A COMPARISON OF TWO SEARCH METHODS FOR DETERMINING THE SCOPE OF SYSTEMATIC REVIEWS AND HEALTH TECHNOLOGY ASSESSMENTS

Louise Forsetlund
Norwegian Knowledge Centre for the Health Services
louise.forsetlund@kunnskapssenteret.no

Ingvild Kirkehei, Ingrid Harboe, Jan Odgaard-Jensen
Norwegian Knowledge Centre for the Health Services

Objectives: This study aims to compare two different search methods for determining the scope of a requested systematic review or health technology assessment. The first method (called the Direct Search Method) included performing direct searches in the Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Reviews of Effects (DARE) and the Health Technology Assessments (HTA). Using the comparison method (called the NHS Search Engine) we performed searches by means of the search engine of the British National Health Service, NHS Evidence.
Methods: We used an adapted cross-over design with a random allocation of fifty-five requests for systematic reviews. The main analyses were based on repeated measurements adjusted for the order in which the searches were conducted.
Results: The Direct Search Method generated on average fewer hits (48 percent [95 percent confidence interval {CI} 6 percent to 72 percent], had a higher precision (0.22 [95 percent CI, 0.13 to 0.30]) and more unique hits than when searching by means of the NHS Search Engine (50 percent [95 percent CI, 7 percent to 110 percent]). On the other hand, the Direct Search Method took longer (14.58 minutes [95 percent CI, 7.20 to 21.97]) and was perceived as somewhat less user-friendly than the NHS Search Engine (−0.60 [95 percent CI, −1.11 to −0.09]).
Conclusions: Although the Direct Search Method had some drawbacks such as being more time-consuming and less user-friendly, it generated more unique hits than the NHS Search Engine, retrieved on average fewer references and fewer irrelevant results.

Keywords: Bibliographic databases, Search engine, Cross-over studies, Evaluation studies

One of the core activities of the Norwegian Knowledge Centre for the Health Services (NOKC) is the compilation and summary of research literature into systematic reviews and health technology assessments. In most instances these reports are commissioned by health authorities and policy makers at varying levels of the Norwegian healthcare system and cover a wide thematic range of topics. Upon receipt of the requests, each research question is classified as an effect of an intervention, a comparison of two or more diagnostic tests or a cost-effectiveness evaluation. Other types of questions are generally deemed outside the mandate of NOKC.

All accepted requests are discussed at an annual workshop. At the workshop, each research question is summarized into a table giving the background of the research question, a precisely-worded version of the research question (3), a summary of the outcomes of the preliminary search and an estimate of the resources necessary to perform the review. The preliminary searches are conducted by NOKC's researchers in preselected information sources for all research questions identified in the requests to determine the scope of each request. The preliminary searches may reveal existing systematic reviews covering the same research question or multiple reviews of interventions for the same condition. If a recent, high quality systematic review is identified, it will be forwarded to the commissioner to determine if it meets the commissioner's needs. If multiple reviews are found, it may be appropriate to perform an overview of the identified systematic reviews. If no reviews are identified we will undertake a review of primary studies. The result of the preliminary search is therefore important for the further handling of the request (1;2).

Our standard procedure has been to directly search the Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Reviews of Effects (DARE), the Health Technology Assessments Database (HTA), the Danish HTA (DACE-HTA), the Finnish HTA (FINOHTA), the Swedish Council on HTA (SBU) and the NOKC websites. Both DARE and HTA are accessed through the Centre for Reviews and Dissemination (CRD). Since our last workshop, the British National Health Service (NHS) has launched a new search engine (NHS Evidence http://www.evidence.nhs.uk/default.aspx). This search engine covers the CDSR and the CRD databases as well as several organizations' websites, for example the WHO Health Evidence Network (HEN), Agency for Health Research and Quality (AHRQ), and the Scottish Intercollegiate Guidelines Network (SIGN). However, it does not cover the Scandinavian HTA websites mentioned previously. Consequently, the content of these two information sources are somewhat different with

the NHS search engine covering more resources but also with a large overlap. In our experience, the main sources resulting in relevant hits are the CDSR and the DARE databases. We anticipated that the NHS Search Engine would return more results than the Direct Search Method, partly because of the additional sources sought and partly because search engines could have other underlying search functionalities than databases. We were unsure how the NHS Search Engine used the functionality of the subordinate databases, CDSR and CRD. For example, it could be more difficult to reduce the number of search results to a manageable level when using the NHS Search Method than when searching the databases directly.

The trial had a pragmatic approach (4). The intention of a pragmatic study is to keep or recreate circumstances that are as close to normal practice as possible. The data sample ought to reflect the heterogeneity of the population in common practice. Usually, the intervention is complex and loosely defined and neither those delivering the intervention nor those participating are blinded with respect to who receives the intervention. The results of this study were intended to inform the decision of which search method to use in future workshops in a typical health technology assessment environment.

Our objective was to compare the results generated by direct searches of the preselected sources with search results generated by a search engine when it retrieves data essentially from the same sources. Our aim was not to compare the quality of the information sources as such, but what practical significance they would have to us. We named the method representing our usual practice the Direct Search Method and the new search method to be tested was named the NHS Search Engine. First, we wanted to compare the number of hits, precision of the search, number of unique hits, and the search time. Second, we wanted to compare the researchers' satisfaction with the two methods, their perception of relevance of the two search sets and whether perceived relevance depended on several factors (such as subject area, researchers' self-reported search experience level or search techniques such as MeSH and the Boolean operators AND/OR). A literature search in MEDLINE and an inquiry to the staff at NHS Evidence did not identify any other previously published studies directly relevant to this objective.

METHODS

Design
We used an adapted cross-over design for the evaluation. The review requests were randomly allocated to search first either the NHS Search Engine or the Direct Search Method. The requests were consecutively numbered. Using a computer program for extracting random numbers, an independent statistician assigned the NHS Search Engine as the initial search method to the identified requests. All remaining requests were assigned the Direct Search Method as the initial search method. The statistician sent the requests to another person not involved

in the study to mark them with their initial search method. They were then returned to the statistician who controlled that the method sequence was correct before forwarding them to the research team.

Sample
Between January 1 and November 25, 2009, NOKC received seventy-five requests for systematic reviews. All of these requests were eligible for inclusion in our study. Requests were included if the research question was focused on an effect of interventions, a comparison of diagnostic tests or a health economic evaluation.

Intervention
The intervention was defined as "searching using the NHS Search Engine" compared with our conventional Direct Search Method. As per our standard procedure, a workshop was convened to evaluate the systematic review requests, perform a preliminary search to determine the scope and summarize the results. However, at this workshop the researchers were asked to duplicate their searches following the sequence given on the data capture form. All NOKC researchers who perform systematic reviews were invited to the workshop. The researchers have varied academic backgrounds including medicine, natural or social sciences and health economics. Due to varying experience in searching, a brief introduction to search functions and techniques is always the initial part of the workshop, this year in both search methods.

The individual researcher independently decided how best to formulate the search queries for the reviews assigned to him/her. However, some attempts at standardization were attempted. The researchers were told that, because we wanted to compare the two methods, it was important to apply the same concepts in the search algorithm. An example was provided in the data collection form. We also specified in the form that they should check for hits in the NHS Search Engine by clicking in the left-hand column and selecting "Systematic Reviews" and "Health Technology Assessments." The researchers were allowed to use features and tools which differ between the two methods, such as the controlled vocabulary (Medical Subject Headings) in the CDSR and CRD. The researchers were instructed to search for systematic reviews, studies of diagnostic methods, and overviews of cost-efficacy or cost-benefit studies, depending on the research question. In accordance with the standard workshop procedure the authors also participated in the workshop, two as librarians available for search guidance (I.K., I.H.) and one as a researcher performing preliminary searches (L.F.).

Data Collection and Outcomes
The standard workshop table was updated to include the new data collection form and designed so that it was easily understood which search method should be used first

(Appendix 1, which can be viewed online at www.journals. cambridge.org/thc2012009). The new form was piloted on four researchers (three of whom had previously participated in the workshops). Two requests for systematic reviews from the previous year were used for the pilot. The initial search method was determined by tossing a coin. The data collection form was tested for comprehensibility, efficiency, and practicality. The form was then amended accordingly. Additionally, the pilot researchers were asked for their opinion regarding the practical feasibility of accomplishing the project in a workshop. All of the pilot researchers considered it possible to perform the evaluation as part of the workshop.

The data were captured by the study researchers in parallel with conducting the searches. For each search they copied their search algorithms as well as all references assessed as potentially relevant for each search method into the data collection form. They also reported the following outcomes for each search method for all review requests: The research question(s) of the requested systematic review; Total number of hits; Number of relevant references; Perceived degree of relevance (scale from 1 to 7, i.e., "a little relevant" to "very relevant"), independent of how many were found; User satisfaction with user interface and ease of use (scale from 1 to 7, i.e., "very dissatisfied" to "very satisfied"); Time: How long did it take to come to a conclusion on relevance; and Previous search experience (scale from 1 to 7, i.e., "completely inexperienced" to "very experienced").

After the workshop, we analyzed the search histories with regard to whether the participants had used the Boolean operators AND and OR, if they had truncated any of the terms or if they had used MeSH in the CDSR or the CRD Databases. We also analyzed whether there were any syntactic errors in the search algorithm and whether they had performed the same type of search for both methods. These outcomes were measured dichotomously as either Yes or No. The number of unique hits for each search method was counted. Finally, we calculated the precision of the search as the ratio between the number of relevant documents and the number of total hits.

Data Analyses

Our main analyses were based on repeated measurements. The differences between the two search methods in the number of hits, the number of relevant hits and the number of unique hits was analyzed using repeated measurements based on a Poisson distribution and log as the link function. The corresponding analyses of the remaining outcomes were performed using repeated measurements, based on the normal distribution and identity as the link function. All analyses performed using repeated measurements were adjusted for the order in which the searches were conducted. The results of Poisson repeated measurements analyses were expressed as ratios: hits from the search by Direct Search Method/hits from the NHS Search Engine, that is, a relative difference. The results from the normal repeated measurements analyses were expressed as absolute differences (Direct Search Method—NHS Search Engine). In the sensitivity analyses, the differences between the two search methods were tested using paired t-tests.

RESULTS

Sample

Before the allocation of the search method sequence assignment, seventeen requests for systematic reviews were excluded from the seventy-five eligible because they did not meet the inclusion criteria. Additionally, three requests were excluded as duplicates (Figure 1). Therefore, fifty-five requests were assessed at the workshop. A flow chart of the procedures for processing requests for systematic reviews and health technology assessments is provided in Figure 2. During the workshop, two requests were merged into one due to subject matter. Another request proved to be an enquiry for methodological cooperation with NOKC rather than a request for a systematic review and was, therefore, excluded. In the end, fifty-three requests were analyzed for this study.

OUTCOMES

Descriptions

The requests for systematic reviews covered six thematic areas: Mental health (eight requests), drugs (seven requests), primary care (ten requests), specialist services (thirteen requests), healthcare organization (eleven requests) and health economics (four requests). Thirty-eight researchers evaluated the requests. As shown in Table 1, there was more or less the same use of Boolean operators and truncations and the same amount of syntax errors for both search methods. The search algorithms used were mostly of a simple character, reflecting the participants' self-reported level of experience in searching as 4.6, measured on a scale from 1 to 7. We assessed the search history for both methods to be identical for thirty-seven of the fifty-three review requests (70 percent).

Unadjusted average values for the different variables that were registered during the workshop are presented in Supplementary Table 1 (www.journals.cambridge.org/thc2012010).

Table 1. Analysis of Search Strategy

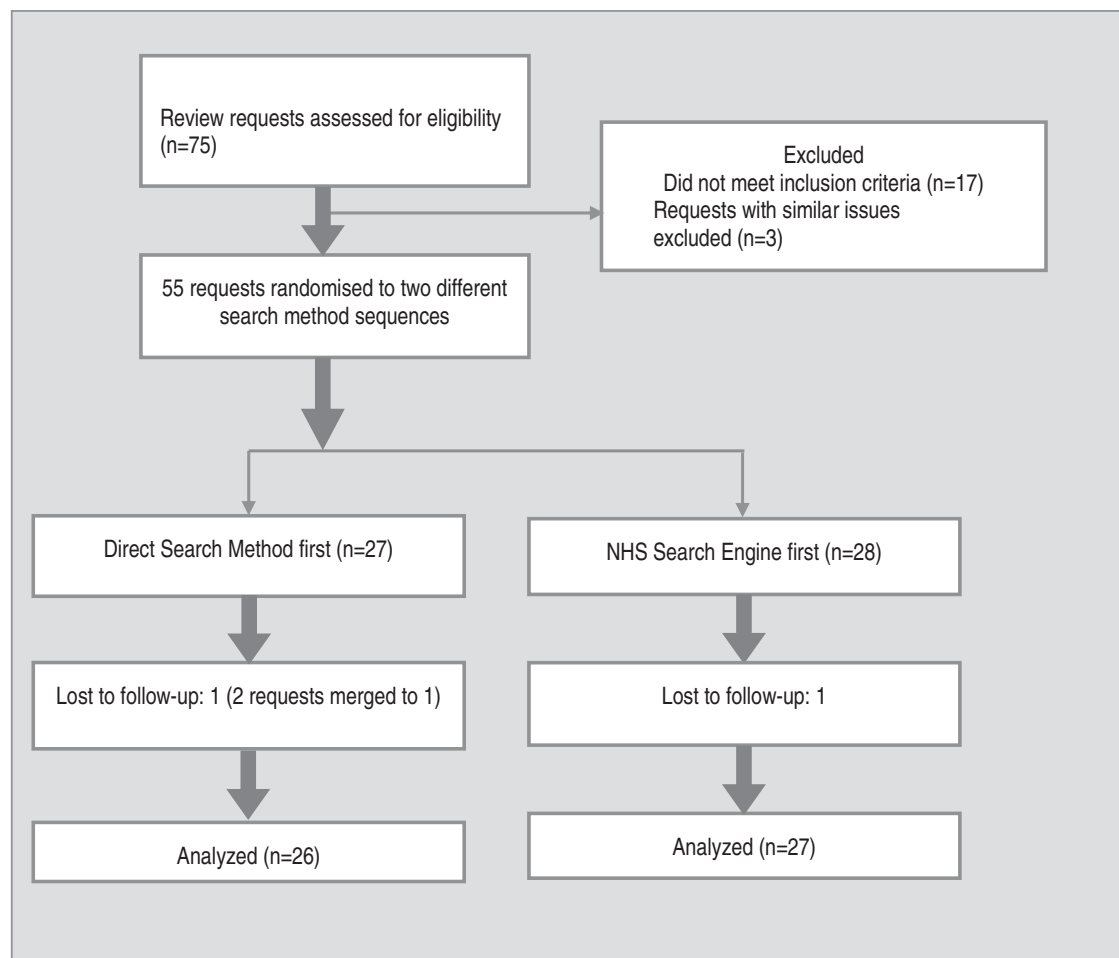| Use of functions/tools in the search strategy	NHS Search Engine $N = 53$ % (n)	Direct Search Method $N = 53$ % (n)
MESH	Not relevant	4 (2)
AND	70 (37)	74 (39)
OR	42 (22)	40 (21)
Truncation	38 (20)	36 (19)
Syntax errors	9 (5)	11 (6)

Figure 1. Flow of requests for systematic reviews or health technology assessments. NHS, National Health Service.

For example, the mean length of time for conducting a pilot search was 46 minutes using the Direct Search Method and 30 minutes using the NHS Search Engine. The precision of search results was rather low for the NHS Search Engine (14 percent) and moderate for the Direct Search Method (35 percent).

Primary Analyses: Differences Between Search Methods

The Direct Search Method generated on average fewer hits (48 percent (95 percent CI, 6 percent to 72 percent)), had a higher precision (0.22 [95 percent CI, 0.13 to 0.30]) and more unique hits (50 percent [95 percent CI, 7 percent to 110 percent]) than when searching by means of the NHS Search Engine (Table 2). On the other hand, the Direct Search Method took longer (14.58 min [95 percent CI, 7.20 to 21.97]) and was perceived as somewhat less user-friendly (−0.60 [95 percent CI, −1.11 to −0.09]). The differences in precision, time to reach a conclusion and perceived user-friendliness were statistically significant also in the sensitivity tests (data not shown).

There was no statistically significant difference between the number of references assessed as relevant between the two search methods. Excluding an outlier with 411 relevant hits did not change this result. There was also no statistically significant

difference in perceived relevance of the identified references, irrespective of the number, for the two methods (−0.30 [95 percent CI, −0.66 to 0.07]).

Secondary Analyses

Precision and perceived degree of relevance did not vary statistically significantly with the subject field, level of experience, the use of MeSH, or with the use of Boolean operators (Supplementary Table 2 [www.journals.cambridge.org/thc2012010]). There was also no statistically significant variation in the number of unique hits with the use of MeSH terms in the direct searches. For the other potentially predictive variables of degree of uniqueness (subject fields, experience level, and use of Boolean operators), the tests for dependency could not be performed as specified due to non-convergence in the estimation of the parameters in the model.

DISCUSSION

Summary of Main Findings

The NHS Search Engine generated more hits, had a lower precision in the search results and fewer unique hits, but by using

Table 2. Primary Analyses: Repeated Measurements - Adjusted for Search Sequence

Variable	N	Ratio Direct Search Method/ NHS Search Engine	95% CI	P-value
Number of hits	106	0.52	0.28 to 0.94	0.030
Number of relevant hits	100	2.35	0.83 to 6.63	0.104
Number of unique hits	96	1.50	1.07 to 2.10	0.021
Number of relevant hits with outlier excluded (outlier with 411 hits from the Direct search method)	99	1.20	0.97 to 1.50	0.094

Variable	N	Difference Direct − NHS	95% CI	P-value
Precision	100	0.22	0.13 to 0.30	<0.000
Length of time to reach a conclusion (minutes)	100	14.58	7.20 to 21.97	0.000
Perceived degree of user-friendliness	106	−0.60	−1.11 to −0.09	0.023
Perceived degree of relevance	92	−0.30	−0.66 to 0.07	0.108
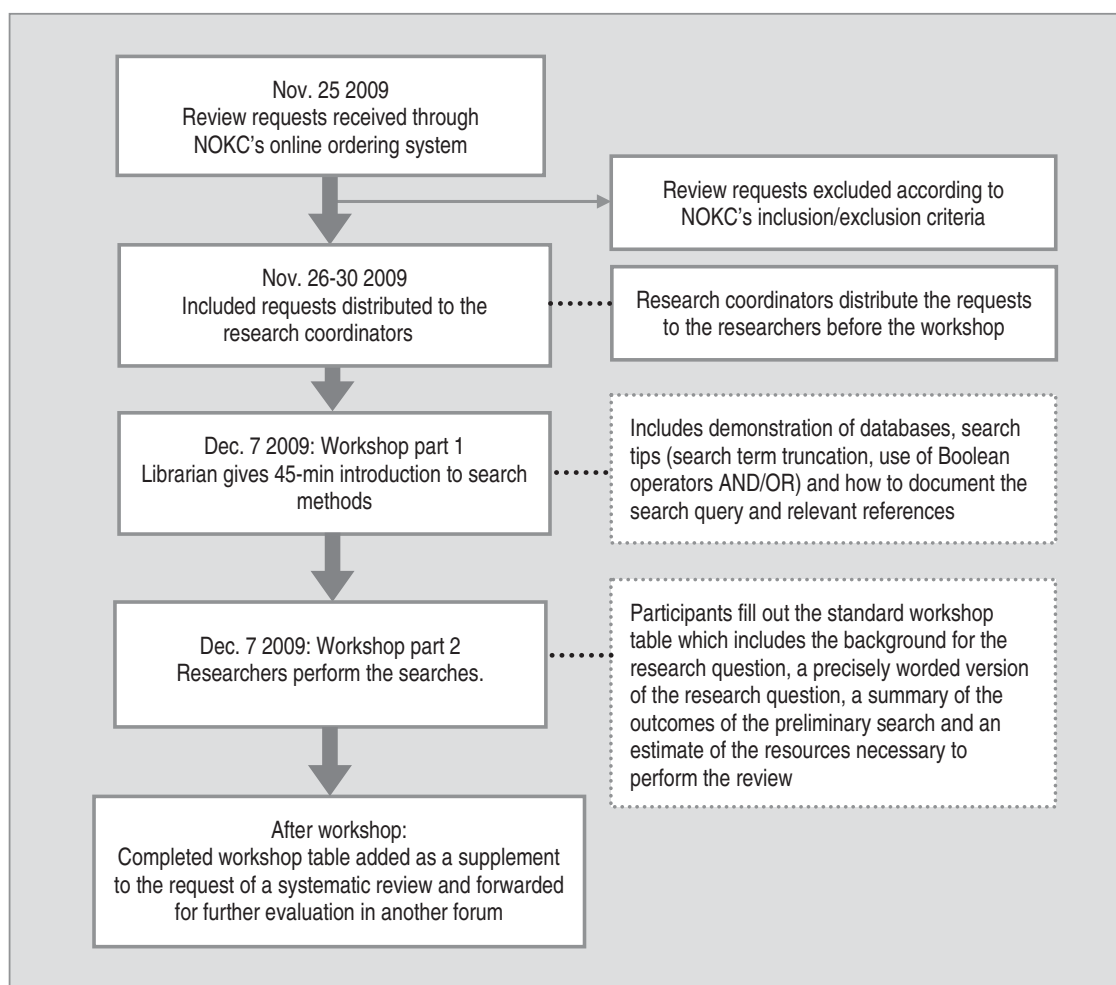
Figure 2. Flow chart of the procedures for processing requests for systematic reviews and health technology assessments. NOKC, Norwegian Knowledge Centre for the Health Services.

the NHS Search Engine it was less time-consuming to generate a search result and identify some relevant references. However, too many references may render it impossible to screen all for relevance, and this may explain that fewer number of unique hits were identified when using the NHS Search Engine. The NHS Search Engine scored higher on user satisfaction while the difference in researchers' perceived relevance of identified references was not statistically significant. Due to a lack of data, we could not draw any conclusions regarding the potential influence of the predictive variables (type of subject, the searcher's level of experience or the search mode) on precision, the number of unique hits or perceived relevance.

Strengths and Limitations

We consider the main strength of this study to be the design we used for comparing the two search methods, that is, the random allocation of the review requests to a search method sequence. The random selection ought to have eliminated potential bias related to the order in which the searches were performed. Also, the pragmatic attitude and approach were planned beforehand. The population was heterogeneous in terms of the participants as well as in terms of the theme of the requests. However, one should remember that each participant and request was its own control, as we have used an adaptive cross-over design for this trial. The chosen method of analysis takes the paired data into account when calculating the estimates of effect, and, furthermore, adjusts for the sequence in which the searches have been performed.

Our findings may not be generalizable to other settings. Others working in similar settings may require other sources to determine the scope of their reviews. In addition, user interfaces and search functions in search engines or databases often change, making replication of this study difficult. Therefore, we believe that the main lesson from our study is how one may methodologically compare different search methods in a systematic manner as part of regular practice.

CONCLUSION

Although the Direct Search Method had some drawbacks, such as being more time-consuming and less user-friendly, it gener-ated more unique hits than the NHS Search Engine, retrieved on average fewer references and gave fewer irrelevant results. Therefore, NOKC decided to continue to use the Direct Search Method for our preliminary searches to define the review's scope until further notice.

SUPPLEMENTARY MATERIAL

Appendix
www.journals.cambridge.org/thc2012009
Supplementary Table 1
Supplementary Table 2
www.journals.cambridge.org/thc2012010

CONFLICT OF INTEREST

All authors report having no potential conflicts of interest.

CONTACT INFORMATION

Louise Forsetlund, PhD, Senior Researcher, Ingvild Kirkehei, Master of Library and Information Science (MLIS), Research Librarian, Ingrid Harboe, Bachelor of Library and Information Science, Research Librarian, Jan Odgaard-Jensen, Statistician, Researcher, Norwegian Knowledge Centre for the Health Services, Oslo, Norway

REFERENCES

1. Bettany-Saltikov J. Learning how to undertake a systematic review: Part 1. *Nurs Stand*. 2010;24:47-56.
2. Higgings JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.0.2. [updated September 2009]. The Cochrane Collaboration; 2009:91. www.cochrane-handbook.org (accessed July 4, 2011).
3. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: A key to evidence-based decisions. *ACP J Club*. 1995;123:A12-A13.
4. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, Oxman AD, Moher D, CONSORT group, Pragmatic Trials in Healthcare (Practihc) group. Improving the reporting of pragmatic trials: An extension of the CONSORT statement. *BMJ*. 2008;337:a2390 doi:10.1136/bmj.a2390. http://www.bmj.com/content/337/bmj.a2390.full.pdf (accessed March 9, 2011).