International Journal of Technology Assessment in Health Care, **26:4** (2010), 431–435.

doi:10.1017/S0266462310000966

THEME SECTION: INFORMATION RETRIEVAL FOR HTA

How much searching is enough? Comprehensive versus optimal retrieval for technology assessments

Andrew Booth
University of Sheffield

Objectives: The aim of this study is to review briefly different methods for determining the optimal retrieval of studies for inclusion in a health technology assessment (HTA) report.
Methods: This study reviews the methodology literature related to specific methods for evaluating yield from literature searching strategies and for deciding whether to continue or desist in the searching process.
Results: Eight different methods were identified. These include using the Capture–recapture technique; obtaining Feedback from the commissioner of the HTA report; seeking the Disconfirming case; undertaking comparison against a known Gold standard; evaluating retrieval of Known items; recognizing the Law of diminishing returns, specifying *a priori* Stopping rules, and identifying a point of Theoretical saturation.
Conclusions: While this study identified a variety of possible methods, there has been very little formal evaluation of the specific strengths and weaknesses of the different techniques. The author proposes an evaluation agenda drawing on an examination of existing data together with exploration of the specific impact of missing relevant studies.

Keywords: Databases, Bibliographic, Technology assessment, Biomedical, Information retrieval, Selection bias, Sensitivity and specificity

The trade-off between rigor and relevance is encountered throughout the health technology assessment (HTA) process (19). Constraints of time and human resource impact upon team decisions as to how thoroughly each stage of the process will be undertaken and how resources will be distributed across all subsequent stages. A key limiting factor relates to the identification of studies because the number of references to be read, abstracts to be sifted and full text articles to be processed holds implications for each subsequent stage of production of the HTA. Indeed, investigations of the resource required for meta-analyses find that, in addition to a fixed amount of time associated with any review, the variable factor is the number of citations to be processed (1). For many years, the acknowledged standard for study identification for HTAs has been the comprehensive literature search, drawing heavily on the classic systematic review model (20). Indeed, the specific concerns above are situated within the wider context of an identified need to speed up the production, and subsequent updating, of systematic reviews (21;22). Is comprehensiveness still an appropriate marker of the quality of an HTA search?

This trade-off between rigor and relevance is perhaps exemplified most clearly in the recent development of methods for conducting rapid reviews (25;26). These typically adopt a

deliberately-circumscribed searching approach in a quest for a faster turnaround time, reduced utilization of resources and, particularly, to satisfy the limited policy windows of opportunity of key stakeholders. Information specialists involved in such activities have been required to meet this challenge by developing, and subsequently justifying, the use of methods that uphold the principles of relevancy and robustness while continuing to recognize the risk averse environment within which technology assessment is carried out. Preliminary mapping of a field for potential review, to identify where efforts can most productively be targeted, is an associated response to such demands (8). Such methods seek to offer a pragmatic alternative to the systematic review model of comprehensiveness.

In reality systematic review processes do actually factor in recognition that it is unlikely to be possible to identify the entire population of studies in a particular topic area. Methods such as Rosenthal's File Drawer calculation (18), which computes how many studies are required to overturn a specific review result, are used to test the robustness of specific reviews to potentially missing studies. More familiar mechanisms such as funnel plots are used to investigate the likely bias resulting from omitting studies that may systematically impact upon review findings (e.g., smaller studies, unpublished studies, studies with non-significant results) (10). However, this willingness to handle uncertainties relating to the population of studies is seldom acknowledged at earlier points of the process, such as at the study identification stage. As a consequence HTA reports that engage with systematic review methodologies embody the implicit value that "big is beautiful"; that rigor in study identification is best attested to by exhaustive searches, sensitive search strategies and long lists of databases searched. The limitations of such an approach are clear; numbers of references retrieved *per se* are not the principal concern of the HTA team but, instead, numbers of studies that, either individually or collectively, could overturn the prevailing assessment result or recommendations. Such studies either represent isolated large studies (which are consequently less likely to be missed), more plentiful studies with either small or no effect (which are less likely to be published anyway), or small studies with large effects (which are less likely to be valid). Furthermore, as search strategies can never be one hundred per cent sensitive, unless one is to sift every single reference on every single database, each search strategy already carries an implicit judgment on what is an acceptable parameter for the sensitivity of searches. Finally, there is no inherent virtue in a lengthy list of databases searched unless each database added to the list holds a high likelihood of yielding unique references (16;24). If this is not the case, then an HTA team is increasing the resources consumed by the review without adding to its value—adding, at best, more duplicates of references already retrieved. For example, MEDLINE consistently delivers an average of 80 percent of included studies for systematic reviews (2;14;24). How much effort should we expend on populating the remaining 20 percent? All the above leads us to conclude that the aspiration of the HTA literature search should not be comprehensiveness but rather the minimization of bias.

Why Perform "Thorough" Searches?

Putting aside for one moment our assertion that we should privilege minimization of bias over comprehensiveness, let us review why we need to perform "thorough" searches (where "thorough" might equally signify comprehensiveness or minimization of bias). There appear to be three main reasons: (i) To maximize the chance of identifying all *relevant* references; (ii) To convince readers that the process underpinning the HTA is *robust*; (iii) To minimize the *risk* of reports being challenged for incompleteness.

We have already seen that the quest to identify *all* relevant references should be viewed in realistic and relative terms because no type of review can conclusively make such a claim. The requirement to be robust would still be fulfilled by constructing evidence-based search procedures as an alternative to the assumptions behind the systematic review method. Finally, existing methods of evidence retrieval already embody strategies to minimize the risk of missing relevant reports and, most importantly, these largely exist independently from strategies for comprehensiveness. For example, through a process of natural selection that we are only just beginning to understand, studies with the greatest potential impact are more likely to be published, more likely to appear in high quality journals, more likely to be covered in multiple databases, and more likely to be cited (27). The chances that such studies will be missed are already relatively slight. Studies that are more likely to be overlooked would thus be best served by casting our information retrieval "lantern" into hitherto dark corners (e.g., searching the gray literature, the Internet etcetera) rather than searching increasing numbers of bibliographic databases (simply because this is where there is light) and retrieving ever-diminishing returns (9).

Methods to Decide When Enough Is Enough

The concept of "information foraging" recognizes that part of any search strategy involves locating, assessing and using a patch (or high concentration) of useful information and, then knowing when to desist searching in that patch (17). How would an HTA team decide when they have conducted enough searching so that the resultant assessment is "fit for purpose" (i.e., in seeking to minimize bias rather than to demonstrate comprehensiveness) and they can desist searching? Basically, this is achieved by recognizing the "cost" (in terms of time spent, resources used, and opportunities lost) of pursuing one particular strategy in preference to others and offsetting this against the value of information to be subsequently gained from that strategy.

Table 1. Methods for Deciding When to "Desist" Searching

Method	Description	Strengths	Weaknesses
Capture-recapture technique	Epidemiological technique for establishing size of unknown population (in this case number of studies) (3;11;12;23); Involves obtaining initial sample by some method of capture, tagging relevant references and then identifying how many tagged records are recaptured in subsequent independent samples	Established method in other contexts	Limited examples within information retrieval
Commissioner feedback	Provide ongoing interim status reports to HTA commissioners; At each reporting point provide resource estimate and ask for decision on next steps	Commissioner responsible for "fitness for purpose"	Intensive on review management resources; may delay process
Disconfirming case	Aggressively seek studies that counter prevailing findings or results (e.g., smaller nonsignificant studies) (15)	Introduces an internal test of "robustness"	Requires iterative approach to searching
Gold standard	At predefined point review your search strategies against report from another HTA agency – what percentage of references are retrieved by your strategies? (6)	Independent validation of strategy	Dependent on existence of, and adequacy of, "gold" standard
Known items	Similar to Gold standard above but uses items retrieved by means of any method as convenience sample to assess adequacy of database searches	Compares retrieval adequacy against all types of sources	Not truly independent sample for comparison
Law of diminishing returns	Ensure sources searched strictly according to likelihood of yield, based on similar reviews. Conclude at predetermined cut-off point (24)	Identification of sources separated from actual searching.	Requires good data and "sameness" of topic types
Stopping rules	Predefine acceptable yield, for example, 1 relevant item per 100 references scanned. Conclude a search route when random sample falls below acceptable yield (5)	Yield is defined *a priori* preventing expediency	Samples must be completely random to protect against bias
Theoretical saturation	For qualitative data: cease searching when new items do not add substantively to understanding of phenomenon (13); Differs from Stopping rules (above) in that the point of termination is discovered empirically and not specified *a priori*	Based on accepted primary research methodology	Difficult to agree when saturation point has been reached

With the assistance of colleagues, I have identified several possible methods for operationalizing this cost/value decision in deciding when to "desist" searching (see Table 1). Readers may be able to suggest additional methods for exploration or even identify methods that they have actually used in their HTA teams.

It should be noted that few methods have been investigated empirically. Chilcott et al. (5) used stopping rules for a methodological review whereby they discontinued searching of a database if the yield of relevant articles fell below 1 percent. Kastner et al. (11;12) used capture-mark-recapture (CMR) modeling to estimate the total number of articles for a review of clinical decision support tools for osteoporosis management using Medline, EMBASE, CINAHL, and EBM reviews. While concluding that the CMR technique could be used to estimate the closeness to capturing the total body of literature on a given topic, they caution that more studies are needed to objectively determine such estimates as a stopping rule strategy. Egger et al. (7) examined the effect of excluding randomized controlled trials that were difficult to locate (e.g., not indexed in MEDLINE) and those of lower quality. They concluded that rather than preventing biases comprehensive literature searches may, in fact, increase them by including studies of lower quality. Royle and Milne (20) examined twenty technology assessment reports and found that searching additional databases beyond the Cochrane Library, MEDLINE, EMBASE, and Science Citation Index (plus BIOSIS for meeting abstracts only), was seldom effective in retrieving additional studies for inclusion. Instead, additional resources would be more productively targeted at searching non-database sources (including submissions from manufacturers, recent meeting abstracts, contact with experts and checking reference lists). A comparative shortage of empirical investigation means that there remains an extensive evaluation agenda.

An Evaluation Agenda

Review teams already collect much data, or could collect extra data with little additional effort, to answer "how much searching is enough?". Where a review team produces PRISMA, formerly QUOROM, flowcharts (http://www.prisma-statement.org/statement.htm) these contain useful data on yield and, importantly on the "conversion rate" of relevant studies from number of database hits for different types of review topic. It should, for example, be possible to build up a comparative picture of typical conversion rates for surgery, diagnostic, public health, and new drug types of topics. Such data would help HTA teams to agree realistic levels of searching for a particular topic and to plan

https://doi.org/10.1017/S0266462310000966 Published online by Cambridge University Press

the time and human resources for particular types of assessment. Furthermore, it is a relatively easy task to document the source of included studies from a completed review; indeed some HTA reports already capture such data. This would again allow us to build up an evidence base on yield from different databases for different topics and, indeed, on sources of duplicate references. For example, if a particular database consistently only yields duplicate references it can either be deleted from the standard list of database sources or costed as an extra for verification purposes only.

When tackling this not inconsiderable evaluation agenda, it will be necessary to factor in not just the consequences of missing a particular reference (so often the preoccupation of information retrieval) but, more importantly the likely consequences (or indeed non-consequences) of missing such a study from an assessment (4). Sensitivity analyses could explore the consequences of including only references from particular database sources for a meta-analysis result. Yes, the results (in terms of number of hits or relevant studies) may be different but do they actually make a difference? Answering such questions will help HTA teams to identify the robustness of their review product against potential bias and thus to identify whether their procedures are "fit for purpose." In this way, the emerging discipline of health technology assessment will be able to make a much-needed evolution from comprehensive toward optimal retrieval.

CONTACT INFORMATION

Andrew Booth, MSc (A.Booth@sheffield.ac.uk), Reader in Evidence Based Information Practice, School of Health and Related Research (ScHARR), University of Sheffield, Regent Court, 30 Regent Street, Sheffield, South Yorkshire S1 4DA, United Kingdom

CONFLICT OF INTEREST

The author reports having no potential conflicts of interest.

REFERENCES

1. Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. *JAMA*. 1999;282: 634-635.
2. Bayliss SE, Dretzke J. Health technology assessment in social care: A case study of randomized controlled trial retrieval. *Int J Technol Assess Health Care.* 2006;22:39-46.
3. Bennett DA, Latham NK, Stretton C, et al. Capture-recapture is a potentially useful method for assessing publication bias. *J Clin Epidemiol*. 2004;57:349-357.
4. Booth A. The number needed to retrieve: A practically useful measure of information retrieval? *Health Info Libr J*. 2006;23:229-232.
5. Chilcott J, Brennan A, Booth A, et al. The role of modelling in prioritising and planning clinical trials. *Health Technol Assess.* 2003;7:iii, 1-125.
6. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ.* 1994;309:1286-1291.
7. Egger M, Juni P, Bartlett C, et al. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol Assess*. 2003;7:1-76.
8. Grant MJ, Booth A. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Info Libr J*. 2009;26:91-108.
9. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *BMJ.* 2005;331:1064-1065.
10. Ioannidis J. Why most published research findings are false. *PLoS Med.* 2005;2:e124.
11. Kastner M, Straus S, Goldsmith CH. Estimating the Horizon of articles to decide when to stop searching in systematic reviews: An example using a systematic review of RCTs evaluating osteoporosis clinical decision support tools. *AMIA Annu Symp Proc*. 2007;389-393.
12. Kastner M, Straus S, McKibbon K, et al. The capture-mark-recapture technique can be used as a stopping rule when searching in systematic reviews. *J Clin Epidemiol*. 2008;62: 149-157.
13. Khan KS, Coomarasamy A. A hierarchy of effective teaching and learning to acquire competence in evidenced-based medicine. *BMC Med Educ*. 2006;6:59.
14. Löhönen J, Isohanni M, Nieminen P, et al. Coverage of the bibliographic databases in mental health research. *Nord J Psychiatry.* 2010;64:181-188.
15. Noyes J, Popay J, Pearson A, et al. Qualitative research and Cochrane reviews. In: Higgins JPT, Green S, eds. *Cochrane handbook for systematic reviews of interventions.* Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. www.cochrane-handbook.org.
16. Ogilvie D, Hamilton V, Egan M, et al. Systematic reviews of health effects of social interventions: 1. Finding the evidence: How far should you go? *J Epidemiol Community Health*. 2005;59:804-808.
17. Pirolli P, Card S. Information foraging in information access environments. In: Proceedings of the Human Factors in Computing Systems, CHI '95. Association for Computing Machinery, 1995.
18. Rosenthal R. The file drawer problem and tolerance for null results. *Psychol Bull.* 1979;86:638-641.
19. Rotstein D, Laupacis A. Differences between systematic reviews and health technology assessments: A trade-off between the ideals of scientific rigor and the realities of policy making. *Int J Technol Assess Health Care.* 2004;20: 177-183.
20. Royle P, Milne R. Literature searching for randomized controlled trials used in Cochrane reviews: Rapid versus exhaustive searches. *Int J Technol Assess Health Care.* 2003;19: 591-603.
21. Sampson M, Shojania KG, Garritty C, et al. Systematic reviews can be produced and published faster. *J Clin Epidemiol*. 2008;61:531-536.

22. Shojania KG, Sampson M, Ansari MT, et al. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med*. 2007;147:224-233.
23. Spoor P, Airey M, Bennett C, et al. Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. *BMJ.* 1996;313:342-343.
24. Stevinson C, Lawlor D. Searching multiple databases for systematic reviews: Added value or diminishing returns? *Complement Ther Med.* 2004;12:228-232.
25. Watt A, Cameron A, Sturm L, et al. Rapid reviews versus full systematic reviews: An inventory of current methods and practice in health technology assessment. *Int J Technol Assess Health Care.* 2008;24:133-139.
26. Watt A, Cameron A, Sturm L, et al. Rapid versus full systematic reviews: Validity in clinical practice? *ANZ J Surg*. 2008;78:1037-1040.
27. Weale AR, Lear PA. Randomised controlled trials and quality of journals. *Lancet.* 2003;361:1749-1750.