A Large-Scale Analysis of the Reasons Given for Excluding Articles that are Retrieved by Literature Search During Systematic Review

Tracy Edinger, ND, MCR, Aaron M. Cohen, MD, MS
Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

Abstract

Objective: *A literature search to identify relevant studies is one of the first steps in performing a systematic review (SR) in support of evidence-based medicine. To maximize efficiency, the search must find practically all relevant studies and retrieve few that are irrelevant; however, this level of precision is seldom attained. Therefore, many articles must be manually examined for relevance. To better understand the limitations of current search tools as applied to SR, we characterized the most common reasons that papers retrieved by SR searches were excluded from the review.* ***Methods:*** *The textual reasons given for retrieved but excluded articles were extracted from 6,743 SRs performed by 54 Cochrane Collaboration review groups. The frequencies of different exclusion reasons were analyzed, and we developed a taxonomy summarizing these reasons.* ***Results:*** *Almost 65% of articles were excluded because the means of comparison were inappropriate. Of these, about 72% were due to the randomized controlled trial (RCT) design being required but not employed by the excluded study. Mismatching interventions and outcomes and incorrect population characteristics were also common reasons for exclusion.* ***Conclusions****: Currently available search methods do not adequately address the most common exclusion reasons for systematic review, even those based primarily on study design.*

Introduction

Systematic reviews (SRs) are literature reviews "designed to locate, appraise, and synthesize the best-available evidence from clinical studies of diagnosis, treatment, prognosis, or etiology, and provide informative empirical answers to specific medical questions."[1] The practice of evidence-based medicine (EBM) is dependent upon clinicians having ready access to the best-available primary evidence applicable to their patients.[2] SRs and meta-analyses (MA) make the available evidence more accessible and usable in clinical practice. SRs inform medical recommendations, guiding both practice and policy, such as in the creation of published practice guidelines.[3]

The Cochrane Collaboration states that an SR:

> "…attempts to collate all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimizing bias, thus providing more reliable findings from which conclusions can be drawn and decisions made."[4]

The process of creating and maintaining SRs is resource- and labor-intensive, typically requiring 6-12 months of effort, the main expense being the time of expertly trained personnel. Review updates take about as much effort as a first time review on a new topic.[5] Once the topic is determined and inclusion criteria are defined, reviewing papers for inclusion is a time-intensive process. Articles in bibliographic databases such as MEDLINE are manually indexed with metadata such as key concepts and study design. Literature searches can utilize these indexes to improve recall and precision. An ideal literature search would retrieve all relevant papers for inclusion and no irrelevant papers that need to be excluded. However, previous research has demonstrated a number of studies that are not fully indexed, as well as a number that are indexed incorrectly.[6] These factors diminish the effectiveness of literature searches by decreasing the accuracy of keyword and index-based search queries. In order to better understand what approaches to improving literature retrieval for EBM might be most useful, we did an analysis of the most common reasons that articles are excluded from an SR.

The Cochrane Collaboration conducts systematic reviews of all relevant studies on many different topics. There are over 50 Cochrane groups that perform reviews in their area of specialty. Reviewers must conduct thorough literature searches to obtain all relevant papers. Search strategies range from simple to fairly complex, utilizing MEDLINE medical subject headings (MeSH) terms, keywords, and complex search logic. The number of articles retrieved depends both on the topic and on the search; some searches will retrieve thousands of articles. Once the list of

papers is retrieved, the studies must be examined for relevance. Several levels of examination are necessary to obtain the final set of relevant papers to include in the review. Some studies can be excluded by reviewing simply the title or abstract. Other studies must have the complete published article pulled for full text review in order to determine relevance. This number varies but can be as large as several hundred per review. Exclusion of studies after full-text review is particularly labor intensive; the ability to exclude them earlier in the process or not retrieve them at all would be a great advantage to the literature review process.

We obtained an XML data set of all Cochrane reviews through August 2011, including a list of studies that were excluded from each of the SRs after the article's full text was pulled and read, along with a textual comment stating the reason for each exclusion. The excluded articles listed in the Cochrane XML files include only those papers that made it through the initial title and abstract screening and were determined to be excluded only after pulling and reviewing the full text article. The current study seeks to determine and characterize the primary reasons these retrieved studies were not included in systematic reviews.

Methods

We obtained data from reviews performed by a total of 54 different Cochrane study groups. Data from each review was stored in an XML file. Each file contained lists of included and excluded articles; inclusion status for each article was based on review of the full text article. Reasons for exclusion were given as free-form text for articles not included in the reviews. The complete dataset contained 6,743 XML files and a total of 83,588 excluded articles.

We used a Python (www.python.org) script to parse the raw XML files, extracting the listed exclusion reasons from the *CHAR_REASON_FOR_EXCLUSION* tag associated with each study ID in each review. Manual inspection of the merged lists showed that many exclusion reasons were common to multiple reviews but appeared slightly different due to differences in spelling or phrasing. To address this, we extended our script to conduct some low-level textual normalization procedures. We first converted all exclusion reasons to lower case and stripped any leading or trailing whitespace. We then split the text entries by periods and semi-colons in order to individually count multiple exclusion reasons applied to a single article in a single SR. We also normalized the spelling of "randomised" to "randomized" and "rct" to "randomized controlled trial." Reasons containing "randomized controlled trial" and the word "not" were normalized to "not a randomized controlled trial."

Exclusion reasons were counted and grouped together in an Excel spreadsheet. For each exclusion reason, the spreadsheet listed the number of articles, the free-text exclusion reason, the number of Cochrane reviews that used that reason, and the names of the XML files for those reviews. In some cases, multiple reasons were listed on one line because they were not identified as separate reasons by our low-level textual normalization. In those cases, we duplicated that line and separated each exclusion reason onto its own line, giving each line the original count as we assumed that each listed free-text reason pertained to all of the papers to which it was applied.

In order to keep the manual review task feasible and focus on the most commonly used exclusion reasons, we filtered out any exclusion reason that was used to exclude fewer than ten papers. This yielded a total of 710 distinct "common" reasons. If the reasons were not specific enough to determine why the articles were excluded, we looked at the review XML file and the full text of the Cochrane review for more detail.

To assist the organization of the exclusion reasons into concepts, we utilized the PICO model, the evidence-based medicine (EBM) framework for clinical questions.[7] This model contains four elements to assist in framing clinical questions: Population, Intervention, Comparison, and Outcome. We chose the PICO framework for several reasons. Study inclusion and exclusion criteria for Cochrane and other SRs are organized by PICO concepts, so reasons for incorrect retrieval would be likely to fall into similar categories. In addition, using this framework would be beneficial in identifying concept categories on which development of innovative natural-language processing (NLP) and advanced information retrieval (IR) techniques could focus in order to improve document-retrieval systems.

We developed a small PICO-based taxonomy to group the exclusion reasons, classifying them as a more detailed specific reason under the top-level PICO categories population, intervention, comparison, and outcome. We added a fifth group for reasons that did not fit into a PICO category; examples were papers that were not available to the SR team, and those articles that contained data already presented in another included paper. We developed the categories iteratively, discussing each version of the taxonomy as a team and adding new categories or refining the current ones as needed to best describe the data while keeping the size of the taxonomy small.

Once the exclusion reason categories were developed, the total number of papers using each exclusion reason was calculated. A tree bubble diagram was constructed grouping categories by PICO code and sizing the bubbles relatively according to the sum of the counts in each category and sub-categories. Because there was a large numeric range between the smallest and largest counts (10-5,498), we used the square root of the count to render a legible figure while maintaining a useful sense of scale.

Results

We found 45,587 total unique text string exclusion reasons, of which 710 appeared as the exclusion reason for at least ten publications. These 710 unique exclusion reasons covered a total of 28,012 individual excluded articles out of a total of 84,229 excluded articles. Of the exclusion reasons that we did not annotate, 39,503 appeared only once. We reviewed these briefly and determined that many of these describe very specific and detailed differences between the clinical protocols used in the publication research versus the desired systematic review inclusion criteria. We believe that the specific nature of these differences makes them less amenable to general analysis and a more difficult target for automated tools than the more common and general exclusion reasons we found. These very specific differences also suggest that these are articles that needed to be examined in full text prior to determining inclusion.

Table 1 details the codes and textual definitions for the final taxonomy we created to characterize the reasons used to exclude articles from inclusion in systematic reviews. Table 2 lists the categories of exclusion reasons, grouped by PICO code, and the number of papers in each category and top-level PICO category; each count includes the count of all subcategories, if any.

Below we review the exclusion reasons and frequencies under each of the top level PICO categories:

Population. About 8% of retrieved papers were not included in reviews because of characteristics pertaining to the study population. Most papers in this category were excluded because of a variation in the disease of the patients to be included in study. In some cases, reviews were of treatment-naïve subjects, and studies of subjects who had previously received a treatment were excluded. Other papers were excluded because the participants did not have the same disease, the same stage of disease (for example, major versus moderate depression, or first malignancy versus recurrence), or the correct combination of disease and comorbidities. Retrieving studies of the correct age group was the next largest issue; most reviews included studies of either adults or children, but not both, and excluded papers with the wrong age group. Some reviews specified the diagnostic process and excluded studies using different methods. Other reviews specified outpatient interventions and excluded studies of inpatients. Other issues included studies of the wrong species, laboratory studies, studies of the wrong sex, and sample sizes that were smaller than the specified minimum.

Intervention. Several aspects of the intervention emerged as problematic areas in document retrieval; this group of exclusion reasons contained slightly less than 15% of excluded papers. In some cases, reviewers required a specific duration for the intervention and excluded those with shorter durations. Some reviews included very specific interventions in combination with specific (or requiring no) other interventions, excluding studies that did not use the correct combination of interventions. Other reviews excluded studies using a different dose than the one specified by the review criteria or a different route of administration. Reviews of procedures excluded studies in which different procedures were used in addition to the one of interest. There were also several reviews that specified who provided the intervention and excluded studies in which the wrong person or practitioner administered the intervention. Examples include reviews of lay- or peer-led groups, nurse-delivered interventions, and self-administered interventions. Reviewers excluded eleven articles because insufficient details were available to determine whether the intervention met the inclusion criteria.

Comparison. Almost 65% of retrieved papers were excluded because of issues related to comparison methods. In many cases, the problem pertained to the study design, with reviewers retrieving studies that were not blind/double blind, not randomized, not controlled, and/or not clinical trials. A number of studies were excluded because they did not use the correct control, or did not use a placebo. Several reviews excluded studies because they used the wrong study design; these included "before and after study," "used an ab rather than an aba design," crossover studies, and prospective observational studies. Case reports and case series were excluded from a number of reviews requiring different study designs such as randomized controlled trials. Review articles and commentaries were excluded from a number of reviews for similar reasons. Overall the largest single reason for an article to be excluded was due to not being deemed a randomized controlled trial when the SR inclusion criteria required it.

PICO Category	Taxonomy Code	Description
P	wrong setting	Something about the research setting was not correct. In most cases, this was an outpatient setting rather than inpatient.
	wrong previous treatment	Participants in these studies had previously received treatments that were excluded in the review article, or they were healthy when the review wanted only participants with a specific condition. For example, several reviews were of new interventions in treatment-naive patients; studies that enrolled participants already using those treatments were not included.
	wrong disease	Participants in these studies had a condition that was excluded, or it was not the same condition being evaluated in the review.
	wrong stage of disease	Participants in these studies had the desired disease but were at the wrong stage.
	wrong diagnostic process	Participants were not diagnosed using the desired criteria. For example, depression diagnosis without using DSM criteria.
	wrong comorbidity	Participants had a comorbid disease that was excluded from the review, or they did not have an included comorbidity.
	wrong age group	Participants were not in the desired age range. In most cases, the desired population was children (or adults) and the excluded study enrolled adults (or children).
	wrong sample size	The sample size was too small for inclusion in the review.
	wrong species/lab study	The study did not evaluate humans, or it was an in-vitro study.
	wrong sex	The study enrolled the wrong gender.
I	wrong duration	The duration of the intervention or the length of the study was not long enough for inclusion in the review.
	wrong intervention	The study used the wrong non-drug intervention.
	multiple interventions	The study used multiple interventions, but the review looked at only one.
	wrong drug intervention	The study used the wrong drug.
	wrong person delivering intervention	The intervention was not delivered by the person specified in the review criteria. For example, a review of nurse-delivered interventions excluded studies of home care performed by the patient.
	wrong procedure	The study did not use the desired procedure. For example, a review of laparoscopic cholecystectomy excluded studies of open cholecystectomies.
	wrong dosing/administration route	The dosing schedule was not the one desired, or the intervention was not administered via the desired route.
	insufficient intervention data	Intervention was not described in enough detail to determine the study should be included.
C	not randomized	The design of the study did not include randomization.
	not controlled	The design of the study did not include a control group.
	not a clinical trial	The study was not a clinical trial
	not an RCT	The design of the study was not a randomized controlled trial.
	not blind/double-blind	The design of the study failed to include required subject or researcher blinding.
	no placebo	The study design did not include a required placebo control.
	wrong control	The study used a control but not the one desired.
	wrong/unspecified design	The design of the study was not the one desired.
	case report/series	This was a case report or series, not a trial.
	review/comment	This was a review or comment, not a trial.
O	wrong outcome	The study used the wrong outcome.
	lack of outcome data	The study reported insufficient outcome data to extrapolate for the review.
	inadequate data analysis	The study utilized inadequate data analysis.
	lost to follow up	Too many participants were lost to follow up.
N	article not available	The staff performing the systematic review were unable to obtain the full text of the article.
	duplicate data/study	Data reported in this study were already reported in another included study.
	wrong language	The study was published in a language not accessible to the reviewers.
	article published in wrong year	The article was published prior to the range of years specified in the review criteria.
	can't determine exclusion reason	The reason for exclusion can not be determined.

Table 1. Final PICO-based (**P**opulation, **I**ntervention, **C**omparison, **O**utcome) taxonomy of reasons used to exclude articles from systematic reviews. **N** = not a PICO-based exclusion reason.

PICO Category Primary Exclusion Reason	Count	Secondary Exclusion Reason	Count
Population	2,145		
Wrong setting	181		
Wrong previous treatment	115		
Wrong disease	1,284	Wrong stage of disease	570
		Wrong comorbidity	74
Wrong diagnostic process	21		
Wrong age group	404		
Wrong population	15		
Wrong sample size	47		
Wrong species/lab study	64		
Wrong sex	14		
Intervention	4,037		
Wrong duration	1,172		
Wrong intervention	1,437	Multiple interventions	117
Wrong drug intervention	605		
Wrong person delivering intervention	442		
Wrong procedure	174		
Wrong dosing/administration route	196		
Insufficient intervention data	11		
Comparison	18,143		
Not an RCT	12,957	Not randomized	4,061
		Not controlled	2,215
		Not a clinical trial	1,184
Not blind/double-blind	328		
No placebo	756		
Wrong control	742		
Wrong/unspecified design	3,360	Case report/series	1,165
		Review/comment	1,573
Outcome	3,072		
Wrong outcome	2,975	Lack of outcome data	1,749
Inadequate data analysis	11		
Lost to follow up	86		
Not a PICO-based exclusion reason	615		
Article not available	203		
Duplicate data/study	135		
Wrong language	10		
Article published in wrong year	21		
Can't determine exclusion reason	246		

Table 2. Article counts of PICO codes, exclusion categories and subcategories.

Outcome. Approximately 11% of incorrectly retrieved articles were discarded because of issues with the outcome. Several reviews required specific outcome data to be present, and excluded studies that did not record sufficient outcome data. In some cases, the data analysis was not of the quality required for inclusion. Some reviews required minimal loss to follow up and excluded studies with high rates of loss to follow up.

Not a PICO-based exclusion reason (other reasons). Exclusion reasons for slightly more than 2% of the articles did not fit into any PICO category. Several reviews excluded articles that presented data already presented in other included articles. In other cases, articles were not available to the reviewers, or the articles were not available in English. Several reviews specified a time frame for studies and excluded those done prior to that time. In about one-third of the studies (246 articles) we placed in this category, not enough information was given to determine the reason for exclusion, even when we looked more closely at the review, abstract, and inclusion/exclusion criteria.

Examples of exclusion reasons in this category are "does not meet inclusion criteria," "non systematic review article," and "guidelines." If more detailed exclusion text were provided, these studies would likely have fit into one of our PICO reasons.

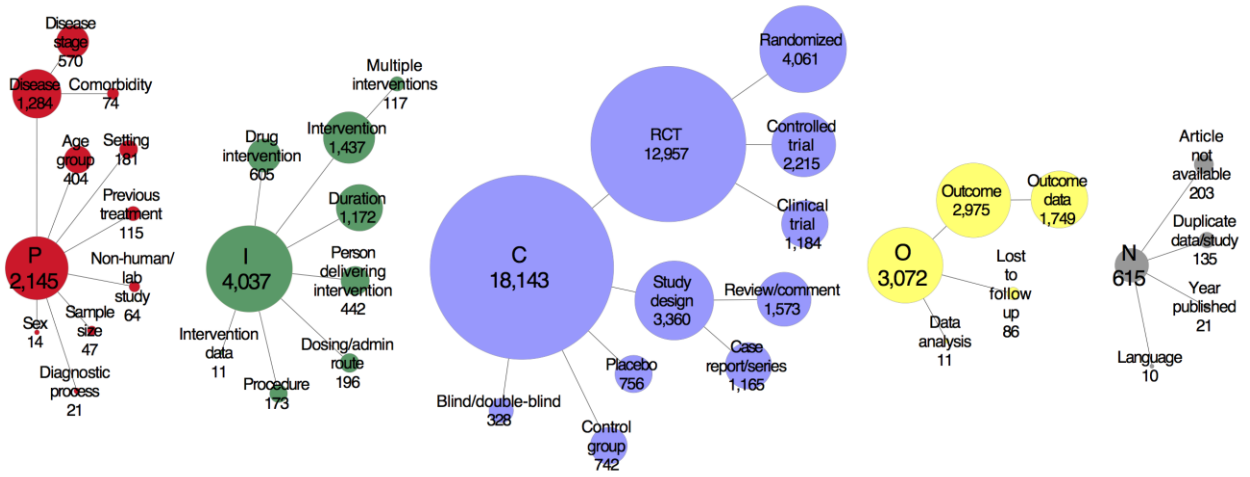

Figure 1. Tree bubble diagram of the relative frequency of exclusion reasons, by top level PICO category. In order to keep the diagram readable we have abbreviated the exclusion reasons listed in Table 2 by removing negation terms such as "not" and "wrong."

Figure 1 depicts the relative number of articles excluded in each category. The size of the nodes reflects the count for each category plus the counts for any lower level categories below (to the right in the figure) that node. Lines connecting higher level nodes to lower level nodes show the parent/child relationship between higher level exclusion reasons and more detailed reasons. Visualizing the relative usage frequency of each exclusion reason and PICO category in this manner makes it easy to compare exclusion reasons and categories. The diagram clearly illustrates that the majority of articles were not included for reasons related to the comparison, and most of these had problems with study design. The intervention was the second largest category, and articles addressing the wrong intervention, drug or non-drug, or the wrong duration of treatment made up the majority. Problems related to the population and outcome were somewhat less common at approximately equal rates. A very small proportion of the exclusion reasons did not fit into our PICO-based taxonomy.

We further examined some of the issues surrounding exclusion reasons in greater detail. Because an RCT is such a common requirement of a SR and randomization and control are important features of a clinical trial, the ability to reliably filter articles on this basis could greatly facilitate narrowing a literature search. We specifically examined the reviews that included RCTs only, as stated in the SR inclusion criteria for that review. First we considered included articles and counted the number indexed with the "Randomized Controlled Trial" publication type in MEDLINE. We found that more than 16% of articles indexed as non-RCTs were, in fact, RCTs as determined by the fact that these studies were included in SRs that specifically required RCTs as part of their inclusion criteria. We then looked at the excluded articles for these SRs where the exclusion reason was given as "not an RCT" and found that more than 12% were indexed as "Randomized Controlled Trial" in MEDLINE. These data corroborates the findings of Wieland et al. who also found that RCT publication type annotation was not consistent enough for SR search filtering.[6] This data is summarized in Table 3.

	RCTs included in review (true RCTs according to SR group)	**Articles excluded for not being an RCT (not RCTs according to SR group)**
MEDLINE Publication Type	Count (%)	Count (%)
RCT	5412 (83.6)	**170 (12.5)**
Not RCT	**1061 (16.4)**	1188 (87.5)

Table 3. Correspondence between article index in MEDLINE with the "Randomized Controlled Trial" publication type and inclusion or exclusion in Cochrane systematic review for being or not being a randomized controlled trial in systematic reviews requiring randomized controlled trial as an inclusion criteria.

We also found that 6% of the articles included in Cochrane reviews were not assigned the MeSH term *Human*. Presumably these articles are enough about humans to include in a systematic review about human disease. Currently it is unclear to us why these articles were not tagged as *Human*. While it is possible that some of these articles are about systematic review methods, we manually inspected several dozen articles in this group and they were all included in SRs relating to clinical interventions, and most of the other articles included in these SRs were tagged with the MeSH term *Human* (see Table 4).

MEDLINE	Count	Percent
Humans[mh]	43508	94%
NOT Humans[mh]	2773	6%
Total	46281	100%

Table 4. Correspondence between articles indexed with the *Humans* MeSH term and inclusion in a Cochrane systematic review. Table shows included articles in systematic reviews, and whether they are assigned the Humans *MeSH* term in MEDLINE.

Discussion

We examined articles that had been retrieved in literature searches, passed the initial screening of title and abstract, and were pulled in full text prior to determining they did not meet inclusion criteria. The results suggest that retrieving studies matching the inclusion requirements maps nicely to the PICO framework, and this framework describes the vast majority of areas of difficulty. Furthermore, of the PICO categories, the issues involving the means of comparison, that is, identifying articles with appropriate study designs, are the most problematic. Previous work has demonstrated significant rates of inaccuracy in MEDLINE indexing and tagging by study type.[6]

The ability to screen for randomization, controls, clinical trials, and RCTs with high accuracy would greatly enhance search efficiency, as would the ability to filter out case reports and commentaries. While this issue created the greatest amount of review exclusion work for systematic reviewers, it is somewhat surprising that this is the case. MEDLINE supports specific publication types and MeSH terms to support retrieval by study design, yet it appears that the searchers when performing an SR are not using these terms in their search criteria. This may be due to the fact that literature search for SR is by necessity a very high recall oriented task. While MEDLINE indexing is very extensive and useful for the vast majority of users, a 15% disagreement on the status of papers being or not being an RCT is very significant to those conducting an SR and needing to collect virtually all of the available evidence. MEDLINE searching via MeSH terms and publication types is extremely useful to a large number of users, however, compared to systematic reviewer experts, most users are likely more interested in high precision, rather than high recall. Identifying a few high-quality studies relevant to a patient's care is a typical high precision search task for a physician. Identifying all of the studies relevant to set of very specific inclusion criteria is the high recall task for systematic reviewers, and is a very different task for an information retrieval system to support.

Exhaustively annotating all publications for all relevant MeSH terms, publication types, and meta-data would be extremely resource intensive. Furthermore, the MEDLINE annotation process has been found to be only about 50% consistent on main MeSH headings across multiple annotators.[8] This implies both that there is some disagreement over which terms are most important as well as correct annotations missing that would have been applied if a different annotator had reviewed the article. Automated means of aiding MeSH annotators have been developed, with best performance centering on proposing 20 annotations per article.[9] This number is likely not enough to exhaustively annotate all articles for relevant terms. Other researchers have also found variability and inconsistency in the assignment of MeSH terms in very specialized domains.[10,11] A more efficient and flexible, in terms of recall/precision tradeoffs, means of assigning annotations (e.g., whether an article is about an RCT) in a subject domain is required for SR users with their very high recall requirements.

Fortunately, at least in some very common cases, identifying practically all the articles meeting a specific study design should be a reasonable problem to solve with text mining and NLP techniques. These approaches could be used to develop a set of automated study design annotation labels for articles in MEDLINE and other databases. Because the characteristics of study designs are fairly well understood and have stable definitions, NLP tools can be developed to accurately identify the factors suggesting a particular study design and provide a confidence level for the identification of that design. We have done preliminary work showing that the RCT study design can be identified with 96% recall and 94% precision using a combination of text classification techniques similar to that used in our prior work on topic-specific classification for SR.[12] According to the work of Wieland et al.,[6] this is a

higher level of accuracy than the MEDLINE publication type annotations. Furthermore, the user can determine how low to go in the confidence of automated predictions to accept that a study is of the appropriate type based on the task, the topic, or the total size of the literature base. Identifying randomized studies, whether controlled or not, could also be a useful automated annotation for SR. The definition of randomization can include different strategies, so this feature might be more difficult to identify accurately in all situations; however, a standard definition could be used to develop an NLP tool.

The second-largest problem area involved characteristics of the intervention. Most of the problems occurred with retrieving studies with the correct non-drug intervention. Exclusion reasons varied widely depending on the SR topic: "not manual therapy," "not exercise training," "other intervention than length of bed rest." Duration of treatment was also problematic, with many articles excluded for durations that were too short. Automated identification and annotation of study and treatment duration is a more complex problem than study design annotation. However, there has been significant prior work in the area of temporal extraction from text, and that work should be applicable here.[13]

Lack of outcome data or the use of the wrong outcome in the article's study were reasons for excluding about 10% of papers. Issues with the population accounted for slightly more than 7% of excluded papers; about half of these were excluded because of the wrong disease or wrong stage of disease. Studies excluded for using the wrong age group comprised about 19% of this category. Certainly automated identification of the age range for study subjects seems tractable and could be a useful improvement for narrowing the results of a search.

Although we found a broad range of exclusion reasons, a small subset of these reasons accounted for the vast majority of papers that were pulled and then excluded. These pulled and then excluded papers account for the majority of extra work required by the review experts in the SR literature collection process. Initial work in modifying information retrieval for systematic review and evidence-based medicine should start with these most common reasons: study design, intervention duration, population age group. Large scale automated tagging in MEDLINE and other literature databases would allow greater flexibility in retrieving articles according to the needs of systematic reviewers.

Limitations

Our study has several limitations. In this work, we utilized a list of reasons for excluding studies from Cochrane reviews and did not consider reasons for excluding studies from reviews done by other organizations. Although we think that the reasons given by the groups of the Cochrane Collaboration were representative of SRs, it is possible that other types of reviews have different reasons for excluding papers or different distributions of reasons. Also, because of time limitations, we only looked at exclusion reasons used for at least ten articles, resulting in a sample that constituted only 1.56% of unique exclusion reasons but a full 33.26% of excluded papers. Thus, our sample contains only a fraction of the exclusion reasons; however, the remaining unexamined reasons each apply to only a very small number of papers. In spite of this, it is likely that our taxonomy does not cover all potential exclusion reasons.

Development of the taxonomy was a group-based iterative process; however, only one author (TE) looked at the original exclusion reasons and topic-specific systematic review criteria to develop the initial list. In some cases, the text reason given by the Cochrane reviewers was very brief. When the meaning seemed clear and unambiguous, we assigned the reason to one of our categories. It is possible that we misinterpreted some of these very brief descriptions.

Conclusions

This study provides insight into opportunities to improve search, retrieval, and text processing systems for EBM and SR. The search strategies were not able to incorporate enough detail to cover the most common exclusion reasons. While simple keyword or metadata search terms were insufficient to accurately filter articles not meeting the most common exclusion reason, not being an RCT, more sophisticated approaches could be applied to significantly reduce the manual work in screening out these studies.

Acknowledgements

The authors wish to acknowledge the Cochrane Collaboration for providing the review process data XML files on which this work is based. The authors also wish to acknowledge Kyle Ambert for his assistance with the pre-processing of the exclusion reason text. This work was supported by grant number 5R01LM010817 from the National Library of Medicine.

References

1. Cohen AM, Ambert K, McDonagh M. Studying the potential impact of automated document classification on scheduling a systematic review update. BMC Medical Informatics and Decision Making. 2012 Apr 19;12(1):33.

2. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71–2.

3. Haynes RB. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. Evid Based Med. 2006;11(6):162–4.

4. Higgins, JPT, Green, S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011] [Internet]. 2011;Available from: www.cochrane-handbook.org

5. Moher D, Tsertsvadze A, Tricco AC, Eccles M, Grimshaw J, Sampson M, et al. When and how to update systematic reviews. Cochrane Database Syst Rev. 2008;(1):MR000023.

6. Wieland LS, Robinson KA, Dickersin K. Understanding why evidence from randomised clinical trials may not be retrieved from Medline: comparison of indexed and non-indexed records. BMJ. 2012;344:d7501.

7. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc. 2006;:359–363.

8. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Libr Assoc. 1983 Apr;71(2):176–183.

9. Huang M, Neveol A, Lu Z. Recommending MeSH terms for annotating biomedical articles. J Am Med Inform Assoc. 2011;18(5):660–667.

10. Portaluppi F. Consistency and accuracy of the Medical Subject Headings thesaurus for electronic indexing and retrieval of chronobiologic references. Chronobiol. Int. 2007;24(6):1213–1229.

11. Wilczynski NL, Haynes RB. Consistency and accuracy of indexing systematic review articles and meta-analyses in medline. Health Information & Libraries Journal. 2009;26(3):203–210.

12. Cohen AM, Ambert K, McDonagh M. Cross-topic Learning for Work Prioritization in Systematic Review Creation and Update. J Am Med Inform Assoc. 2009;16(5):690–704.

13. Li Z, Liu F, Antieau L, Cao Y, Yu H. Lancet: a high precision medication event extraction system for clinical text. J Am Med Inform Assoc. 2010 Sep 1;17(5):563–567.