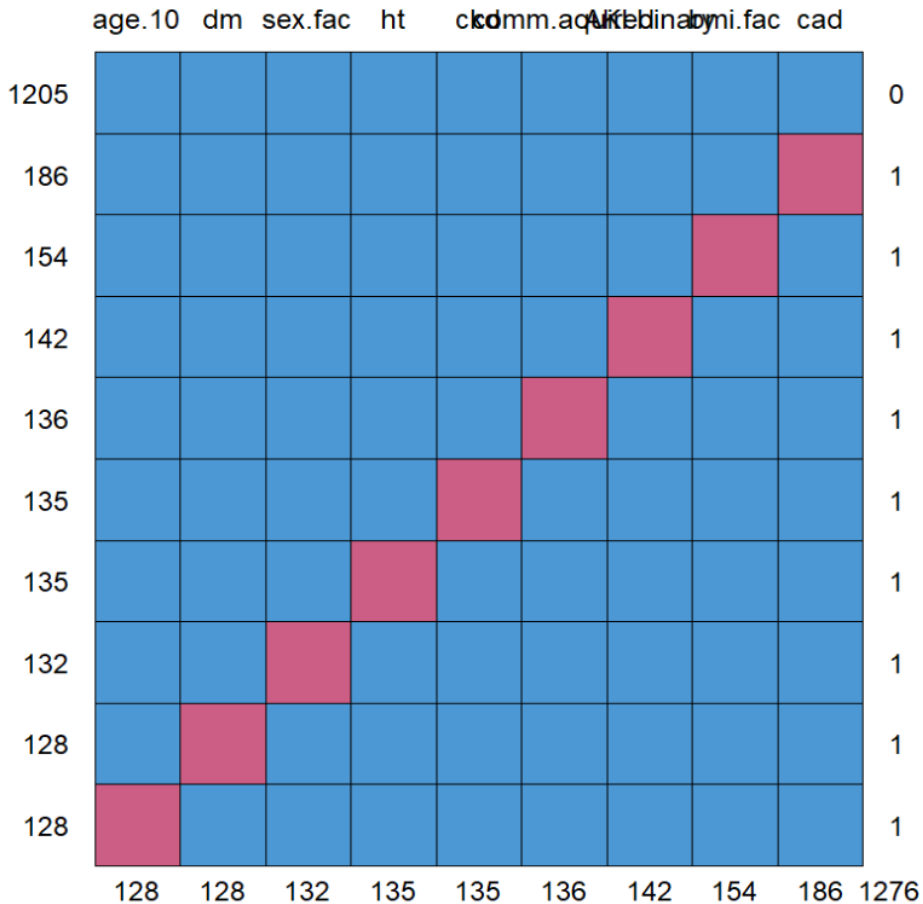
Original data didn't have too many missing values, so I generated them (using the `ampute` command in `R::mice`)

Variable	Missing values
Any missing (NOT complete case)	1276
SiteID, PatientID and Record_day	0
AKI.binary	142
Sex	132
Age	128
BMI	154
Community-acquired(y/n)	136
HTN	135
DM	128
CAD	186
CKD	135

We can see we have 2481 - 1276 = 1205 complete cases, and missing values (per variable) range from about 4.5% to 6.5%.

Missing values patterns

Do missing values 'co-occur'?



No. Looks great (Not a surprise I did generate them to be MAR)

A litle of the `mice` package (and R)

We have gotten to the stage where I can't avoid showing you the R code used to conduct our imputations. Again, I am using the `mice` package in R.

Initializing `mice`

```
# Start with default mice imputation methods
initialize <- mice(Rama2023.NA.df, maxit = 0)
my.methods <- initialize$meth
my.methods
```

Note: I am providing the full R script file to anyone who wants it.

```
> my.methods
patient_number           site    record_day       sex.fac       age.10
            ""             ""            ""         "pmm"         "pmm"
      bmi.fac  comm.aquired            ht            dm           cad
        "pmm"          "pmm"         "pmm"         "pmm"         "pmm"
        ckd     AKI.binary
      "pmm"          "pmm"
```

Imputation methods avaiable in `mice`: Cross-sectional

MANY available!!!

Code	Variable type	Method
""	any	Do nothing
`pmm`	any	Predictive mean matching
`midastouch`	any	Weighting predictive mean matchin
`sample`	any	Random sample (non-missing value
`cart`	any	Classification and regression trees
`rf`	any	Random forest imputation
`mean`	numeric	Unconditional mean imputation
`norm`	numeric	Baysian linear regression
`normal.nobs`	numeric	Linear regression ignoring model err
`norm.boot`	numeric	Linear regression using bootstrap
`norm.predict`	numeric	Linear regression, predicted values
`lasso.norm`	numeric	Lasso linear regression
`lasso.select.norm`	numeric	Lasso select + linear regression

Imputation methods avaiable in `mice`: Cross-sectional

Code	Variable type	Method
`quadratic`	numeric	Imputation of quadratic terms
`ri`	numeric	Random indicator for non-ignoral
`logreg`	binary	Binary Logistic regression
`logreg.boot`	binary	Logistic regression with bootstra
`lasso.logreg`	binary	Lasso logistic regression
`lasso.select.logreg`	binary	Lasso logistic regression with boo
`polr`	ordered	Proprtional odds ordinal logistic
`polyreg`	unordered	Nominal logistic regression
`lda`	unordered	Linear discriminant analysis

Imputation methods avaiable in `mice`: Correlated data

The following methods can be used for correlated data (such as longitudinal data)

Code	Variable type	Method
`2l.norm`	numeric	Level-1 normal heteroscedastic
`2l.lmer`	numeric	Level-1 normal homoscedastic
`2l.pan`	numeric	Level-1 normal homoscedastic, pan
`2l.bin`	binary	Binary logistic mixed effect model
`2lonly.mean`	numeric	Level 2 class mean
`2lonly.norm`	numeric	Level 2 class normal
`2lonly.pmm`	numeric	Level 2 predictive mean matching

As far as I know, imputation models for ordinal or nominal longitudinal (or otherwise correlated) have not been implemented (probably no 'theoretical' barrier to somebody doing this)

`pan` is a multivariate imputation method for panel (or clustered data). See the R library, `pan` for details

Missing value patterns

For our dataset, I am going to use the following imputation models

Variable	Imputation model
SiteID, PatientID and Record_day	none ("")
AKI.binary	`2l.bin`
Sex	`lasso.logreg`
Age	`lasso.norm`
BMI*	`polr`
Community-acquired(y/n)	`lasso.logreg`
HTN	`lasso.logreg`
DM	`lasso.logreg`
CAD	`lasso.logreg`
CKD	`lasso.logreg`

Note:

1. That I am treating BMI class as an ordinal variable (monotonic?)
2. Remember that AKI.binary should be the only 'correlated' variable

Specifying our imputation models

Recall:

```
> my.methods
patient_number          site     record_day       sex.fac       age.10
            ""            ""             ""         "pmm"         "pmm"
      bmi.fac  comm.aquired             ht            dm           cad
        "pmm"         "pmm"          "pmm"         "pmm"         "pmm"
         ckd    AKI.binary
       "pmm"         "pmm"
```

mice: Specify imputation model for each variable

```
# Binary baselines
my.methods[c(4, 7:10)]<-"lasso.logreg"
# Normal baselines
my.methods[5]<-"lasso.norm"
# Ordinal baselines
my.methods[6]<-"polr"
# Binary longitudinal
my.methods[11]<-"2l.bin"
```

Our imputation models

So we have:

```
> my.methods
patient_number          site     record_day         sex.fac          age.10
            ""            ""             "" "lasso.logreg"   "lasso.norm"
       bmi.fac   comm.aquired             ht              dm             cad
        "polr" "lasso.logreg" "lasso.logreg" "lasso.logreg" "lasso.logreg"
           ckd     AKI.binary
"lasso.logreg"       "2l.bin"
```

NOW we can run our imputations

mice: Specify imputation model for each variable

```
#Now generate imputaions (m=5)
my.imp <- mice(Rama2023.NA.df,
meth=my.methods, print = FALSE)
```

Incidentally, these 5 imputation sets took 2 hours to run

Good idea to try your model out on a single imputation first.

Comparing our results

So here we have the complete case and (pooled) imputed analysis results based on a Binary logistic mixed effect regression:

effects	OR	L95	U95	OR.imp	L95.imp	U95.imp
record_day	1.063	0.985	1.147	1.062	1.013	1.113
comm.aquired	1.069	0.194	5.888	1.631	0.474	5.614
sex.fac	0.964	0.189	4.927	0.736	0.350	1.544
age.10	0.850	0.526	1.376	0.878	0.618	1.248
bmi.fac2	0.196	0.013	3.004	1.011	0.268	3.810
bmi.fac3	0.355	0.070	1.804	0.761	0.228	2.536
bmi.fac4	1.489	0.026	85.654	2.653	0.218	32.328
ht	1.668	0.254	10.964	1.749	0.453	6.751
dm	0.667	0.077	5.789	1.381	0.309	6.175
cad	0.745	0.050	11.174	0.788	0.151	4.108
ckd	0.930	0.089	9.747	1.814	0.373	8.809

For effects that were not even close to significant, there was a large change in both ORs and the 95%CIs. **Annoyingly, GLMM not implemented in mice's `pool()` function (did itmyself)**

MCAR, MAR and MNAR

- Most model-based imputation methods are really only for MCAR and MAR missing data, so **what should we do if our data are mssing not at random (MNAR)**
- It is a tricky issue, but I have read in a few simulation studies (studies that examined the strengths and weaknesses of imputation methods under experimental conditions), that naive methods (e.g. mean replacement for numerical, and mode replacement for categorical) are probably the best approach, as these methods tend to be conservative.
- It is also important to note that just because one of your variables is MNAR, doesn't mean they are all are (so you can use a mixed methods approach)

'Machine-learning' based imputation

- As with many areas involving prediction, there is considerable work being done on machine learning-based methods
- For example, several simulation studies have shown random forests imputation seems to perform well (also available in `mice`)
- By all means, use these methods, but just be careful. As with many machine learning prediction algorithms, people (falsely) assume that because they can't understand them, they must be good (counter-initiative when you think about it).
- I am a big fan of machine-learnign methods, but be aware of their strong propensity to overfit (particular neural networks and it's relatives). Many people train ANNs on datasets that are WAY TOO small for them (leading to great within-sample prediction but TERRIBLE generalizability)

Concluding remarks

Imputation (which techniques) and whether it should be done at all is very controversal. **So what should you do when presented with missing data????**

What do I do??? Conduct three analyses:

1. A missing value analysis: Try to find systematic patterns in your missing data (that might indicate that your data in MNAR)
2. A complete case analsis: A "do nothing" analysis, your statistics software does this by default
3. Multiple imputations: Paying particular attention to how my βs (ORs, RRs etc) move around for each individual dataset. A forest plot containing both the individual (per imputation set) and pooled estimates is a good way of gauging this.

THANK-YOU