USE OF MCMASTER PLUS FOR UPDATING SYSTEMATIC REVIEWS

ASSESSING THE USE OF MCMASTER PREMIUM LITERATURE SERVICE IN THE EFFICIENT UPDATING OF SYSTEMATIC REVIEWS

BY

ABHA H. ATHALE

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the

Requirements for the Degree Master of Science

Department of Clinical Epidemiology and Biostatistics

Health Research Methodology Program

McMaster University MASTER OF SCIENCE (2015) Hamilton, Ontario

(DEPARTMENT OF CLINICAL EPIDEMIOLOGY AND BIOSTATISTICS)

TITLE: Assessing the use of McMaster Premium LiteratUre Service in the efficient updating of Systematic reviews

AUTHOR: Abha Athale B.A.Sc (McMaster University)
SUPERVISOR: Dr. Alfonso Iorio

Pages: xii, 50

Lay Abstract:

Systematic reviews (SR) summarize all available evidence about a treatment question, and are often used to guide treatment decisions. To remain current, new information should be promptly added to SRs. Adding new information is time-consuming and can prevent timely updating of SRs. The purpose of this thesis was to see if using McMaster PLUS, a database including only pre-appraised, high-quality studies, and Clinical Query filters- specialized search filters- could increase the efficiency of SR updating. We identified 92 SRs that were updated, and had a change in the conclusions. We found the studies that were newly added, and searched for them in the PLUS database. We found that 21.2% of these trials were found in PLUS. The CQ filters found 96.2% of the studies not found in the PLUS database. This project shows that using the PLUS database and CQ filters may help increase the updating of systematic reviews.

Abstract:

Background

Systematic reviews (SRs) of treatment effect are evidence syntheses that inform clinical practice decisions and healthcare policy. To maintain validity, SRs should be regularly updated to include novel research. In reality, updating practices are irregular, with resource and time constraints often cited as major barriers. The McMaster Premium LiteratUre Service (PLUS) is a database of high quality, pre-appraised evidence, which may be of potential help in efficient updating of SRs.

Objective

To determine the utility of McMaster PLUS to increase the efficiency of systematic review updating

Methods

Updated Cochrane reviews published from January 2012-January 2013 with changed conclusions were identified. Using the PubMed IDs of references in the updated review, which were not present in the previous version, we looked for the presence of these references in the PLUS database. Further, using Clinical Query (CQ) filters on PubMed, we identified the references not found in PLUS.

Results

Eight hundred fifty-four unique trials, reported in over 1498 references were used to drive a change in conclusion in the 92 included reviews. Of the 854 unique trials, 180 (21.1%) were found in the PLUS database. All of the newly added trials were in PLUS for 8 of 92 reviews, and none of the newly added trials were in PLUS for 26 of 92 reviews. Of the 834 references not found in PLUS, there were 728 unique PubMed IDs. Using the sensitive CQ filter, 701 (96.2%) of these trials were identified.

Conclusion

PLUS included 21.1% of trials used to drive a change in conclusion in 92 Cochrane reviews. Furthermore, the CQ filters performed admirably in the retrieval of articles not found in PLUS. These alternate search methods should be considered when updating SRs to help increase the efficiency of the update process. These methods should be further tested prospectively.

Acknowledgements

I would first and foremost like to acknowledge and sincerely thank my supervisor, Dr Alfonso Iorio for his mentorship, guidance and opportunities throughout my master's degree and beyond. His support and direction has given me the confidence to nurture my scientific curiosity and has allowed me to build foundation of knowledge and experience for my future research endeavours.

To my committee members, I would like to thank you for your guidance, and patience and feedback.

To Linda Sheridan, I would like to extend my thanks for all of your support in scheduling and facilitating various meetings.

To Nancy Wilczynski, for your enthusiasm, insight and advice in designing the Clinical Queries section of this project.

To HiRU technology group for facilitating the technical aspects of this project, and for your patience with my many questions.

To the faculty and students in the Health Research Methodology Program for teaching me new skills, and challenging me to become a better researcher.

Most importantly I would like to thank my parents, Uma and Harish, for their unwavering support and faith in me. And for always encouraging me to reach for the stars.

Table of contents

List of figures

List of tables

List of abbreviations

BMJ	British Journal of Medicine
Cc	Change in conclusion
CCT	Controlled clinical trial
CI	Confidence interval
CQ	Clinical Queries
CSDR	Cochrane Database of Systematic Reviews
DOI	Digital object identifier
EBM	Evidence based medicine
JAMA	Journal of the American Medical Association
HKR	Health Knowledge Refinery
NNR	Number needed to read
Ns	New search
PLUS	McMaster Premium LiteratUre Service
RCT	Randomized control trial
SR	Systematic review

Declaration of Academic Achievement

I, Abha Athale, am the primary author of this thesis and the material included within. I substantially contributed to the conception, design, development and writing of the presented work.

Dr. Iorio supervised me in the completion of this work.

SECTION 1: BACKGROUND

1.1. IMPORTANCE OF UP-TO-DATE SYSTEMATIC REVIEWS

Medical research is an important avenue to inform clinical decision-making. Aside from technological advancements, modern day medicine is largely a product of clinical research. For the most part, the research process begins with work in the basic sciences, progresses to translational research and safety and efficacy studies, and culminates with long-term follow-up studies. Research has allowed us not only to better understand disease processes and incidence, but has also allowed us to answer many clinically-relevant questions, such as identifying curative substances, pinpointing their dosages and potential toxicities. With the increasing prevalence of evidence-based medicine (EBM), which is the practice of using high-quality evidence from research studies to inform medical decisions, the practical utility of clinical research studies becomes increasingly apparent.

However, the high publication volume of primary medical literature can make it difficult for an individual to comprehensively access important information that may affect clinical decision-making [1]. A 2005 study by Druss and Marcus, which investigated the growth of medical literature showed that a mean of 398,778 articles were published each year in MEDLINE between 1994 to 2001, which was a dramatic increase from the 272,344 articles published yearly between 1978-1985. This same study showed, that the percent of randomized control trials published between these two time frames had tripled (1.9% in the first time period to 6.2% in the second)[2]. These numbers, while large, still do not include conference abstracts, posters, unpublished works, which a reader may need to take into account to make a clinical-decision.

A systematic review (SR) of the literature is a collation of primary research information about a specific research question. SRs often include qualitative and quantitative syntheses, such as meta-analyses, to effectively summarize all evidence on a topic. Because of their inclusiveness and rigorous methodology, SRs are considered the highest form of evidence, and are used to make healthcare decisions, create clinical practice guidelines, and inform health care policy[1].

Because of their use as a foundation for clinical decision-making, it is important that the information presented and conclusions drawn in a systematic review reflect the most current evidence. The repository of health care research is very dynamic, and new evidence is continually produced. The systematic reviews, which are based on health care research, are also dynamic and need to frequently incorporate new evidence to reflect the evolution of scientific knowledge. If new evidence is not incorporated in a systematic review, then the validity of the review may be compromised. Past studies have shown that, depending on how quickly evidence becomes available in a particular field, systematic reviews can become old or out of date within a couple of years of publication. In fact, one survival analysis, which looked at the time between publication of a review and the occurrence of an update signal, showed that 23% of reviews in their cohort had an indication for update within 2 years of publication, and 15% had an indication within 1 year of publication[3]. Moreover, past studies had documented that reviews may become out of date even during the time of completion and publication[4,5] To ensure that their results reflect current evidence, systematic reviews should be regularly updated.

1.2. Current Updating Processes and the Cochrane Collaboration

For the most part, the frequency and process of updating systematic reviews varies a great deal. A recent review by Moher et al identified four strategies, one technique and two statistical approaches that have been used to update systematic reviews. However, these methods have not been empirically tested or compared to one another to assess efficiency [1]. The Cochrane Collaboration is an international network, which has developed a wide collection of methodologically sound systematic reviews

with the intention of supporting clinical-based decisions. Keeping systematic reviews current is one of the Collaboration's ten key principles[6]. In line with this, the Collaboration recommends authors to update systematic reviews every two years or justify why the update should take place at a later time point[6].

A recent study identified that over 85% of Cochrane reviews failed to meet the target update time within two years, with a median time to update of six years [7]. As discussed above, with some reviews having a signal for an update within two years of publication[3], having a median time of update of six years may mean that certain reviews remain out of date for many years, until new information is incorporated. However, the same study showed that the majority of reviews only had a signal for the need for update about 5.5 years following initial publication[3]. Hence, the implications of having a median survival time of six years may be more impactful for some reviews in comparison to others. Expanding further, we challenge the paradigm of requiring an update in an allocated period of time. Given that there will be a subset of reviews, for which new research is not frequently conducted and published, having an allocated time (ie two years) within which an update needs to occur would likely not be an optimal use of human and financial resources. Instead perhaps, it may be useful to implement a centralized surveillance system, which could identify and dispatch evidence as required to authors of relevant systematic reviews.

1.3. BARRIERS TO UPDATING REVIEWS

Past research has identified different barriers to timely and efficient updating of systematic reviews. The predominant issues precluding efficient updating are time and resource constraints [8]. Arguably, an update of a systematic review is akin to completing a new review, as updating generally includes a new literature search, assimilation of any new information in both qualitative descriptions and quantitative analysis, and updating conclusions as necessary. This process, while necessary to complete a comprehensive and methodologically sound update, is very time and resource intensive. Without specific resources dedicated to this process, completing an update becomes unfeasible for many review authors [8]. As well, the limited scope for academic credit stemming from an

updated review in comparison to conducting a new review was cited as a barrier to efficient updating[8]. Further, the lack of a universally accepted definition of what constitutes an update is another barrier to consistent SR updating [9].

1.4. McMaster Premium LiteratUre Service

The McMaster Premium LiteratUre Service (PLUS) is a database created by the McMaster Health Knowledge Refinery (HKR). PLUS is an online database, with a search engine tool, which contains high-quality evidence, including studies and reviews, which have been pre-selected and appraised for sound methodology, relevance and 'newsworthiness'. Through a critical appraisal process, which is unique to PLUS, studies that are poorly conducted are filtered out, thus resulting in a subset of high-quality studies. Details about the quality appraisal process can be found at the following website: http://hiru.mcmaster.ca/hiru/HIRU_McMaster_PLUS_Projects.aspx.

The studies selected for inclusion in PLUS on the basis of quality are subsequently scored by practicing health care providers, in the specific discipline(s), for their clinical relevance (readiness for clinical application) and newsworthiness (i.e. likelihood for the evidence provided by the study to change practice) on a scale ranging from one (low end) to seven (high end). Only studies scoring at least a three on both scales are entered in PLUS. This method makes PLUS a repository of high-quality and highly relevant clinical knowledge[10]. However, it should be noted that PLUS only reviews about 115 journals and some other sources of reviews (like the Cochrane Library and HTA sources), rather than all possible journals. Notably, these journals were selected from over 800 journals, based on high yield of articles that meet inclusion criteria. Included journals are annually appraised, and journals with low yield (fewer than one or two articles per year) are removed from the list. New journals are added as the need to expand a research field arises. All new journals must conform to inclusion criteria, as found on the website above[10].

1.5. Clinical Query Filters

Clinical Query (CQ) filters are empirically derived, highly specialized search filters that are designed to identify and retrieve from broad databases articles belonging to

specific study categories. Currently, there are filters created and calibrated to identify articles reporting results for therapy, economics and diagnosis- to name a few. The filters are available to search in MEDLINE, Embase and PsychINFO.

Cochrane reviews typically include only randomized controlled trials (RCTs), controlled clinical trials (CCTs) or quasi-randomized trials. For this reason, we decided to explore the performance of the therapy filter(s), which selectively retrieved randomized clinical trials. There are three types of therapy filters:

1. Sensitive - maximizing filters comprehensiveness, reducing the proportion of relevant articles that are missed.
2. Specific - maximizing filters selectivity, reducing the proportion of irrelevant articles found
3. Combined filters that balance and maximize both sensitivity and specificity.

Historically these filters have performed very well [11]:

Table 1: Past performance of CQ therapy filters

Filter	Sensitivity	Specificity
Max Sensitivity	99.2	70.1
Max Specificity	94.0	97.5
Combined	96.5	95.1

Given the past performance of these filters, we hypothesized that the use of the clinical queries filters after PLUS would allow us to capture the articles that were not captured in PLUS.

As well, past research has shown that the use of these filters can reduce the number needed to read by 10 times- meaning that an individual would have to read 10 times fewer articles to identify one relevant article[12]. The CQ filters could be used in the search stage to reduce the number of articles that would initially be needed to be screened.

1.6. Scope of using PLUS and CQ filters for Systematic review updating

A recent study by Hemens et al showed that PLUS captured major articles that were used for a systematic review update[7]. However, in the sample of their study, PLUS was not able to capture all articles used for an update in many of the reviews. As well,

there were a proportion of reviews for which PLUS found no articles. Based on this past study and the inherent nature of PLUS, as a database with a selection of studies from the overall health research repository, we recognized that since retrospectively PLUS could not identify all articles used to update a review, it would be highly unlikely for PLUS to prospectively capture all articles to update a review.

For the present study, we looked at a subset of updated reviews that had experienced a change in conclusion, supposedly based on the availability of new evidence. Our leading hypothesis was that a change in conclusion of a clinically oriented systematic review would most commonly be driven by novel or significantly incremental research results, and we aimed to concentrate our resources and efforts, in assessing how PLUS would perform in providing relevant evidence to prompt updates of systematic reviews, on this specific subset of reviews, instead OF on all updated studies, as had been done before[7].

Given that in previous research PLUS captured about one in four articles[7], and selectively captured more of those ending in driving a change in conclusions, we hypothesized that PLUS a) will be capturing more articles in this subset of reviews b) will prove useful as an "update prompting mechanism", selectively capturing those studies prompting the update of SR with subsequent change in conclusion and c) would still not capture all of the articles incorporated in the update. For this reason, we added a second evaluation step, and, building on the performance of the clinical query filters, we secondarily aimed to see how the CQ filters would perform in updating the SR literature search. Our hypothesis was that CQ filters could be used to capture all the articles not found in PLUS, decreasing the number needed to read for review authors updating a systematic review without loss of relevant information.

In summary, should our two hypotheses be satisfactorily proven true, PLUS and CQ might be used together prospectively in a model as follows:

1. Subscribe to PLUS's specific alerts generated when content potentially relevant for a specific review is entered in the database, or periodically search PLUS

- To be alerted/see if there are relevant articles either that would prompt an update or to be used in an update

2. When conducting the update, using CQ filters to update the search, to decrease the number of search results to be screened.

Searching a variety of databases and filtering through results for relevant articles is quite time consuming and can add to the resource and time burden that is frequently cited as a barrier to updating systematic reviews. Under the hypothesis that searching the PLUS database would yield all impactful articles that would be needed to accurately update a systematic review, the use of PLUS may be a way to reduce the burden of searching databases when updating systematic reviews. With the potential to reduce time and resources needed to search various databases, the updating of systematic reviews may become more timely if PLUS is regularly used in the update process. The primary aim of this investigation is to investigate the role of McMaster Premium LiteratUre Service in the updating of systematic reviews.

SECTION 2: OBJECTIVES

2.1. OVERALL OBJECTIVE

To determine whether the use of McMaster Premium LiteratUre Service (PLUS) database can enable efficient updating of systematic reviews.

2.2. PRIMARY OBJECTIVE

To determine the performance of PLUS in locating studies that drove a change in conclusions in Cochrane Reviews

2.3. SECONDARY OBJECTIVES

1. For articles not found in PLUS, investigate if they can be found using Clinical Queries (CQ) search filters in PubMed
2. Characterize reasons for changed conclusions in Cochrane Reviews
3. Characterize implications for changed conclusions

SECTION 3: METHODS

The Cochrane Database of Systematic Reviews (CDSR) is a database that includes protocols for systematic reviews, as well as new and updated systematic reviews. As per the Cochrane mandate, each review is to be updated every two years, to help ensure that the review findings remain current and pertinent to practice.

According to Cochrane guidelines, the update process of a review must start from a new search. The next course of action will depend on the results of the new search. When there are no new eligible articles found, there may be no further work required beyond reporting the date of the new updated search. When, however, there are new studies found, review authors will need to incorporate the new data with existing data, which may include qualitative interpretation, about the quality of the study and results, or quantitative analysis of the data, including any meta-analyses completed. With new data to interpret, the new review may have the same conclusion, or may have different conclusions than the previous review. In the Cochrane database of systematic reviews, tags to help clarify the type of review demarcate all of these changes. For instance, a review is tagged with a dark blue tag, "Review" to distinguish it from a protocol, and a new review is tagged with an orange, "new". For reviews with new searches conducted, the review tag is accompanied by, the "Ns" tag, and for updated reviews with their conclusions changed, the review and new search tag is accompanied by a "Cc" tag to indicate the changed conclusion.

For the purposes of this project, we were interested in identifying those systematic reviews that underwent an update and had a change in conclusions (ie those marked with Cc). We used to following methods to identify eligible updated systematic reviews.

3.1. METHODS TO IDENTIFY POTENTIALLY ELIGIBLE REVIEWS

All records from the Cochrane Database of Systematic Reviews from January 1st 2012 to January 31st 2013 were screened to identify reviews with changed conclusions following an update. At the time of the search, the CDSR advanced search limit to identify reviews with changed conclusions did not capture all reviews tagged to have a change in conclusion. Hence, all records were hand-searched. Reviews with the changed conclusions tag were then assessed using the criteria below to determine eligibility. Screening for eligible reviews was completed by a single reviewer. Reviews

where eligibility was not apparent were discussed with another author until consensus was reached.

3.1.1. CRITERIA TO DETERMINE ELIGIBLE REVIEWS

- Indication of 'Conclusions Changed' status in publication history
- At least two published versions of the review, one prior and one current version (with the current version published between Jan 2012 to Jan 2013)
- Indication that a 'new search' was conducted to inform the current version
    - A new search being a search that differs with respect to the date or databases searched from the prior version
- At least one new or updated trial, which was identified from the new search, was added to the included studies of the updated review
- The reason indicated for conclusions changed was driven by new evidence (i.e. basis for changed conclusions is because of new data found from the new search, not because of other factors, such as typing errors in prior versions)
- Only randomized (RCT), quasi-randomized or controlled clinical trials (CCT) were included in both versions of the review

We did not restrict by review, year of the previous version or subject area of the review. We included reviews whose authors had changed from the previous version to the current version. We included reviews whose protocol had been modified (e.g. updated version had different primary outcomes, or different population parameters than the previous version).

After the updated reviews were identified, the previous version (the version from which the update was based) was acquired, and the two versions of the review were linked together and given a unique ID. The DOI, year of publication, author list, and Cochrane review group were identified for both versions of each included review.

3.2. CLASSIFICATION OF THE CHANGE IN CONCLUSION

With the inclusion of new information, we were interested in classifying the type of statistical change in conclusion listed by the authors. Looking at the results for the analysis or meta-analyses for the indicated primary outcome or first listed outcome, where a primary outcome was not listed, we categorized the change in conclusion according to the following four criteria:

1. Change in statistical significance status (statistical significance being $p<0.05$ or confidence interval for pooled statistic that does not cross null value) of a primary outcome.
    a. The treatment effect for the primary outcome is statistically significant where it was not previously
    b. The treatment effect for the primary outcome is not statistically significant where it was previously
2. Change in magnitude of treatment effect, given that direction of treatment effect stays the same
    a. Magnitude of the treatment effect for the primary outcome increased in any amount in the same direction
    b. Magnitude of the treatment effect for the primary outcome decreased in any amount in the same direction
3. Availability of new evidence for one or more outcome(s) which were empty in the previous version of the review

Many reviews included meta-analyses, and conclusions for both the primary and secondary outcomes. For the purposes of this investigation, we categorized the change in conclusion based on the primary outcome only. For reviews where there were multiple primary outcomes and reviews where there was no defined primary outcome, we categorized the change in conclusion for the first listed outcome. For reviews where the primary outcome had multiple comparisons, we identified all types of changes in conclusions for the different comparisons.

We were also interested in classifying whether the change in conclusions was clinically relevant or not. To do so, we considered the review authors' conclusions and their recommendations for a change in clinical practice. We considered the change in conclusion to be clinically relevant if the incorporation of new evidence led to a change in the recommendation for clinical practice (e.g. incorporation of new evidence in the update changed a recommendation to not adopt an intervention to a recommendation to adopt the intervention or vice versa). We considered the authors of each respective review to be experts in their clinical field, and hence being the best equipped to identify the clinical significance of the review results.

3.3. IDENTIFYING NEWLY ADDED REFERENCES

After both versions of eligible reviews were acquired, we wanted to identify the references new to the updated version. To do so, we manually compared the reference lists of the included trials between the different review versions and categorized each reference into one of the following categories:

1. Newly added trial: Trials and accompanying references, which were completely new to the updated review (ie were not in the previous version of the review)
2. Excluded trial: Trials and accompanying references, which were included in the previous version of the review, but were not included in the updated version of the review
3. Updated trial: A trial that was present in the previous version of the review, but had further references included in the updated version of the review that were not present in the previous version of the review
4. Trial in both versions: Trials and accompanying references that were present in both the previous and updated versions of the review and remained completely unchanged (ie no newly included or excluded references) through the update

For the references of all newly added and updated trials, the following information was acquired:

- Title of reference

- Type of reference (ie journal article, abstract at conference, protocol, or other)
- Year of publication, as indicated by reference list
- Journal of publication, as indicated by reference list
- Pubmed ID (if available)
    - Pubmed ID was provided in the review reference list for some references. For those references where the Pubmed ID was not supplied by review authors, we searched for individual references in Pubmed by title in attempt to acquire the correct Pubmed ID. If the search by title did not yield the reference, then we searched by first author and year of publication in further attempts to acquire the Pubmed ID for the reference.

3.4. CLASSIFICATION OF ARTICLES FOUND IN A JOURNAL INCLUDED IN PLUS VERSUS NOT INCLUDED IN PLUS

For each reference, we also identified if the journal of publication was indexed by PLUS or not. To do so, we compared the journal of publication to a list of PLUS journals (found at http://hiru.mcmaster.ca/hiru/journalslist.asp). If the journal of publication was present on the list, then we considered it to be indexed in PLUS. We also came across a few references that were not listed as published in a PLUS journal but were still retrieved in PLUS. We searched these articles by title in the PLUS database to see if they had been also published in a PLUS journal.

3.5. ASSESSMENT OF EXTRACTION CORRECTNESS

A single reviewer completed the extraction and subsequent categorization of articles. A sample of 30 randomly chosen reviews was reassessed to ensure accuracy of extraction. Even though the initial extraction and reassessment were completed by the same reviewer, we believed that it would have been difficult for the reviewer to recall details of the original extraction. Hence we used an unweighted Cohen's kappa statistic to calculate agreement[13].

3.6. SEARCH METHODS TO IDENTIFY ARTICLES IN THE PLUS DATABASE

The next step was to identify if a newly added or updated trial was included in the PLUS database. To increase the efficiency of the search process, we used the PubMed ID's to search for the articles, rather than other identifiers (e.g. article title). Using the PubMed IDs for individual studies, we searched the PLUS database for individual references via a Health Knowledge Refinery search page. The search function of this page identified which ID, from the inputted list of PubMed IDs, was associated with an article in the PLUS database. The identified PubMed IDs were then matched back to the article with which they were associated.

We considered a trial to be included if at least one of the references of the trial was found in the PLUS database. For instance, if a specific trial had four reported references, and at least one of them was found in the PLUS database, then we considered the trial to have been found in PLUS.

3.7. SEARCH METHODS TO IDENTIFY ARTICLES WITH PUBMED IDS BUT NOT FOUND IN THE PLUS DATABASE

Since the articles PLUS database had been previously screened for methodological quality, the PLUS database is not a comprehensive collection of all published articles in the included journals. Hence, we did not expect all articles used to update the eligible systematic reviews to be found in the PLUS database. To identify the studies that had PubMed IDs but were not found in the PLUS database, we used the Clinical Query filters found in PubMed. The CQ filters are specialized search filters that can be accessed through the PubMed search bar and are designed to retrieve articles within specific scopes of research (eg treatment, diagnosis, etc). The purpose of these filters is to increase the efficiency of evidence retrieval. Past studies, which empirically tested the use of the filters in study retrieval, showed that the filters had a maximum sensitivity of 99.3% (95% confidence interval (CI) 98.7% to 99.8%), and a maximum specificity of 97.4% (95% CI: 97.3% to 97.6%)[12].

Since the reviews in our investigation were limited to randomized, quasi-randomized or clinical trials, we used the sensitive, specific and combined 'therapy' filters to find new or updated trials, with PubMed IDs that were not found in the PLUS database.

In the search bar of PubMed, we inputted the terms for the filter that was being tested and used the Boolean operator 'AND' to combine the filter terms with a list of PubMed IDs that were connected with the Boolean operator 'OR'. The final search term was similar to the following ((Filter terms) AND (PubMed ID1 OR PubMed ID2 OR …)). By combining the search terms using the Boolean operators in this way, we would receive results of articles with PubMed IDs that were captured by the filter. The composition of the individual clinical query filters can be found online at the following address: http://hiru.mcmaster.ca/hiru/HIRU_Hedges_MEDLINE_Strategies.aspx. While we were most interested in the performance of the sensitive filter, we also checked the performance of the specific and balanced filters.

3.8. Statistical Analysis

Given the nature of the proposed study, we anticipated many of the results to be descriptive and qualitative in nature. Hence, we did not have any planned statistical analyses. We tabulated results and calculated proportions, frequencies and other numerical values to represent our findings.

3.9. Sample Size

We did not have a pre-determined sample size. Instead we chose to collect a convenience sample of all reviews published in the pre-defined time range of January 2012 to January 2013. We anticipated that the number of reviews with changed conclusions would amount to approximately 100, which we believed was a manageable yet thorough sample. By restricting our search to a particular time frame, we tried to ensure that the reviews that would comprise our final sample would have similar publishing standards and hence would be comparable.

3.9.1. POST HOC POWER CALCULATION

We performed a post-hoc sample size calculation as outlined below:
We referred to sample size calculation for proportions, setting alpha at 0.05 and beta at 0.8, and we used our findings that Hemens that 21% of trials were retrieved in PLUS.

We used the formula $N = (4z_{\alpha}^{2}(P(1-P)))/W^{2}$, where z_{α} is the confidence value in a normal distribution, P is the expected proportion of trials which are included in PLUS, and W is the width of the full confidence interval. Using this formula, the sample size would be 255 newly added trials found in PLUS (n=[4(1.96^2)(.21*.79)]/(.1^2)).

3.10. PERFORMANCE MEASURES

Outcome measures included:

1. The proportion of new and updated studies included in PLUS with the following denominators
    a. All identified studies
    b. Studies from journals indexed in PLUS
2. Classification and significance of changed conclusions in included studies

3.11. POST HOC ANALYSES

3.11.1: INVESTIGATION OF CLINICAL QUERY FILTER PERFOMANCE

We wanted to quantitatively explore how the clinical queries (CQ) filters might have reduced the search burden. To do so we selected a random sample of five reviews, in which the review authors had provided the search strategy for Medline and had applied some type of ad-hoc filter to limit the results by study type (ie clinical trials). In OVID, we recreated the search strategies of the selected reviews, and looked to see how many articles were retrieved using a) the content terms and b) the content terms plus the author's filter terms. We then c) applied the Clinical Queries therapy filter to the content terms from the original search strategy. Through this method we were able to identify the total number of articles retrieved by: 1) searching with only the content terms, 2) searching with content terms and any filters that the authors used, 3) searching with content terms and with the CQ filters, using either the sensitive or the specific setting. We subsequently manually checked how many of the newly added trials for the

individual reviews were found in the search results. Finally, we calculated the number needed to read (NNR) for each search strategy. NNR was calculated as 1/precision, where precision was calculated as the number of relevant articles found from the total search results over the total number of articles retrieved.

3.11.2. INVESTIGATING THE EFFECT OF TYPE OF STATISTICAL CHANGE IN CONCLUSION ON PLUS PERFORMANCE

We conducted post-hoc analyses to see if there were any differences in the number of newly added trials, trials found in PLUS and trials not found in PLUS by the type of change in statistical conclusion. We conducted three different ANOVAs with the following dependent outcomes: 1) All newly added trials, 2) Newly added trials found in PLUS and 3) Newly added trials not found in PLUS. We categorized the type of change in conclusions into the following five categories: 1) Gain in statistical significance, 2) Loss of statistical significance, 3) Increase in magnitude, 4) decrease in magnitude and 5) new evidence added to a previously empty review.

We originally classified the statistical change in conclusions for all comparisons for the primary outcome. Hence in Table 9, the total for the number of reviews is higher than the 92 reviews that are included. For the purposes of the ANOVA we considered only the first comparison for the primary outcome so that each review would only be counted once. Our hypotheses for our ANOVA were as follows:

H_0: Mean number of trials is the same across all groups.

H_A: The mean number of trials is different in at least one group.

Had our initial ANOVA showed a significant difference, we would have further explored the data using a Tukey's test to elucidate the groups that were differing.

3.11.3. INVESTIGATING THE EFFECT OF REVIEW GROUP ON PLUS PERFORMANCE

We were interested in identifying predictive factors that might indicate subsets of reviews for which PLUS may have better performance. We hypothesized that one such factor may be the Cochrane group in which a review was published. Ideally, we would

have considered comparing review groups in the area of internal medicine/general practice, areas to which PLUS is particularly geared, but we had to balance this consideration with the sample size of updated reviews available for each of group having had updates in the year of the study. Consequently, we planned to explore the hypothesis of a content-dependent contribution of PLUS by planning to compare the groups that had >5 reviews included to groups with <5 reviews included to see if there was a difference in the number of trials found in PLUS, to explore if there was a significant difference in PLUS performance, defined as the number of articles retrieved in PLUS over the total number of possible articles, between those groups with more reviews included and those with fewer reviews included. The comparison was performed with a post-hoc Analysis of Variance (ANOVA), as outlined below We planned to explore any difference by discipline on the same ANOVA table, if at all possible.

ANOVA outline: Comparing groups with >5 review updates to groups with ≤3 review updates included.

> The purpose of this comparison was to determine if there was a difference in PLUS performance between the review groups with the most and least number of included review updates. In looking at this subset of review groups, we were hoping to determine if there was a difference in PLUS performance between groups with more reviews published and those with fewer reviews published. In this exploratory ANOVA, we did not include groups with four or five reviews to try and discern if there was a difference in the extreme values.

By not including the groups with four or five reviews and comparing groups with the highest and lowest number of reviews, we tried to create a sample that would have been most likely to show a difference. Had the ANOVA above showed a significant difference in PLUS performance between groups with the most and least number of groups, we would have conducted a second ANOVA to include the groups with four or five reviews that were not included in the above ANOVA. The purpose of this analysis would have been to ascertain if there a variability in the outcome depending on specific characteristics of the review groups.

To create the comparison groups, we gave each review in groups with >5 reviews a unique code, and the remaining reviews were grouped together to form a single comparison group. Since this was an exploratory post-hoc analysis, we explored using the outcome of proportion of new trials found in PLUS (ie #new trials found/total # of new trials) and outcome of absolute number of new trials found.

The null and alternate hypotheses guiding this analysis are the same as those presented in section 3.11.2. All analyses were conducted with IBM SPSS version 20.0.0. A p-value of ≤ 0.05 was considered significant for all analyses.

SECTION 4: RESULTS

4.1. SEARCH RESULTS

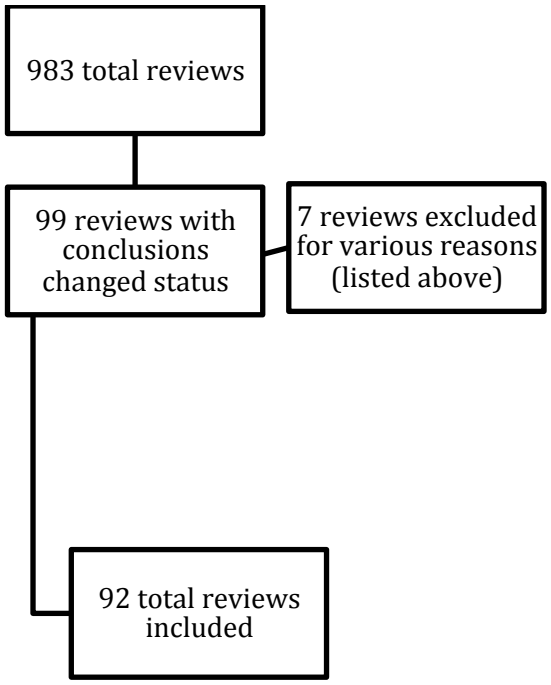
From the dates of January 1st 2012 to January 31st 2013 there were 1549 published articles in the Cochrane Database of Systematic reviews, of which 983 were full Cochrane reviews and 566 were published protocols. Of these, 884 reviews were excluded because they were not marked as having a change in conclusion. The remaining 99 reviews were screened for eligibility according to the criteria previously provided. Of the 99 reviews, seven were excluded for the following reasons:

Table 2. Reason for exclusion of reviews

	Review Title	Publication month, year	Reason for Exclusion
1	Interventions for preventing falls in older people in health care facilities and hospitals	December, 2012	Change in conclusion status that was previously given was revoked as noted in the 'what's new' and 'history' sections of the review.
2	Rehabilitation for Hamstring Injuries	December, 2012	No new trials were added from previous version to updated version. Change in conclusion status was based only on the exclusion of a previously included trial, rather than the addition of new information for newly added trials.
3	Hyperbaric Oxygen therapy for promoting fracture healing and treating fracture non-union	November, 2012	No new trials were added from the previous version to the updated version of the review. Change in conclusions status was based on three newly-identified ongoing studies. These studies, while not included in the current version of the review were basis for alterations in the 'Implications for Research' Section. Due to these changes, the review was marked as having a change in conclusion.

4	Sonothrombolysis for acute ischaemic stroke	October, 2012	No new trials were added from the previous version of the review to the updated version of the review. The change in conclusions status was based on the correction of a data error, as noted by the authors in the 'What's new' and 'History' sections of the review.
5	Colony stimulating factors for prevention and treatment of infectious complications in patients with acute myelogenous leukemia	June, 2012	The change in conclusions status for the updated version of the review was given due to a comment added from feedback to the review rather than because of the inclusion of new evidence from newly added or updated trials.
6	Surgical removal versus retention for the management of asymptomatic impacted wisdom teeth	June, 2012	No new trials or updated trials were present in the updated version of the review. The change of conclusions status given was based only on the exclusion of a previously included trial. According to authors, this trial was excluded because the methodology in the report was not appropriate for the standards of the updated review.
7	Recombinant human activated protein C for severe sepsis in neonates	April, 2012	No new trials or updated trials were included in the updated version of this review. The intervention compared in the review has been withdrawn from the market due to risk of mortality. Because of this, 'Implications for Research' section was updated to indicate that no new trials should be conducted in this area. This alteration was the basis for the change in conclusion for the review.

Figure 1. Flowchart of screening results



The final sample included 92 reviews, belonging to 31 of the 52 (59.6%) Cochrane Review Groups. Table 3 shows the distribution of the included reviews across review groups. The review groups to have the largest number of included reviews were the Cochrane Pregnancy and Childbirth Group and Cochrane Airways Group with 13 and 10 reviews, respectively.

Table 3. Included Reviews by Review Group

Cochrane Review Group	**Number of reviews [n(%)]**
Cochrane Airways Group	10(10.8)
Cochrane Anaesthesia Group	1(1.1)
Cochrane Bone, Joint and Muscle Trauma Group	2(2.2)
Cochrane Consumers and Communication Group	1(1.1)
Cochrane Dementia and Cognitive Improvement Group	1(1.1)
Cochrane Developmental, Psychosocial and Learning Problems Group	1(1.1)
Cochrane Ear, Nose and Throat Disorders Group	1(1.1)
Cochrane Eye and Vision Group	2(2.2)
Cochrane Fertility Regulation Group	1(1.1)

Cochrane Gynaecological Cancer Group	3(3.3)
Cochrane Haematological Malignancies Group	2(2.2)
Cochrane Heart Group	5(5.4)
Cochrane Hepato-billiary group	3(3.3)
Cochrane Incontinence Group	4(4.3)
Cochrane Inflammatory Bowel Disease and Functional Bowel Disorders Group	3(3.3)
Cochrane Injuries Group	2(2.2)
Cochrane Menstrual disorders and Subfertility Group	5(5.4)
Cochrane Movement Disorders Group	1(1.1)
Cochrane Multiple Sclerosis and Rare diseases of the Central Nervous System Group	1(1.1)
Cochrane Neonatal Group	6(6.5)
Cochrane Neuromuscular Disease Group	3(3.3)
Cochrane Occupational Safety and Health Group	1(1.1)
Cochrane Oral Health Group	1(1.1)
Cochrane Pain, Palliative and Supportive Care group	1(1.1)
Cochrane Peripheral Vascular Disease group	1(1.1)
Cochrane Pregnancy and Childbirth Group	13(14)
Cochrane Renal Group	4(4.3)
Cochrane Skin group	1(1.1)
Cochrane Stroke Group	4(4.3)
Cochrane Tobacco Addiction Group	5(5.4)
Cochrane Wounds Group	3(3.3)

Table 4 shows the distribution of reviews by publication month, December seems to be the month with the most published reviews in our sample with 15 (16.3%) reviews.

Table 4. Included reviews by month

Month and Year	Number of reviews [n(%)]
January 2012	4(4.4)
February 2012	2(2.2)
March 2012	7(7.6)
April 2012	8(8.7)
May 2012	11(12.0)
June 2012	4(4.3)
July 2012	4(4.3)
August 2012	8(8.7)
September 2012	11(12.0)
October 2012	5(5.4)
November 2012	6(6.5)
December 2012	15(16.3)
January 2013	7(7.6)

4.2. IDENTIFYING NEWLY ADDED REFERENCES

From the 92 reviews, there were a total of 2166 trials reported over 3924 references. This corresponds to an average (±SD) of 23.5 (±26.8) [95%CI: 22.35-22.65] trials and 42.7 (±50.0) [95%CI:41.10-44.30] individual references per review. Of these trials, there were 854 newly added trials reported over 1498 references. This corresponds to an average of 9.3 (±10.3) newly added trials and 16.3 (±22.9) newly added individual references per review. There was a moderate cohen's kappa of of 0.74 [95%CI:0.68-0.79] between the initially extracted and reassessed reviews.

Of the 854 new trials, 180 (21.1% 95% CI: [18.3-23.8]) were found in the PLUS database, and 674 (78.9%) were not found in the PLUS database.

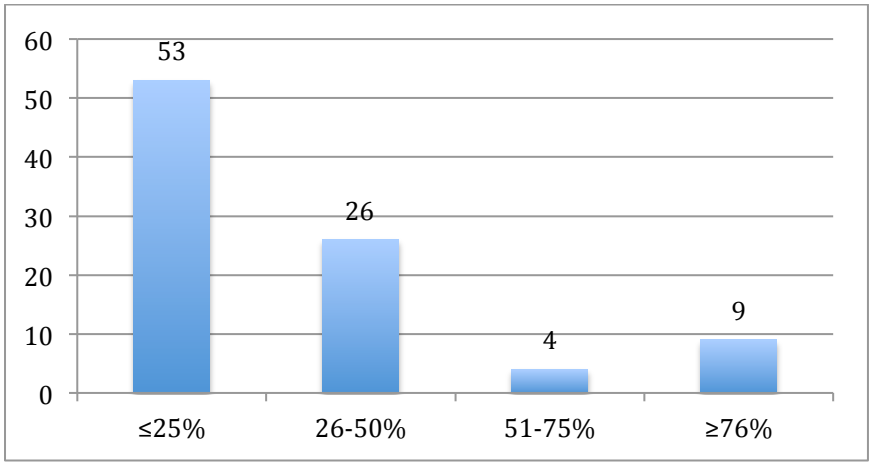
Of the 1498 unique references, 447 (29.8%) did not have a PubMed ID. Since their inclusion in PLUS database was verified by searching using PubMed IDs, the presence of these 447 references was not verified in the PLUS database. Hence, the remaining analysis is based on the 1051 references, with PubMed IDs, which were easily searchable in the PLUS database. Of these 1051 references, 217 (20.6% 95% CI: [18.2-23.1]) unique references were found in the PLUS database, and 834 (79.4%) were not found in the PLUS database.

Table 5. Distribution of newly included trials found and not found in PLUS

#of included reviews	# new trials	Trials in PLUS		# new references	PM ID status of references		References with PM ID in PLUS n(%) [95%CI]
		In n(%) [95%CI]	Out n(%) [95%CI]		+PM n(%) [95%CI]	-PM n(%) [95%CI]	
92	854	180 (21.1%) [18.3-23.8]	674 (78.9%) (76.2-81.7)	1498	1051 (70.2%) [67.8-72.5]	447 (29.8%) [27.5-32.1]	217 (20.6%) [18.2-23.1]

Table 5 provides a summary of the trials included in the update that were and were not found in PLUS, by the month of publication of the review.

Figure 2: Percentage of newly found trials in PLUS by review



From graph 1, we can see that in 53 of 92 (57.6%) reviews there were 25% or fewer newly added trials in the PLUS database. None of the newly added trials were found in PLUS by 26 of 92 reviews (26.2%). Conversely, in 9 of 92 (9.8%) of reviews, 76% or more newly added trials were found in the PLUS database. In 8 of 92 reviews, 100% of the trials were found in PLUS. Interestingly, 6 of the 8 (75%) reviews, of which 100% of the trials were found in PLUS only included one newly added trial in the update. This can be compared to 6 of 26 (23%) of the reviews with no newly added trials found in PLUS having only one newly added trial in the update. The remaining 20 of 26 reviews with no newly added trials found in PLUS had 2 or more newly added trials to the update.

4.3 Articles found in PLUS journals

At the time or investigation, the PLUS database indexed articles from approximately 120 journals. Of the 1498 newly added references included in the reviews used in this investigation, 458 (30.6%) were from a journal indexed by the PLUS database, and 1040 (69.4%) were from a journal not indexed by the PLUS database. Of the 458 articles from a PLUS journal, 193 (42.1%) were found in the PLUS database, and 265 (57.9%) were not found in the PLUS database. Table 6 summarizes the citations that were published in a journal indexed by PLUS versus those that were not indexed by PLUS. As well, table 6 further categorizes the citations with PubMed IDs that were and were not published in a journal monitored by PLUS.

Table 6: References published in and not published in journals monitored by PLUS

Year	Month	# of reviews	All citations with PLUS Journals		Citations with PM in PLUS Journals		# all Citations in PLUS Journals NOT in PLUS	# all Citations in PLUS journals in PLUS
			In	Out	In	Out		
2013	Jan	7	64	53	54	49	41	23
2012	Dec	15	163	298	100	169	108	55
2012	Nov	6	13	49	13	42	2	11
2012	Oct	5	13	37	12	27	9	4
2012	Sept	11	24	56	23	50	11	13
2012	Aug	8	19	72	18	28	9	10
2012	July	4	4	27	3	23	2	2
2012	June	4	10	51	10	34	3	7
2012	May	11	59	204	52	158	33	26
2012	April	8	17	44	16	25	7	10
2012	March	7	19	50	12	40	10	9
2012	Feb	2	29	51	17	38	19	10
2012	Jan	4	24	47	15	23	11	13
Total		**92**	**458**	**1040**	**345**	**706**	**265**	**193**

4.4. Performance of Clinical Query filters in identifying articles not found in PLUS

We used the clinical query filters on PubMed to search for the 834 citations not found in the PLUS database. Of these 834 references, there were 728 Unique PubMed IDs. Table 7 shows the recall rate of the sensitive, specific and combined Clinical Query filters in Medline, in identifying the 834 references.

Table 7. Performance of MEDLINE Clinical Query filters

Clinical Query filter	# articles of 728 retrieved n(%)	95%CI
Sensitivity	701 (96.2)	94.9-97.7
Specificity	606 (83.2)	80.5 -86.0
Combined	629 (86.4)	83.9-88.9

There are several reasons that articles reporting trials deemed to be eligible for inclusion in a systematic review are not captured by CQ filters in Medline, among which the most likely is that the random allocation of participants to comparison groups is not reflected in the study design and not declared in the abstract[10].

Of the three filters, the sensitive filter performed best, and identified 96.2% of articles that had PubMed IDs but were not found in PLUS.

After accessing the articles that were not found by the CQ filter, we were able to categorize the reason why they were not included (summarized in table 8). Five of these 27 articles were not accessible as full-text articles. Hence, we could not categorize the reasons why they were not found in by the CQ filters.

Table 8. Reasons why articles were not captured by the Medline CQ filters

Reason		**Number [n (%)]**
Article not accessible		5 (18.5)
Study not randomized	Article reporting non-randomized study associated with RCT	9 (33.3)
	Subgroup or secondary analysis of an RCT	4 (14.8)
	Comment on RCT	2 (7.4)
	Controlled clinical trial	5 (18.5)
RCT with less than 80% follow-up		2 (7.4)
Analysis not consistent with study design		0 (0.0)

The most frequent reason that an article was not captured by the CQ filters was that it was not reporting a randomized control trial.

4.5. CLASSIFICATION OF THE CHANGE IN CONCLUSION

4.5.1. CHANGE IN STATISTICAL SIGNIFICANCE

Table 9, below, shows the distribution of types of change in statistical significance for the primary or first outcome listed for each of the included 92 reviews. Some of the outcomes analyzed had multiple comparisons, and we classified the change in statistical significance for all included comparisons. For this reason, the total number of reviews listed in the table exceeds the total number of eligible reviews.

Table 9. Change in statistical significance

Type of change in conclusion	**Number of reviews**
Increased magnitude in the same direction	50
Decreased magnitude in the same direction	8
Gain of statistical significance	45
Loss of statistical significance	14
Previously empty outcome to at least one trial included	7

The inclusion of new information seemed to have the most effect on increasing the magnitude of the treatment effect or causing a gain of statistical significance where there was none previously. Interestingly, addition of new information did not frequently decrease the magnitude of the treatment effect or cause a loss of statistical significance where there was one previously.

Table 10. Cross-tabulation of trials found and not found by PLUS according to typology of change in conclusion

Type of change in conclusion	# of reviews	Newly added trials			Trials in PLUS			Trials not in PLUS		
		Mean (SD) [95%CI]	Min	Max	Mean (SD) [95%CI]	Min	Max	Mean (SD) [95%CI]	Min	Max
Increased magnitude in the same direction	50	10.2(9.1) [7.6-12.8]	1	32	2.1(2.6) [1.5-2.9]	0	10	8.0(7.4) [5.9-10.1]	0	27
Decreased magnitude in the same direction	8	10.5(7.9) [4.9-16.1	1	22	1.4(1.5) [0.3-2.5]	0	4	9.1(7.1) [4.1-14.1]	1	18
Gain of statistical significance	45	9.3(11.6) [5.8-12.8]	1	69	2.1(2.2) 1.44-2.76]	0	8	7.2(10.4) [4.1-10.3]	0	61
Loss of statistical significance	14	7.6z(7.1) [3.8-11.4]	1	27	1.3(1.3) [0.6-2.0]	0	4	6.3(6.4) [2.9-9.7]	0	23
Previously empty to full	7	2.7(2.2) [1.0-4.4]	1	7	1.1(1.3) [0.1-2.1]	0	3	1.6(1.7) [0.3-2.9]	0	4

Table 10 shows the average number of trials found and not found by PLUS categorized by the type of change in conclusion for the primary outcome of the review. On average, there seemed to be more newly added trials to the reviews that had a change in magnitude (either increased or decreased) than those reviews that had a loss or gain in statistical significance. Interestingly, we found on average that there were more trials found in PLUS for reviews that had an increased magnitude of statistical significance in the same direction and reviews where there was a gain of statistical significance in comparison to reviews where there was a and reviews that had a loss of statistical significance. However, given the breath of the confidence intervals, these differences are likely not statistically significant.

4.5.2 CHANGE IN CLINICAL SIGNIFICANCE

Table 11 shows that there was a clinically significant change results of the updated review in 32 (34.5%) of reviews, as indicated by review authors.

Table 11. Change in clinical significance

Change in clinical significance	Number of reviews [n (%)]	95% CI
Yes	32 (34.8)	25.1-44.5
No	60 (65.2)	55.5-74.9

4.6. POST HOC ANALYSES

4.6.1: CLINICAL QUERY FILTER PERFORMANCE

Table 12: Search results by type of filter

Study ID	# Articles retrieved by filter: Content terms only	Content terms and author filters	Content terms and CQ sensitive filter	Content terms and CQ Specific filter	Proportion reduced by filter*: Author filter n%(95%CI)	CQ sensitive filter n%(95%CI)	CQ specific filter n%(95%CI)
12Nov2	862	299	134	48	65.2 (62.0-68.4)	84.3 (81.9-86.8)	94.4 (92.9-96.0)
12Mar3	5280	497	190	86	90.6 (89.8-91.4)	96.4 (95.9-96.9)	98.4 (98.0-98.7)
12Apr1	3163	453	579	271	85.7 (84.5-86.9)	81.7 (80.3-83.0)	91.4(0.5-92.4)
12May7	6301	1085	256	40	82.8 (81.8-83.7)	95.9 (95.4-96.4)	99.4 (99.2-99.6)
12July4	4478	1217	447	235	72.8 (71.5-74.1)	90.0 (89.1-90.9)	94.8 (94.1-95.4)
Average					79.4 (70.3-88.5)	89.8 (83.9-99.7)	95.7 (92.8-98.6)

*calculated as [(#retrieved by content terms-# retrieved by filter)/#retrieved by content terms]

As seen in the table above, the use of any type of search filter reduced the number of retrieved articles by over 50%. In all cases but one, use of any CQ filter reduced the number of retrieved articles by a larger amount than non-CQ filters used by the authors. When comparing the CQ filters, we found that the CQ specific filters performed better than the CQ sensitive filters with respect to reducing the number of search results. After conducting a t-test to determine the differences in proportional reduction of the author and CQ sensitive filter, we found the difference to not be

statistically significant (p=0.09; data not shown). We found a significant difference in the proportional reduction between the author and specific filters (t(8)= 3.38,p=0.0009)

Table 13: Proportion of newly added trials retrieved by search filter

Study ID	Relevant article retrieval Total possible	Content terms n(%)	Author filter n(%)	CQ sensitive filter n(%)	CQ specific filter n(%)
12Nov2	3	3 (100)	3 (100)	3 (100)	3 (100)
12March3	4	4 (100)	4 (100)	4 (100)	3 (75)
12April1	4	4 (100)	4 (100)	4 (100)	4(100)
12May7	2	2 (100)	2 (100)	2 (100)	2 (100)
12July4	1	1 (100)	1 (100)	1 (100)	1 (100)

We then accessed the results for each search, and identified if the newly added trials for the individual reviews were found in the search results. As can be seen in the table above, all newly added trials were by definition present in the search results using the content terms, and author filters (which is the search strategy actually used to retrieve them). All of the articles were also retrieved when using the CQ sensitive filter. When using the CQ specific filter, all except for one newly added trial was found in the search results.

Table 14: Number Needed to Read (NNR) and reduction in NNR using search filters

Study ID	Number needed to read to find one new trial*				Proportion Reduction in NNR** n%(95%CI)		
	Content terms only	Content terms and author filters	Content terms and CQ Sensitive filter	Content terms and CQ Specific filter	Content to Author	Content to CQ Sensitive filter	Content to CQ Specific filter
12Nov2	**288**	**100**	**45**	**16**	**65.3 (59.8-70.8)**	**84.4 (80.2-88.6)**	**94.4(91.8-97.1)**
12March3	**1320**	**125**	**48**	**29**	**90.5 (89.0-92.1)**	**96.4 (95.4-97.4)**	**97.8(97.0-98.6)**
12April1	**791**	**114**	**145**	**68**	**85.6 (83.1-88.0)**	**81.7 (79.0-84.4)**	**91.4(89.4-93.4)**
12May7	**3151**	**543**	**128**	**20**	**82.8 (81.4-84.1)**	**95.9 (95.2-96.6)**	**99.4(99.1-99.6)**
12July4	**4478**	**1217**	**447**	**235**	**72.8 (71.5-74.1)**	**90.0 (89.1-90.9)**	**94.8(94.1-95.4)**
Average					**79.4 (70.3-88.5)**	**90.6 (84.3-96.9)**	**95.6 (92.8-98.4)**

*all values have been rounded up to the nearest integer
**calculated as [(# of articles retrieved by content only NNR – # of articles retrieved by Filter NNR)/(# of articles retrieved by Content NNR)]

As can be seen in the table above the use of any search filters reduced the NNR. Using a t-test, we found that the difference in NNR between the author and sensitive filter was not statistically significant (p=0.15). However, the difference of NNR between author specific filter was significant (t(8)=3.39, p=0.01). These findings should be evaluated in a larger cohort.

4.6.2. Effect of Type of change in statistical conclusion on PLUS performance

To see if there was a difference in PLUS performance in retrieving trials between reviews that had an overall different type of statistical change from the prior version to the present version, we conducted an ANOVA, for which the results are below, and categorized by the outcome variable (ie total number of newly added trials).

Table 15: ANOVA table for PLUS performance by type of statistical change

Outcome	F-statistic	Significance level
Total number of newly added trials	0.083	0.987
Number of newly added trials found in PLUS	0.135	0.969
Number of newly added trials not found in PLUS	0.110	0.979

As we can see in table 15 above, the p-value for all of the ANOVA analyses were greater than 0.05 and were not statistically significant. This indicates that we cannot reject our null hypotheses, and there were no differences in the mean number of trials for any of the groups of change in statistical conclusion for any of the three dependent variables that we explored.

4.6.3. Effect of review group of publication and PLUS performance

In our sample, there were three groups with >5 reviews:

1. Cochrane Pregnancy and Childbirth group (n=13)
2. Cochrane Airways group (n=10)
3. Cochrane Neonatal group (n=6)

The reviews in the Pregnancy and Childbirth group had 189 total newly added trials with 37 (19.6%) in PLUS. The reviews of the Airways group had 81 total newly

added trials with 14 (17.3%) found in PLUS. Finally, the reviews of the Neonatal group had 43 total newly added trials, with 15 (34.8%) found in PLUS.

Further, there were three groups with 5 reviews (Cochrane Heart, Menstrual disorders and Subfertility and Tobacco addiction groups) and 3 groups with 4 reviews (Cochrane Incontinence, Renal and Stroke groups), which were excluded. Hence, our final sample included 65 reviews representing 25 review groups.

Table 16. ANOVA table for PLUS performance by review group

Comparison group	Outcome	F-statistic	Significance level
Groups with >5 reviews vs with ≤3 reviews	Number of newly added trials	1.199	0.318
Groups with >5 reviews vs with ≤3 reviews	Proportion of newly added trials	1.304	0.281

Given that none of the significance levels were below 0.05, we cannot reject the null hypothesis for either of our analyses. Hence, there does not seem to be a difference in the performance of PLUS given the review group. However, given the small sample sizes, we likely did not have sufficient power in this analysis to detect a change if there was truly one present.

SECTION 5: DISCUSSION

In this investigation, we explored how the McMaster PLUS database performed in capturing the studies that were used to drive a change in conclusion of a selection of updated Cochrane reviews. A total of 854 new trials were used to update and drive a change in conclusions of the 92 reviews included in this investigation. Of these, 180 (21.1%) were found in the PLUS database. All of the newly added studies that drove a change in conclusion were found in 9 of 92 (9.8%) of reviews. Conversely, none of the newly added studies used to drive a change in conclusion were found in 26 of 92 (26.2%) of reviews.

The Clinical Query filters in PubMed performed well in retrieving the references not found in the PLUS database. The sensitive filter retrieved 701 (96.2%) of the articles not found in PLUS. There were 27 articles that were not captured by the CQ filters. As

Table 7 showed, these articles were describing non-randomized trials or were RCTs with less than 80% follow-up. Since the CQ filters were designed to capture articles describing RCTs with more than 80% follow-up, the subset of 27 articles would not have been found by the CQ filter. This indicates that the CQ filters found all of the possible articles, which they were designed to find.

With the advent and implementation of evidence-based medicine, there is an increasing need for persons in the healthcare field to remain up-to-date with new knowledge. However, this task is becoming increasingly difficult. A recent study noted that to evaluate newly released information, a primary care physician would have to spend an average of 627.5 hours per month[14].

Systematic reviews of literature qualitatively and mathematically summarize available data on a specific research topic, while assessing included evidence for quality and bias. By removing the need for authors to read individual primary care articles and individually summarize the results, systematic reviews can help reduce the time needed for health care professionals to synthesize data and remain up-to-date on medical literature.

Of systematic reviews, Cochrane Reviews have been recognized to be of high methodological quality, due to the rigorous and meticulous review and peer-review processes that are required to publish a Cochrane systematic review. The utilization and access of Cochrane Reviews has been steadily increasing. For instance, the impact factor of the Cochrane Database of Systematic Reviews increased from 5.785 in 2012 to 5.939 in 2013. As well, the number of citations of the Cochrane Database of Systematic reviews increased from 29,593 in 2011, to 34,230 in 2012, to 39,856 citations in 2013[15,16]. This trend indicates the increased use of the Cochrane Database of Systematic reviews and perhaps hints at the increasing future use of the Cochrane Database of Systematic reviews.

With the high level of utilization of systematic reviews and publication volume of clinical trials, the need for systematic reviews to stay up to date, and consequently valid,

will become increasingly important. However, as a report by Hemens et al in 2012 noted, more than 87% of the Cochrane reviews in their sample were not updated in the two-year time frame, despite the requirement that is mandated by the Cochrane Collaboration[7].

To complete an update of a systematic review, authors generally have to complete a literature search, screen the results for relevant data, extract the data from eligible studies, incorporate the new data into qualitative syntheses or meta-analyses, as appropriate, and interpret the results and draw conclusions as required. All of the reviews in our sample included at least one new trial; the maximum number of new trials included in one review was 69, which were reported over 113 unique references. Even for reviews where no new trials are found through the literature search, the authors will still have to complete the search and screen eligible articles. All of these steps can be very time-consuming, and contribute to the time and resource consumption that has been cited as a barrier to timely updating of systematic reviews[8].

Working under the concept that searching for articles in a database where included articles were pre-assessed for methodological quality and newsworthiness would yield the most influential articles that would drive a change in conclusions for a systematic review update, we empirically tested the performance of PLUS in capturing the articles that drove a change in conclusion. PLUS captured approximately one fifth of all trials used to drive a change in conclusions. This value was similar to that found by Hemens et al[7].

5.1. Change in Conclusion

PLUS retrieved none of the newly added trials in 26 of the 92 reviews. However, only in seven of these 26 reviews (26.9%) did authors report a clinically relevant change in conclusions based on the update. This low number may indicate that, while new evidence is being synthesized in certain subject areas, the comparisons being made are not sufficiently novel to cause a clinically significant change. Articles in the PLUS database are pre-appraised for 'newsworthiness'. The topics of these 26 reviews may have been in areas, wherein RCT-based research that is newsworthy is not published

frequently. This fact may be the reason why none of the trials in these 26 reviews were found in PLUS. Conversely, PLUS retrieved all of the newly added trials in 8 of the 92 reviews. Of these eight reviews, there was a clinically relevant change in conclusions in 4 reviews (50.0%). Of the remaining 58 reviews, where at least one trial was found in PLUS, 21 (36.2%) had a clinically relevant change in conclusion, whereas the remaining 37 (63.2%) did not have a clinically relevant change in conclusion.

We further looked at the distribution of trials found or not found by PLUS with respect to the type of statistical change in conclusion. Reviews either where the primary outcome had an increased or decreased magnitude of effect without a change in direction had, on average, the highest number of newly added trials. Among these, however, the reviews where there was an increased effect size or where a previously non-significant effect became significant had, on average, the highest number of newly added trials found in PLUS. While this trend was not statistically significant, it may indicate the presence of publication bias due to higher impact journals perhaps being less likely to publish studies reporting a null result.

As expected, the reviews that had found no studies reporting on their primary outcome in the previous version, and had data found for the update, had a fewer number of total newly added trials. Subsequently, these reviews had fewer trials found in PLUS. This might reflect the fact that orphan or rare diseases, for which few trials are performed and published, might not reach the journals appraised by PLUS, or not being of enough high quality to be included.

We considered whether there would be any difference in how PLUS performed depending on what type of change in statistical conclusion there was between the original and updated version of the review. From the results of our post-hoc ANOVA, we found that there was no difference in the recall of newly added trials by the type of statistical change in conclusion. Further, we also discovered that there was no significant difference in the number of newly added trials across the five predetermined groups. In efforts to avoid double counting a review in our analysis, we only categorized the statistical change in conclusion for the first comparison for the primary outcome with

available evidence. Hence, it is possible that we incorrectly categorized the change in conclusion for the overall review.

5.2. CLINICAL QUERY FILTERS

We used the clinical query filters designed for Medline to locate the articles that had Pubmed IDs but were not found in PLUS. The filters performed admirably in retrieving the articles not found in PLUS. This was especially true for the filter designed to maximize sensitivity. The sensitive filter retrieved 96.2% of articles not found in PLUS. There were 27 (3.8%) of articles not found by the sensitive filter. Five (18.5%) of these articles were not available as full-texts, and the information in the abstract, when given, was not sufficient to categorize them into one of the three reasons why the article would not be captured by the CQ filters. Of the remaining 22 articles, 20 (74.0%) were not reporting a randomized trial, and 2 (7.4%) were reporting a randomized trial with less than 80% follow-up. While all of the included reviews only used RCTs, CCTs and quasi-randomized studies, many of the included trials were reported over multiple unique citations. However, not all of the listed citations were describing the results of the RCT. Some references were protocols, secondary analyses, or descriptions of tools or scales used to measure outcomes. Hence, all of the trials not captured by the CQ filters were trials that would not have been captured by the filters.

The results from investigating the use of the CQ filters indicate that the CQ therapy filters, either those maximized for sensitivity or those maximized for specificity, individually have the potential to drastically decrease the number of search results and NNR that would apply to a Cochrane search without paying a price in terms of missed papers. Empirically, we showed that use of the specific filter caused the highest proportional reduction (95.6% (± 3.1%)) in NNR. However, the use of the specific filters failed to capture one newly added trial out of 14. In comparison, the sensitive filters, which an average proportional reduction of 90.6% (± 7.0%), were able to capture all of the newly added trials. A recent survey about updating practices identified 70% of respondents identified lack of time and resources as a reason to not commence an update[8]. Testing the use of the CQ filters in this small subset of reviews shows empirically that the CQ filters on average decrease the number of articles screened, and

they seem to perform better than author derived filter terms. Based on these data, it may be warranted to conduct further comparative studies on a larger set of studies, to more consistently define how the filters perform in comparison to other proposed filters.

In using the CQ filters, authors can decrease the time investment at the beginning of the search, and subsequently time needed to screen the search results. The CQ filters capture articles based on the methodological aspects of the trial (ie they search for RCTs). By using the CQ filters, authors will be able to focus on creating the content terms of the search, while being confident that the filters will screen and capture trials based on relevant methodology. In our sample, 50 reviews reported the number of unique articles retrieved by their initial search. These 50 reviews reported a median of 873 articles (range: 107 to 22,012) that needed to be initially screened. These 50 reviews finally included a median 17 articles (range: 1 to 98). Looking at how many articles review authors needed to read to find one included article (ie total number of articles undergoing initial screening/final number of articles included), these 50 reviews reported a median of 62.5 articles (range 6.4 to 1966). Lack of randomization was a common reason for exclusion of many of the articles. The broad filter that was created to maximize sensitivity has shown to have 99.3% sensitivity in past studies, which means that the filter was able to capture 99.3% of relevant trials. Moreover, use of the sensitivity filter was able to reduce the number needed (NNR) to read to 10[17]. The NNR, which is defined as the inverse of precision, is a performance measure that indicates how many articles one would have to read to locate one relevant article[18,19]. With an NNR of 10, use of the sensitive CQ filter would require a reader to read 10 articles to find one relevant article. Hence, use of the CQ filters has the potential to reduce the number of articles needed to be screened to identify relevant trials, thus increasing the efficiency of the update process. The use of the more specific filter, though paying a 3% reduction in sensitivity, cut down the NNR to 3.

5.3 Effect of publication in PLUS journal

PLUS indexes articles from only a subset of all medical journals. Journal inclusion is based on impact factor of the journal, and the yield of articles from each journal that meet the rigorous pre-assessment criteria for inclusion into the PLUS

database among other factors[10]. Since systematic reviews aim to comprehensively review all available literature, we anticipated that some articles used to update the review would have been published in a journal not indexed by PLUS.

Of the 1498 newly added articles, 458 (30.6%) were published in a PLUS journal, of which 193 (42.1%) were found in the PLUS database. The lower number of articles found in PLUS, of the ones published in a PLUS journal is indicative of the rigorous pre-screening process that each article undergoes before inclusion in the database. The low number of articles that was found in a PLUS journal but not included in PLUS may indicate that some reviews are including studies that were deemed of insufficient quality according to the PLUS inclusion criteria. The lower quality of these included studies may mean that the reviews that include them would have less definitive results, as the conclusions based on these results would likely have to be downgraded for increased risk of bias due to poorer methodological quality. In terms of decision-making, the implication of inclusion of these articles, which were published in PLUS journals but not found in PLUS, may be that these reviews cannot give a definitive direction for choosing one treatment intervention over another. Of the 92 reviews, 19 (20.7%) had no articles published in a PLUS journal, and 25 (27.1%) had only one article published in a PLUS journal. The maximum number of articles, for a single review, published in a PLUS journal was fifty-one.

5.4. FACTORS AFFECTING PLUS PERFORMANCE

We statistically explored the hypothesis that PLUS performance would differ depending on the Cochrane group, in which a review was published. When looking at the groups in our sample that had the most reviews versus those that had the fewest reviews we did not find any statistically significant difference between review groups in terms of PLUS performance. However, the most relevant determinants of a different contribution would have been the area of interest of review groups, a hypothesis that we could not formally test for the small sample size of individual groups. However, we assessed separately the contribution of PLUS to the three largest groups, and we found that there was no difference in PLUS performance by review group. However since this was a post-hoc analysis, it could have been the case that our sample of reviews was not large

enough to detect a difference between groups, and we cannot completely discount the fact that there may be a difference in PLUS performance by review groups. Furthermore, since we combined review groups with few reviews, it is possible that we missed any differences that may have been present in these groups. However, given our current sample, we thought that exploring differences in review groups with few reviews would not have given us a reliable sample. Incidentally, it is worthy noting that Hemens et al had a similar distribution of reviews by Cochrane editorial groups[7].

While we did not find review group to be a significant review-based factor that would predict PLUS performance, we hypothesized multiple potential factors that may influence this performance. For instance, using PLUS for reviews that included older studies in the update, perhaps due to a change in protocol for the update, may be less fruitful than in reviews that use more recent evidence to update the review. Articles published prior to the implementation of PLUS are not included in the database[10]. As well, older evidence may be less indicative of current medical practice, or be held to different publishing standards which would preclude them from being included in the PLUS database[20]. Similarly, using PLUS in subject areas where there are many potential participants, allowing for large treatment effects that can affect a large portion of the population may be more likely to be included in PLUS. This effect could be due to two reasons; first, these studies may be more likely to be published in higher impact journals, which comprise a large portion of the journals indexed by plus. As well, their higher 'newsworthy' potential may make them more likely to be included in the PLUS database, in comparison to studies that have lower 'newsworthiness'.

Using PLUS in research areas where there are more specialized journals indexed in the PLUS database may be more successful than using PLUS for reviews where there are fewer field-specific journals. While PLUS indexes numerous general medical journals (eg *BMJ, JAMA, Lancet*, or *PLoS medicine*), many other journals are more field-specific. However, some fields are more represented than others. For instance, there are three journals specifically for diabetic-related research (*Diabetes Obesity Metabolism, Diabetic Care,* and *Diabetic Medicine*); yet there is only one journal specifically for the field of neonatology (*Archives of Disease in Childhood, Fetal and Neonatal Edition*). Hence,

articles in these medical fields may have a lower chance of being published in a journal indexed by PLUS, and PLUS might be less beneficial to use when updating these reviews.

5.5. COMPARISON TO PAST RESEARCH IN THE AREA

The following sections give an overview of the similarities and differences between our work and a previous investigation by Hemens et al in 2012, which sought to investigate the performance of PLUS in retrieving articles needed to update a systematic review.

Our approach was very similar to that of Hemens et al. Both studies identified a subset of reviews from the Cochrane database of systematic reviews, with similar inclusion criteria. Further, both studies extracted data and classified included trials in the same way. As well, both investigations had similar definitions for what constitutes an included trial.

Our reviews were specifically those with changed conclusions, while Hemens' work used reviews with updates without specification of changed conclusions. The reasoning behind using reviews with changed conclusions was that a change in conclusion may have been due to the publication of novel research of high methodological content, which may have been more likely to be found in PLUS. Hemens et al also included a lower bound on the date of publication for new trials. The PLUS database began operation in 2003, and does not index articles that were published prior to the start date. In our investigation we did not include such a restriction for trials that we searched in the database. Because of this, we may have included trials in our search that would not be found in PLUS due to their publication data. This may be in theory one reason that accounts for the lower proportion of trials found in PLUS in our investigation as compared to Hemens'. Since our cohort assessed reviews updated in 2012 and 2013, the majority of newly added trials were published after 2003. However, the oldest newly added trial was published in 1979.

Because it was out of the scope of our investigation we did not attempt to measure the performance of other databases, such as EMBASE and MEDLINE as Hemens did. Hence, we were not able to empirically compare the performance of PLUS

to other databases for our sample of reviews. However, Hemens did not find any difference between MEDLINE and EMBASE[7].

With regard to the CQ filters, while both investigations used the CQ sensitive filter for therapy, there were some differences in how the filters were used. In Hemens' study, they sought to determine the overall recall performance of the CQ filters, and used the CQ filters to search for all newly added trials. In comparison, we used the CQ filters to identify the subset of trials which were not found in PLUS. Our intention for the CQ filters was to see how well it could be used in the search schema described in the background section- namely the use of PLUS to prospectively identify trials, followed by the use of CQ filters to aid completing the search update. While we also could have searched for all newly added trials with the CQ filters, we assumed that the trials from Cochrane reviews that were found in PLUS would have also been included with the CQ-sensitive therapy filters, as they were all randomized trials. Moreover, they were already found in PLUS, and in a real life implementation of the process it would be needed to retrieve them. The consequence of our choice is that the evidence we provide cannot be used to support using or not using CQ filters alone, but in tandem with PLUS. Again, we only searched MEDLINE, but in future investigations would also search EMBASE and PsychInfo, to ensure that all possible trials are found.

Finally, we did not determine the effect on the review meta-analyses of excluding trials that were newly added in the update but not found in PLUS. We thought this statistical approach, though sound, was not within the scope of our project. Rather We preferred to propose and pursue a different less elegant but more pragmatic approach to elucidate the effects of added trials in the update, i.e. we characterized the change in statistical conclusion and potential clinical implications of the updated review based on the author judgement and planned to investigate the association of this change with the amount of evidence provided by PLUS.

Being Hemens' study was the first and only in the field, and we considered it important to simply confirm or deny their study results. We too often forget that the statistical interpretation of a study assumes that the results would fall 95% of the cases

in the confidence interval for the effect size, but that by the same assumption, 5 significant studies in 100 are definitely wrong. Until you repeat the study, there is very little certainty about its results, and this is the foundation for the existence of the Cochrane Collaboration itself. Moreover, we introduced three aspects that are novel to our approach. We empirically tested the use of the CQ filters to determine how they would impact the number of articles retrieved from the search strategy given by the author when undertaking an update. We assumed the authors would have received PLUS alerts, or would consider PLUS broadcasted references as a first step. We further compared the performance of CQ filters to results obtained using no filters (ie the content terms) and the search filters used by authors. Finally, in a post-hoc assessment, we attempted to test how review group, as a predictive factor, might modulate PLUS performance.

5.6. LIMITATIONS OF OUR APPROACH

The original scope of our study was to descriptively explore PLUS performance in Cochrane reviews with changed conclusions. We did not aim to statistically assess any differences from chance. Like other investigations, ours was not without limitations. Given the results of the Hemens' analysis, it was possible for us to *a priori* calculate the sample size needed to prove/exclude a statistically significant difference. Traditionally, a sample size calculation is conducted prior to commencing an investigation, and is calculated by taking into account the desired level of significance, the expected event rate, effect size and the margin of error. The calculated sample size then determines how many unit of analysis are needed in the study to identify a difference between two populations, if it exists, with the desired power. Indeed, having an adequately sized sample will decrease the risk of committing a Type I or Type II error, and allows an investigator to confidently make conclusions based on the data.

At the time of the proposal, we did not anticipate conducting any statistical hypothesis testing, so instead we decided to use a convenience sample of all reviews with changed conclusions published in a one-year time frame. Based on previous reports, we anticipated that within a year, there would be approximately 100 reviews with changed conclusions. By including an entire year of reviews with changed conclusions, we hoped that our sample would have been uniform with regard to publishing standards

and would represent the general demographic of reviews that were likely to have changed conclusions.

In our sample, we found 180 of 854 new trials in PLUS over the 92 reviews. This constitutes 71% of the projected sample size we would have needed to detect a difference in proportions. The lower than expected sample size decreased the precision of our estimates. We found that on average there were 9.3 (SD: 14.1) new trials added per review, of which about 20% were found in PLUS. Given these values, we would have needed approximately 40 more reviews to reach the targeted sample size.

We did not perform the extraction of included studies in duplicate, rather the same reviewer re-extracted data from a subset of 30 reviews. Duplicate extraction is generally used for two main purposes: first to increase the accuracy of the data extraction and decrease the chance of random error, and second to mitigate any biases that an individual extractor may have had and decrease the chance of systematic error. In this investigation, the data that were extracted included the study title, year and journal of publication of the article. After initial extraction, the trials were classified into the categories of: newly added, included, excluded and present in both.

Extracting the information about the included trials did require minimal judgement or interpretation, and hence likely would not have been subject to systematic error. While it was possible that there may have been some extraction error in the trial information, the calculated Cohen's kappa statistic to quantify agreement of 0.74 [95%CI:0.68-0.79] between the initially extracted and reassessed reviews suggests that these errors were minimal. However, given that the second component of the extraction-namely the categorization of trials- required some judgement, it is possible that this step would have been prone to bias and some trials may have been misclassified.

Given that the reviews in this investigation comprised 2166 individual trials reported over 3924 references, the time needed to extract the required information was significant, whereas the accuracy of the initial extraction seems quite high. In future similar investigations it would be important to employ duplicate extraction to decrease

the risk of bias; however, it may be pragmatic to limit the duplicate extraction to areas where there is more judgement involved. For instance, a single individual could have extracted the trial information, and a second reviewer could verify the results, then the portion where trials would have been classified could have been completed in duplicate. A past study showed that while extraction in duplicate was associated with fewer errors, it also took considerably longer than single extraction with verification[21]. Hence, for future investigations it may be prudent to use both single and double extractions to balance accuracy and resource use.

We classified change in statistical significance from the original to the updated version in three ways, as stated above. While change in statistical significance and statistical significance based on addition of new information are self-explanatory, we thought that the criteria of 'change in effect size' could have been better defined and elucidated. In our initial design, we quantified a change in effect size as any increase or decrease. This methodology was quite ambiguous and did not consider the clinical significance of this change in magnitude nor did it quantify what magnitude of change would be significant. To evaluate how the change in magnitude could have clinical implications, we may have used measures such as surveying experts in the respective fields to see if the magnitude of change would have any implications on their practice.

Statistically, we might have implemented a cut-off proportion by which the relative risk reduction should have changed from the original to the update. We might have also considered a criterion by which the confidence interval of the estimate should have decreased form the original to the update. To calculate these cut-offs, we might have compared the changes in a subset of updated reviews that had new data added to the comparisons but did not have a change in conclusions to a subset of reviews, in which new data addition did result in a change in conclusion to see if there was a difference in magnitude that would have initiated a change versus not initiated a change. We might have also searched the literature to identify any minimally important clinical differences that are commonly used, and applied those as cut-offs. However, a limitation to all of these approaches is that the clinical areas presented in the reviews are very diverse, hence a decrease of a certain magnitude in one area which would be significant might

not be so in another clinical area. Hence, it would be prudent to try and customize any rule that would be applied for the clinical area of interest.

Despite the potential strategies to more stringently categorize the significance of change in conclusions, as listed above, we chose not to employ such methods in our investigation. Given the small number of reviews in certain areas, we could not have been certain that any findings were not independent from chance.

Another limitation of this investigation was in classifying whether the change conclusion was clinically significant. In our methods section, we stated that we used the review authors' judgements to classify if a change in conclusion was clinically significant or not. To do so we accessed the authors' conclusions for implications for practice, and compared the text between the versions. If the recommendations to use or not use an intervention described in the review had changed, then we considered that a change in clinical implication. However, the review authors could have drawn biased conclusions. For instance, if a review author was also an author on a newly included study, then they may have an intellectual conflict of interest and be inclined to report a conclusion similar to their trial conclusion, to show that their trial changed the conclusions of the review when indeed it did not. While the Cochrane Collaboration has methods in place to assess and present these biases in a transparent way, we cannot discount the possibility that these biases may have influenced the conclusions made by the authors and hence our categorization of change in conclusion.

To further assess the clinical implications in changed conclusions, we might have referenced recent clinical practice guidelines to see how congruent the current guidelines were to the proposed clinical implications, As well, for those reviews that we deemed did have a change in conclusion but were not yet included in clinical practice guidelines, we might have estimated if the proposed conclusions would have changed guidelines if considered by the panel. These methods, would have allowed us to confirm the clinical implications of the review conclusions.

We categorized the change in conclusion for the primary outcome, where listed, and the first given outcome where a single primary outcome was not listed. In doing so, we may have disregarded how a change in conclusion would have been categorized for

other outcomes. However, since the primary outcome is the one that often dictates study procedures, sample size calculation and final conclusions, we thought that categorizing the change in conclusions based on the primary outcome would allow us to assess the primary objective of the review. In future versions, we might consider also looking at secondary conclusions, as these can also have a large impact on clinical practice.

We used PubMed ID numbers to search for articles in the PLUS database. In doing so, we were unable to search for grey literature (abstracts for conference proceedings or protocols) or articles not indexed in Medline with a PubMed ID. However, while searching the 'grey' literature is still an important component of the systematic review process, articles without PMID would likely not have been included in the PLUS database, and hence would not have affected our results.

5.7. AREAS FOR IMPROVEMENT

If this investigation were to be done again, aside from mitigating the limitations as mentioned above, there would be some other changes that we would incorporate. There were a few areas to which we would have expanded the scope of our project. Both our and Hemens' analyses found that PLUS identified about a quarter of trials used to update Cochrane reviews. When doing this investigation again, we would have expanded the scope to include an exploration of factors that may indicate where PLUS would have better performance. In a post hoc analysis of this investigation, we considered how review group would have affected PLUS performance, but did not have any significant findings. However, this could have been either due to there being no difference between review groups or that our sample was not large enough to detect a difference. In completing this analysis again, we would have increased the sample size, and included more reviews from groups that we would assume to be best covered for content need by PLUS content (eg internal medicine, primary practice, etc). Indeed, a potential predictive factor we had no chance to explore might be the amount of relevant content in the area of the review. PLUS indexes both general and area-specific journals, and some areas have more journals indexed in PLUS than others. Hence it may be possible that PLUS will better work for areas that have more journals indexed. Since

PLUS grades articles that are newsworthy and highly relevant, it might be the case that PLUS would perform better in areas where there was more novel research being conducted. This characteristic would be difficult to define, but perhaps one could look at number of articles published in high-impact journals to identify fast growing clinical areas. If preliminary analyses for these factors proved to be significant, it would be interesting to conduct a regression analysis with the outcome of trials found in PLUS to see how different predictive factors contribute to determining PLUS performance.

Some useful information, which we could gather if we were to do this investigation again, would be to elucidate the average timeline between publication and inclusion in a Cochrane review of a trial report. Anecdotal evidence from a study being currently conducted in authors of the Cochrane Musculoskeletal and Upper Gastrointestinal and Pancreatic Disease groups receiving PLUS alerts targeted to their own SR shows that one in two to three of the selected references for the MSK and UGPD groups would have been sufficient to trigger an update of the SR. Then, it would be interesting to measure the time elapsed between inclusion in PLUS and inclusion in the review update. Further it would be interesting to explore if there are certain study characteristics that more often trigger updates. One might conduct a time-to-event analysis to see if there are any significant factors that trigger review updates. Currently the Cochrane Collaboration recommends a review every two years, but it may be the case that some areas should be updated more quickly and some more slowly. Results of an investigation such as the one proposed may help identify trends in updating process, and identify areas that require updating more quickly or slowly.

Investigation of the clinical queries filters in the five reviews investigated showed that use of the CQ filters for the most part reduced the number of search results retrieved without loss in accuracy. Further, the CQ filters for the most part performed better than other non-CQ filters utilized by the review authors. If we were to conduct this analysis again, we might enlarge the number of systematic reviews used to test the clinical query filters against other search filters currently in use to see their performance. Should the finding that the CQ filters perform as well as, or better than, other filters be confirmed in a well sized sample of reviews, one might make the case to use them more extensively in

review searches. Two additional benefits of universally utilising a filter would be consistency of results and predictability of resource requirements.

Outside of expanding the scope of the project, there were some methodological steps, which we would have conducted differently. In this investigation, we did not verify or double extract the information about the statistical change in conclusions or clinical implication of the change in conclusions. While any questionable cases were discussed with another reviewer, the lack of verification could have led to incorrect classification. Should this or similar investigations be completed, it would be important to classify and interpret the information for changed conclusion in duplicate.

Further, to identify articles for the CQ filters we searched only MEDLINE using the MEDLINE filters. However, it is common practice to also search other databases such as the Cochrane database, Embase or PsychInfo, and, indeed, there are CQ filters optimized for Embase and PsychInfo. Were we to do this investigation again, we would have included searches in these two databases, to ensure that we retrieved all possible studies using the CQ filters, and thus most accurately demonstrated their performance. However, at least for the 5 reviews analyzed in detail, all trials included were retrieved in PubMed using the sensitive filters.

5.8. CONCLUSIONS

In our investigation, we tested how PLUS and the CQ filters would perform in retrieving articles needed to drive a change in conclusion in Cochrane reviews. We also explored different factors that may have affected PLUS performance. Our results showed that PLUS captured one in five articles that was used to drive a change in conclusion. Given that this value was similar to the one found in the investigation by Hemens et al, we can infer that PLUS would not be a comprehensive database to that would contain all articles needed to update a systematic review. However, PLUS does seem to have more utility as an alert system to provide new data to authors as it becomes available.
Further, the CQ filters seem to perform well in recalling articles not found in PLUS, but used to drive a change in conclusion of systematic reviews. Prospective studies using

PLUS as an alert system and CQ filters to search should be conducted to prospectively test their use in updating systematic reviews.

Using a system where PLUS was used to alert authors of new data and CQ filters were used when conducting the update would potentially have a two-fold impact on increasing the efficiently of systematic review updates. First, if review authors are prompted to conduct an update based on the availability of new data, then there would be decreased time and resource utilization for conducting an update before it is necessary. For instance, Cochrane mandates a uniform update time of two years. However, for certain subject areas, where perhaps the turnover for studies from conception to publication is greater than two years, or the volume of overall research is low, the two-year update time may be too short. Conversely, for areas where publication of novel research is quicker than two years, the mandated update time may be too long. Reviews falling under both scenarios would benefit from a more tailored indication of when an update is necessary, as might be provided from the system proposed. Secondly, use of the Clinical Query filters may reduce the amount of articles needed to be filtered to find pertinent review articles. The use of this framework would need to be prospectively studied.

SECTION 6: WORKS CITED

1. Moher D, Tsertsvadze A, Tricco A, et al. When and how to update systematic reviews. *Cochrane Database Syst Rev*. 2008;23(1).

2. Druss BG, Marcus SC. Growth and decentralization of the medical literature: implications for evidence-based medicine. *J Med Libr Assoc*. 2005;93(4):499-501. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1250328&tool=pmcentrez&rendertype=abstract. Accessed April 20, 2015.

3. Shojania K, Sampson M, Ansari M, Ji J. Updating Systematic Reviews. Technical Review No. 16. *Prep by Univ …*. 2007;(16). http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Updating+Systematic+Reviews--Technical+Review+16#3. Accessed April 3, 2013.

4. Beller EM, Chen JK-H, Wang UL-H, Glasziou PP. Are systematic reviews up-to-date at the time of publication? *Syst Rev*. 2013;2(1):36. doi:10.1186/2046-4053-2-36.

5. Unalp A, Tonascia S, Meinert CL. Presentation in relation to publication of results from clinical trials. *Contemp Clin Trials*. 2007;28(4):358-369. doi:10.1016/j.cct.2006.10.005.

6. Collaboration TC. Our Principles|The Cochrane Collaboration. *Web Page*. http://www.cochrane.org/about-us/our-principles.

7. Hemens BJ, Haynes RB. McMaster Premium LiteratUre Service (PLUS) performed well for identifying new studies for updated Cochrane reviews. *J Clin Epidemiol*. 2012;65(1):62-72.e1. doi:10.1016/j.jclinepi.2011.02.010.

8. Garritty C, Tsertsvadze A, Tricco AC, Sampson M, Moher D. Updating systematic reviews: an international survey. *PLoS One*. 2010;5(4):e9914. doi:10.1371/journal.pone.0009914.

9. Moher D, Tsertvadze A. Systematic reviews: when is an update an update. *Lancet*. 2006;18(387):881-883.

10. Health Information Research Unit - HIRU ~ McMaster PLUS. 2015. http://hiru.mcmaster.ca/hiru/HIRU_McMaster_PLUS_Projects.aspx. Accessed October 26, 2014.

11. Unit H research information. Health Information Research Unit - HIRU ~ Search Strategies for MEDLINE in Ovid Syntax and the PubMed translation. 2015. http://hiru.mcmaster.ca/hiru/HIRU_Hedges_MEDLINE_Strategies.aspx. Accessed June 30, 2015.

12. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ*. 2005;330(7501):1179. doi:10.1136/bmj.38446.498542.8F.

13. Hunt RJ. Percent Agreement, Pearson's Correlation, and Kappa as Measures of Inter-examiner Reliability. *J Dent Res*. 1986;65(2):128-130. doi:10.1177/00220345860650020701.

14. Alper BS, Hand JA, Elliott SG, et al. How much effort is needed to keep up with the literature relevant for primary care? *J Med Libr Assoc*. 2004;92(4):429-437. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=521514&tool=pmcentrez&rendertype=abstract. Accessed October 25, 2014.

15. Revised Impact Factor announced for Cochrane Database of Systematic Reviews | The Cochrane Collaboration. http://www.cochrane.org/news/news-events/current-news/revised-impact-factor-announced-cochrane-database-systematic-reviews. Accessed October 26, 2014.

16. 2013 Impact Factor released for Cochrane Database of Systematic Reviews | The Cochrane Collaboration. http://www.cochrane.org/news/tags/authors/2013-impact-factor-released-cochrane-database-systematic-reviews. Accessed October 26, 2014.

17. Wilczynski NL, Mckibbon KA, Haynes RB, Information H. Search Filter Precision Can Be Improved By NOTing Out Irrelevant Content. :1506-1513.

18. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Inform Assoc*. 9(6):653-658. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=349381&tool=pmcentrez&rendertype=abstract. Accessed November 23, 2014.

19. Toth B, Gray JAM, Brice A. The number needed to read-a new measure of journal value. *Health Info Libr J*. 2005;22(2):81-82. doi:10.1111/j.1471-1842.2005.00568.x.

20. Patsopoulos NA, Ioannidis JPA. The use of older studies in meta‐analyses of medical interventions: a survey. 2009;3(2):62-68.

21. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. *J Clin Epidemiol*. 2006;59(7):697-703. doi:10.1016/j.jclinepi.2005.11.010.