Version 6, 12/10/2009

Evidence based urology in practice: heterogeneity in a systematic review meta-analysis

Mari Imamura[1], Jonathan Cook[2], Sara MacLennan[1], James N'Dow[1] and Philipp Dahm[3] for the Evidence Based Urology (EBU) Working Group

[1] Academic Urology Unit, University of Aberdeen, UK

[2] Health Services Research Unit, University of Aberdeen, UK

[3] Department of Urology, University of Florida, College of Medicine, Gainesville, Florida

Running head: Heterogeneity

Work count: 1476

Funding: None

MeSH headings: Evidence-Based Medicine, meta-analysis, review, data interpretation

Correspondence:
Mari Imamura, PhD
Research Fellow
Academic Urology Unit
University of Aberdeen
Health Sciences Building
Foresterhill
Aberdeen AB25 2ZD, UK
Phone: (44) 1224 554877
Email: m.imamura@abdn.ac.uk

Case scenario

You are seeing a 62 year male patient in your clinic with longstanding lower urinary tract symptoms (LUTS) refractory to medical management. His international prostate symptom score (IPSS) is 22/35 and he is very unhappy about his current condition. A digital rectal exam is benign with no appreciable nodules. Given his overall good health and degree of symptoms, you recommend a transurethral resection of the prostate (TURP) which you consider the most effective treatment for symptomatic benign prostatic hyperplasia (BPH). The patent is willing to undergo surgery but asks about whether he could undergo one of more 'modern' minimal invasive treatment forms that he has heard good things about. You decide to update your knowledge by searching for the current best evidence on the surgical treatment of BPH.

Clinical question

In patients with BPH (population), how effective are minimally invasive treatments (interventions), when compared with TURP (comparator), in improving symptom scores after surgery (outcome)?

Finding the best evidence

You decide to look for a well-designed systematic review on this topic using PubMed.[1] Two sets of searches using the terms 'benign prostatic hyperplasia' and 'minimally invasive surgery' that you combine with the AND function yields 266 citations (date of search: October 12, 2009). Using the Limits tub to apply a 'systematic review' filter from 'Subset/Topic' limits the search results to 18 review articles. Here, you identify a systematic review entitled 'Minimally invasive treatments for benign prostatic enlargement: systematic review of randomised controlled trials' that addresses your research question.[2]

In this review, 22 trials met the review inclusion criteria. Of these, eight trials reported data on symptom scores 12 months after surgery (primary outcome). Of these, three trials compared TURP with transurethral microwave treatment (TUMT), one with transurethral needle ablation (TUNA) and four with laser coagulation. You

also notice that the review produced a quantitative summary of results across studies (a meta-analysis) for each comparison. The individual trial results and the summary result are shown visually in a 'forest plot' figure (Figure 1).

Evaluating the methods

Criteria for assessing validity of review articles have been previously described.[3] You decide to use the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist,[4] and are satisfied that the methods of the systematic review were strong. Turning to the results, you now assess whether the results are similar (consistent) from study to study; in other words, whether there is any heterogeneity among studies included in the systematic review.

What is heterogeneity?

In interpreting the results of a meta-analysis, it is important to consider not just the summary result but the degree to which the individual studies assess the same clinical question and that their results agree. Where the included patients and application of the interventions broadly similar? Are the results consistent (homogeneous) or do they vary so much that the summary is not reliable? Variability in the study results beyond that which could be expected by chance is called (statistical) heterogeneity. To some extent, it is inevitable that the estimated effects of a specific intervention will vary from study to study since two studies are never identical in design and conduct.[5] However, assessment of heterogeneity is not purely a statistical issue but also involves clinical judgement as to whether it is sensible to combine individual studies or not. Such considerations should occur *prior* to conducting a meta-analysis.

The first step for identifying heterogeneity is to visually inspect the forest plot. Is the direction and magnitude of effect consistent across studies? Do the confidence intervals for individual trials tend to overlap? A large difference in the point estimates with a lack of overlap in confidence intervals would indicate the presence of heterogeneity.

The second step is to look at a statistical test of homogeneity (i.e. no heterogeneity).,The result of such a test is commonly shown in forest plots and it formally assesses whether there is evidence of heterogeneity. A statistically significant result from this test (conventionally assessed using a p-value of 0.10 as opposed to 0.05 due to low sensitivity) indicates systematic variability in study results beyond chance. Another way of expressing this information is through the I^2 statistic that quantifies the inconsistency of study estimates. [5,6] The I^2 ranges from 0% to 100% with larger values indicating a greater proportion of variability is attributable to heterogeneity. As a rough guideline, I^2 values higher than 50% may indicate a moderate to severe heterogeneity that requires caution.[6] Such statistical approaches are no panacea; they are influenced by the number and size of studies included in the meta-analysis and the underlying level of heterogeneity.[5,7] The magnitude of heterogeneity (τ^2 value) could also be considered. When heterogeneity should be considered problematic is debatable.

In our example for the outcome of reduction in symptom scores after surgery, of the four trials comparing laser coagulation with TURP, three trials have point estimates (mean difference) greater than 0 (suggesting TURP is better than laser coagulation), whereas the other has point estimates less than 0 (suggesting laser coagulation is better; see Figure 1). Looking at the confidence intervals for the same comparison, the intervals are quite diffuse. The p-value for the χ^2 test is less than 0.10 signifying statistical evidence of heterogeneity. The corresponding I^2 is over 80% suggesting a high level of heterogeneity. Caution is therefore required in interpreting such results. Similarly for the comparison of TUMT with TURP, the confidence interval of one of the studies (de la Rosette) does not overlap with the other two; Again, this is reflected in a high I^2 value of 84%.

Causes of heterogeneity

We distinguish the potential sources of heterogeneity (clinical and methodological differences) from observing (statistical) heterogeneity in the results. Variation in the estimated intervention effects may reflect differences in patient characteristics, the study setting, or how the interventions were implemented in the included studies

(clinical heterogeneity). Differences in findings may also result from methodological differences (methodological heterogeneity) such as failure to introduce important methodological safeguards against bias such as allocation concealment and blinding.

In our example, the review authors postulate that differences in prostate size and symptom score at baseline may be important sources of heterogeneity observed when looking at the results after surgery. For laser coagulation, the authors also noted variation in operative technique and treatment protocols between trials (e.g. power settings, temperature and site or duration of laser application). They noted that it was difficult to assess their possible impact on intervention effects, as many trials did not describe the technologies used in sufficient details, a problem regularly encountered when conducting systematic reviews.

What can we do when there is heterogeneity?

In the presence of large unexplained heterogeneity investigators may choose not to pool the study results and only report the results of the individual studies. Investigators may decide to conduct a random effects meta-analysis as was done for our example. Use of a random-effects model does not remove heterogeneity but formally allows for a degree of variability between individual trial results (random effects) while still assuming overall coherence. When there is heterogeneity, the confidence intervals around the summary (overall) estimate in the random-effects analysis will be wider than the corresponding one from a fixed-effect analysis. Many therefore see the random-effects approach as the more conservative (and arguably the default) option.[8] However, compared to a fixed-effect meta-analysis, smaller studies will be given more influence in a random-effects meta-analysis. Particular care is required when trial size varies greatly or there are few studies. This difference can be seen in Figure 2 where the random effects analysis of the same studies is conducted and the point estimate and the confidence interval for the summary estimate differ.

Applying the results to the care of your patient

Having applied the outlined methodology to assess the impact of heterogeneity in the case study, you take a cautious approach. Current evidence does not support

minimally invasive treatments over TURP in terms of symptoms at 12 months, let alone does there appear to be any high quality evidence on the long-term outcomes in the published literature. Indeed, current best available evidence supports TURP over TUNA and tends to favour TURP over the other two minimally invasive treatments. Given this, you explain to your patient that the evidence suggests TURP to be at least as good as minimally invasive treatments and possibly the superior treatment of LUTS secondary to BPH in terms of symptom control at 12 months. You would also want to apply the same approach to the other important outcomes reported such as peak urine flow, the need for reoperation and possible adverse effects.

Conclusion

Urologists are often faced with systematic reviews of individual clinical trials with inconsistent results. In this case, it is important to assess the extent and magnitude of heterogeneity and ask whether investigators have sufficiently explored clinical and methodological sources of heterogeneity, which was the focus of a previous BJUI article.[9] Uncritical interpretation of meta-analysis can be misleading and misguide clinical practice. The ability to recognize and interpret heterogeneity is therefore critical to evidence-based practice of urology.

Conflict of interest

None declared.

Acknowledgement

We thank Tânia Lourenço (University of Aberdeen) for providing the original ‘forest plot’ figures and helpful comments.

References

1. Krupski TL, Dahm P, Fesperman SF, Schardt CM. How to perform a literature search. Journal of Urology 2008; 179:1264-1270.
2. Lourenço T, Pickard R, Vale L, Grant A, Fraser C, MacLennan G et al. Minimally invasive treatments for benign prostatic enlargement: systematic review of randomised controlled trials. BMJ 2008; 337:a1662.
3. Tseng TY, Dahm P, Poolman RW, Preminger GM, Canales BJ, Montori VM. How to use a systematic literature review and meta-analysis. Journal of Urology 2008; 180:1248-56.
4. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Medicine 2009: e1000097.
5. Higgins JP. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. International Journal of Epidemiology 2008; 37(5):1158-1160.
6. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327(7414):557-560.
7. Gavaghan DJ, Moore RA, McQuay HJ. An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. Pain 2000; 85(3):415-24.
8. Poole C, Greenland S. Random-effects meta-analyses are not always conservative. American Journal of Epidemiology 1999; 150:469-75.
9. Wang S, Ou Y, Cheng C, Dahm P. Evidence-based urology in practice: when to believe a subgroup analysis? BJU International, in press.