Early stopping of clinical trials: impacts on treatment effects, meta-analyses and cost-effectiveness

Ian Marschner

NHMRC Clinical Trials Centre
University of Sydney

Australian Pharmaceutical Biostatistics Group (APBG)
Sydney, December 2019

Joint work with Manjula Schou and Lisa Askie

Truncated clinical trials

- A randomised controlled trial (RCT) that stops earlier than initially planned is called a **truncated RCT**
- There are various reasons why an RCT might be truncated:
  - Benefit: Experimental treatment superior
  - Safety: Experimental treatment has unacceptable adverse effects
  - Futility: Little prospect of showing superiority
- **In this talk we will only discuss truncation due to benefit**
- Early stopping requires a sequential monitoring rule that appropriately controls the type I error for multiple testing e.g. O'Brien-Fleming, Haybittle-Peto, Pocock *etc.*

Early stopping bias

- There has been concern that early stopping of randomised clinical trials (RCTs) due to benefit of the experimental treatment over the control leads to **overestimation** of the treatment effect

- This bias may arise because random fluctuations in the direction of the experimental treatment can lead to early stopping of the RCT

- Some researchers have been particularly concerned about the effect of early stopping on systematic reviews and have advocated sensitivity analyses assessing the impact on meta-analyses

Debate

- **Early stopping biases systematic reviews:**
  - *Clinicians should view results of RCTs stopped early for benefit with skepticism* (Montori *et al.*, *JAMA* 2005)
  - *If reviewers do not note truncation and do not consider early stopping for benefit, meta-analyses will report overestimates of effects* (Bassler *et al.*, *JAMA* 2010)
  - *Pooled effects from meta-analyses including truncated RCTs are likely to overestimate effect ... Such circumstances call for sensitivity analyses omitting truncated RCTs.* (Bassler *et al.*, *SMMR* 2013)

- **Early stopping does not bias systematic reviews:**
  - *Systematic reviews are not biased by results from trials stopped early for benefit* (Goodman, *J Clin Epi* 2008)
  - *Early termination . . . does not lead to substantive bias in the estimation of treatment effects* (Berry *et al.*, *JAMA* 2010)
  - *The pooled effects from meta-analyses that include truncated RCTs do not show a problem of upward bias* (Wang *et al.*, *Clin Trials* 2016)

Bias in treatment effect estimates

- When we talk about bias in treatment effects from sequentially monitored studies we need to distinguish between two types of bias

- **Unconditional bias:**
  - Average difference between estimated effect and population effect, averaged over stopping stages that *could* occur

- **Conditional bias:**
  - Average difference between estimated effect and population effect, for the particular stopping stage that *did* occur

Conditional effect of truncation

- The observed treatment difference in a **truncated RCT** will tend to **overestimate** the actual treatment difference
- That is, the estimate $\hat{\theta}$ (MLE) of the treatment difference θ is *conditionally* biased

$$E(\hat{\theta}|\text{truncation}) > \theta$$

- This arises because random fluctuations favouring the experimental treatment can lead to early stopping that "locks in" a higher estimate

Conditional effect of non-truncation

- The observed treatment difference in a **non-truncated RCT** will tend to **underestimate** the actual treatment difference
- That is, the estimate $\hat{\theta}$ (MLE) of the treatment difference θ is *conditionally* biased

$$E(\hat{\theta}|\text{non-truncation}) < \theta$$

- This arises because we are conditioning on having less extreme treatments so that the study makes it through to the final analysis

Does early stopping always lead to overestimation?

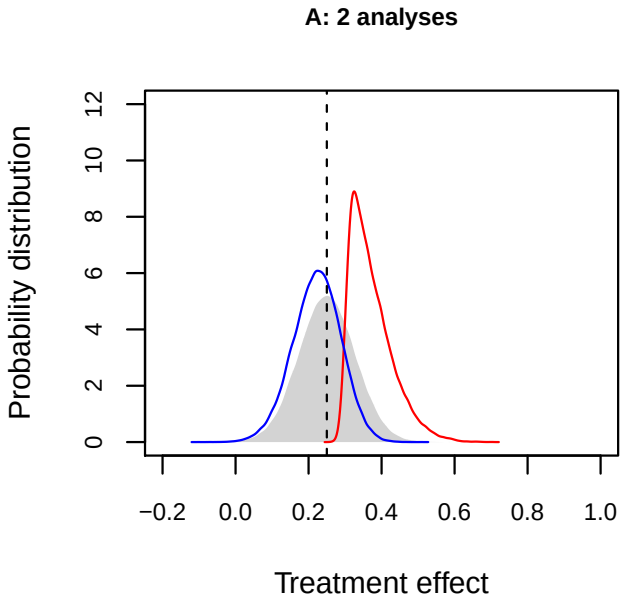
- The debate around impacts of early stopping usually assumes early stopping always leads to overestimation. Using theory and simulations we can see that the situation is more complex.

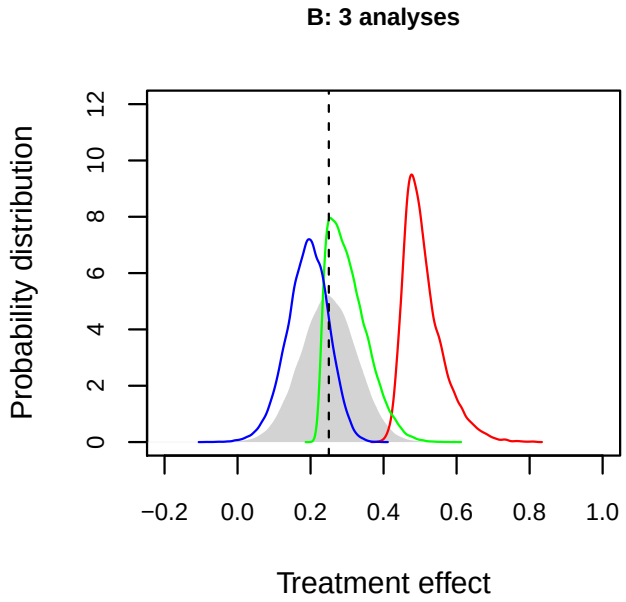
  e.g. Fan, DeMets and Lan *J Biopharm Stat* 2004; Schou and Marschner *Stat Med* 2013; Marschner and Schou *SMMR* 2019

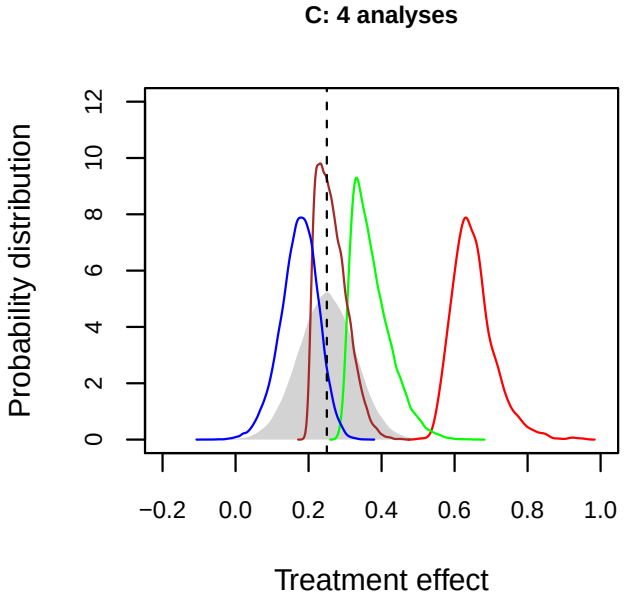
- Conditional on the stopping stage:
  - **First interim analysis:** overestimation
  - **Final analysis:** underestimation
  - **Intermediate interim analyses:** Not predictable – could be overestimation, underestimation or unbiased estimation
  - **Earlier analyses:** greater tendency for overestimation
  - **Later analyses:** greater tendency for underestimation

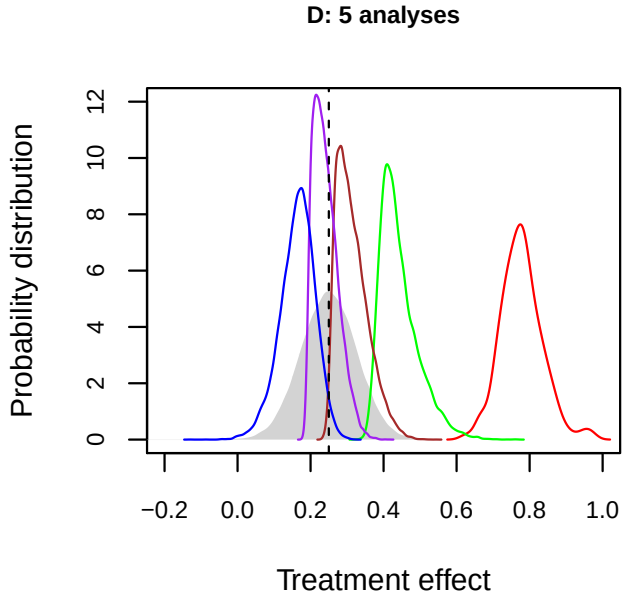
- Unconditionally (averaging over all stopping stages): unbiased

Red: First analysis Blue: Final analysis Shaded: Unconditional









Impact on clinical trials guidelines

- The extensive debate about **truncation bias** in high impact journals has led to an impact on guidelines for conducting clinical trials and systematic reviews
    - CONSORT Statement for transparent trial reporting
    - GRADE Guidelines for rating the quality of evidence
    - PRISMA guidelines on reproting of systematic reviews and meta-analyses
    - ICH E9 statistical principles for clinical trials
- This impact has centred on overestimation due to early stopping with little discussion about **non-truncation bias** that leads to underestimation

CONSORT Statement

- Reporting standards for RCTs: Sections 7b and 14b address RCTs that stopped early
- *"Readers will likely draw weaker inferences from a trial that was truncated in a data-driven manner versus one that reports its findings after reaching a goal independent of results"*
- Reviews empirical evidence supporting the view that early stopping leads to overestimation of treatment effects
- Makes recommendations on information that should be reported: *"timing of interim analyses, what triggered them, how many took place, whether these were planned or ad hoc, and whether there were statistical guidelines and stopping rules in place a priori"*

GRADE Guidelines

- A methodology for assessing the quality of evidence and for rating the strength of recommendations
- A key element of the methodology is consideration of the risk of bias [GRADE Handbook, Section 5.2.1; Guyatt et al. *JClinEpi*, 2011]
- *"Limitations in the study design and execution may bias the estimates of the treatment effect"*
- Early stopping is a study limitation carrying the risk of bias: *"Substantial overestimates are likely in trials with <500 events"*
- *"Systematic reviews should provide sensitivity analyses of results including and excluding studies that stopped early for benefit; if estimates differ appreciably, those restricted to the trials that did not stop early should be considered the more credible"*

PRISMA Statement

- Guidelines for transparent reporting of meta-analyses and systematic reviews
- Does not explicitly refer to early stopping of RCTs
- Assessment of the risk of bias is a key element of the PRISMA reporting checklist
- Systematic reviewers may therefore investigate the impacts of early stopping as part of the PRISMA reporting process, using the GRADE sensitivity analysis approach

ICH E9

- Original version pre-dates debate on early stopping bias and focuses on issues such as control of type I error and integrity of interim analysis process
- Recent addendum focuses on estimands and "intercurrent events", as well as sensitivity analysis to assess the impact of intercurrent events on the ability to estimate the estimand
- Although the ICH E9 does not mention interim monitoring bias directly, early stopping (or indeed not stopping) may be considered an intercurrent event that affects interpretation of the study endpoint
- Sensitivity analysis and/or adjustment for interim monitoring may be considered as part of the framework covered by ICH E9

Adjustment for bias

- Since most commentary has focused the bias associated with truncated studies, we considered the behaviour of non-truncated studies
- In particular, we considered the impact of restricting sensitivity analyses to non-truncated studies
- **Adjustment for bias**: Underestimation in non-truncated studies can be adjusted for using *conditional maximum likelihood estimation*
- *If truncated studies are excluded from sensitivity analysis, it may be better to first adjust the non-truncated studies for underestimation*

Conditional statistical inference

- A general analysis method that interprets the observed treatment effect using knowledge of the specific design that *did* occur, rather than with reference to all designs that *could* have occurred
- The fact that a particular stopping stage has occurred, gives us information about how the observed treatment effect will behave
- By using knowledge of how treatment effects tend to behave for particular stopping stages, we can adjust for conditional bias
- **Conditional maximum likelihood estimation** can be used to remove the conditional bias associated with standard analyses

Sequential framework

- $N_1 < N_2 < \cdots < N_K < N_{K+1}$ are the total sample sizes at K interim analyses and 1 final analysis
- θ is a treatment effect size on an appropriate scale
- Z_k is a standardised test statistic at analysis k with

$$Z_k \sim \mathsf{N}\big(\theta\sqrt{N_k}, 1\big)$$

- At analysis $k \leq K$ continue if $Z_k \in [-C, C]$ else stop
- At analysis $K+1$ stop regardless of the value of Z_{K+1}
- If T is the (random) stopping stage then the MLE is

$$\hat{\theta} = \frac{Z_T}{\sqrt{N_T}}$$

- MLE performs well *unconditionally*, averaged over all stopping stages

Conditional MLE

- Conditioning on the stopping stage may be better for an individual study where we know the stopping stage
- Conditional likelihood

$$L(\theta|T=k) = \frac{\exp\left[-0.5(Z_k - \theta\sqrt{N_k})^2\right]}{\Pr(T=k|\theta)}$$

- Conditional score equation

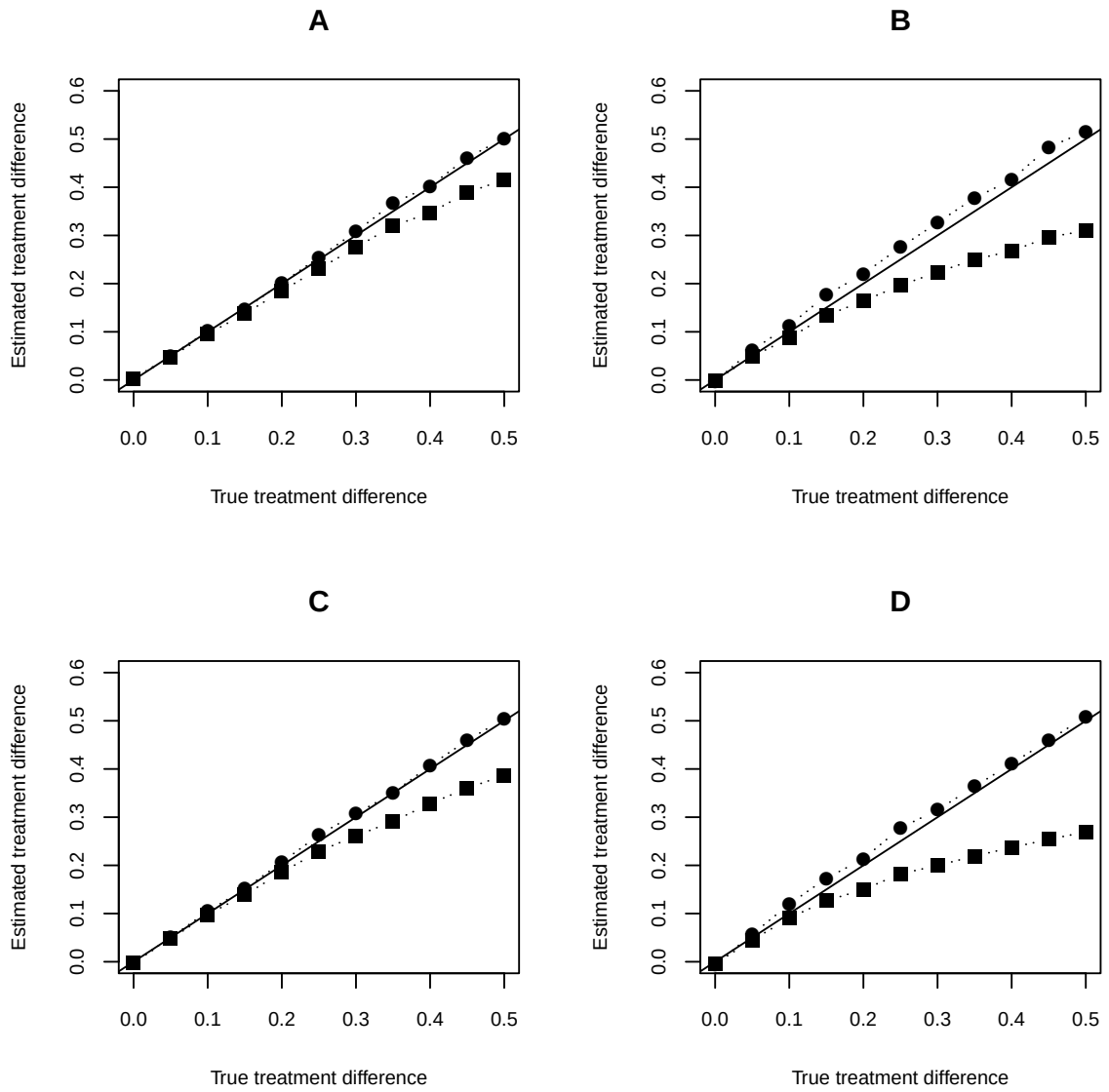
$$S(\theta) = Z_k - \theta\sqrt{N_k} - P_k(\theta) = 0$$

where

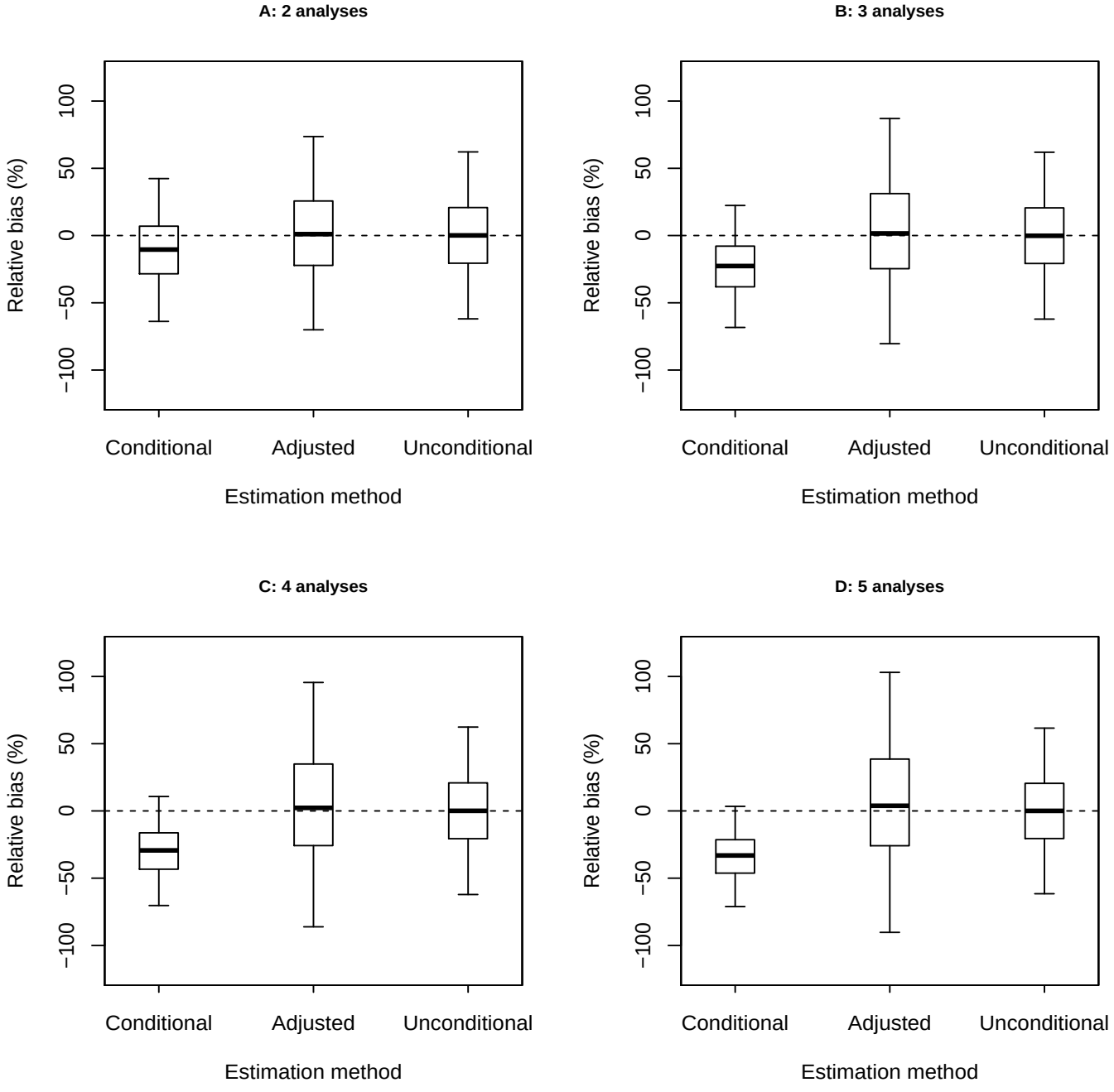
$$P_k(\theta) = \frac{(d/d\theta)\Pr(T=k|\theta)}{\sqrt{N_k}\Pr(T=k|\theta)}$$

- Conditional MLE is the solution of $S(\hat{\theta}_c) = 0$

Conditional versus unconditional: non-truncated studies



Non-truncated studies: conditional vs adjusted



Confidence intervals

- Fisher information associated with conditional log-likelihood can be used for confidence interval construction

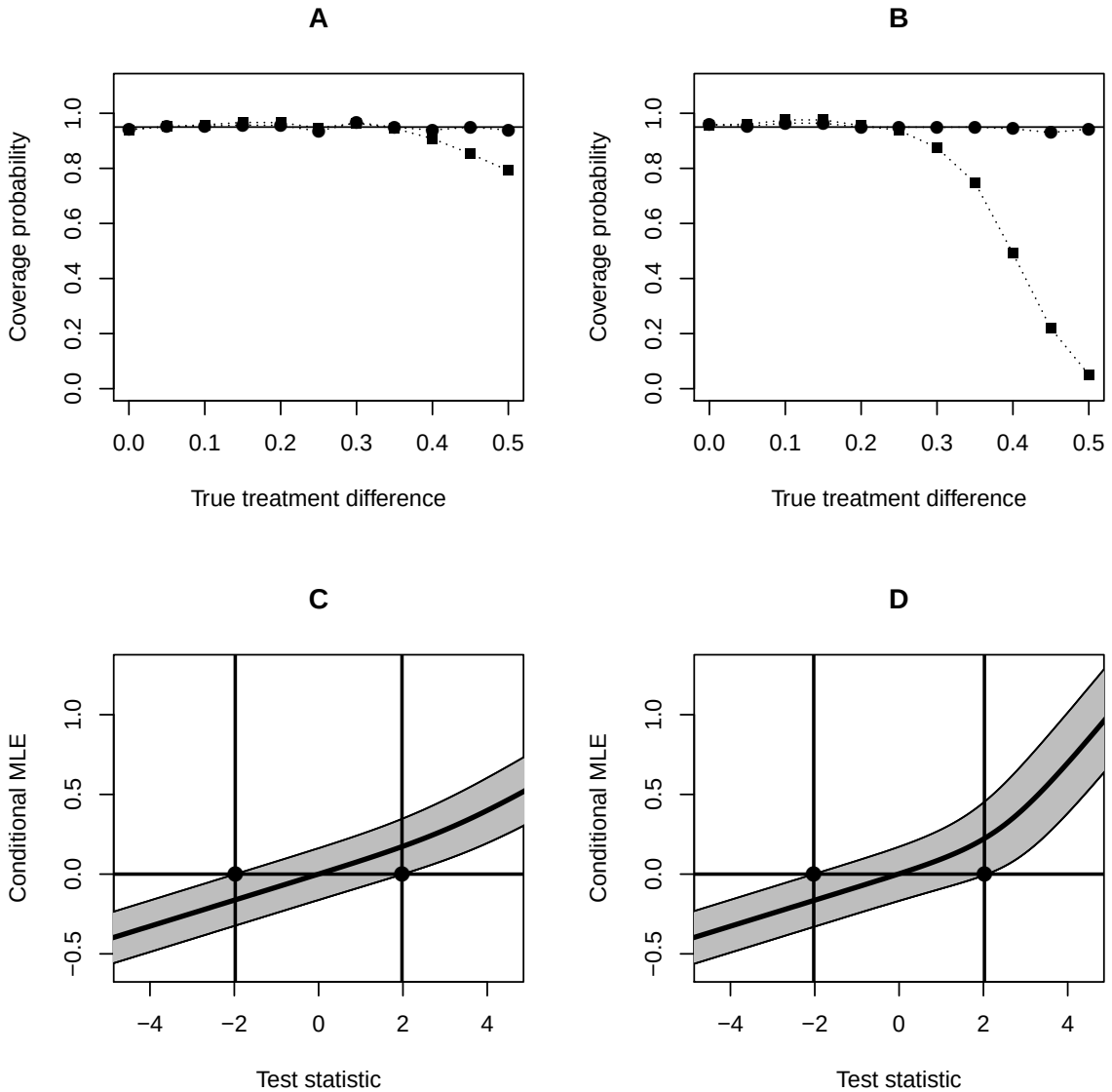
$$I_c(\theta) = -\frac{d^2 L_c(\theta)}{d\theta^2} \qquad SE(\hat{\theta}_c) = \sqrt{I_c(\hat{\theta}_c)^{-1}}$$

- Confidence interval:

$$[\theta_{Lc}, \theta_{Uc}] = \left[\hat{\theta}_c - \frac{q_{1-\alpha/2}}{\sqrt{I_{Kc}(\hat{\theta}_c)}}, \hat{\theta}_c + \frac{q_{1-\alpha/2}}{\sqrt{I_{Kc}(\hat{\theta}_c)}}\right]$$

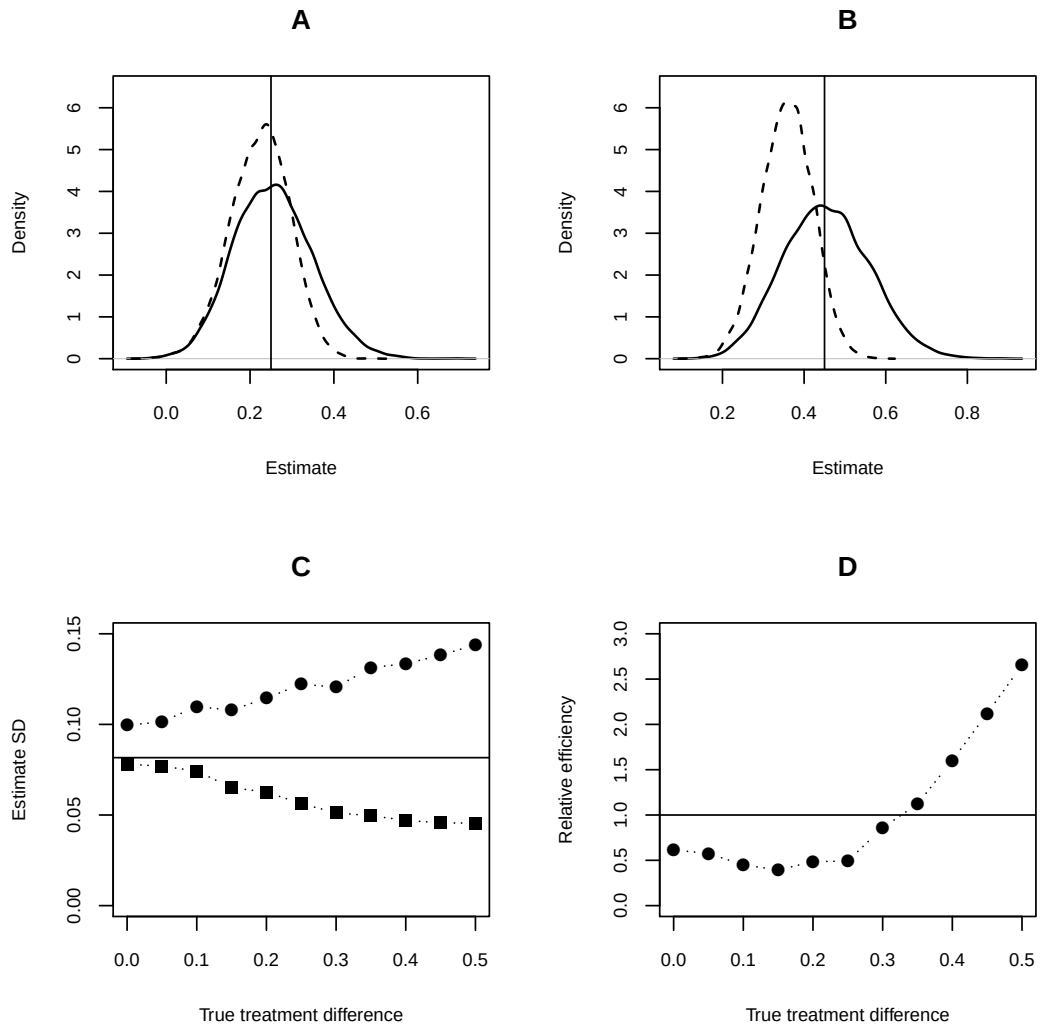
- Main issues:
  - Coverage probability of confidence interval
  - Consistency of confidence interval and sequential hypothesis test

Coverage and testing



Loss of information

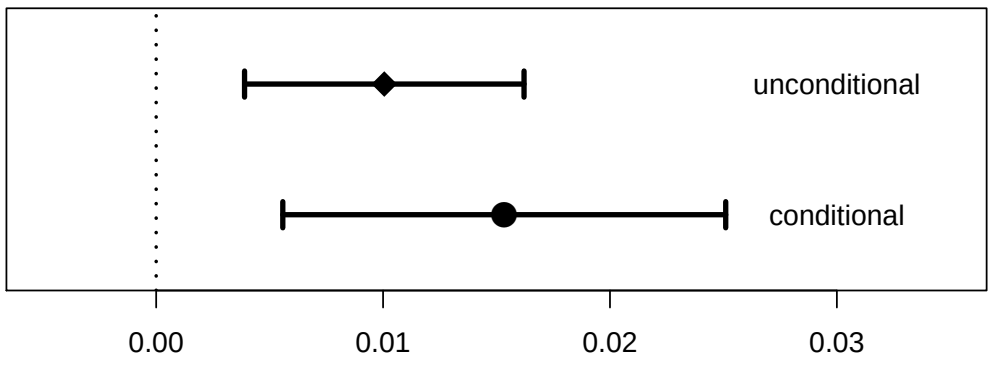
A problem with the conditional MLE is that T is not an ancillary statistic



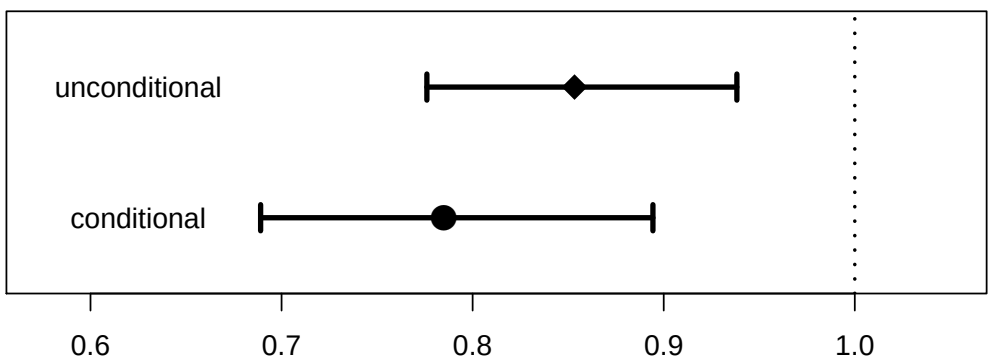
Example: GUSTO study

- Streptokinase vs tPA in acute myocardial infarction (heart attack)
- Primary outcome 30-day mortality
- 3 interim analyses at approx 25%, 50% and 70% of planned information (OBF stopping rule)
- Proceeded through to final analysis with no early stopping
- Study was significant in favor of tPA but treatment effect was small
- Generated controversy due to small treatment effect
- Magnitude of treatment effect may have been underestimated due interim monitoring
- This may have important implications, particularly for cost-effectiveness analyses

Example: GUSTO study



Risk difference and confidence interval



Odds ratio and confidence interval

Cost-effectiveness

- The GUSTO study generated substantial controversy because the mortality reduction of 1.01% (7.31% on streptokinase compared to 6.31% on tPA) was considered not worth the substantial cost

- Incremental cost-effectiveness ratio (ICER)

$$\text{ICER} = \frac{\Delta\text{cost}}{\Delta\text{risk}} \approx 99 \times \Delta\text{cost}$$

- The adjusted mortality reduction of 1.53% has a direct impact on the cost-effectiveness

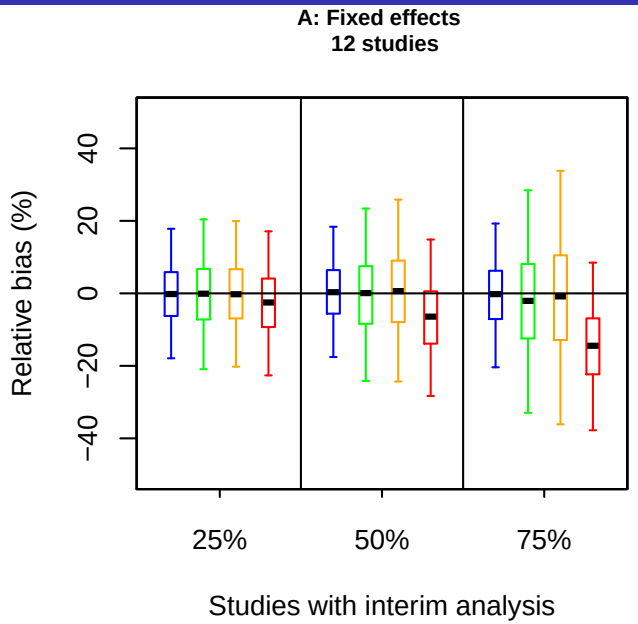
$$\text{adjusted ICER} = \frac{\Delta\text{cost}}{\text{adjusted } \Delta\text{risk}} \approx \frac{2}{3} \times \text{ICER}$$

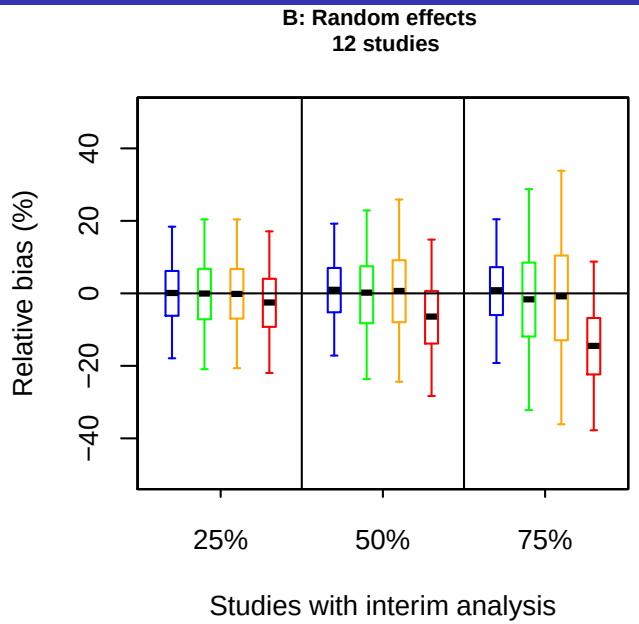
- This may have an important effect on the assessment of cost-effectiveness

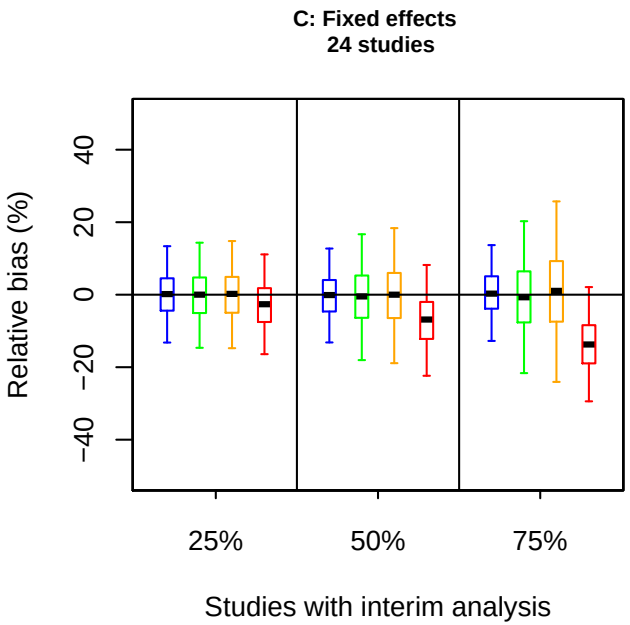
Meta-analysis strategies

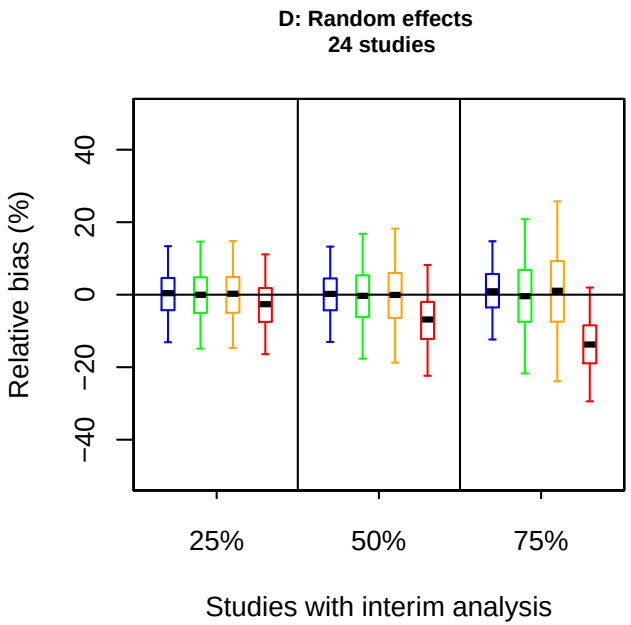
- **All-study strategy**: a standard meta-analysis of all available studies, including studies that stopped early
- **Crude sensitivity analysis**: a standard meta-analysis excluding all studies that stopped early
- **Adjusted sensitivity analysis**: a meta-analysis excluding studies stopped early, but first adjusting non-truncated studies for underestimation
- **Restricted sensitivity analysis**: a standard meta-analysis restricted to studies that were not subject to interim monitoring

Blue: all studies Green: adjusted Yellow: non-sequential Red: non-truncated









Conclusions

- For clinical trials with potential early stopping, underestimation of the treatment effect in non-truncated studies may be just as important as overestimation in truncated studies
- Conditional estimation procedures may be used to validly inflate the treatment effect estimate and adjust the confidence interval if a study proceeds through to its final analysis
- This is important in contexts where the magnitude of the treatment effect is crucial, such as cost-effectiveness analyses
- Conditioning has the potential to lose some of the information in the sample because the stopping stage is not an ancillary statistic
- Research is ongoing to investigate the usefulness of augmenting the conditional MLE with information that has been lost through the conditioning process

Implications for GRADE guidelines

- GRADE Guidelines for rating the quality of evidence

  "Systematic reviews should provide sensitivity analyses of results including and excluding studies that stopped early for benefit; if estimates differ appreciably, those restricted to the trials that did not stop early should be considered the more credible"

- **This recommendation would lead to underestimation of treatment effects. It should be re-evaluated and modified.**

Recommendations – Reporting

- In adhering to Section 7b and 14b of the CONSORT reporting standards, report all information that would be required to undertake a conditional adjustment of the treatment effect estimate
- When supplemented with standard final analysis information (Section 17), the specific information required is:
  - **Early stopping:** did the study stop earlier than planned? Regardless of whether it stopped early, was there interim monitoring that *could* have led to early stopping?
  - **Timing:** what proportion of the planned total sample size was available at *each* of the interim analyses? It is important to interpret the word "timing" in CONSORT in terms of statistical information.
  - **Stopping rule:** what was the stopping rule used at each interim analysis and was it pre-specified? This can be expressed in terms of a stopping boundary or a well-known name e.g. O'Brien-Fleming.

Recommendations – Sensitivity analysis

- In adhering to the GRADE guidelines on assessing the impact of early stopping on systematic reviews:
- **All-study strategy**: A standard meta-analysis of all studies, including those that stopped early, is the most efficient and unbiased method, and should be the primary analysis in a systematic review
- **Crude sensitivity analysis**: A standard meta-analysis excluding studies stopped early will underestimate the treatment effect and should not be conducted
- **Adjusted sensitivity analysis**: A meta-analysis excluding studies stopped early, but adjusting non-truncated studies for underestimation, is unbiased and is the preferred strategy
- **Restricted sensitivity analysis**: Restricting a meta analysis to only those studies that had no interim monitoring is also unbiased but may be quite inefficient