# Understanding the reasons for missing data/dropout

James Carpenter & Mike Kenward

Department of Medical Statistics
London School of Hygiene & Tropical Medicine

James.Carpenter@lshtm.ac.uk
https://missingdata.lshtm.ac.uk

June 2005

# Table of Contents



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

# Introduction

This sections reviews why it is important to understand the reasons for missing data, and discusses some graphical and model-based approaches for exploring the causes of dropout.

# Why is this important?

Suppose our data consists of two observations $(Y_1, Y_2)_i$ on $i = 1, \ldots, n$ units. Suppose $Y_1$ is observed on all units. As before, for unit $i$, let $Y_{Oi}$ denote the observed data and $Y_{Mi}$ the missing data. So, for some units $Y_{0i} = (Y_1, Y_2)_i$ and $Y_{Mi}$ is empty, while for others $Y_{Oi} = Y_{1i}$ and $Y_{Mi} = Y_{2i}$.
Let $R_i$ be an indicator for observing $Y_{2i}$ so that

$$R_i = \begin{cases} 1 \text{ if } Y_{2i} \text{ observed} \\ 0 \text{ otherwise.} \end{cases}$$

As we have $n$ units, the full data consists of $(Y_O, Y_M, R)_i$, $i = 1, \ldots n$. We have already seen that when we have missing observations it is helpful to consider joint distribution of $[R|Y_O, Y_M]$, and how it simplifies.

# Missingness mechanism

Here we consider the distribution $[R|Y_O, Y_M]$. For simplicity, we assume the same model applies to the whole data, and hence it is sufficient to consider a single unit. There are three cases:

1. $[R|Y_O, Y_M] \sim [R]$
   - the reason for missing data doesn't depend on the observed or unobserved data. Missing observations are *Missing Completely At Random (MCAR)*.

2. $[R|Y_O, Y_M] \sim [R|Y_O]$
   - the reason for missing data can be explained by the observed data; after accounting for this, there is no further information in the unseen data. Missing observations are *Missing At Random (MAR)*.

3. $[R|Y_O, Y_M]$ does not simplify.
   - Even after considering the information in the observed data, the reason for ,missing observations depends on the unseen observations. Missing observations are *Not Missing At Random (NMAR).* In this case, a joint model for the dropout indicator, $R$, and the rest of the data is required. Such models are often not straightforward to fit.

# Why we need this classification

The data alone do not tell us which of the above three cases we are in, because they can't rule out NMAR. Sometimes, however, other sources of information point to a particular mechanism. Even if this is not the case, the classification clarifies the assumptions made by analyses.

Further, if we are carrying out a MAR analysis, we need to attempt to identify the covariates conditional on which the unseen data are MAR (recall the example about income and property band in the document *understanding common jargon*).

# Why it's good to look at your data

In the light of this we now consider some graphical and modelling approaches for understanding reasons for dropout.

Underlying this is the principal that it's good to look at your data carefully. For example, a colleague noticed the majority of his missing data had adjacent serial numbers. He was then able to locate the box of questionnaires the data entry team had overlooked!

We illustrate some approaches using examples.

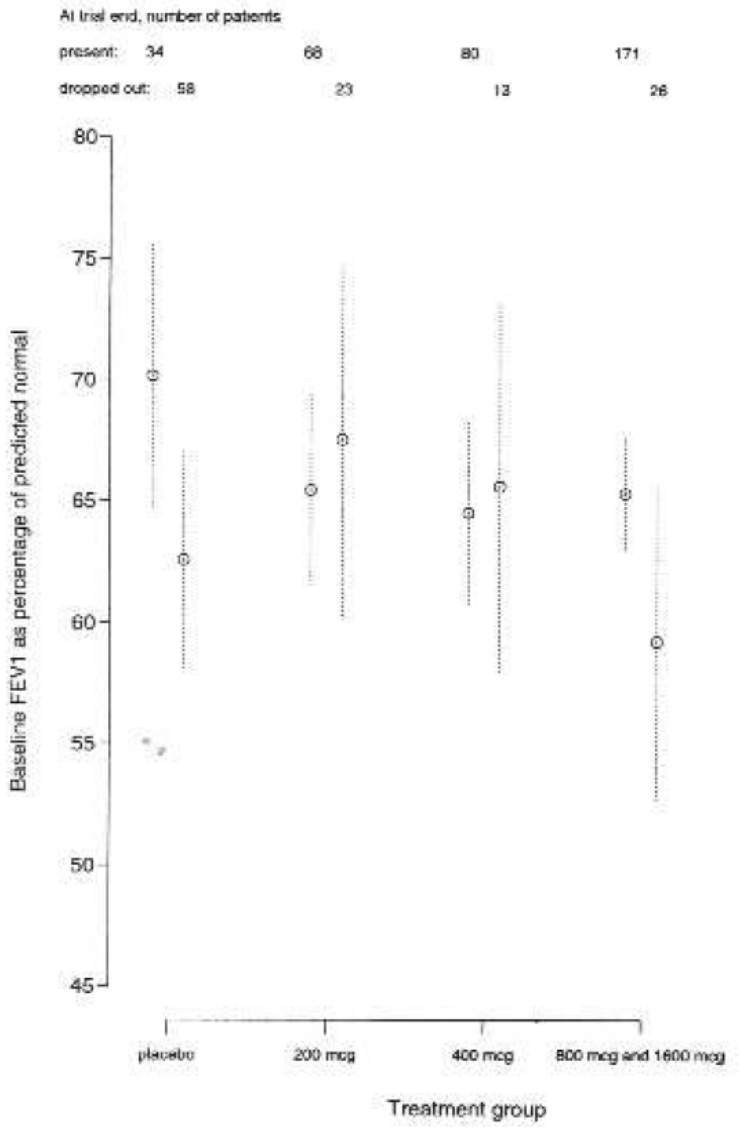
# Example: asthma clinical trial

Longitudinal clinical and social studies have similar issues, as participants tend to be lost to follow-up over time.

In this trial 473 patients with chronic asthma were randomised to placebo or one of four increasing does of budesonide and followed up for 12-weeks. Several measures of lung function recorded, we consider $FEV_1$. This is the maximum volume of air a patient can exhale in 1 second. We consider it as a percentage of that expected for an adult of that age, sex and height. All our readings are below 100% as this patient population is quite ill.
We focus on the placebo and lowest active dose groups. In the placebo arm, 27% patients were lost to follow-up by week 2; in the lowest treatment arm, 12% patients were lost to follow-up by week 2.

# Example: asthma clinical trial

Loss to follow-up is clearly treatment related, but are there other things going on? The following graph shows dropout by treatment.

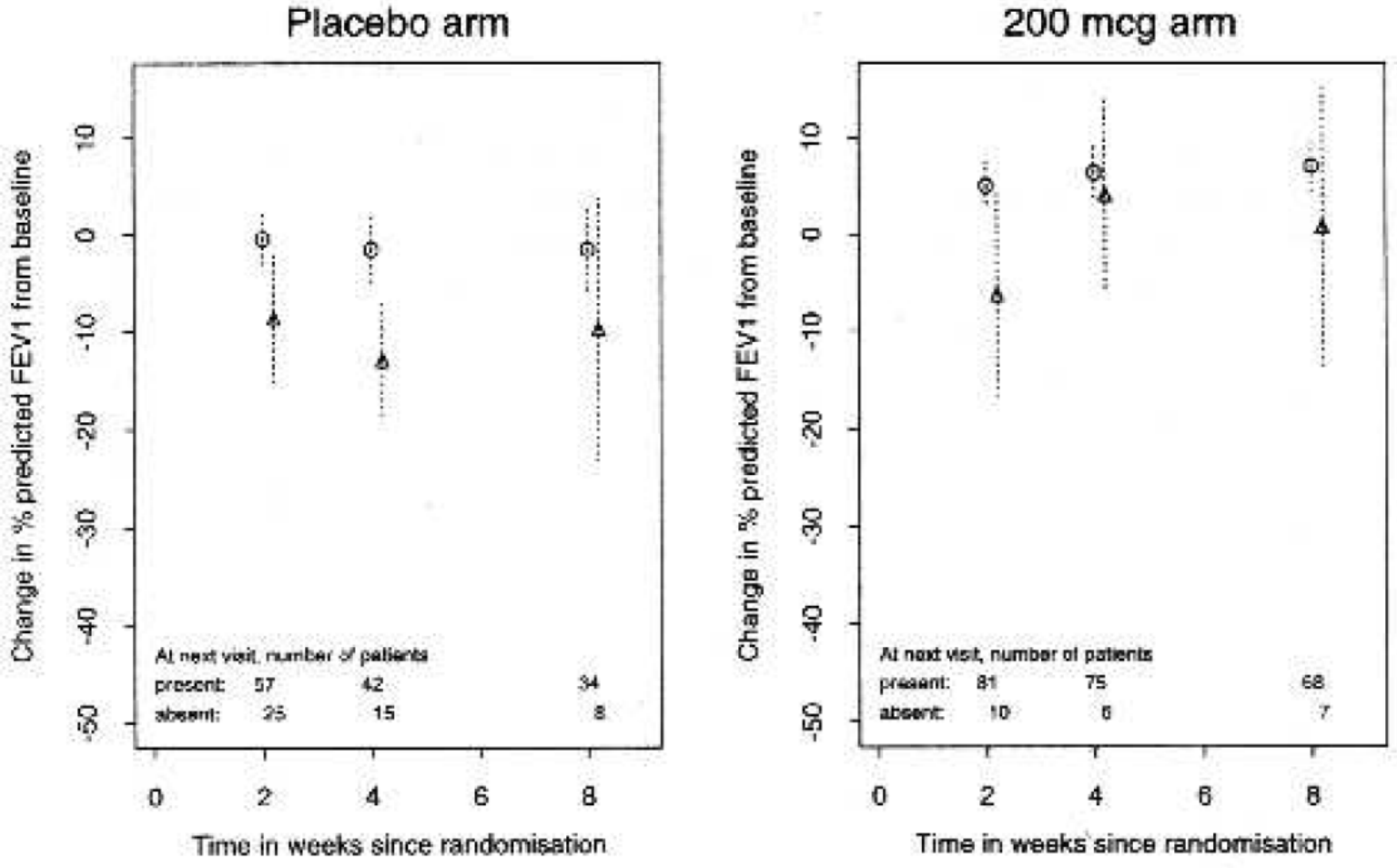


# Example: asthma clinical trial

The next graph shows dropout by treatment and occasion. At each follow-up visit, the graph shows the mean $FEV_1$ (+/- 2 std err) for
(i) those patients who attend the next visit and
(ii) those who do not attend the next visit.

# Example: asthma clinical trial

We see that those do not attend the next visit (i.e. those who are about to dropout) have a markedly lower $FEV_1$ :



# Modelling $R$

If we have one partially observed variable, define the 'missingness indicator', $R_i$ as before, and construct a logistic model:

$$\text{logit}\,\Pr(R_i = 1) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots$$

We can compare models using standard methods, and so select a final model for dropout. We should consider interactions if we suspect different mechanisms are causing missing observations in different data subgroups.

Such models are not only useful guides to interpreting analyses, they also indicate which variables we should include for our models to be valid under missing at random (MAR) and provide estimates of the weights for methods that use inverse probability weights.

# Modelling $R$

We can generalise this approach to cope with the situation were we have two partially observed variables, and the second is always unobserved when the first is (i.e. loss to follow-up):

1. Construct a logistic model for the probability of the first variable being observed.
2. *For those units for which the first variable is observed*, construct a logistic model for the probability of the second variable being observed.

# Modelling $R$

Then

$$\begin{aligned} &\Pr(\text{second variable observed}) && (1)\\ = &\Pr(\text{second variable observed given first variable observed}) && (2)\\ &X\ \Pr(\text{first variable observed}) && (3) \end{aligned}$$

This approach can be extended for additional follow-up visits.

# Example: asthma trial

The Table below shows the odds-rations in a model for dropout by week 4 of the asthma trial. Two variables are important: whether or not a patient is in the placebo or lowest dose group, and last observed lung function.

| Covariate | Odds ratio for dropout |
|---|---|
| Last lung function value | 1.22 |
| Last lung function value $^2$ | 0.999 |

# Survival models

As an alternative to the methods above, with longitudinal studies it is sometimes useful to build a survival model, where the event is loss to follow-up.
In other words, the *hazard* is

$$\Pr\left(\begin{array}{c}\text{Lost to follow up at time } t \text{ given} \\ \text{lost to follow up at time} \geq t\end{array}\right) = h_0(t)exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots)$$

# Summary

- Analyses with missing data make assumptions about the reasons the data are missing.
- It is therefore sensible to look at why the data are missing
  - in conjunction with other researchers on the project
  - using graphical methods (useful for discussions with the project team)
  - using logistic or survival modelling
- Among other insights, this identifies the covariates we need to include in models that assume missing at random
  - useful whether models use likelihood, weighting or some other approach.
- It also provides the starting point for sensitivity analysis using NMAR models.