Statistical Issues

Multiplicity

Missing data

Peter Fayers

1495 UNIVERSITY OF ABERDEEN

What is multiplicity?

1. Subgroups
2. **Multiple outcomes**
3. Repeated assessments (follow-up)

Subgroup analyses

- Assume there the results of a clinical trial are suggestive of a difference, but the effect is not statistically significant.
- Divide the patients into two subgroups – e.g. those mildly ill, and those severely ill.
- One subgroup will show a smaller effect, the other a larger effect.
- To obtain a significant result, keep forming different subgroups until you are lucky!

Subgroup analyses

A subgroup analysis claiming a qualitative interaction—in which the treatment is beneficial in one subgroup but harmful in another—is unlikely to be true in a clinical trial …

The overall ‘average’ result of a randomised clinical trial is usually a more reliable estimate of treatment effect in the various subgroups examined than are the observed effects in individual subgroups.

Subgroup analyses

Within a complex table reporting subgroup analyses of the odds of vascular death after streptokinase, aspirin, both, or neither for acute myocardial infarction, the first "presentation feature" given is astrological birth sign. For people labouring under the star signs Gemini and Libra, aspirin was no better than placebo. For others, aspirin had a strongly beneficial effect.

(Not PROs, but the point is universally true anyway)

ISIS-2 infarction. *Lancet* 1988; **ii:** 349–60.

Subgroup analyses

"Investigators should be cautious when undertaking subgroup analyses. Subgroup findings should be exploratory, and only exceptionally should they affect the trial's conclusions. Editors and referees need to correct any inappropriate, overenthusiastic uses of subgroup analyses."

Subgroup analysis and other (mis)uses of baseline data in clinical trials
Susan F Assmann, Stuart J Pocock, Laura E Enos, Linda E Kasten
Lancet • Vol 355 • March 25, 2000 pp1064-1069

Editorial pp 1033-1034

Multiplicity

If statistical tests performed on several independent outcomes,
each at 5%,
the chance of at least one false positive is:

1	5%
2	10%
5	23%
10	40%

Type I error, p-value, significance

Multiple Outcomes

ICH E9 Guidelines:

Generally, clinical trials should have ONE primary outcome variable

Bonferroni Adjustments

- Ultra conservative – assumes all tests are on independent outcomes
- If the total number of significance tests is *N* use $\alpha^* = \alpha / N$
  - E.g. For 10 tests with an overall type I error of 0.05 (i.e. 5% p-value), test each outcome and reject the null hypothesis if the p-value < 0.05 / 10 (i.e. use p<0.005)

Bonferroni Adjustments

- Alternatively can report adjusted p-values.
  - E.g. Suppose 9 outcomes are being tested. If the calculated p-value from one particular *t*-test is p=0.03, the "adjusted" *p*-value is 9x0.03=.27

Not very impressive!

Adjustments

- Holm procedure is better (more powerful)
- Alternative methods are available, many based on “spending” α (the overall type I error)
- MUST be pre-specified in the protocol
- These procedures assume independence of outcomes!

Alternatives to adjusted p-values

Declare a single “primary outcome” in the protocol

- All other outcomes are secondary and hypothesis generating for future studies
- E.g. In HRQL studies, sometimes “overall QoL” from a global question.

Use a summary measure or statistic

- E.g. the average benefit, the maximum toxicity, area under the curve

Global tests

- Multivariate tests – Hotelling’s T, MANOVA
- Complex to carry out, difficult to interpret, frequently inefficient.

Special cases

No need for adjustment

- Sometimes two (or more) primary outcomes are *both* required to be significant.
  - E.g. in CPMP Guidance – treatment of Alzheimer’s disease.
  - The testing procedure inflates the type II error, which is considered acceptable in this situation

Two or more primary variables ranked for clinical relevance

- E.g. Reduction in mortality in MI followed by prevention of serious effects.
- Confirmatory claims ONLY allowed for variables ranked below other ***significant*** p-values

Other issues

- Safety variables
  - For adverse effects, p-values ae of limited value as substantial differences (e.g. large OR) suffices to raise concern.

For other special cases that, many of which apply to PROs, and a useful general discussion, see:

CPMP/EWP/908/99 Points to consider on multiplicity issues in clinical trials.

Conclusions

- Chance of a spurious positive p-value increases with multiple testing.
- Regulatory authorities are concerned about the opportunity that this offers for selecting a favourable result from the findings of many statistical tests.
- Therefore rigorous ***pre-specification*** of procedures for handling multiplicity are necessary ***in protocol***.
- No method ideal.
- Many methods cannot provide confidence intervals.
- Best is to define a single primary outcome!

Missing data

Missing data is a major problem in trials reporting PROs.

… especially in trials with extended follow-up of patients.

The problem: arguably, it may be likely that patients with the worst HRQL are those most prone to stop completing questionnaires.

Why does missing data matter?

1. Bias

If the proportion of data missing is not small then:

are the characteristics of patients with missing data different from those for whom complete data are available?

Why does missing data matter?

2. Power

A study loses power if data are missing – a larger sample size is required

Note that increasing the sample size will compensate for the loss of power, but will *not* reduce the bias

Always try to minimise the amount of missing data!

Compliance

Many clinical trials have poor compliance with QoL assessment

It is common for less than 2/3 patients to return QoL assessments during or after treatment

Which patients fail to complete the questionnaires? The most ill …?

How can we interpret the results of the study if there is possible bias?

What is an acceptable rate of loss to follow-up?

"Only one answer, 0%, ensures the benefits of randomisation."

"Obviously, this is unrealistic at times. Some researchers suggest a simple five-and-20 rule of thumb, with fewer than 5% loss probably leading to little bias, greater than 20% loss potentially posing serious threats to validity, and in-between levels leading to intermediate levels of problems. .... they opine, and we agree, that a trial would be unlikely to successfully withstand challenges to its validity with losses of more than 20%."

What is an acceptable rate of loss to follow-up?

“Indeed, some journals refuse to publish trials with losses greater than 20%.”

“Although the five-and-20 rule is useful, it can oversimplify the problem in situations with infrequent outcomes. Expectations for loss to follow-up depend on various factors, such as the topic examined, the outcome event rate, and the length of follow-up.”

Schulz KF, Grimes DA. Sample size slippage in randomised trials: exclusions and the lost and wayward. Lancet. 2002; 359: 781-785.

Patterns of missing data

Missing Completely at Random (MCAR)

- When the probability of response at time *t* is independent of both the previously observed values and the unobserved values at time *t*.

Not Missing At Random (NMAR)

- When the probability of response at time *t* depends on the unobserved values at time *t*.

Missing At Random (MAR)

- When the probability of response at time *t* depends on the previously observed values but not the unobserved values at time *t*.

Missing items within a form

Methods have been developed to impute the most likely value for these missing items.

Missing forms tend to be a far more serious problem than missing items.

Forms are more frequently missing, and if a form is missing, so are all the constituent items on the form.

Methods for missing forms

When a whole QoL assessment is missing the imputation procedure must use information from other "similar" patients, values from previous and/or later assessments by the same patient, or a mixture of both.

If items are used only as components of the scale, it may not be necessary to impute values for those items, only for the scale score itself.

Last Value Carried Forward (LVCF)

The values that were recoded by the patient at the last previously completed QoL assessment are carried forward.

Simple Mean Imputation

This is usually the replacement of missing QoL scores by the mean score calculated for patients who did complete the assessment.

Reduced Standard Deviation (SD)

Mean imputation:

- The mean of the augmented dataset remains the same as Ā the mean for the original data.
- The estimate of the SD will be reduced artificially.
- This can lead to distorted significance tests and falsely narrow CIs.
- The SD can be corrected or equivalently use the SD of the non-missing values.

Horizontal Mean Imputation

Unlike LVCF, mean imputation ignores the longitudinal nature of the data.

An alternative is to impute the missing value from the mean of the patients own previous scores.

This method is termed horizontal as it takes into account the longitudinal nature of the QoL data.

It reduces to the LVCF method if there is only one previous assessment available or if there was no change in QoL score over time.

Markov Chain (MC) Imputation

In the methods described so far, the imputed values will be the same for any two patients with the same profile of successive non-missing values.

MC imputation allows these two patients to have different imputed QoL values.

It assigns, for a patient in a particular QoL state at one assessment ***transition*** probabilities of being in each of the possible states, including the same at the next assessment.

Hot Deck (HD) Imputation

Selects at random from patients with observed QoL data, the QoL score from one of these and substitutes this as the imputed value for the patient with the missing QoL assessment.

The hot deck refers to the deck of responses of patients with observed data from which the missing QoL score is selected.

The particular deck chosen may be restricted to those patients that are similar to the patient with the missing QoL score.

Multiple imputation

If deterministic methods such as LVCF are used then the augmented dataset will be unique.

If a random element is included in the choice of missing values, the augmented dataset will be just a single random one out of many potential datasets.

The idea of multiple imputation is that many alternative "complete" datasets can be created.

The analysis can be repeated for each dataset and then combined into a final summary analysis (Rubin 1987).

Analytical methods

Some analytical methods do not involve explicit imputation of values –
but inevitably they make assumptions about the nature of the missingness.

There is no panacea for missing data.

Which method is best?

Many investigators have used simulation methods.
This approach is fundamentally flawed.

- the missing data mechanism is known
- the missing data mechanism may be unrealistic

Aberdeen RECORD trial

Trial of refracture in the elderly

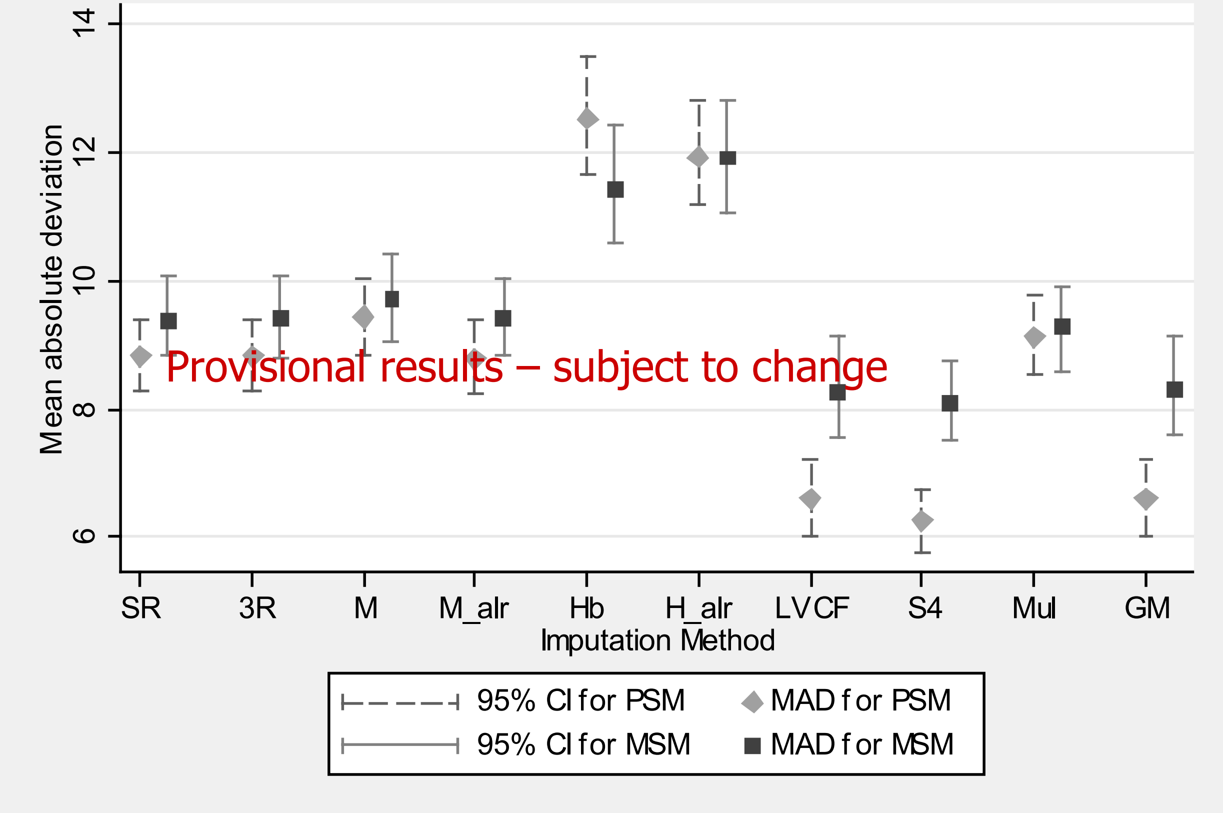
Strenuous efforts to recover QoL data that was initially missing,

by issuing repeated reminders

By offering to interview patients

Unique opportunity to explore the performance of imputation methods

Figure 4: MAD with 95% CI for the PSM and MSM



Provisional results – subject to change

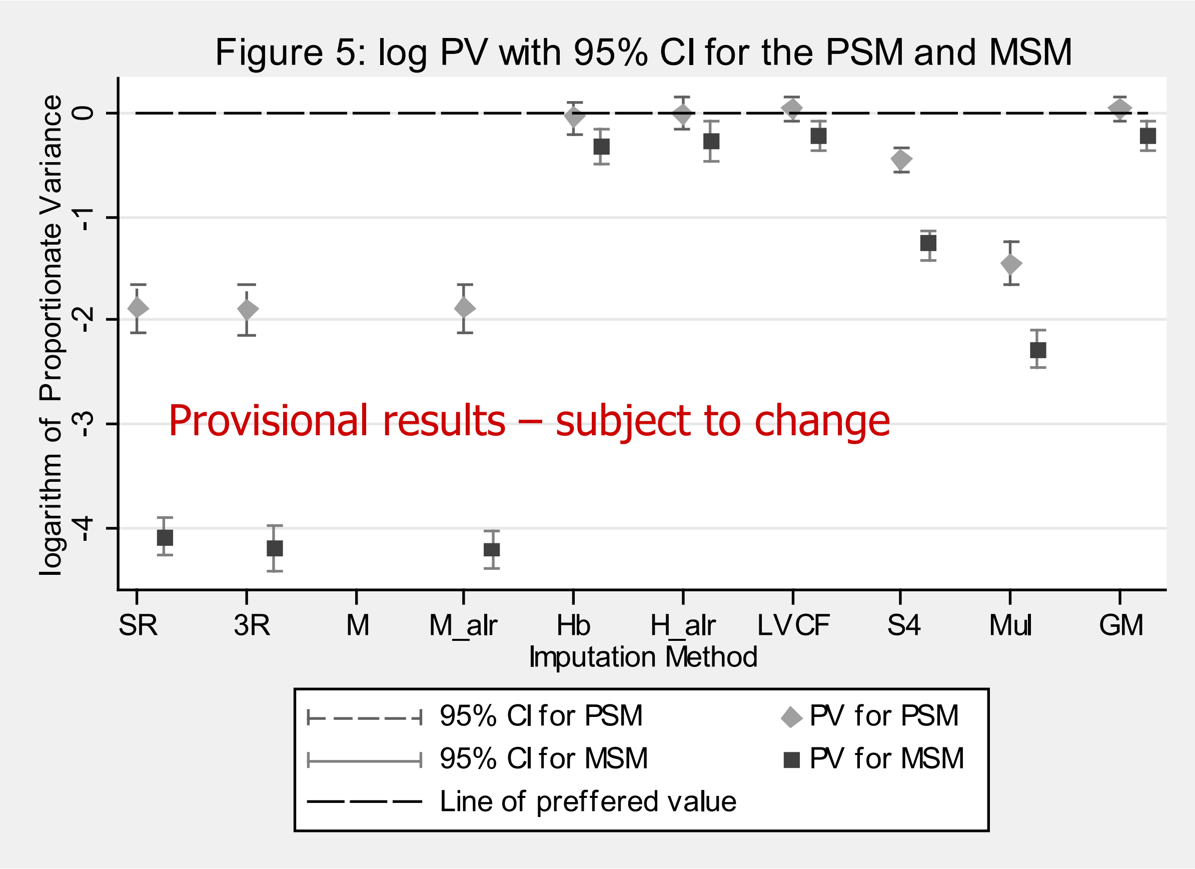


Figure 5: log PV with 95% CI for the PSM and MSM

Provisional results – subject to change

- For this data, LVCF performed best.
- LVCF is one of the simplest methods to implement.
- Others have found that LVCF is often the best method.

BUT it would be inappropriate for patients who change over time –

- Groups of patients known to be deteriorating over time
- Cancer patients when assessed AFTER cytotoxic chemotherapy!

Conclusions I

Many investigators are suspicious about using imputation techniques, because of the assumptions overtly involved.

However, *not* imputing missing data makes the assumption that patients failing to respond are similar to those who do.

Imputation tries to use available information to make better allowances for patients with missing data.

Conclusions II

Markov Chain & Hot Deck imputation are efficient as they take additional patient information into account and preserve the magnitude of the SDs so the CI can be correctly calculated.

Methods may be specific to the individual items or scales concerned as well as the assessment sequence.

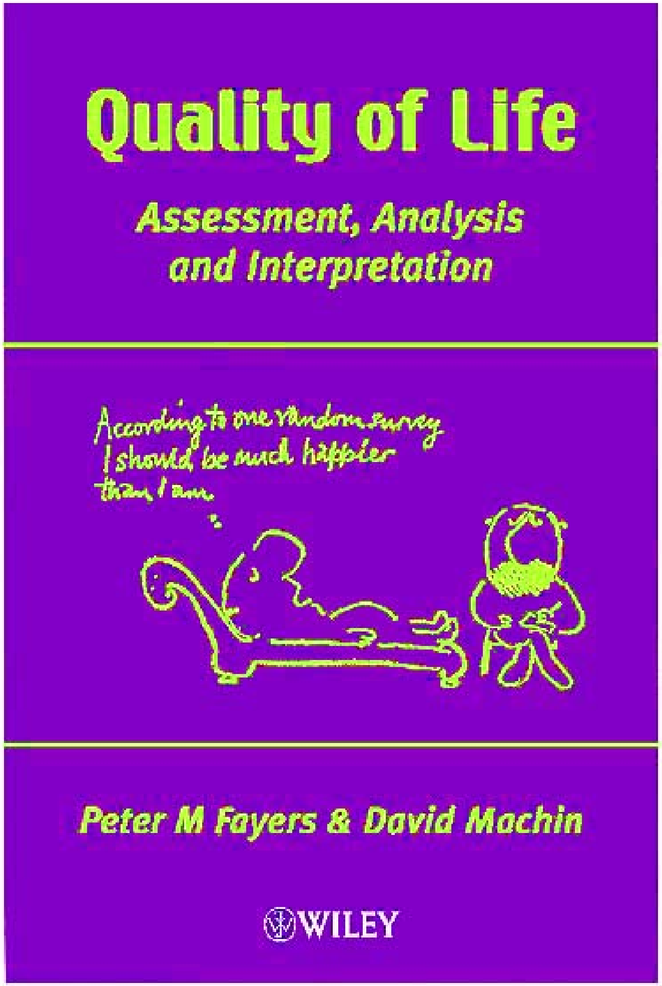
Decide the method of imputation in advance.

The QoL scales which are major endpoints should be the focus for determining the imputation process.

Conclusions III (final!)

- Sophisticated imputation methods are no substitute for the real data.
- On cannot create data from nothing! Imputation is a salvage job.
- The only way to be confident that there is no bias is to ensure good compliance.
- A study with very poor compliance will remain unconvincing and unpublishable, no matter how carefully the data is analysed.

Always aim for 100% compliance!



Quality of Life

Assessment, Analysis and Interpretation

Peter M Fayers & David Machin

WILEY

Fayers,PM & Machin,D
(2000)

Quality of Life: Assessment, Analysis and Interpretation.

J Wiley & Sons Ltd:
Chichester.

ISBN: 0-471-96861-7