Extending the assessment of outcome reporting bias to harms (ORBIT II workshop)

Facilitators: Dr Jamie Kirkham
University of Liverpool (UoL)

Acknowledgements: Pooja Saini (UoL)
Professor Doug Altman (University of Oxford)
Professor Carrol Gamble (UoL)
Dr Yoon Loke (University of East Anglia)
Professor Paula Williamson (UoL)

(MRC Research Grant: MR/J004855/1)

UNIVERSITY OF LIVERPOOL

MRC Medical Research Council

The Cochrane Collaboration
Working together to provide the best evidence for health care

How do we select harm outcomes to include in our reviews?

- Small groups
  - Primary versus secondary harms
  - Pooled harms versus specific harms

Considerations for reporting harms data within a review?

- Meta-analysis versus tabulation/descriptive

The Cochrane Collaboration
Working together to provide the best evidence for health care

Cochrane Handbook

"There should in general be no more than three primary outcomes and they should include at least one desirable and at least one undesirable outcome (to assess beneficial and adverse effects respectively)."

ORBIT II study

- Identified 234 intervention reviews
  - [Issue 9, 2012 to Issue 2, 2013]

- 39 (17%) reviews specified no harm

- 103 (44%) specified only pooled harms

- Can we assess selective reporting for ‘pooled’ harms?

Outcome Reporting Bias (ORBIT I)

- Definition: Selection of a subset of the original recorded outcomes on the basis of the results, for inclusion in publication*

- Bias results from:

  - Results reported as p>0.05 only (NS)
  - Negative results suppressed altogether

*BMJ 2010;340:c365

Potential mechanisms for selective reporting of harm outcomes

- Assessment could be the same as for efficacy outcomes
  - Bias could be associated with non-significant results ($p>0.05$)

- BUT assessment could also be more complex
  - Harms are measured very differently
    - Specific testing/questioning for a particular harm
    - Open questions (e.g. have you experienced an AE?)
    - Combination of both

- Risk of bias will be influenced by what is known about the harms that are reported
  - Bias could also result from significant harm results ($p<0.05$)
  - Or an undesirable outcome

Examples: selective reporting of harms

- Empirical evidence
  - Reporting of harms data is worse than efficacy (Chan 2004)
  - Interviews with trialists (Smyth 2011)

"We didn't bother to report it, because it wasn't really relevant to the question we were asking. That's a safety issue thing; there was nothing in it so we didn't bother to report it. It was to keep ethics committee happy. It is not as if we are using a new drug here, it is actually an established one, just an unusual combination, so if we are using new things we report all that sort of stuff, so it's not that experimental"

"When we looked at that data, it actually showed an increase in harm amongst those who got the active treatment, and we ditched it because we weren't expecting it and we were concerned that the presentation of these data would have an impact on people's understanding of the study findings"

Is the mechanism for assessing ORB in harms the same as for efficacy outcomes? [ORBIT II]

Selective reporting of harms: your experiences

- Knowledge of the trial protocol as the trialist, statistician, etc.

- Experienced ORB in harms
  (exclude legitimate protocol changes)

- Other experiences you have heard about
  - Dubious reporting of harms data / poor reporting practice
  - Spontaneous reporting?

Example: Metabolic and Endocrine Disorders

- Primary Harm Outcome: Hypoglycaemia
- 6 studies (1450 individuals), 3 (1064 (73%)) included
- 3 trials with no data
- For two trials: *"There were no clinically important differences in the incidence of hyperglycemia or hypoglycemia between treatment groups"*

Clear that the primary harm was measured and compared → result p>0.05

Example: Metabolic and Endocrine Disorders

- In the third trial: *"A total of 14 patients (40%) reported ≥1 AE: 8 (47%) in the colesevelam group and 6 (33%) in the placebo group (Table 2)…There were no deaths, serious AEs, or other significant AEs."*

Specific primary harm outcome not mentioned/reported → all 14 AEs listed in Table 2 for each treatment group → likely no hypoglycaemia events

Example: Epilepsy

- Primary Harm Outcome: Skin irritation
- 5 studies (975 individuals), 1 (281 (29%)) included
- In one trial: *"The most common side effect necessitating a change or cessation in therapy was acute allergic skin rash. Rashes occurred in 28 patients who were treated with antiepileptic drugs."*

Clear that the primary harm was measured but NOT compared → results reported globally

Example: Epilepsy

- In a second trial: *"Four patients treated with phenytonin showed a hypersensitivity skin reaction, occuring between 6 days and 4 weeks after the start of treatment".*

Clear that the primary harm was measured but NOT compared → results reported for one treatment arm only

NOTE: some clinical consideration may need to be made about known AEs for comparator groups. For example, is it possible to observe the specific harm in all groups? E.g. Surgical Infections, where comparator is a non-surgical intervention

Example: Epilepsy

- In a third trial: *"Overall they reported 28/129 adverse effects in the ZNS group and 30/126 in the PB group".*

Pooled AEs measured and compared→ some of the AEs could be the specific harm of interest

- In two other trials no AE data were mentioned, however knowledge of clinical area suggests data would be collected routinely

Example: Anaesthesia

- Primary Harm Outcome: Biochemical (pH) [continuous]
- 13 studies (646 individuals), 7 (450 (79%)) included
- 6 trials did not mention/report on the primary harm outcome
- Clinical judgement says 4 of these would have measured the outcome
  - Trials were in major surgery (pretty routine practice)
  - Methods suggested blood samples were taken and analysed
  - Lactate was measured and reported / good indicator pH measured
- Clinical judgement says 2 of these were unlikely to have measured the outcome
  - Both trials were not done in major surgery
  - 1 trial was also conducted 30 years ago which probably didn't reflect current practice to measure specific harm of interest

Assessment within review

Similar criteria for benefit outcomes:

- Exclusion criteria should not include ‘did not report outcome data of interest’

- Number of eligible trials > number included in MA/ ***adequately**** reported in the text

* Event rates reported on specific harm of interest for each treatment arm / clear that there were ‘no events’ for the specific harm for all treatment groups

NOTE: measuring harms on a continuous scale is rare although special consideration needs to be made when considering ‘adequately reported’ in such cases

ORBIT II Classification Categories

- Explicit the specific harm outcome was measured and compared across groups

  - Result $p>0.05$ only
  - Result $p<0.05$ only
  - Insufficient reporting for MA/full tabulation (continuous instrument scales)

- Explicit the specific harm outcome was measured but **not** compared across groups

ORBIT II Classification Categories

- Explicit outcome was measured, not clear whether compared or not
  - Measured but no results reported
  - Result reported globally (across all groups)
  - Result reported from some groups only

ORBIT II Classification Categories

- Outcome not explicitly mentioned, likely measured, not clear whether compared or not
  - Pooled AEs reported (some of which may be the specific harm of interest)
  - No harms mentioned/reported (not even pooled AEs), clinical judgement says likely measured

ORBIT II Classification Categories

- Outcome not explicitly mentioned, likely measured (no events)
  - Specific harm not mentioned but all other specific harms reported *adequately* *
  - No mention of specific harm, likely to be no events (consider similar trials)

*NOTE: Be careful of reporting thresholds, e.g. specific harms reported only if observed in >=3% patients.

ORBIT II Classification Categories

- Outcome not explicitly mentioned, unlikely measured
  - No harms mentioned/reported (not even pooled AEs), clinical judgement says unlikely measured
- Explicit the harm outcome was not measured
  - Report clearly specifies the data on the specific harm of interest was not measured. E.g. *"Non-serious AEs such as,…,pain at vaccination site were not collected."*

Group Work

Pre-operative aspirin on bleeding
(European Heart Journal 2008)

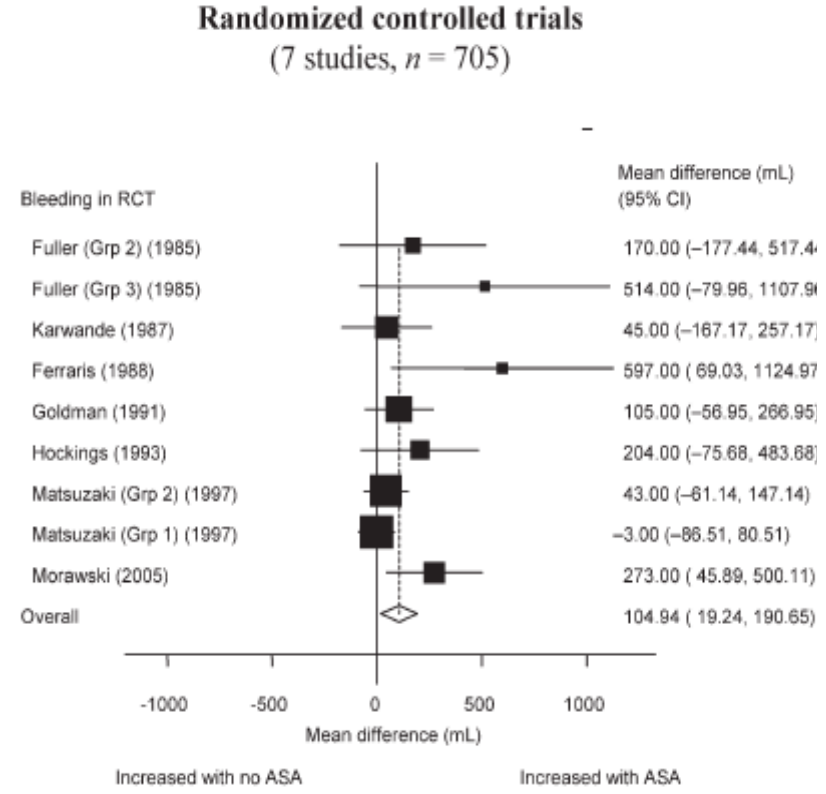
- Primary Harm Outcome: Post-operative bleeding

- 22 studies identified (8 RCTS, 14 observational)
  - 5 studies did not report post-operative bleeding
  - 5 additional studies identified as 'NROD'
  - Mean difference
    - *104.94 mL [95% CI (19.24, 190.65)] from 7 RCTs*
    - *113.59 mL [95% CI (45.16, 182.02)] from 10 NRSs*

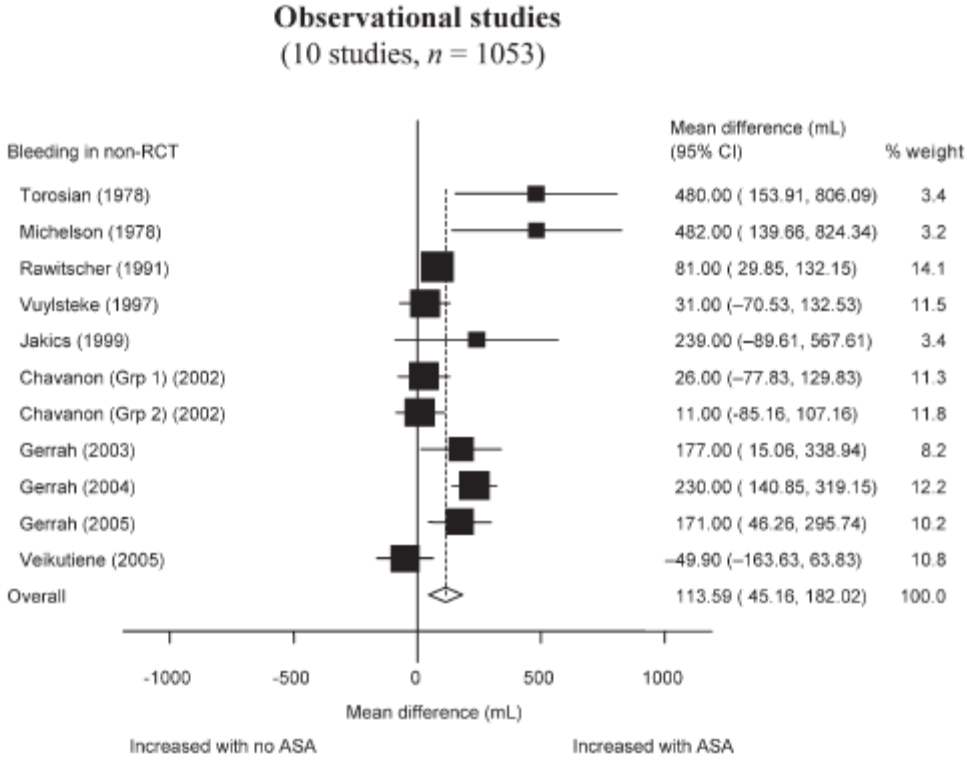
- Author's conclusions: "Pre-operative aspirin increases post-operative bleeding"

Pre-operative aspirin on bleeding

(European Heart Journal 2008)



Overall effect: $p = 0.016$
Heterogeneity $\chi^2 = 13.60$ ($df = 8$), $P = 0.093$
$I^2 = 41.2\%$



Overall effect: $P = 0.001$
Heterogeneity $\chi^2 = 34.81$ ($df = 10$), $P < 0.01$
$I^2 = 71.3\%$

1 missing RCT (n=100)

4 missing NRS (n=2389)

Task

- In groups, assess the potential reasons for the non-reporting/partial reporting of the primary harm outcome in the following studies

  - Kallis 1994 (RCT)
  - Reich 1994 (NRS)
  - Weightman 2002 (NRS)
  - Karwande 1987 (Excluded Study)

ORBIT CLASSIFICATION SYSTEM

P1	States outcome analysed but reported only that p-value > 0.05
P2	States outcome analysed but reported only that p-value <0.05
P3	Insufficient reporting for MA/full tabulation (more for continuous)
Q	Clear outcome was measured but not compared
R1	Clear that outcome was measured but no results reported
R2	Result reported globally across all groups
R3	Result reported from some groups only
S1	General AEs reported (some of which may be SPHO)
S2	No harms mentioned/reported (not even general AEs), clinical judgement says likely measured
T1	SPHO not mentioned but all other specific harms fully reported
T2	No description of specific harms, likely to be no events (consider similar trials)
U	No harms mentioned/reported (not even general AEs), clinical judgement says unlikely measured
V	Report clearly specifies the data on the specific harm of interest was not measured
FULL	Full reporting

Results for ORBIT II Study

Classification	Cochrane (n=170)		Harms cohort (n=316)	
P1: Analysed *p*>0.05	6	3.5%	12	3.8%
P2: Analysed *p*<0.05	0	0%	4	1.3%
P3: Compared – insufficient reporting	0	0%	1	<1.0%
Q: Measure – not compared	0	0%	0	
R1: Measured – no results reported	5	2.9%	37	11.7%
R2: Measured – global reporting	7	4.1%	38	12.0%
R3: Measured – reported some groups only	7	4.1%	9	2.8%
S1: General AEs reported only	2	1.2%	16	5.0%
S2: No harms mentioned – *likely* measured	31	18.2%	74	23.4%
T1: All specific harms reported in full	8	4.7%	5	1.6%
T2: *Likely* no events	35	20.6%	37	11.7%
U: Not mentioned / *unlikely* measured	30	17.6%	25	7.9%
V: Explicitly not measured	0	0%	4	1.3%
FULL REPORTING	39	22.9%	54	17.0%

Benefit - Risk

An example - Gastro-intestinal bleeds

- Two systematic reviews comparing aspirin vs. placebo:
  - Gastro-intestinal (GI) bleeding (harm)
    - McQuaid & Laine, 2006 (22 studies)

RR 2.07 (95% CI 1.61, 2.66) [placebo]

  - Prevention vascular events (efficacy)
    - Herbert & Hennekens 2000 (4 studies)

RR 0.87 (95% CI 0.81, 0.95) [aspirin]

Benefit-Harm ratio (NNT/NNH)

- Using methods of Loke, 2002 (risk adjusted):
  - Per 10,000 patients aspirin therapy for 1-year
    - Prevent **65** cardiovascular events (95% CI 25,95)
    - Cause **32** GI bleeds (95% CI 18,50)

- Taking aspirin suggests twice as many vascular events prevented compared to harms observed (GI bleeds)

BUT

- Only 14/22 studies contributed data to the meta-analysis of GI bleeds

Benefit-Harm ratio (NNT/NNH)

- Using methods of Loke, 2002 (risk adjusted):
  - Per 10,000 patients aspirin therapy for 1-year
    - Prevent **65** cardiovascular events (95% CI 25,95)
    - Cause **32** GI bleeds (95% CI 18,50)

- Taking aspirin suggests twice as many vascular events prevented compared to harms observed (GI bleeds)

BUT

- Only 14/22 studies contributed data to the meta-analysis of GI bleeds

Outcome reporting bias?

- Eight studies not reporting on GI bleeds
  - Clear that complications and bleeding were measured
  - No data on GI bleeds presented

- Were data suppressed because they suggested a disadvantage for aspirin?
  - If YES, this would have introduced bias
  - True results being even more favourable towards placebo.

- How does this affect the Benefit-Harm ratio?

Sensitivity analysis

- Applying the sensitivity analysis (Williamson & Gamble,2007)
  - Adjusted RR for GI bleeds:

RR 2.55 (95% CI 1.98, 3.28) [placebo]

- Revised risk adjusted Benefit-Harm ratio
  - Prevent **65** cardiovascular events (95% CI 25,95)
  - Cause **47** GI bleeds (95% CI 29,68) [+ 15 events per 10,000]

- Does this difference tip the balance on whether to treat?

Discussion points

- Prevalence of non-reporting of harms is higher than benefit outcomes
  - Cochrane (55 vs. 86% reviews affected)
  - Cochrane (31 vs. 76% eligible studies)

- Are any of the reasons (classifications) for missing harms data biased?
  - Or are they just poor reporting issues (R+S classifications (57%))
  - Excluded studies / Partial reporting / Omission of data

- Implications for systematic reviewers / researchers
  - Incomplete data. How can this be resolved?

- Implications for clinicians and patients
  - Poor reporting / bias creates difficulty in judging harm (benefit/risk)

- Reporting terminology
  - "All Adverse Events" – avoids making causal link between intervention