Improving Outlier Detection in Subgroup Analysis using Bayesian Predictive Cross-validation Models

Wilmar Igl, PhD

Veranex

Regulatory Guidance

EMA, 2019, Guideline on Subgroups:

"Evidence is considered to be more ***robust if treatment effects across the trials in the application,*** *as well as* ***in relevant subgroups within one trial … are consistent*** *and substantiate the claim to be made for the experimental treatment."*

EMA, 2001, Points to Consider on Meta-Analysis:

"Plan for evaluation of ***consistency and robustness****. This [meta-analysis] should aim to demonstrate* ***similar effects in alternative analyses of different endpoints, different sub-populations, different subsets of studies*** *…"*

Issues

- **Decision Errors (Precision, Reliability)**
  - Type 1 Error: multiple testing
  - Type 2 Error: insufficient sample size/power
- **Bias (Validity, Accuracy):** eg, selection of (positive) studies/subgroups
- **Separation of sources** of treatment heterogeneity:
  - (unsystematic) measurement error
  - random overall treatment effect
  - true differences between treatment effects caused by differences in study design (including patient population)

Examples

- **Example 1 (2009 – 2024):**
  - Indication: Acute Coronary Syndrom
  - Treatment: Ticagrelor (Brilinta, US; Brilique, EU)
  - Controversial data: PLATO RCT
  - Subgroup of interest: US

- **Example 2 (1984 – 2009):**
  - Indication: Acute Myocardial Infarction
  - Treatment: Intravenuous Magnesium Sulfate (MgSO4)
  - Controversial data: Meta-Analysis of 16 RCTs
  - Study of interest: ISIS-4 trial

Methods

- **Many methods** proposed (see reviews and tutorials on explorative subgroup analysis by Lipkovich, Dmitrienko, D'Agostino, 2017; Lipkovich, Svensson, Ratitch, Dmitrienko, 2024)

- **Bayesian Hierarchical Models** (BHM, "hierarchical models"), eg Ruberg et al. (2023)

- **Bayesian Predictive Cross-Validation Models** (BPCVM, "predictive models"), eg Diaz et al (2011/12c), Igl & Constant (2024)

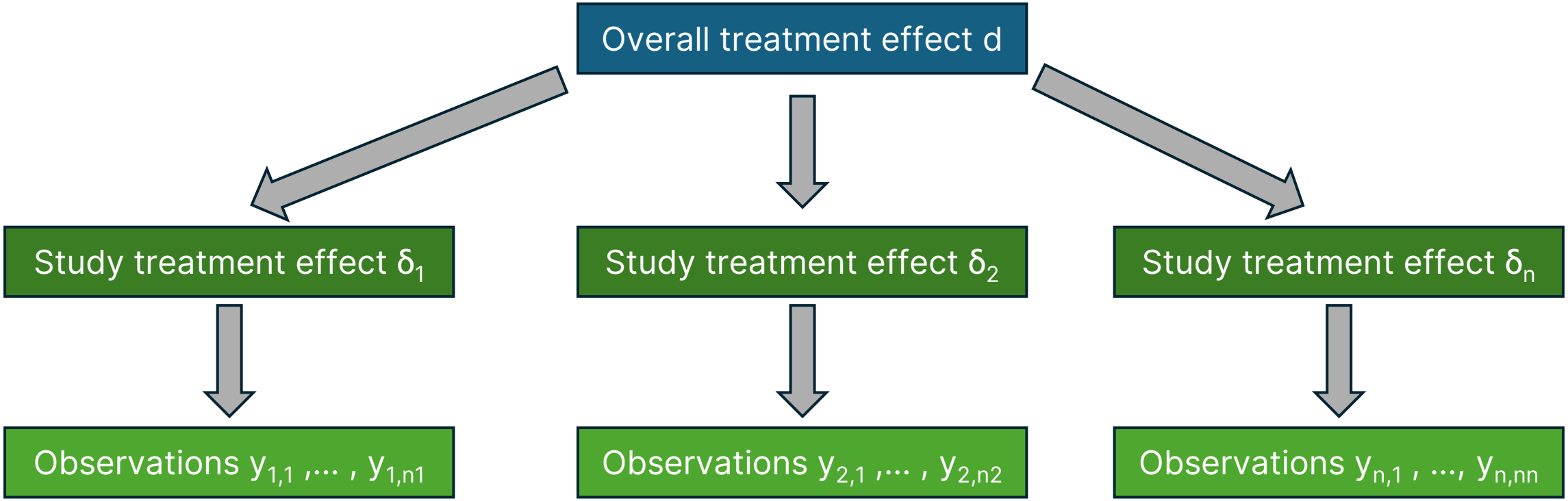
Some Recommendations

- **Bayesian Hierarchical Models promoted for subgroup analysis** (outlier detection) without addressing problematic assumptions (cf, Ruberg et al, 2023; Igl & Constant, 2024; Ruberg, 2024; Nature Reviews Drug Discovery)

- *"The appropriate tool for examination of single trials in a meta-analysis is **cross-validation based on a "leave one out"** approach."* (Dias, Sutton, Welton, & Ades, 2011/2012c, NICE Decision Support Unit)

- *"Plea for routinely presenting **prediction intervals** in meta-analysis."* (IntHout, Ioannidis, Rovers, & Goeman, 2016, BMJ Open).

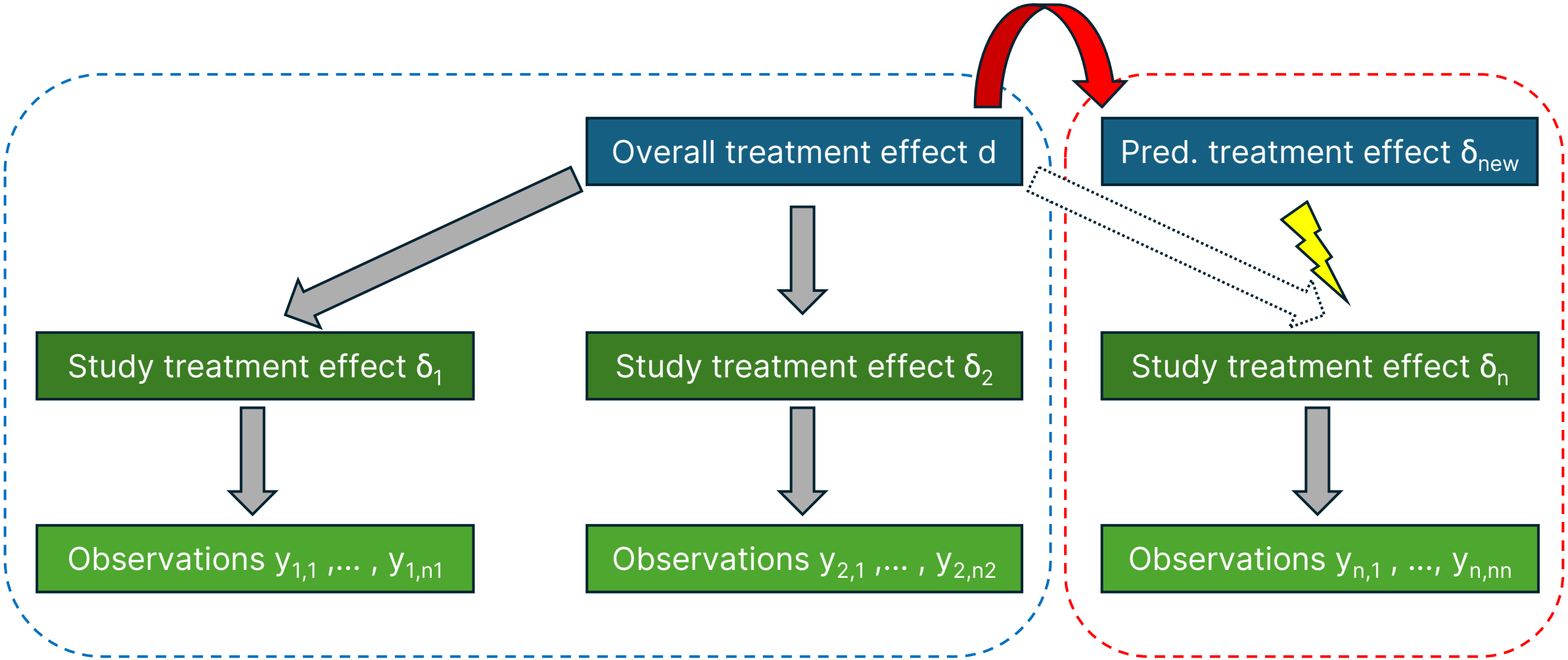
Bayesian Hierarchical Model



Bayesian Hierarchical Model

- **Basic Method:**
  - assumes overall treatment effect (including study of interest)
  - "borrows strength" between subgroups & "shrinks" estimates
  - estimates overall mean and its credible interval
- **Pros:**
  - widely used
  - available code
- **Cons:**
  - Assumption not appropriate for subgroup outlier detection
  - Overall mean and its credible interval less relevant
  - Crude interpretation: Overlap of CrI of treatment effect of trial & overall treatment effect

Bayesian Predictive Cross-Validation Model



Bayesian Predictive Cross-Validation Model

- **Basic Method:**
  - does not assume an overall treatment effect (incl. study of interest)
  - predicts the treatment effect in future subgroup like the study of interest based on all other subgroups.
- **Pros:**
  - (plausibly) assumes overall treatment effect only in other subgroups
  - Predictive mean & interval for future study like study of interest
- **Cons:**
  - less used
  - other assumptions still required, eg, exchangeability, representativeness

Example: Pairwise Meta-Analysis of treatment of Acute Myocardial Infarction

- **Disease**: Acute myocardial infarction
- **Treatments**: Magnesium (IV) vs Placebo
- **Studies**: 16 studies , incl. ISIS-4 "mega-trial"
- **Design**: Pairwise meta-analysis of randomized controlled trials

- **Controversially discussed treatment and indication** with randomized controlled trials and meta-analysis between 1984 and 2009, see reviews by Egger, Davey-Smith (1995), Higgins & Spiegelhalter (2002), Li et al. (2007/2009)

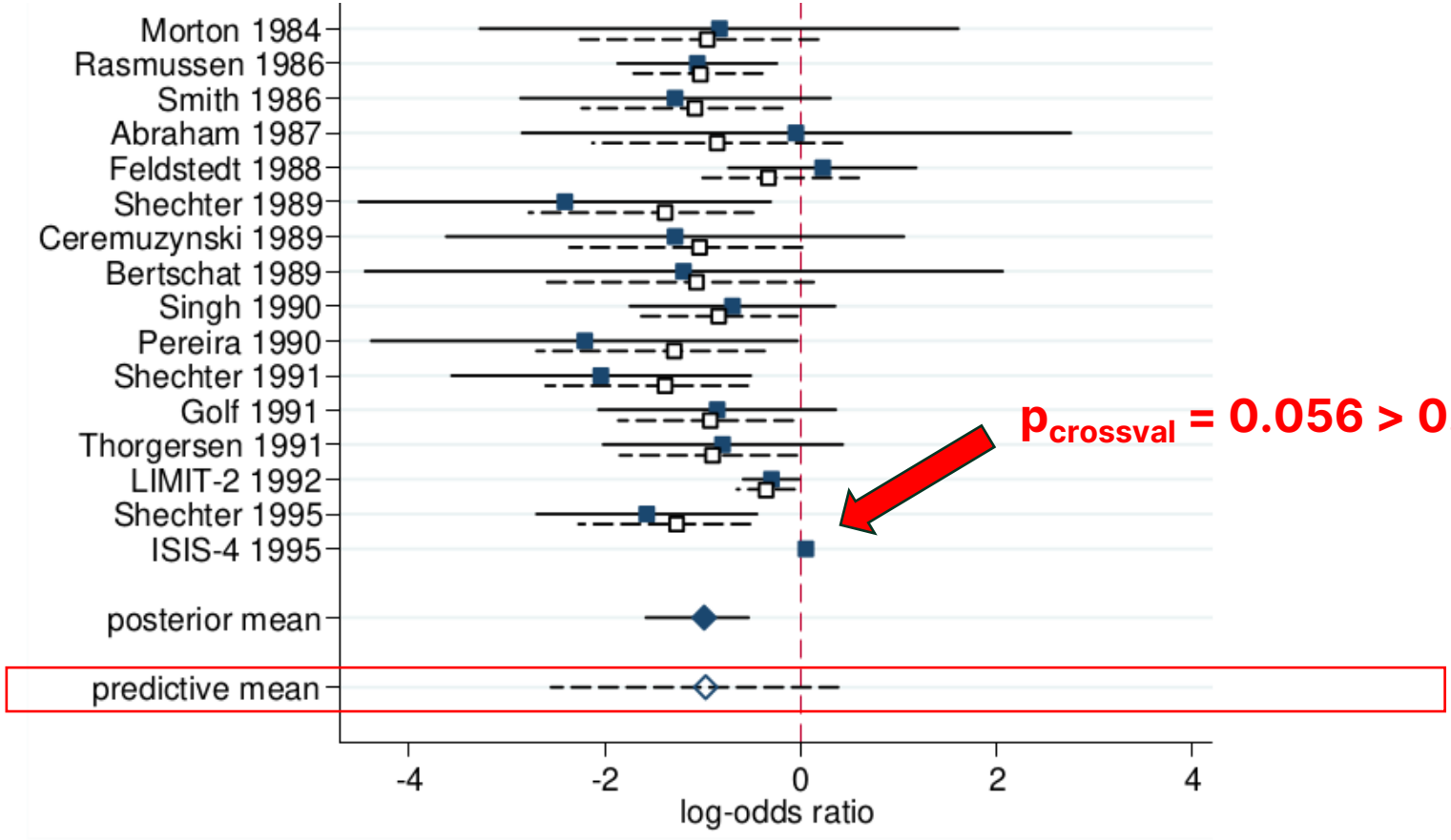
Meta-Analysis of 16 trials: Magnesium (IV) vs Placebo

Table 1 Number of deaths out of the total number of patients for 16 trials of intravenous magnesium against placebo, for patients with acute myocardial infarction.[27]

			Placebo		Magnesium	
Trial ID	Trial Name	Year	Deaths	Total	Deaths	Total
1	Morton	1984	2	36	1	40
2	Rasmussen	1986	23	135	9	135
3	Smith	1986	7	200	2	200
4	Abraham	1987	1	46	1	48
5	Feldstedt	1988	8	148	10	150
6	Shechter	1989	9	56	1	59
7	Ceremuzynski	1989	3	23	1	25
8	Bertschat	1989	1	21	0	22
9	Singh	1990	11	75	6	76
10	Pereira	1990	7	27	1	27
11	Shechter1	1991	12	80	2	89
12	Golf	1991	13	33	5	23
13	Thorgersen	1991	8	122	4	130
14	LIMIT-2	1992	118	1157	90	1159
15	Shechter2	1995	17	108	4	107
16	ISIS-4	1995	2103	29039	2216	29011

Diaz et al., 2011, p. 19

BPCVM with 15 Trials & ISIS-4



$p_{crossval} = 0.056 > 0.025$

Figure 3 Magnesium Example: Crude log-odds ratios with 95% CI (filled squares, solid lines); posterior mean with 95% CrI of the trial-specific log-odds ratios, "shrunken" estimates, (open squares, dashed lines); posterior mean with 95% CrI of the posterior (filled diamond, solid line) and predictive distribution (open diamond, dashed line) of the pooled treatment effect, obtained from a RE model excluding the ISIS-4 trial.

Conclusion

- **Methods:**
  - BPCVM conceptually superior to BHM
  - Technical barrier: WinBUGS system files (.odc) no longer available & WinBUGS no longer available

- **Example: Magnesium (IV) vs Placebo**
  Treatment effect as extreme as ISIS-4 is unlikely, but still possible, ie no outlier (cf, $p_{crossval} = 0.056$ > all two-sided decision thresholds)

- **Limitation:** Application of hierarchical models limited to scenarios were treatment effects are estimated in multiple subgroups, e.g. meta-analysis, less useful for analysis of few subgroups in a single clinical trial


THANK YOU
THANK YOU
Bedankt
Bedankt
BEDANKT
BEDANKT
Danke
Danke
DANKE
MERCI
GRATIAS TIBI
GRATIAS TIBI
HVALA VAM
Hvala vam
ΕΥΧΑΡΙΣΤΩ
ευχαριστώ
СПАСИБО
СПАСИБО
СПАСИБО
Спасибо
Teşekkürler
Teşekkürler
Teşekkürler
TEŞEKKÜRLER
TERIMA KASIH
TERIMA KASIH
gracias
gracias
MULŢUMESC
MULŢUMESC
ありがとう
ありがとう
ありがとう
谢谢
谢谢
谢谢
धन्यवाद
धन्यवाद
धन्यवाद
ধন্যবাদ
ধন্যবাদ
ขอบคุณ
ขอบคุณ
நன்றி
მადლობთ
მადლობთ
متشكرم
متشكرم
متشكرم
شكرًا لك
شكرًا لك

References – BHM vs BPCVM

- Ruberg, S. J. et al. Application of Bayesian approaches in drug development: starting a virtuous cycle. Nat. Rev. Drug Discov. 22, 235–250 (2023).
- Igl, W., Constant, J. Bayesian approaches in drug development: continuing the virtuous cycle. Nat Rev Drug Discov 23, 962–963 (2024).
- Ruberg, S.J. Reply to 'Bayesian approaches in drug development: continuing the virtuous cycle'. Nat Rev Drug Discov 23, 964 (2024). https://doi.org/10.1038/s41573-024-01054-8
- IntHout, J., Ioannidis, J. P. A., Rovers, M. M. & Goeman, J. J. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open, 6, e010247 (2016).
- Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. Clinical Trials 2011; 8(2): 129–143.
- Henderson NC, Louis TA, Wang C, Varadhan R. Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. Health Services and Outcomes Research Methodology 2016; 16(4): 213–233.
- Berger JO, Wang X, Shen L. A Bayesian approach to subgroup identification. Journal of Biopharmaceutical Statistics 2014; 24(1): 110–129.

References – BHM vs BPCVM

- Dias, S., Welton, N.J., Sutton, A.J., Ades, A.E. (2011/2012a). NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making. 2011; last updated April 2012; available from http://www.nicedsu.org.uk
- Dias, S., Welton, N.J., Sutton, A.J., Ades, A.E (2011/2012b). NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis. Decision Support Unit, ScHARR, University of Sheffield, 2011; last updated April 2012 (see Appendix/Program 1a), https://www.sheffield.ac.uk/media/34176/download?attachment
- Dias, S., Sutton, A. J., Welton, N. J. & Ades, A. E (2011/2012c). NICE DSU Technical Support Document 3: Heterogeneity: Subgroups, Meta-Regression, Bias and Bias-Adjustment. Decision Support Unit, ScHARR, University of Sheffield, 2011 (see Appendix/Program 1), https://www.sheffield.ac.uk/media/34179/download?attachment
- NICE Decision Support Unit (2025). Full list of technical support documents (TSDs). University of Sheffield. https://www.sheffield.ac.uk/nice-dsu/tsds/full-list

References – Magnesium (IV) & ISIS-4 trial

- Li, J., Zhang, Q., Zhang, M., Egger, M. (2007/2009). Intravenous magnesium for acute myocardial infarction. Cochrane Database of Systematic Reviews. Issue 4: Art. No.: CD002755
- Higgins, J.P.T., Spiegelhalter, D.J. (2002). Being sceptical about meta-analyses: a Bayesian perspective onmagnesium trials in myocardial infarction. International Journal Of Epidemiology; 31(1):96-104.
  Note: see Appendix 3 for BUGS code for meta-regression approach adjusted for sample size
- Egger, M., Davey-Smith, G. (1995). Misleading meta-analysis. British Medical Journal 1995; 310:752-754.
- ISIS-4 Collaborative Group (1995). ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction. Lancet 1995;345:669-85.
- Yusuf S, Koon T, Woods K (1993). Intravenous magnesium in acute myocardial infarction. An effective, safe,simple, and inexpensive intervention. Circulation;87:2043-6.

References – Reviews on Subgroup Analysis

- Lipkovich I, Svensson D, Ratitch B, Dmitrienko A (2024). Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. Stat Med. 2024 Sep 30;43(22):4388-4436. doi: 10.1002/sim.10167. Epub 2024 Jul 25. PMID: 39054669. Pre-Print: https://arxiv.org/pdf/2311.14889v2
- Lipkovich I, Dmitrienko A, B. D'Agostino Sr. R (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Statistics in Medicine 2017; 36: 136–196.
- Brand KJ, Hapfelmeier A, Haller B. A systematic review of subgroup analyses in randomised clinical trials in cardiovascular disease. Clin Trials. 2021 Jun;18(3):351-360. doi: 10.1177/1740774520984866.

Appendix

Bayesian Hierarchical (Logistic) Model

Likelihood Distribution

$$r_{ik} \sim \text{Binomial}(p_{ik}, n_{ik})$$

- r = #deaths, n = total #patients
- k = treatment arm, i = number of trial

Linear Predictor

$$\text{logit}(p_{ik}) = \mu_i + \delta_{i,1k} I_{\{k \neq 1\}}$$

- mu = treatment effect in placebo group
- delta = relative treatment in experimental group

Common Distribution

$$\delta_{i,12} \sim N(d, \sigma^2)$$

- d = overall treatment effect in experimental group

Prior Distributions

$$d \sim N(0, 100^2)$$
$$\mu_i \sim N(0, 100^2)$$
$$\sigma \sim U(0,5)$$

- Vague prior for overall treatment effect d
- Vague prior for treatment effect in control group mu
- Vague prior for between-study variance sigma^2

Dias et al. 2011/2012c

Sampling from Predictive Distribution

Predictive treatment effect in experimental group

$$\delta_{new} \sim N(d, \sigma^2)$$

- delta_new = treatment effect of **future trial of same size as study of interest**

Predictive treatment effect in placebo group

$$p_{base} \sim \text{Beta}(a, b)$$

- a = number of deaths
- b = number of non-deaths

Predictive absolute treatment effect in experimental group

$$\text{logit}(p_{new}) = \text{logit}(p_{base}) + \delta_{new}$$

- p_new = predictive probability of mortality in experimental treatment arm in a future study

Predictive number of deaths in experimental arm in future trial like study of interest

$$r_{new} \sim \text{Binomial}\,(p_{new}, n_{new})$$

- p_new = predicted number of deaths in experimental group in study of interest
- n_new = future (=observed) number of patients in experimental group in study of interest

Evaluation of Outlier

Probability of observing a result as extreme result as observed in study of interest, ie $\mathbf{p_{crossval}}$

$$\Pr(r_{new} \geq r_{observed})$$

- r_{new}: Number of predicted deaths in experimental arm in observed study of interest
- $r_{observed}$: Number of observed deaths in experimental arm in observed study of interest

Evaluation Threshold:

- **Use conventional, nominal threshold**: $p_{crossval} \leq 0.05/2 = 0.025$
- **Use Bonferroni Method**: $p_{crossval} \leq 0.05/(2*n)$

BHM with all Trials (incl. ISIS-4)

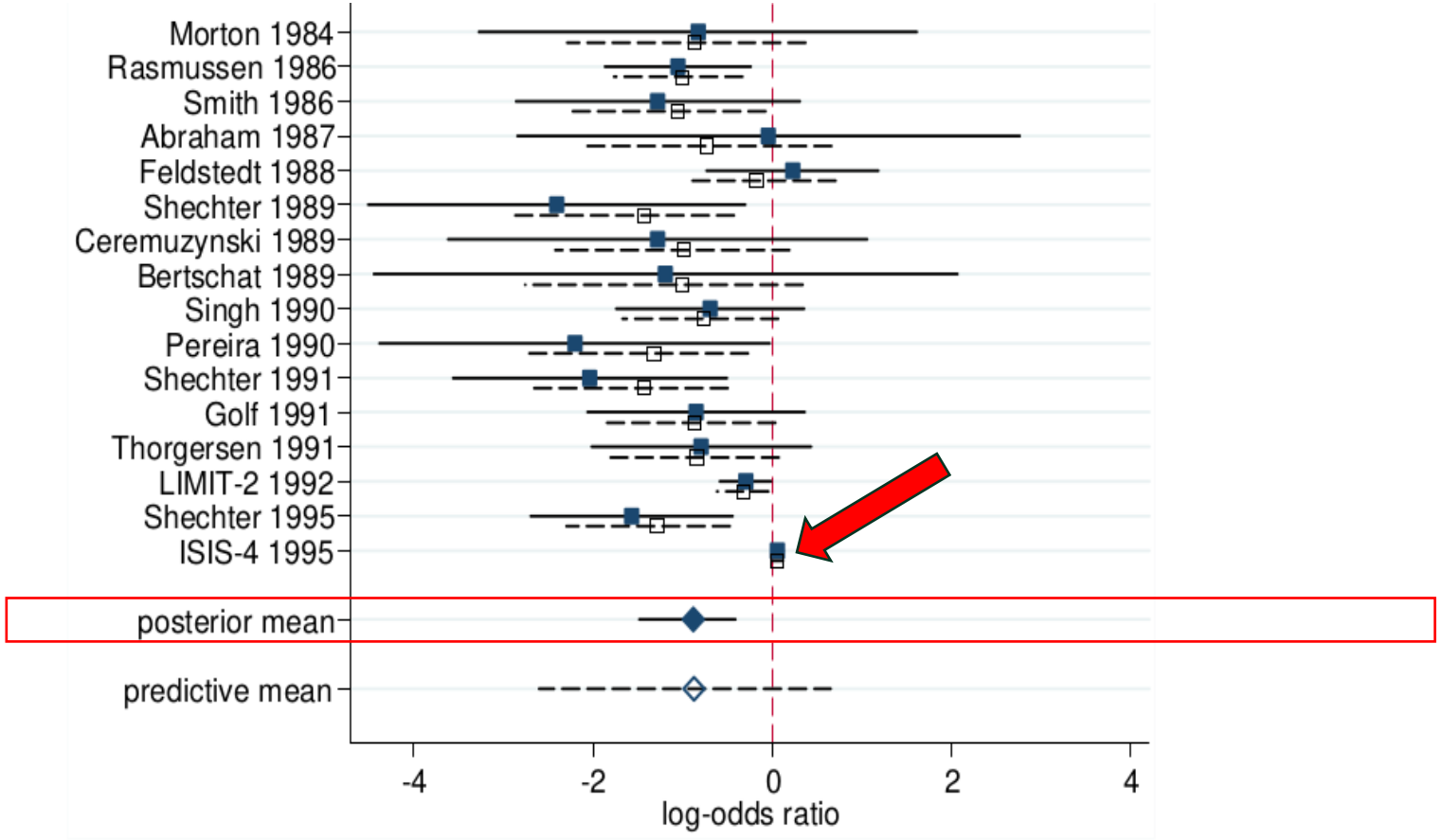


Figure 2 Magnesium Example: Crude log-odds ratios with 95% CI (filled squares, solid lines); posterior mean with 95% CrI of the trial-specific log-odds ratios, "shrunken" estimates, (open squares, dashed lines); posterior mean with 95% CrI of the posterior (filled diamond, solid line) and predictive distribution (open diamond, dashed line) of the pooled treatment effect, obtained from a RE model including all the trials.

Veranex