SMG advanced workshop, Cardiff, March 4-5, 2010

Missing data

Julian Higgins
MRC Biostatistics Unit
Cambridge, UK

with thanks to Ian White, Fred Wolf, Angela Wood, Alex Sutton

MRC Biostatistics Unit

THE COCHRANE COLLABORATION

Potential sources of missing data in a meta-analysis

- Studies not found
- Outcome not reported
- Outcome partially reported (e.g. missing SD or SE)
- Extra information required for meta-analysis (e.g. imputed correlation coefficient)
- Missing participants
- No information on study characteristic for heterogeneity analysis

Concepts in missing data

1. 'Missing completely at random'

- As if missing observation was randomly picked
- Missingness unrelated to the true (missing) value
- e.g.
  - (genuinely) accidental deletion of a file
  - observations measured on a random sample
  - statistics not reported due to ignorance of their importance(?)
- OK (unbiased) to analyse just the data available, but sample size is reduced - which may not be ideal
- Not very common!

Concepts in missing data

2. 'Missing at random'

- Missingness depends on things you know about, and not on the missing data themselves
- e.g.
  - older people more likely to drop out (irrespective of the effect of treatment)
  - recent cluster trials more likely to report intraclass correlation
  - experimental treatment more likely to cause drop-out (independent of therapeutic effect)
- Usually surmountable

Concepts in missing data

3. 'Informatively missing'

- The fact that an observation is missing is a consequence of the value of the missing observation
- e.g.
  - publication bias (non-significant result leads to missingness)
  - drop-out due to treatment failure (bad outcome leads to missingness)
  - selective reporting
    - methods not reported because they were poor
    - unexpected or disliked findings suppressed
- Requires careful consideration
- Very common!

Strategies for dealing with missing data

- Ignore (exclude) the missing data
  - Generally unwise
- Obtain the missing data
  - Contact primary investigators
  - Compute from available information
- Re-interpret the analysis
  - Focus on sub-population, or place results in context
- Simple imputation
  - 'Fill in' a value for each missing datum
- Multiple imputation
  - Impute several times from a random distribution and combine over results of multiple analyses
- Analytical techniques
  - e.g. EM algorithm, maximum likelihood techniques

Strategies for dealing with missing data

- Do sensitivity analyses
  - But what if it makes a big difference?

Studies not found

Missing studies

- If studies are informatively missing, this is publication bias
- If:
  1. a funnel plot is asymmetrical
  2. publication bias is assumed to be the cause
  3. a fixed-effect of random-effects assumption is reasonable
- then trim and fill provides a simple imputation method
  - but I think these assumptions are difficult to believe
  - I favour extrapolation of a regression line

Moreno, Sutton, Turner, Abrams, Cooper, Palmer and Ades. *BMJ* 2009; **339**: b2981

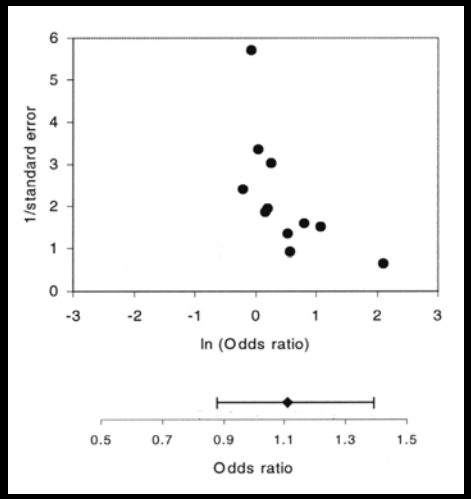
The trim and fill method

- Formalises the use of the funnel plot
- Rank-based 'data augmentation' technique
  - estimates and adjusts for the number and outcomes of missing studies
- Relatively simple to implement
- Simulations suggest it performs well

Duval and Tweedie. *Biometrics* 2000; 56: 455–463

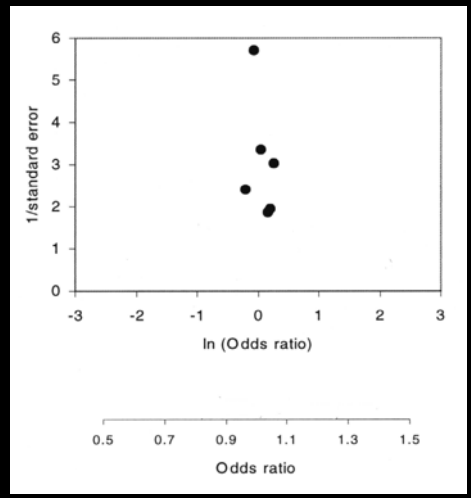
The trim and fill method (cont)

- Funnel plot of the effect of gangliosides in acute ischaemic stroke



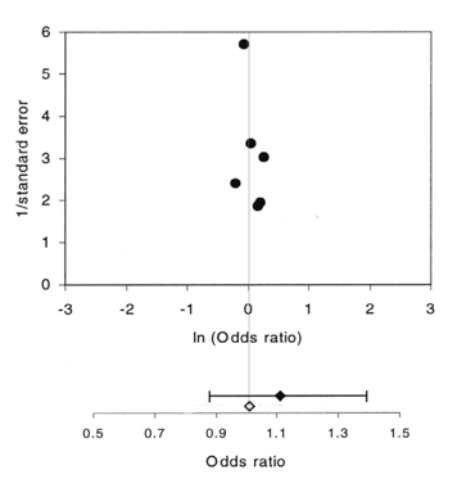
The trim and fill method (cont)

- Estimate number and **trim** asymmetric studies



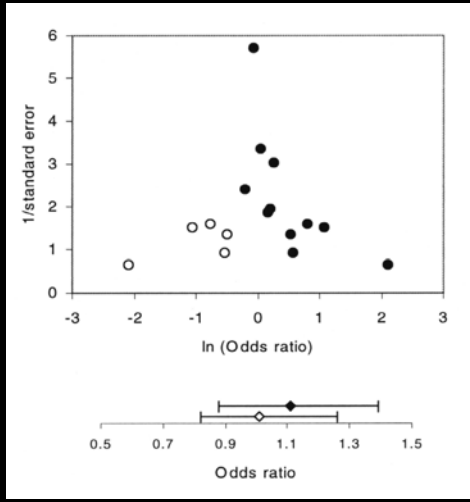
The trim and fill method (cont)

- Calculate ‘centre’ of remainder

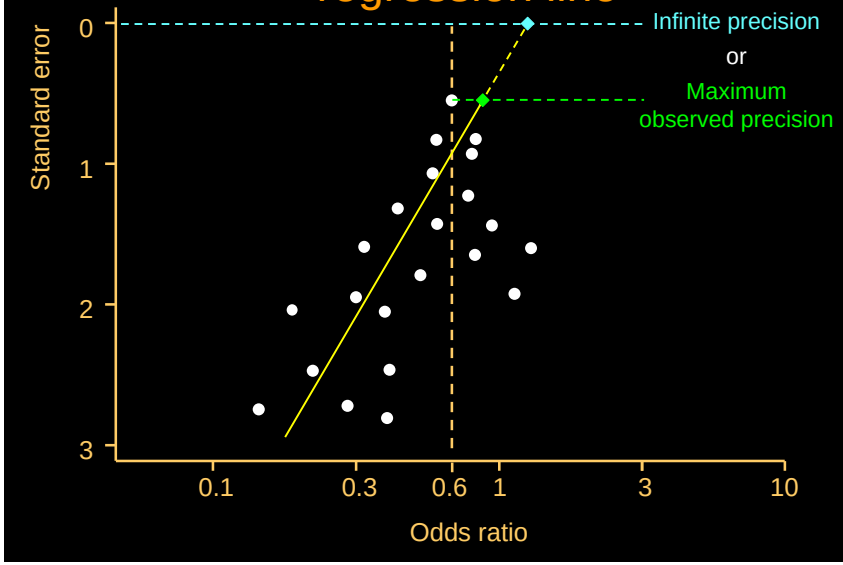


The trim and fill method (cont)

- Replace the trimmed studies and **fill** with their missing mirror image counterparts to estimate effect and its variance



Extrapolation of funnel plot regression line



Outcome not reported

- Wait until tomorrow

Outcome partially reported

Missing standard deviations

- Make sure these are computed where possible (e.g. from T statistics, F statistics, standard errors, P values)
  - For non-exact P values (e.g. P < 0.05, P > 0.05)?
- SDs are not necessary: the generic inverse variance method in RevMan can analyse estimates and SEs
- Where SDs are genuinely missing and needed, consider imputation
  - borrow from a similar study?
  - is it better to impute than to leave the study out of the meta-analysis?

Missing standard deviations

"It was decided that missing standard deviations for caries increments that were not revealed by contacting the original researchers would be imputed through linear regression of log(standard deviation)s on log(mean caries) increments. This is a suitable approach for caries prevention studies since, as they follow an approximate Poisson distribution, caries increments are closely related to their standard deviations (van Rijkom 1998)."

Marinho, Higgins, Logan and Sheiham, *CDSR* 2003, Issue 1. Art. No.: CD002278

Sensitivity analysis

Asthma self management & ER visits *BMJ* 2003, from Fred Wolf

N = 7 studies (n=704)
(no missing data studies)

- SMD (CI)
  -0.25 (-0.40, -0.10)
- Larger effect size
  (over estimation?)
- Less precision
  (more error?)

N = 12 studies (n=1114)
(5 with imputed data)

- SMD (CI)
  -0.21 (-0.33, -0.09)
- Smaller effect size
- Greater precision
- > statistical Power
  (> # studies & subjects)

Other possible sensitivity analyses:
vary imputation methods & assumptions

Extra information required for meta-analysis

Missing correlation coefficients

- Common problem for change-from-baseline / ANCOVA, cross-over trials, cluster-randomized trials, combining or comparing outcomes / time points, multivariate meta-analysis
- If analysis fails to account for pairing or clustering, can adjust it using an imputed correlation coefficient
  - Correlation can often be computed for cross-over trials from paired and unpaired results
    - then lent to another study that doesn't report paired results
  - For cluster-randomized trials, ICC resources exist:
    - Campbell et al, *Statistics in Medicine* 2001; **20**:391-9
    - Ukoumunne et al, *Health Technology Assessment* 1999; **3** (no 5)
    - Health Services Research Unit Aberdeen www.abdn.ac.uk/hsru/epp/cluster.shtml

Missing correlation coefficients

- Guidance on change from baseline, cross-over trials, cluster-randomized trials available in the Handbook
- See also Follmann (JCE, 1992)

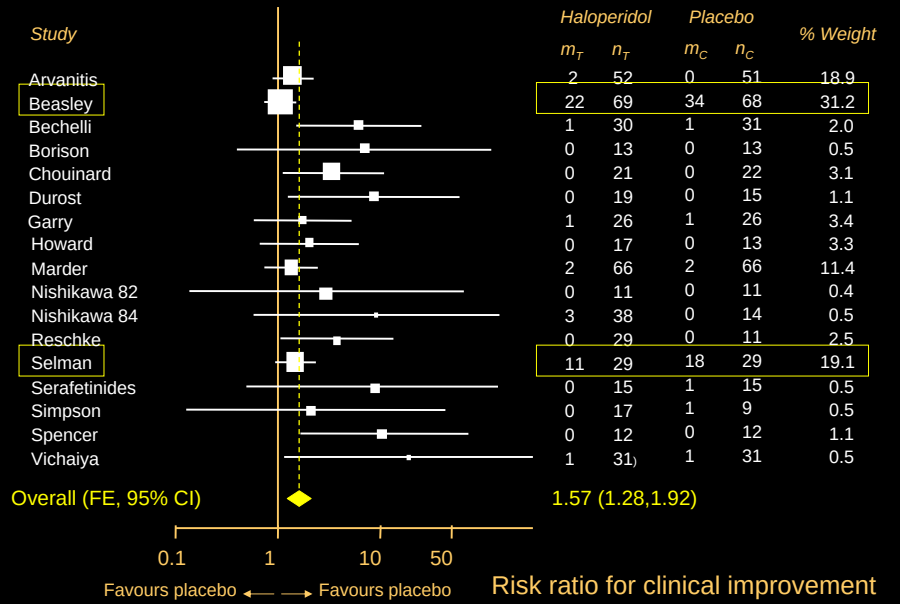
Missing participants

Missing outcomes from individual participants

- Consider first a dichotomous outcome
- From each trial, a 3×2 table

	Success	Failure	Missing	Total
Treatment	r_T	f_T	m_T	n_T
Control	r_C	f_C	m_C	n_C

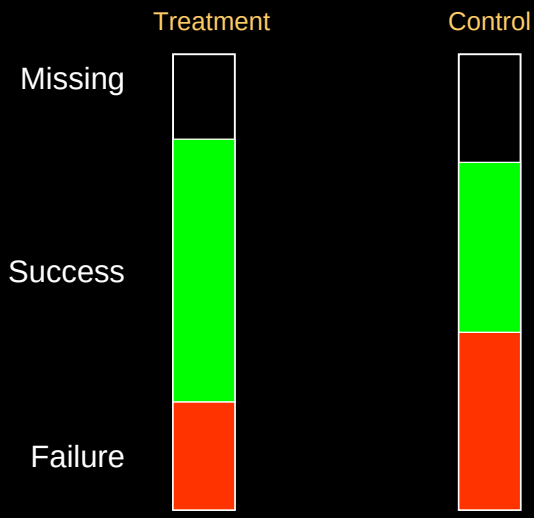
Haloperidol for schizophrenia (Cochrane review)



Typical practice in Cochrane reviews

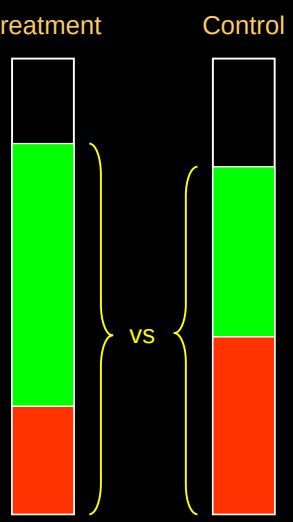
- Ignore missing data completely
  - Call the above an available case analysis
  - often called complete case analysis
- Impute success, failure, best-case, worst-case etc
  - Call this an imputed case analysis
  - analysis often proceeds assuming imputed data are known

Graphical interpretation



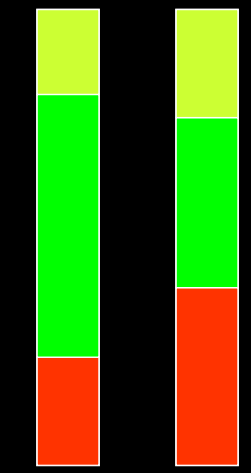
Available case analysis

- (a.k.a. complete case analysis)
- Ignore missing data
- May be biased
- Over-precise
  - Consider a trial of 80 patients, with 60 followed up. We should be more uncertain with 60/80 of data than with 60/60 of data

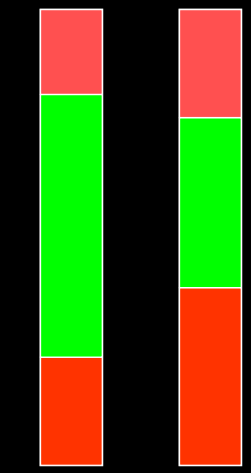


Imputation strategies (1/5)

- Impute all successes

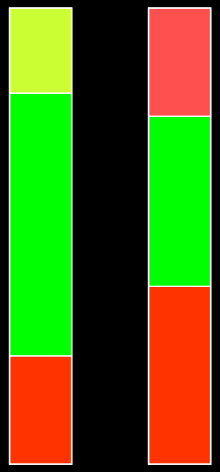


- Impute all failures

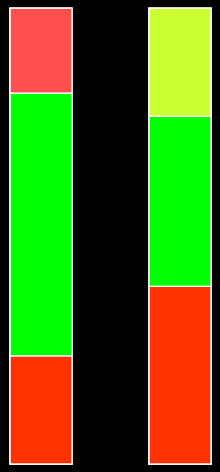


Imputation strategies (2/5)

- Best case for treatment

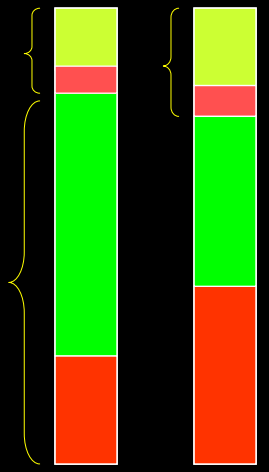


- Worst case for treatment

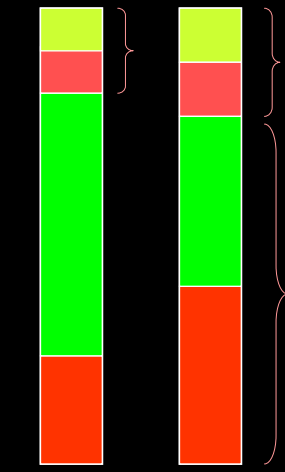


Imputation strategies (3/5)

- Impute treatment rate

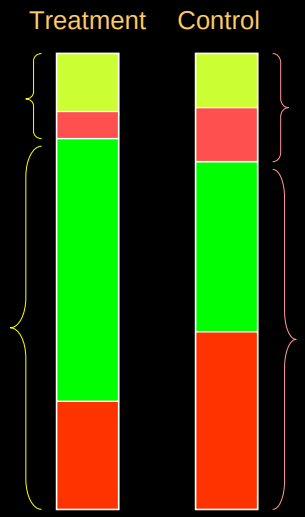


- Impute control rate



Imputation strategies (4/5)

- Impute group-specific rate



Using reasons for missingness

- Sometimes reasons for missing data are available
- Use these to impute particular outcomes for missing participants
- e.g. in haloperidol trials:
  - Lack of efficacy, relapse: impute failure
  - Positive response: impute success
  - Adverse effects, non-compliance: impute control event rate
  - Loss to follow-up, administrative reasons, patient sleeping: impute group-specific event rate

Higgins , White and Wood. *Clinical Trials* 2008; 5: 225–239

Application to haloperidol

Imputation	*Fixed-effect meta-analysis*	
nothing (available case analysis)	1.6 (1.3, 1.9)	Q=27 (16 df)
missing = success (1)	1.2 (1.0, 1.3)	Q=40
missing = failure (0)	1.9 (1.5, 2.4)	Q=22
best case scenario for treatment	2.4 (1.9, 3.0)	Q=22
worst case scenario for treatment	0.9 (0.8, 1.1)	Q=62
according to observed control event rate, p_C	1.4 (1.2, 1.6)	Q=31
according to observed treatment event rate, p_T	1.3 (1.1, 1.5)	Q=34
according to observed group-specific event rate	1.5 (1.2, 1.7)	Q=31
incorporating available reasons for missing data	1.8 (1.4, 2.1)	Q=22

Application to haloperidol

Imputation	*Beasley* *RR (weight)*	*Selman* *RR (weight)*
nothing (available case analysis)	1.0 (31%)	1.5 (19%)
missing = success (1)	0.9 (36%)	1.1 (47%)
missing = failure (0)	1.4 (25%)	2.4 (10%)
best case scenario for treatment	2.5 (30%)	4.0 (11%)
worst case scenario for treatment	0.5 (33%)	0.7 (26%)
according to observed control event rate, p_C	1.0 (37%)	1.3 (27%)
according to observed treatment event rate, p_T	1.0 (25%)	1.1 (51%)
according to observed group-specific event rate	1.0 (36%)	1.5 (32%)
incorporating available reasons for missing data	1.3 (28%)	1.8 (22%)

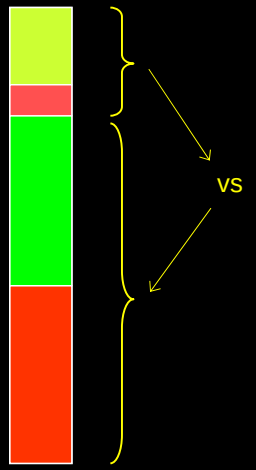
Generalization of imputing schemes

- Consider control group only
- To reflect informative missingness, specify odds ratio comparing

  event rate among missing participants

  vs

  event rate among observed participants
- Call this informative missingness odds ratio ($IMOR_C$)
- Similarly for treatment group ($IMOR_T$)



Connections

Imputation	$IMOR_T$	$IMOR_C$
missing = success (1)	∞	∞
missing = failure (0)	0	0
to create best case scenario for treatment	0 [or ∞]	∞ [or 0]
Impute to create worst case scenario for treatment	∞ [or 0]	0 [or ∞]
all according to observed control event rate, p_C	$\frac{p_C(1-p_T)}{(1-p_C)p_T}$	1
all according to observed treatment event rate, p_T	1	$\frac{p_T(1-p_C)}{(1-p_T)p_C}$
according to observed group-specific event rate	1	1
incorporating available reasons for missing data	$\frac{p^M_T(1-p_T)}{(1-p^M_T)p_T}$	$\frac{p^M_C(1-p_C)}{(1-p^M_C)p_C}$

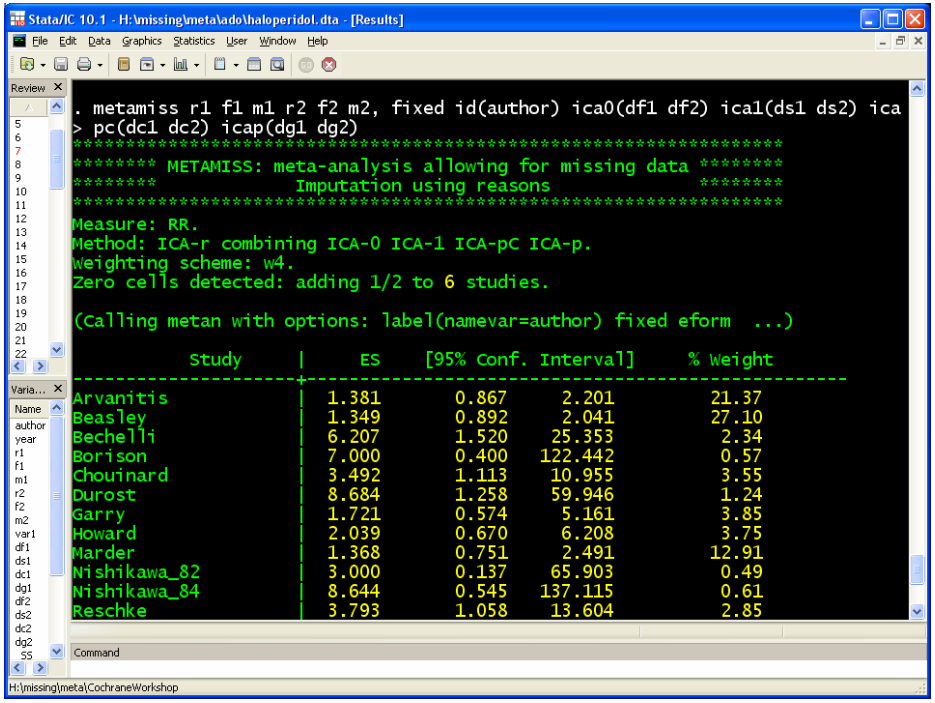
Weighting studies

- Suppose we were to treat imputed data as known
  - Standard errors will tend to be too small and weights too big
- Some simple alternative weights might be used
  - use the available case weights
  - use an effective sample size argument, but with revised event rates
- We have derived more suitable weights based on IMORs: an analytic strategy

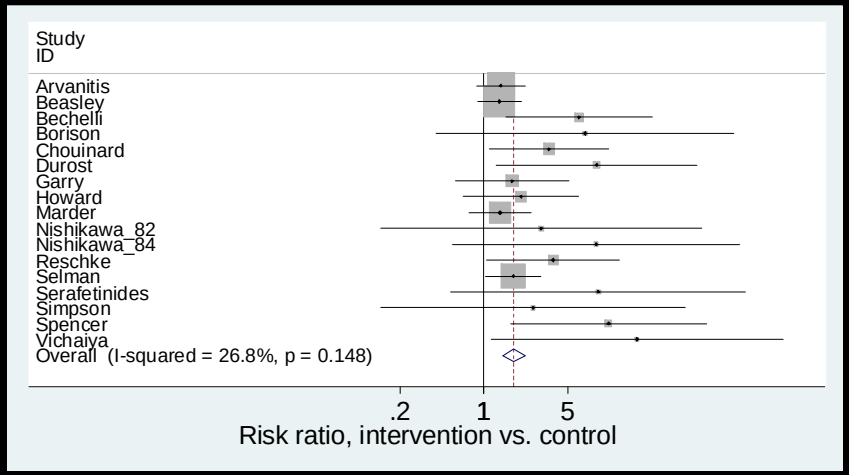
`metamiss` in Stata

- All the above strategies are implemented in our program **`metamiss`**
- Download it within Stata using
  `ssc install metamiss`
- or download the latest version using
  `net from http://www.mrc-bsu.cam.ac.uk/BSUsite/Software/pub/software/stata/meta`

White and Higgins. *The Stata Journal* 2009; **9**: 57-69.



Haloperidol meta-analysis: using reasons

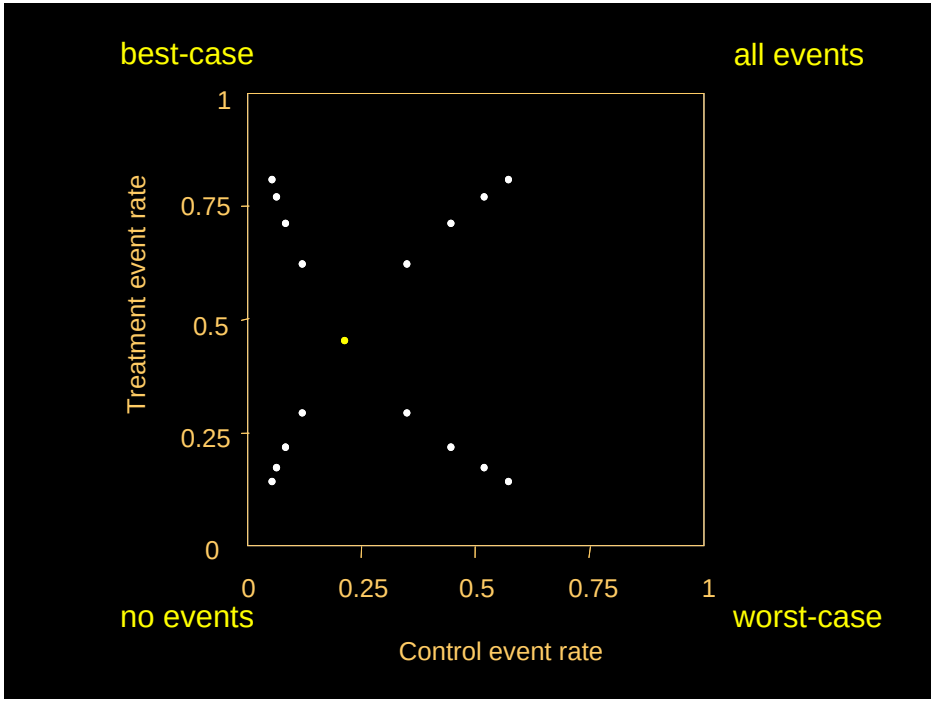


Implications for practice

- We desire a strategy for primary analysis and for sensitivity analyses
- Primary analysis
  - present available case analysis as a reference point
  - make use of available reasons for missingness
  - consider using IMORs based on external evidence or heuristic arguments

Implications for practice

- Sensitivity analysis
  - Best-case/worst case scenarios are fine but probably too unrealistic
  - We suggest to use IMORs to move *towards* these analyses, using clinically realistic values
  - Should also evaluate impact of changing weights



Application to haloperidol

Imputation		*Fixed-effect meta-analysis*
Available case	1.6 (1.3, 1.9)	Q=27 (16 df)
$IMOR_T = 2$, $IMOR_C = 2$	1.4 (1.2, 1.7)	Q=31
$IMOR_T = 1/2$, $IMOR_C = 1/2$	1.7 (1.4, 2.1)	Q=25
$IMOR_T = 1/2$, $IMOR_C = 2$	1.4 (1.1, 1.7)	Q=34
$IMOR_T = 2$, $IMOR_C = 1/2$	1.8 (1.4, 2.2)	Q=22

Conclusion: no important impact of missing data

Missing data introduce extra uncertainty

- Even if we make assumptions about the missing data, we don't know if they are correct
- For example, we don't know if IMOR is 0 or 1.
- Allowing for missing data should introduce extra uncertainty
  - increase the standard error
  - hence down-weight trials with more missing data in the meta-analysis
- We do this by allowing for uncertainty in the IMORs
  - sdlogimor(1) option in **`metamiss`** specifies a sensible degree of uncertainty about the IMOR

White, Wood and Higgins. *Statistics in Medicine* 2008; **27**: 711–727.

sdlogimor(): technical details

- We place a normal distribution on the log IMOR
- For example, we might give the log IMOR mean -1 and standard deviation 1
  - our "best guess" is log IMOR = -1 (IMOR = 0.37)
  - we are 68% sure the log IMOR lies between -2 and 0 (IMOR lies between 0.14 and 1)
  - we are 95% sure the log IMOR lies between -3 and +1 (IMOR lies between 0.05 and 2.7)

Continuous data

- Many methods available to primary trialists, e.g.
  - last observation carried forward
  - regression imputation
  - analytic approaches
- Imputing is much more difficult for the meta-analyst
- You can
  - Impute using the mean (corresponds to increasing sample size to include missing people)
  - Impute using a specific value
  - Apply an 'IMMD' or 'IMMR' ('informative missingness mean difference' or '…mean ratio')
- Methods to properly account for uncertainty are not yet developed

Remarks

- Many imputing schemes are available, but these should *not* be used to enter 'filled-out' data into RevMan
- However, the point estimates from such analyses may be used with weights from an available case analysis (requires analyses outside of RevMan)

Remarks (ctd)

- Informative missingness odds ratios offer advantages of
  - a generalization of the imputation schemes
  - can reflect 'realistic' scenarios
  - statistical expressions for variances
  - ability to incorporate prior distributions on IMORs
- Sensitivity analyses are essential, and should ideally address both estimates and weights
- See: Higgins JPT, White IR, Wood AM. Missing outcome data in meta-analysis of clinical trials: development and comparison of methods, with suggestions for practice. (*Clinical Trials*)

No information on study characteristic for heterogeneity analysis

General recommendations for dealing with missing data

- Whenever possible, contact original investigators to request missing data
- Make explicit assumptions of methods used to address missing data
- Conduct sensitivity analyses to assess how sensitive results are to reasonable changes in assumptions that are made
- Address potential impact of missing data (known or suspected) on findings of the review in the Discussion section