INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Université d'Ottawa · University of Ottawa

The Influence of 'Grey' Literature on Meta-Analysis

by

© Laura M. McAuley

Thesis submitted to
the School of Graduate Studies and Research
in partial fulfillment of the requirements for the
MSc degree in Epidemiology

University of Ottawa

October 1999

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-46592-6

Canada

Abstract:

Introduction

The impact of the inclusion/exclusion of grey literature in meta-analysis (MA) is unclear.

Objectives

To investigate, in a sample of published MA; the prevalence of grey literature, the quality of reporting at the MA and trial levels, and the impact of grey literature on the point estimate and precision of the results.

Methods

Analysis of Variance and regression models were used to consider the quality of reporting and the impact of grey literature on estimates of efficacy.

Results

Grey literature was included in 33% of the MA. Grey inclusive MA tended to be of higher quality than those that excluded it. At the trial level, grey literature was of lower quality than published literature. The exclusion of grey literature led to increases in both the reported effectiveness of the intervention and the precision of the results.

Conclusions

MA that exclude grey literature run the risk of producing biased estimates of intervention effectiveness. Grey literature must be sought and included when it meets pre-defined inclusion criteria.

Acknowledgements:

I would like to thank my husband Michael for reading various drafts and spotting errors I had missed, but especially for acting as a support, for always encouraging me and for always knowing I would succeed even when it was not clear to me. Without his patience and understanding this work would not have been achieved. Our son Spencer for providing me motivation to complete what I had started and for taking naps so mom could work. My family for always believing in me and helping me to believe in myself, for encouraging me to set ambitious goals and then strive to meet them. My in-laws for their encouragement and understanding that sometimes my thesis had to come first.

My thesis advisors David Moher and Dr. Peter Tugwell for supporting and encouraging my ideas, providing me with guidance and valuable feedback and for helping to keep me motivated and focused. For encouraging me to present my work in progress and for helping to support, both moral and financial, such endeavors at various arenas.

Ba Pham for his statistical support and interest in my project, for his patience and ability to help me find the answers to my questions. For reading my analysis plan and helping refine the analysis, and finally for reading drafts and ensuring I had correctly stated my statistics.

Alison Jones for being both a friend and a resource person. For helping with the retrieval and blinding of the trials and for countless other areas where she helped out.

Dr. Ian Graham for encouraging me, being willing to discuss and listen to ideas and problems. For helping me to think things through and mold my ideas.

Dr. Assendelft, Dr. Antonio Saenz, Dr. Rokuro Hama, and Elena Telaros for helping with translation and quality assessment of non English RCT's.

Irene and Susan at the Ottawa Civic Hospital library for helping me track down papers.

All the meta-analysts and methodologists who responded to my request and provided me with grey literature.

Elaine Paice for helping with formatting and proof reading.

And finally all of the people who have expressed interest in my research.

Table of Contents:

List of Tables

List of Figures and Illustrations

List of Abbreviations

MA	meta-analysis or meta-analyses
SR	systematic review
RCT	randomized controlled trial
OR	odds ratio
ROR	ratio of odds ratios
MEDLINE	Medical Literature Analysis Retrieval System On-Line
EMBASE	Excerpta Medica Data Base
SAS	Statistical Analysis System
SPSS	Statistical Package for the Social Sciences
ANOVA	analysis of variance
I	implicit exclusion of grey literature
E	explicit exclusion of grey literature
CI	confidence interval
JAMA	Journal of the American Medical Association
BMJ	British Medical Journal

I Introduction

I-1 Meta Analysis

There are over 20,000 biomedical journals that publish more than 2 million articles annually <1>. This volume makes it difficult for health care providers, researchers and policy makers to keep up-to-date. This abundance of new research evidence has led to the increasing prevalence of systematic reviews and meta-analyses as methods to efficiently integrate information and make it more accessible (Figure 1).

Meta-analyses (MA) and systematic reviews (SR) differ only in the use of statistics to combine the results of the included studies. A systematic review is a review in which evidence is systematically identified, appraised and summarized according to a pre-defined explicit methodology. Meta-analysis is the process of using statistical methods to combine the results of at least two independent studies. The aim is to integrate the findings, statistically pool the data, and report the overall results. Systematic review and meta-analysis differ from traditional reviews in their rigor. They require data be provided on which their conclusions are based. Evidence must be provided to assure the reader that all relevant data was included, and details and reasons for omissions must be clearly stated <2>.

I-1-1 History

The use of meta-analysis to integrate research findings is not new. The earliest known example of its use was in 1904, when Karl Pearson reviewed the evidence for the efficacy of the typhoid vaccine <3>. He combined the results of 5 studies on the typhoid vaccine

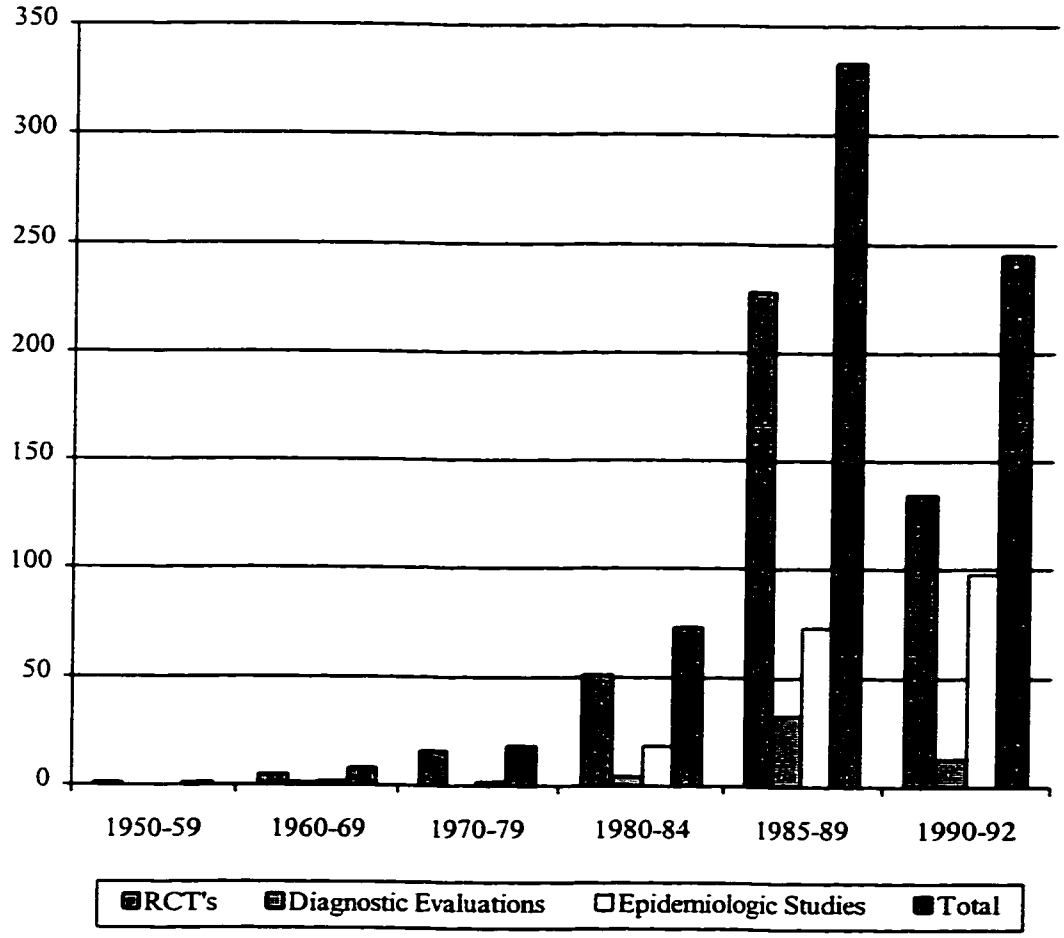


Figure 1: The Evolution of Meta-Analysis in the Medical Literature <2>
This figure depicts the exponential increase in the publication of meta-analyses since the 1980's.

and immunity, and 6 studies on inoculation status and mortality among those who contracted the disease. Pearson's results were published in the British Medical Journal <4>. Although this early meta-analysis is in medicine, the technique had an earlier popularity in psychology and education <5,6>. It has been suggested that the technique gained its popularity in the social sciences as researchers, often faced with hundreds of studies on a given topic, had more pressure to refine summarization techniques to generate specific simple recommendations <7>. In the 1970's, researchers in psychology and education, as in other fields, began to question the existing methods of synthesis and integration of findings <6>. These methods included narrative reviews, box score analysis, and secondary replicate analysis <8>. Narrative reviews tend to be opinion-based with no specific data or methodology provided. Box score analysis is a simple tally method; it is a comparison of the number of studies that confirm the hypothesis to the number that refute it. And finally, secondary replicate analysis involves the collection of original data, which is then pooled for greater statistical power. In 1971, Light and Smith <9> recognized the value of the information gained from combining studies with variation in outcomes, provided the variation was treated with care. Although early meta-analyses in medicine exist, it was not until the late 1980's that it has really emerged as a technique in the medical literature <2>. As depicted in Figure 1, the use of meta-analysis as a technique in medicine has grown exponentially since the 1980's, with meta-analyses of randomized trials leading the way <2>. As a relatively new technique for medical research, meta-analysis has many critics <10,11,12,13> and many proponents <1,14,15,16>. In fact, the debate over the merit of meta-analysis has been mirrored in the lay literature <5,17,18>.

I-1-2 Advantages

MA provides a compilation of large amounts of information. This allows policy makers and providers and users of health care to remain abreast of the evidence around a given intervention without the time commitment required to review all the original research.

Where randomized controlled trial (RCT) evidence is available meta-analyses can provide important answers to research questions without the high cost or time requirements of further RCTs<15>. By avoiding the repetition of RCTs for a given intervention, the effective medical intervention can often be implemented more quickly <1>. This is exemplified by the cumulative meta-analysis of endoscopic treatment of bleeding peptic ulcers. From 1980 to 1990, 24 trials were done comparing endoscopic treatment versus standard treatment. The cumulative meta-analysis clearly demonstrates that endoscopic therapy was shown to be superior to standard treatment by 1982, after only 4 trials. This means that 1,425 patients were enrolled in trials after a clear benefit for endoscopic therapy had been established <19>.

Meta-analysis can also avoid the necessity of further trials because the act of pooling data leads to increased statistical power and precision, which can result in conclusive answers where previously there were none. This is illustrated in a MA<20> published in 1990 investigating the efficacy of corticosteroids given to women expected to give birth prematurely. Of the seven trials in this MA, only two had statistically significant results. On pooling the results, the sample size increases as does the statistical power, and it becomes convincingly clear that corticosteroids reduce the risk of babies dying from complications of prematurity.

MA may afford greater generalizability than individual trials. Pooling distinct trials allows subtle variations in populations and drug doses, for example, to be incorporated in the outcome. These variations are hard to achieve in a single RCT. At the same time as enhancing generalizability, MA provides an opportunity to investigate inconsistencies and conflicts in data when present. Their systematic approach may allow sources of inconsistencies to be discovered and resolved.

A final advantage of MA and SR is the ability of readers to judge their value. The rigorous and explicit methodology used increases the ability to replicate the results, as well as understand the results and conclusions <1>.

I-1-3 Limitations

Cases have been reported where the results of MA do not agree with subsequent large trials <21>. This has led to some concern over the validity and usefulness of meta-analysis. However, proponents of meta-analysis argue that it is a compilation of evidence and should be valued, as trial size does not guarantee generalizability, quality, or freedom from bias <22>. The most important limitation is that MA is a retrospective exercise and as such has all the associated biases (e.g.: selection bias).

Many of the concerns seem to stem from the trials that are pooled. If there is great variation in the way studies have selected patients, administered treatment, or measured outcomes then pooling may reasonably not be valid <10,15>.

The validity of any MA or SR is dependent, among other things, on the studies included. There is debate over which studies merit inclusion in meta-analysis. This debate includes issues of study design (RCT, cohort, case-control), study quality, peer review status

(whether the research has been formally peer-reviewed or not), and related to peer review status, publication status <23>.

I-2 Study Design

In the past, most of the methodological work in meta-analysis focused on RCTs because they are considered the gold standard for evaluation of the effectiveness of most interventions. RCTs are less prone to bias than other study designs. The RCT has been the subject of much methodologic investigation. Guidelines have been established for the reporting of RCTs, and several instruments, tools, and checklists exist to assess the quality of reporting of RCTs <24,25>. Within the RCT, there are parallel arm trials and cross over studies. The parallel arm design randomizes patients at the beginning of the study and they are followed on the intervention to which they were randomized until an end point is reached. In a cross over study, as the name implies, the patients are randomized to one intervention and then at some point before the end of the study they are crossed over to the other intervention. When properly conducted, cross-over design studies have advantages. However, the available evidence suggests that they tend to produce higher estimates of treatment effect <26>. Khan et al <26> analyzed the association between study design and treatment effect in 9 MA in infertility research. They used logistic regression and controlled for intervention and trial quality. They found cross-over trials produced estimates of treatment effects on average 74% (95% Confidence Interval (CI) 2%, 197%) higher than trials of parallel design.

Some research questions are not amenable to being studied in a randomized design. It is encouraging that the methodology of observational data MA is starting to be considered <27>, however it will take some time before it reaches the level of MA of RCTs.

I-3 Methodological Quality

Methodological quality can be defined as 'the extent to which the design, conduct, analysis and presentation of a study are likely to be free of bias' <28>. In considering the quality of reporting of either meta-analyses or RCTs, it is important to realize that what is being assessed is the quality of the report. It is very difficult to assess the actual quality of a study. The purpose of the prepared document, page limitations, and editorial decisions may all influence what actually appears in a study write-up or manuscript. For example, an internal report from a pharmaceutical company may not need to contain details about random allocation or method of double blinding if these things are detailed in other company documents. However, work by Liberati et al <29> suggests that the quality of a report, in published research, is generally a good surrogate for quality of the study. This group assessed the quality of 63 reports of randomized trials in breast cancer. Using a scale with a maximum score of 100 points, they found the average quality of reporting was 50% (95% CI 46%, 54%). They then interviewed the authors of 62 of these reports to determine if information in the manuscript had been removed prior to its publication. Additional information obtained from the authors increased the quality scores by a mean of 7% (95% CI 3%, 9%).

There is concern that unpublished studies, because they have not been peer reviewed, may be methodologically weak. If poor quality is the reason a study does not appear in

the literature, then the peer-review process is working and unpublished literature should not be included. However, Easterbrook et al <30> found no methodological differences between unpublished and published studies (no specific data was reported) suggesting that the inclusion of the latter, in peer reviewed journals, does not guarantee scientific quality <30>. The study by Abby et al <31> also demonstrated that articles peer reviewed and rejected by one journal do get published by other journals. This suggests that either the subsequent publication is not justified, or that peer review and editorial decisions are not based entirely on scientific quality. The second notion is more plausible given the limited number of pages in any one journal.

I-4 Peer Review

One argument against including unpublished material and other sources of grey literature (reports that are unpublished, have limited distribution, and/or are not included in bibliographic retrieval system <32>) in reviews is that they have not been peer reviewed. Peer review is defined as the "process of review of research proposals, manuscripts submitted for publication, abstracts submitted for presentation at scientific meetings, whereby these are judged for scientific and technical merit by others in the same field" <32>.

Peer review, as a system is not transparent or explicit. There is little information on if and how peer review systems work, and what criteria reviewers use to make their decisions. There is no set standard and journals tend to set their own policies, which may not be clear to authors or reviewers.

The irony of the situation is detailed by Squires <33>, who writes in his editorial "it is ironic that this mainstay of validation of scientific information before publication should itself have escaped scientific scrutiny until relatively recently".

A study by Abby et al <31> suggests that peer review is successful, as the proportion of rejected papers that remain unpublished is high. Among North American authors, they found that of 248 manuscripts submitted to and rejected by the American Journal of Surgery in 1989, 50.4% (125/248) were not published by a core medical journal within 3 years. Although not discussed in the main body of this paper, 30% (36/119) of the manuscripts were rejected because they did not adhere to the journal format, which has little to do with the quality of the research presented.

This study also details the issue of duplicate publication. They note six authors had a total of 139 similar publications during the eight-year period. One author had 83 publications on the same theme. Three of these were so similar to the rejected manuscript that only one word in the 18-word title was different. The methods, results and tables were virtually identical, however there was no reference to the similar work. The fact that the authors with the most similar publications did not get their rejected manuscript published may suggest that peer review screens originality of published data to some extent. However, it is alarming to see the amount of duplicate publishing that occurs. Egger et al <34> found a 31% duplication rate in their comparison of trials published in German and English. Although the authors note that peer review prevented a fourth publication of a manuscript, it is unclear why it was ever published more than once with no reference to previous publications. Did peer review fail? Is this a role of peer-review?

In preparing a review, duplication can be as damaging as exclusion. Duplicate or redundant publication has been defined as the presence of the same information or data more than once in the medical literature <35>. What is damaging is the fact that authors usually are not explicit about the duplication. Explicit or overt duplication exists when authors make reference in subsequent publication to the previously published work; this is in contrast to covert publication, where no such referencing occurs. The evidence suggests that prevalence of duplicate or redundant publication is between ten and twenty percent <35>. In a meta-analysis of Ondansterone, inclusion of redundant data led to an increased estimate of treatment effect by 23% <36>. The problem for meta-analysts is that it is often very difficult to determine what is redundant and what is not <37>. Whether duplication is intentional or not <38>, meta-analysts must be aware of its existence and be cautious with data that appears similar.

I-5 Publication Status & Grey Literature

The debate over peer-review status deals with, in part, the role of grey literature in meta-analysis. 'Grey' literature has been defined as "reports that are unpublished, have limited distribution, and/or are not included in bibliographic retrieval system"<32>. For the purpose of this thesis, unpublished studies, abstracts and conference symposia, graduate theses, book chapters, company reports and applications, letters and manuscripts in press were considered sources of grey literature. This definition was reached after correspondence with various individuals with experience in library sciences and the area of publication bias (Ann McKibbon, Jesse McGowan, and Kay Dickersin). It was felt to be quite comprehensive, perhaps too comprehensive. There may be some debate as to

whether abstracts, conference proceedings, and book chapters are truly grey literature. Abstracts and conference proceedings may or may not be peer reviewed, and may be published in journals or journal supplements. Book chapters are peer reviewed and published. The reason these three groups were categorized as grey literature was accessibility and retrievability. Some abstracts and conference symposia are accessible through electronic databases and published in core journals while some are not. In any case, it is clear that there is much controversy as to the merit of inclusion of abstracts and unpublished studies.

A previous study <40> suggests only 31% of published meta-analyses include unpublished studies. Cook et al <40> surveyed editors, meta-analysts and methodologists about their attitudes concerning the inclusion of unpublished material in a scientific review. There is a general sense (among meta-analysts and methodologists) that unpublished literature should be sought and included. However, all journal editors do not consider this accepted practice. Thirty percent of editors surveyed would not publish a review that included unpublished material, even if it received favourable review.

The nature of grey literature makes its exclusion more convenient for investigators undertaking a systematic review or meta-analysis; it is difficult to retrieve, it is often incomplete, its quality may be difficult to assess, and it may not have been through any formal peer review process. Meta-analysts spend considerable time searching for relevant evidence including several electronic databases, hand searching, and consulting available trial registries. The aim of this exercise is to accumulate the universe of evidence, not merely a representative published sample. There is concern that the

"universe" of grey literature on a given topic may not be uncovered, and more importantly that the identifiable grey literature would not be representative of this universe. The concern exists, as it does with published literature, that the inclusion of an incomplete sample of research may lead to biased estimates.

I-6 Publication Bias

Meta-analysts complete extensive searches for relevant literature to avoid the effects of publication bias, and to produce the least biased, most complete and up-to-date synthesis possible. Publication bias is the tendency towards publishing studies that report statistically significant or positive (in favour of treatment) results <41>. Several studies, including a meta-analysis, have shown this bias exists <30,39,42>. Publication bias appears to be due to authors not submitting studies with null findings rather than editors refusing to publish them <39,43>. Authors may expect such studies to be rejected by editors and therefore become unwilling to spend the time necessary to produce the manuscript. Additional effort may be required to publish negative findings. A study of topical vitamin C to prevent radiation induced dermatitis was initially rejected due to the findings, but it was later published after the author responded with a discussion of publication bias <44>. Regardless of where publication bias originates, studies with significant findings are 1.2 to 2.5 times more likely to be published compared to those with neutral or negative findings <42,45>. Research by Mary Lee Smith <46> demonstrates that effect sizes among theses, dissertations and books are also smaller than those of journal published studies. Findings reported in journal articles tend to be 33% more favourable towards the interventions than findings reported in theses and

dissertations. A further problem with negative and null studies is that those that do get published tend to take almost twice as long to get published <47>. A systematic review that includes only published material could, therefore, bias the estimates of the effectiveness of an intervention.

I-6-1 Registries and Databases

One proposed way to overcome and eliminate this concern is through the establishment of prospective trial registries. The effort involved in the retrospective identification of unpublished trials limits its usefulness <48>. In 1997, after their first inaugural meeting, the World Association of Medical Editors (WAME) created an amnesty for unpublished trials - Medical Editors Trial Amnesty (META). A call for registration of all unreported trials appeared in many of the prominent clinical journals (JAMA, Lancet, BMJ, etc) <49,50>.

Several registries and databases of 'unpublished' and 'grey' literature are now available. For example, SIGLE (System for Information on Grey Literature in Europe, http://www.sba.unige.it/erl/sieng.html) is produced by a consortium of leading libraries in Europe and published by Silver Platter. It contains documents from 1980 forward. Fourteen percent of the contents of SIGLE are in the field of medicine. It contains over 534,000 records and the source is indicated in each record. However, the records are all European based documents. Similar efforts are needed to capture non-European based information. Hopefully other pharmaceutical companies will follow the lead of Glaxo-Wellcome and make their research results available. There are also some journals being dedicated to negative studies. For example, NOGO (the Journal of Negative

Observations in Genetic Oncology). This electronic journal allows users to search by key word, view volumes page by page, and submit entire studies or citations.

All of these efforts suggest that there is concern about the ramifications of omitting grey literature from meta-analyses and systematic reviews. Aside from the evidence that publication bias exists, there is currently no empirical evidence about the impact of excluding this literature from meta-analysis.

II Aims and Objectives

The three objectives of this thesis are:

1) To characterize sources of grey literature found in meta-analyses and its prevalence.

2) To examine the quality of reporting of meta-analyses and studies included in meta-analyses. Specifically:

2-1. to compare the quality of reporting of meta-analyses that include grey literature to a sample which does not.

2-2. to compare the quality of reporting of 'published' and grey sources of literature included in the identified meta-analyses.

3) To determine the impact grey literature has on the point estimate, the precision, the statistical significance, and the statistical heterogeneity of meta-analyses of RCTs.

III Methods

The aims of this research were accomplished using a sample of published meta-analyses. These meta-analyses fell into three groups: (1) those that included grey literature; (2) those that explicitly excluded grey literature (i.e. it was stated that grey literature did not meet the eligibility criteria and was not included); and (3) those that implicitly excluded grey literature (i.e. either there was no mention of grey literature, or the eligibility criteria indicated it was acceptable, but none contributed to the quantitative analysis). Meta-analyses from the two groups that excluded grey literature were only used as comparators for quality of reporting of meta-analyses. Meta-analyses that included grey literature were replicated and repeated to investigate the impact of grey literature on the statistical results.

III-1 Database used in Sample Identification

The database from which my sample was drawn was established in 1996. It was assembled through MEDLINE searching from 1966 to 1995. A detailed search strategy (Appendix B) was assembled with the aid of a librarian, to identify all meta-analyses of randomized controlled trials. The search yielded 1467 references. The references were retrieved and coded as systematic review, meta-analysis, methodology paper, or editorial. The identified articles were reviewed and coded by two reviewers independently. Differences were discussed and consensus was reached. In cases where the author of the paper called the work a meta-analysis, it was retained even if the reviewers disagreed with this classification. The resulting database contains 455 articles coded as meta-

analyses of randomized controlled trials. For details about the other 1012 (1467 - 455) references see Figure 2.

III-2 Sample Identification

Given the lack of data on the impact of grey literature on meta-analysis, no a priori sample size calculation could be performed. A sample size of 40 MA was chosen for reasons of feasibility. This was done with the intention of doing a post hoc power calculation if the results were not significant. As the available evidence <40> suggests that 31% of meta-analyses catalogued on MEDLINE include unpublished literature, an initial sample size of 135 was chosen because it was expected to provide about 40 meta-analyses that included grey literature. Using SAS PLAN procedure (Appendix C) a list of 135 random numbers between 1 and 455 was generated. Each number corresponded to the reference number of a single meta-analysis.

III-3 Sample Eligibility Criteria

A MA was eligible for inclusion in the study if after review of the text and/or references, the included studies could be identified, it was deemed to be a meta-analysis (included pooled analyses of the results of at least two independent primary studies), and it could be concluded that at least one item of grey and one item of 'published' literature was used in the generation of a summary statistic of the treatment effect. For reasons of feasibility, only MA that included binary outcomes and at most 100 RCTs were considered. Binary outcomes were required for the calculation of odds ratios (OR) and ratio of odds ratios (ROR) in the assessment of the impact of grey literature. The cut point of 100 RCTs

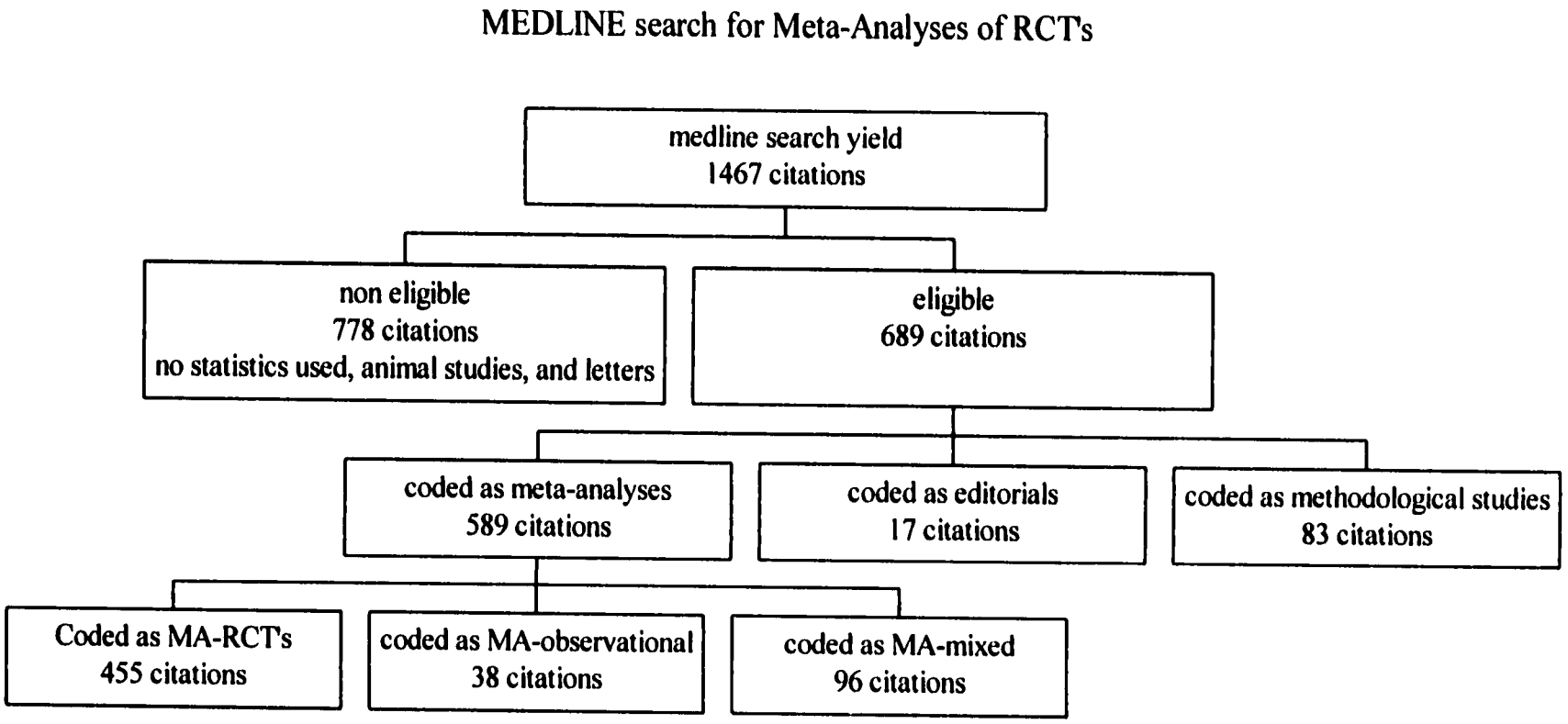


Figure 2: The development of the database of meta-analyses of randomized trials from which a random sample of 135 MA's was drawn.

was set as it was felt that retrieval of the RCTs could become too time consuming and costly. This resulted in the exclusion of seven MA (4 had ≥ 100 RCTs, 3 had non-binary outcomes).

For the comparator sample, a random sample of 40 MA (Appendix D) without grey literature was identified. Although ideally the comparator sample should have the same inclusion / exclusion criteria, for this sample, number of RCTs and type of outcome were not considered as eligibility criteria. Neither including less than 100 RCTs, nor type outcome (binary vs. continuous) was expected to influence the quality assessment of the meta-analysis, as these are not items on the Oxman / Guyatt index and neither has been empirically demonstrated to influence quality. Although there was no reason to exclude MA with more than 100 RCTs, as these RCTs were not being retrieved, none of the MA in this comparator group had more than 100 RCTs. The outcome was not required to be binary as the outcome data was not to be used in the regression modeling. There were seven meta-analyses with continuous outcomes included in this group. A chi-square test was used to ensure that the inclusion of these MA with continuous outcomes did not influence the quality of this group.

III-4 Sample Retrieval

All 135 meta-analyses were retrieved and reviewed to identify the subset that included grey literature (Appendix E). A meta-analysis was deemed to include grey literature if any of the included trials fell into the previously set definition of grey literature.

The original documents were collected for each meta-analysis that included grey literature. Studies not accessible through library services were sought by contacting the corresponding author of the meta-analysis. (Appendix F)

III-5 Data Abstraction

The following information was abstracted from each MA, as required: language and year of publication, disease area (Appendix I), quality of reporting information, number of included studies, number of included subjects, summary statistic, method used to generate it and outcome data for each of the included RCTs.

From each RCT the following data was abstracted; outcome data from the primary outcome, quality of reporting information, year and language of publication, number of subjects randomized, and publication status.

Outcome data for each included study was abstracted from the meta-analysis where possible. If the data was not found in the MA report, then the data abstracted from the RCT was used. The decision to use the information presented in the MA was made because there were cases where this information could not be found in the RCT.

The main outcome was defined as the one stated as such by the authors, or if there was no such statement, the most clinically relevant (i.e. mortality would be selected over morbidity). If there was not one more clinically relevant outcome, then the one contributing the most patients was selected.

III-5-1 Language

For MA and RCTs published in languages other than English and French, I sought help from colleagues with fluency in the required languages. At the MA level, help was sought in determining if an MA included grey literature, as well as in abstracting the needed data and quality assessment. At the RCT level, help was sought with data abstraction when the data was not available from the MA, and with quality assessment. To ensure consistent quality assessment, the 'translators' were not asked to complete the quality assessment, but were asked several questions so that quality could be assessed.

III-6 Objective 1

III-6-1 Objective 1-1

What proportion of published meta-analyses include some form of grey literature?

For this objective the sample of 135 MA was used. Using the pre-stated definition of grey literature, and the inclusion criteria, the number of MA that include some form of grey literature was recorded.

III-6-2 Objective 1-2

What types of grey literature are found in published meta-analyses and what is their prevalence?

Considering the sample of MA found to contain grey literature, descriptive statistics were used to report the sources of grey literature and the prevalence of each source.

III-7 Objective 2

Quality of reports of both the meta-analyses and the RCTs included in the meta-analyses were assessed. All quality assessments were completed under masked conditions with potentially biasing information covered with black marker or removed (i.e. authorship, author affiliation, journal, and references). Masked assessment was done to minimize bias in the assessment. The existing evidence <2,51,52,53> of the impact of masked assessment is inconsistent. One study <51> suggests that masked assessment produces lower and more consistent scores, while another <52> suggests higher scores under masked conditions. The other two studies <2,53> suggest that masked assessment has no impact on the scores. It was felt that given the existing information, and the uncertainty about the effects of masking, that masked assessment should be used. Masking was used in an effort to reduce the bias in the assessment of quality of reporting. In many cases it was not possible to mask the publication status. For example, it was impossible to mask the fact that the document being assessed was an abstract.

III-7-1 Meta-Analyses

For assessment of quality of reporting at the MA level (n = 73 (Appendices D, E)), the Oxman / Guyatt index (Appendix G-1) was chosen. A recent systematic review of the literature <54> has identified 26 instruments (scales and checklists) to assess the quality of meta-analyses and systematic reviews. Eighteen of the 26 were published at the time this work was started. The Oxman / Guyatt index was one of two scales available. A scale differs from a checklist in that each item is scored numerically and an overall score is generated. The Oxman / Guyatt index was chosen as it is the only one to have been

developed with standard instrument development methods <55>. In the development of this index, the construct being measured was defined, discriminatory power of items was measured and reliability studies were undertaken <56>.

The index consists of nine items pertaining to individual aspects in the reporting of a meta-analysis (e.g.: were the search methods used to find evidence (original research) on the primary question(s) stated?). The items are scored on a three point index (yes, partially / can't tell, or no). The index also contains a final summary question, "How would you rate the scientific quality of the overview?" with the score ranging from a low of 1 to a high of 7. The scoring of this summary item incorporates the assessment of the first nine items. As the Oxman / Guyatt index was developed with no specific content area in mind, it is therefore a generic instrument, which should not be affected by content area.

III-7-2 Randomized Controlled Trials

The quality of reports of the RCTs (n = 429) included in the meta-analyses were assessed using the Jadad scale <51> (Appendix G-2). This scale was one of 25 scales available for the assessment of quality <25>. In its development, a large number of items were narrowed down to the final version using standard scale development techniques <55>. A description is provided on how the items were selected, and how the final items came to be included. Inter-rater-reliability has been assessed, as well as its ability to discriminate between trials of different quality <51>.

The scale consists of three items pertaining to descriptions of randomization, knowledge of treatment assignments (double blinding), and inclusion of randomized patients

(withdrawals and dropouts). The scale ranges from zero to five. The first two items (randomization and double-blinding) each have a maximum score of 2 (1 point is given for those who describe the item [i.e. randomized], and a second bonus point is given if the process is detailed and is deemed appropriate [i.e. using a table of random numbers]). The final item withdrawals and dropouts is scored out of one. The Jadad scale was specifically developed for 'pain' literature; however, it has been widely used in other clinical areas.

In addition to the items on the scale, information was also collected on allocation concealment and study design. Allocation concealment refers to keeping the random numbers, once they have been generated, hidden from all RCT participants (patients, investigators, and outcome assessors) until the time of randomization. Previous research <52,57> has shown that inadequate allocation concealment, compared to when it is adequately performed, results in higher estimates of treatment effects by an average of 37-41%. Study design refers to the use of parallel group or cross over designs. Using a logistic regression model, similar to the one used in objective 3, Kahn et al <26> demonstrated an overestimate (i.e.: higher estimates) of the treatment effect in cross over studies as compared to parallel studies by an average of 74% in the infertility literature. This result was achieved after having adjusted for confounding factors such as allocation concealment, and blinding.

III-7-3 *Inter-rater Agreement*

III-7-3-1 *Training*: Each individual item on both of the two quality assessment instruments was discussed before the instruments were applied to any articles.

The items on the Jadad scale tended to be more concise and less prone to interpretation than those on the Oxman / Guyatt index. After reviewing and discussing each item, agreement was reached on definitions and clarifications for those items that were thought to be less transparent (Appendices G-1-2, G-2-2 for items with points of clarification). Once confident that the interpretation of the items was consistent, calibration was considered. The training, calibration, and intraclass correlation process was done by myself and one of my thesis advisors (David Moher [DM]).

III-7-3-2 Calibration: Five RCTs and 5 MA not included in the study sample were retrieved and masked for agreement training for quality. A sample of five was selected as it was felt that this sample would enable us to be confident that the training exercise had been successful. We independently assessed the quality of all studies using the Jadad or Oxman / Guyatt instruments. All discrepancies were reviewed and the items to which they corresponded were clarified.

III-7-3-3 Intraclass Correlation: Fourteen RCTs and 14 MA not included in the study sample were retrieved from current issues of journals known to have such articles. The papers were masked, and assessed independently. A sample of 14 was selected, as it was the number necessary to detect an agreement greater than 0.60, which indicates substantial agreement, with appropriate statistical power <58>. The scores were used to calculate intraclass correlation's (ICC) for each item. The ICC's were calculated using a 2 way mixed effects layout, with a fixed effect being used for the raters and a random effect being used for the number of articles <59>.

III-7-4 Objective 2-1

Is the scientific quality of reporting of meta-analyses that include grey literature different from those that do not?

For this objective the unit of analysis is the meta-analysis. Thirty-three meta-analyses that included grey literature and 40 that did not were included. Of the 40 meta-analyses that did not include grey literature, two distinct groups are found, those that explicitly exclude grey literature (E, Appendix A) (n = 17) and those that implicitly exclude grey literature (I, Appendix A) (n = 23).

III-7-4-1 Overall Quality (Q10): A one-way analysis of variance (ANOVA) model was used to relate the overall quality score (question 10 (Q10)) of the Oxman / Guyatt scale to the inclusion or exclusion of grey literature. Initially, year of publication and the content area of the meta-analyses were also to be included, however, on examination of the data, the year of publication and content area did not differ between the two groups (grey literature included and grey literature excluded).
A second analysis was done, again using a one-way ANOVA to compare grey literature included, and the two 'grey literature excluded' groups, (E vs. I).

III-7-4-2 Individual Items (Q1-9): Using a logistic regression model, features of reporting of the MA assessed in questions 1-9 of the Oxman / Guyatt scale were considered using the grey included group as the reference group. Logistic regression models:

logit (O/G item) = $\beta_0 + \beta_1$ (explicitly exclude) + β_2 (implicitly exclude) + ε.

As the rating scale is ordinal (i.e. "yes", "partially", and "no") for items 1,3,5,7,9, the scale was grouped, so "yes" was compared with "no" or "partially". A sensitivity analysis was completed grouping "yes" and "partially" and comparing them to "no". Logistic regression models (as above) were also used for the features assessed by the remaining questions (2,4,6,8) with grouping of the scale outcome (e.g. "yes" versus "no" and "can't tell"). Again, a sensitivity analysis was done where "yes" and "can't tell" were grouped and compared to "no".

With these analyses, the effect sizes due to grey literature were summarized for each of the first nine items assessed by the Oxman / Guyatt index.

III-7-5 Objective 2-2

Is the overall scientific quality of reporting of published RCTs different from grey RCTs?

For this objective the unit of analysis is the RCT. Each RCT will only be included once, even if it is present in more than one MA.

An ANOVA model was used to determine if 'publication status' was related to the overall quality score from the Jadad scale. The Jadad scale is an interval scale with scores ranging from 0 to 5.

Again, it was my initial intention to consider year of publication and content area as well as both allocation concealment and study design (parallel groups vs. a crossover design) in the ANOVA. However, content area did not differ by publication status and there were

too few studies (30/399) that reported allocation concealment, and too few cross over studies (9/434) to make this feasible.

As year of publication differed between published and grey literature, the model included publication status, year and an interaction terms for publication status and year.

III-8 Objective 3

Does the inclusion of grey literature in MA influence the estimate (in terms of magnitude, direction, or statistical significance) of the intervention effect?

Outcome data (i.e. number of unwanted events and total patients in the treatment and control groups) were extracted from the published MA or from the original trials if necessary. Using the same meta-analytical method, the published MA results were replicated. This step was essential to ensure the data extraction was accurate.

The Peto OR (OR = $\exp [\Sigma^{k}_{i=1} (O_i-E_i) / \Sigma^{k}_{i=1} V_i])$ <60,61> was generated to standardize comparisons. It was also used to visualize the effects of grey literature on the treatment effectiveness. The Peto OR was chosen as it has previously been shown to be robust, particularly to sparse data sets <62>.

After completion of the replication, each MA was repeated using the Peto OR, with all grey items removed. This was done only to provide a visual comparison of the impact of grey literature on meta-analysis.

The effect of grey literature on estimates of intervention effect was evaluated using a logistic regression model done at the trial intervention level. The log odds of unwanted

events experienced by each treatment group were related to trial intervention (to allow for variation in outcome among trial and force intervention comparisons to be made within a trial), meta-analysis, and grey literature.

The model used was as follows:

- $\log OR = \beta_0 + \beta_1 [\text{trial } i^{th}] + \beta_2 [\text{intervention } j^{th}] + \beta_3 [\text{MA } k^{th}] +$

  $\beta_4 [\text{publication status } l^{th}] + \beta_5 [\text{intervention } j^{th} * \text{MA } k^{th}] +$

  $\beta_6 [\text{intervention } j^{th} * \text{publication status } l^{th}] + \varepsilon$

Variation in intervention effect across meta-analyses was accounted for by an interaction term between intervention and meta-analysis. I included an interaction term between publication status (grey vs. published) and intervention (treatment vs. control) to capture the potential that grey literature could modify estimates of intervention effectiveness. The exponent of the coefficient associated with this interaction term represents the ratio of odds ratios (ROR) for the two comparison groups <26, 57> (Appendix H). Using this modeling convention, an odds ratio less than 1 indicated that the intervention was more effective than the control in preventing an unwanted event. Consequently, an ROR between grey and published literature greater than 1 indicated that on average, estimates of intervention effectiveness from published literature were greater than their corresponding estimates from grey literature.

Relative to published literature, the potential effect of various sources of grey literature was evaluated: all grey literature, and grey literature excluding abstracts. The second analysis was of interest, as abstracts make up the largest proportion of the sample of grey literature, and there is debate over whether they should be considered grey literature. Due

to small numbers in the various categories of grey literature, we were unable to consider further sub-sets of grey literature.

I also evaluated the potential effect due to grey literature in the presence of variation in trial quality. This was done by including a main effect for quality, i.e. low quality trials (Jadad score ≤ 2) versus high quality trials (Jadad score > 2), and an interaction term between quality and intervention in the model.

- $\log OR = \beta_0 + \beta_1$ [trial i^{th}] + β_2 [intervention j^{th}] + β_3 [MA k^{th}] + β_4 [publication status l^{th}] + β_5[jadad score p^{th}] + β_6 [intervention j^{th} * MA k^{th}] + β_7 [intervention j^{th} * publication status l^{th}] + β_8[intervention j^{th} * jadad score p^{th}] + β_9 [publication status l^{th} * jadad score p^{th}] + ε

Standard residual diagnostics were used to assess the model goodness of fit. ROR and its 95% confidence intervals were derived from the fitted model.

To consider the impact of grey literature on the test statistic of the no intervention effect hypothesis, Z scores were derived (i.e.: intervention effect size divided by its standard error) from MA with and excluding grey literature. For visual comparison, the Z-scores of the grey inclusive and the grey excluded estimates were plotted against one another. They were then compared statistically using a paired t-test.

IV Results

IV-1 Sample

IV-1-1 Meta-analyses

IV-1-1-1 Retrieval: All of the meta-analyses except one (935) were accessible through normal library retrieval methods. MA #935 was acquired through communication with the lead author. Of the 135 meta-analyses reviewed, 33 were found to contain both grey and published literature (Figure 3). Seventy-five MA had no grey literature, and 27 were excluded (poor referencing (11), not MA (8), included greater than 100 RCTs (4), outcome not binary (3), only grey literature (1)). The 8 excluded because they were not MA were in the original database as the authors considered them MA, however they did not employ statistical methods to combine the results of two or more independent studies. Details of the 33 MA that include grey literature, and the random sample of 40 (drawn from the pool of 75) that do not include grey literature can be found in Table 1.

IV-1-1-2 Language: In the original sample of 135, one of the MA was published in Italian. It was determined that it did not contain any grey literature.

IV-1-2 RCTs

IV-1-2-1 Retrieval: The sample of 33 published MA that include grey literature contains 478 RCTs. Of these, 373 were published as journal articles, and the remaining 105 were grey reports. In the published literature, 8 RCTs were excluded (Table 2) leaving a sample of 365.

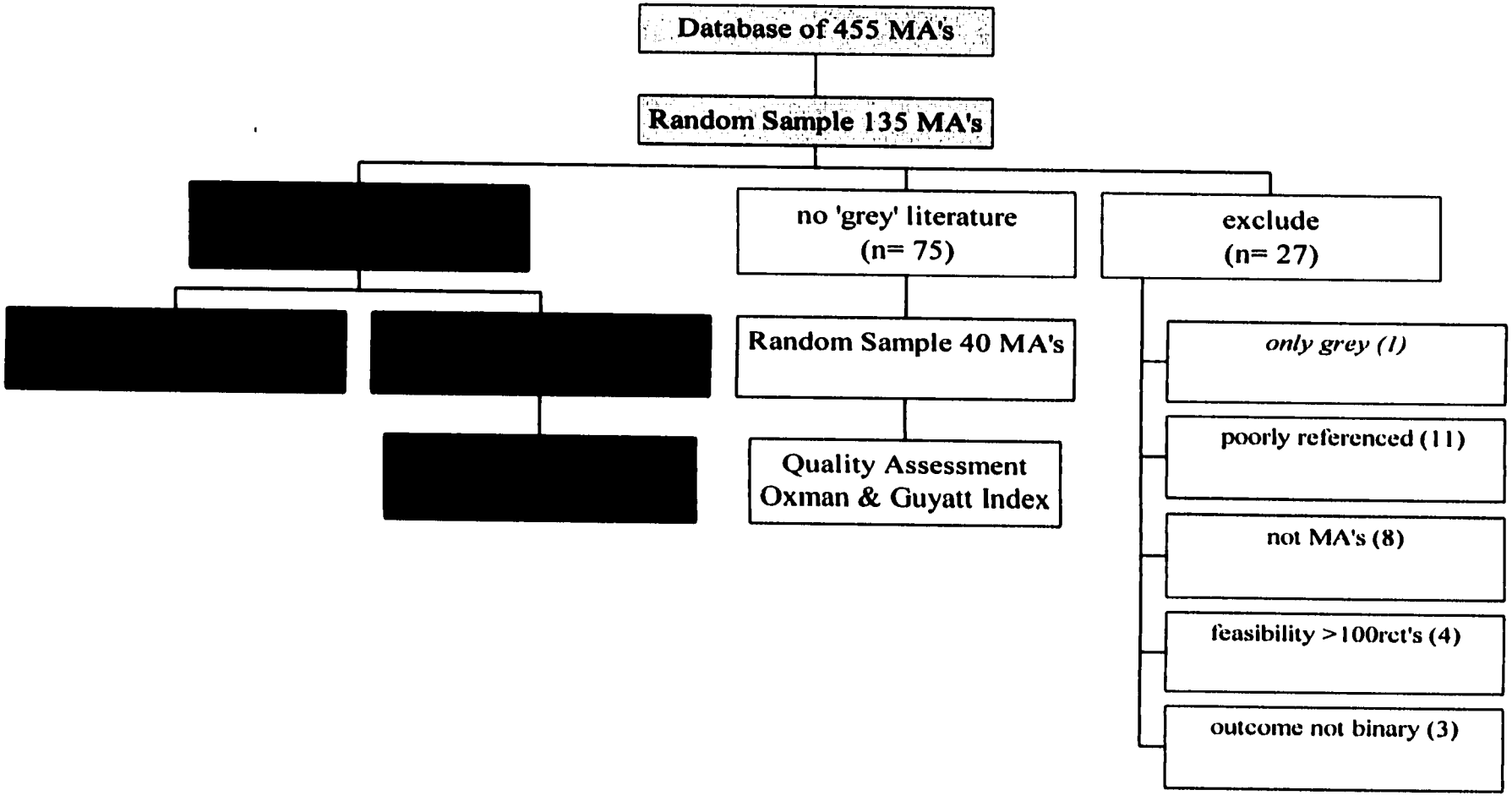


Figure 3: Flow sheet of the identification of the sample of meta-analysis.

Table 1: General Characteristics of Meta-analyses.

	Overall n = 73	MA with grey literature n = 33*	MA without grey literature n = 40	p value** Mann Whitney test
Total number of RCTs	1080	478 (467[x])	626	/
Number of RCTs per MA Median Interquartile range.	 10 6, 20	 10 6, 19	 10 6, 21	 p = 0.916
Number of patients per MA Median Interquartile range	na	 1463 1120, 3163	na	na
Clinical area (frequency) (top 6)) gastrointestinal cardiac circulatory infection reproduction cancer	 20.5% 15.1% 15.1% 11.0% 8.2% 8.2%	 24.2% 21.2% 9.1% 12.1% 12.1% 0%	 17.5% 10.0% 20.0% 10.0% 5.0% 15.0%	 p = 0.013
Total number of grey items	102	102	0	/
Median number of sources of grey literature per MA Interquartile range	/	1 1, 2	/	/
Median number of grey items per MA Interquartile range	/	2 1, 3	/	/
Year of publication of MA Interquartile range	1993 1991, 1994	1993 1991, 1994	1993 1990, 1994	p = 0.75

* number of MA which include grey literature is 41 (in 33 published reports)

** the test statistic was used for descriptive purposes only.

x eleven RCTs were excluded from the sample (Table 3).

na: data was not abstracted.

Table 2: List of Excluded RCTs with Reasons for Their Exclusion.

MA #	RCT #	**Reason For Exclusion**
406	30	no outcome data reported in MA
634	13	removed in original MA by meta-analysts due to heterogeneity
1251	26	seemed to be duplicate or early versions of other included studies, and did not contribute to the summary statistic in the published MA
	28	seemed to be duplicate or early versions of other included studies, and did not contribute to the summary statistic in the published MA
	31*	seemed to be duplicate or early versions of other included studies, and did not contribute to the summary statistic in the published MA
	32	seemed to be duplicate or early versions of other included studies, and did not contribute to the summary statistic in the published MA
1353	17*	outcome of interest not reported in the abstract, and no raw data available in the MA
1418	1	seemed to be related to another included RCT
	2	seemed to be related to another included RCT
	4	seemed to be an early version of another included RCT
	9*	overlaps with another included RCT for the outcome of interest.

*denotes a grey item

RCTs which did not report the outcome of interest, and were used in the meta-analyses for secondary outcomes are not listed here.

Similarly, 3 reports were excluded from the grey literature leaving 102 reports. Not all of the RCTs were available through normal library services. Forty-five grey reports and 5 published ones had to be sought through communication with the authors (Appendix F). For the grey literature, 19 of the 45 documents were retrieved. For the published literature, 4 of the 5 documents were retrieved. Figure 4 gives a breakdown of how the requests for documents were received.

IV-1-2-2 Language: Twenty-eight RCTs were published in languages other than English. Of those, 11 were published in French, which I was able to read. The remaining 17 were published in German (9), Spanish (4), Italian (3), and Japanese (1). In all cases, outcome data was available from the published MA and with help I was able to assess quality.

IV-2 Exclusions

No MA, which met the inclusion criteria, were fully excluded. In some cases duplication of RCTs among the MA occurred. When duplication occurred, steps were taken to ensure that trials were not included more than once in any analysis. This resulted in the exclusion of MA from the regression model for objective 3. For a description of how these were dealt with see the following section on duplication. Within the included meta-analyses, some of the RCTs were excluded. These are listed in table 3 with the reasons for exclusion.

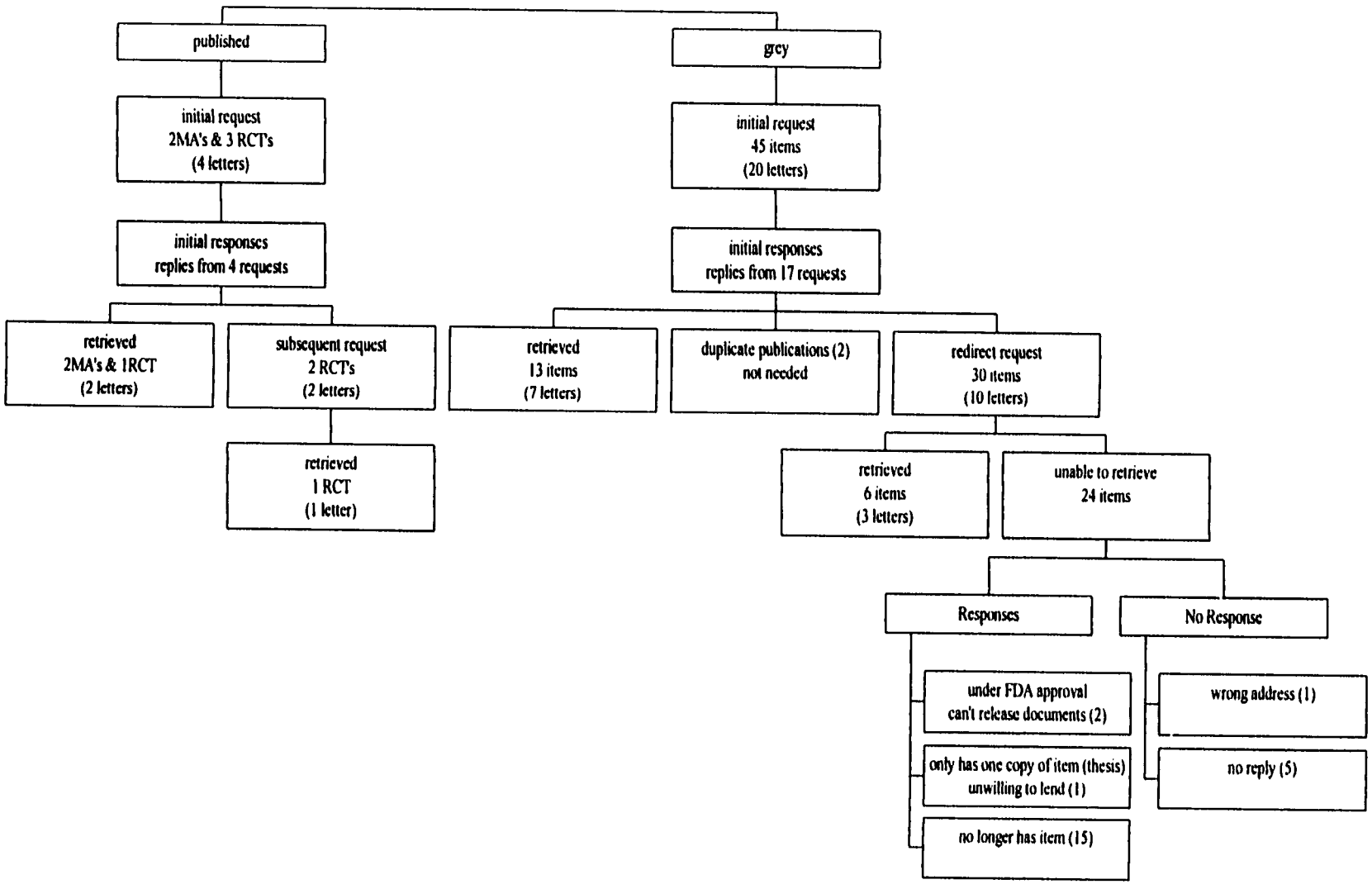


Figure 4: Catalogue of the efforts made to retrieve the 'difficult to get' literature.

Table 3: General Characteristics of RCTs Included in the 33 Published Meta-analyses with Grey Literature

	Overall	grey literature RCTs	'published' RCTs	p value* Mann Whitney test
Total number of RCTs	467	102	365	
Total number of patients	217427	23286	194141	
Number of patients per RCT				p = 0.03
Median	106	83.5	113	
Interquartile range	55, 223	48, 190	59, 228	
Number of RCTs with null or negative results	157 (34.6%)	32 (32.0%)	125 (35.3%)	p = 0.54
Year of publication of RCT	1988	1989	1988	
Median &Interquartile range	1985,1990	1986, 1990	1984 1990	p = 0.01
Number of RCTs in languages other than English	28 (6.4%)	3 (3.5%)	25 (7.1%)	
French	11 (2.4%)	2 (2.0%)	9 (2.5%)	
German	9 (2.0%)	0	9 (2.5%)	
Spanish	4 (0.9%)	1 (1.0%)	3 (0.8%)	
Italian	3 (0.7%)	0	3 (0.8%)	
Japanese	1 (0.2%)	0	1 (0.3%)	

*the test statistic was used for descriptive purposes only.

IV-3 Duplications

Duplication refers to the inclusion of one or more RCTs in more than one meta-analysis in the sample (n = 33). Duplications had to be dealt with, as it was critical that no RCT be included in the analysis (objectives 2-2 and 3) more than once. In three cases the same trial was included in more than one meta-analysis. These duplications were handled in the following ways. The first method was to consider only one treatment comparison per published meta-analysis. For example, MA#1210 and MA#1351 both include trials about beta-blockers for ulcers. MA#1351 also has an analysis for sclerotherapy for ulcers. In this case, as there was overlap with the beta-blocker trials, only sclerotherapy was considered for MA#1351 and beta-blockers for MA#1210. This was done to maximize the number of RCTs and the number of grey items that remained in the sample. The second method was used when there was only one treatment comparison (MA#451 and MA#508). One of the two MA (508) was used for the ROR, the other MA (451) was maintained for all analysis at the MA level (objectives 1 and 2-1), but not used for any analysis at the RCT level (objectives 2-2 and 3). In the third case of duplication (MA#634 and MA#935) only MA#935 was used, as it contributed more grey literature for the overlapping outcome (nicotine patch) than did MA#634, and the grey literature for MA#634 was not retrieved.

Due to these exclusions and duplications in the sample, only 1 published RCT, and 24 grey items were not retrieved.

Details of the RCTs included in the 33 published MA which include grey literature can be found in Table 3.

IV-4 Objective 1 *Characterization and Prevalence of Grey Literature*

From the original 135 MA identified, it was not possible to determine if grey literature was included in 12 cases (11 poor referencing, 1 not retrieved). An additional 8 MA were excluded, as they did not fit the definition of an MA. Of the remaining 115, 38 (33%) were found to contain some grey literature. For all subsequent investigations, the sample of 33 MA with grey literature that met the eligibility criteria was used. When included, the grey literature accounts for between 4.5% and 75% (median 25%, interquartile range 16.7% - 33.3%) of the studies in a meta-analysis. For the sample as a whole, the grey literature contributed 22% (105/478) of the studies. These grey studies contributed 10.7% of the total patients. None of these MA included more than 3 sources of grey literature (Figure 5). The majority only included one source, and the most common source was abstracts, which comprised 61% of the grey literature. The second most prevalent source of grey literature was unpublished papers and from the references there was no indication as to whether a publication was in preparation. Reports, which comprised 5% of the sample included items such as internal organization documents (e.g.: World Health Organization Documents). Symposia could have been grouped with abstracts; however, this category was maintained, as it was the one used by the meta-analysts in their referencing. The two theses in the sample were both at the doctorate level. A detailed breakdown of the sources of grey literature in the sample can be found in Figure 6.

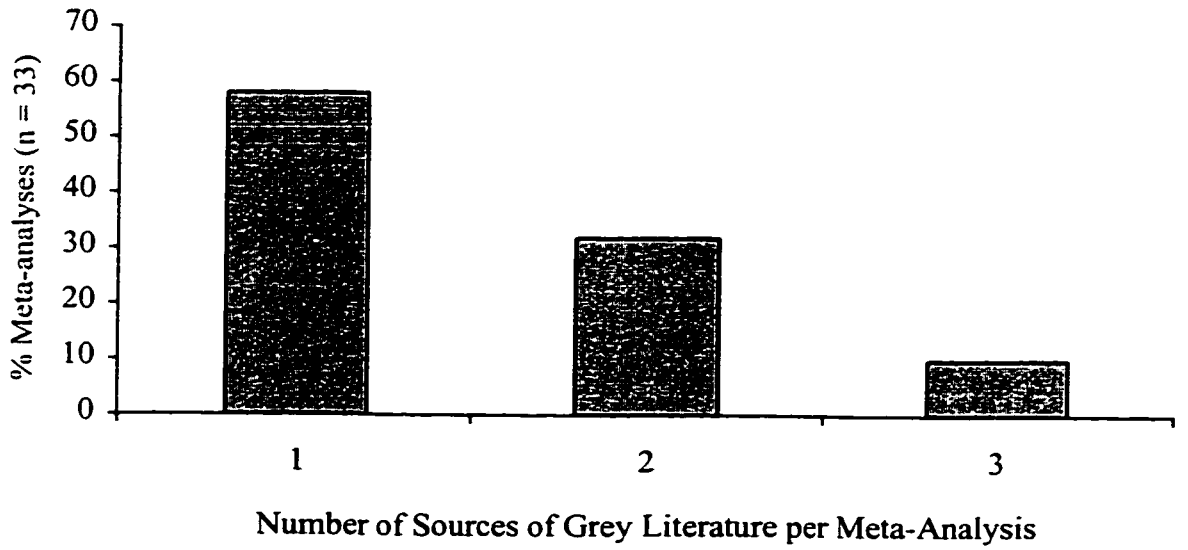


Figure 5: Number of Different Sources of Grey Literature Included in Each Meta-Analysis.

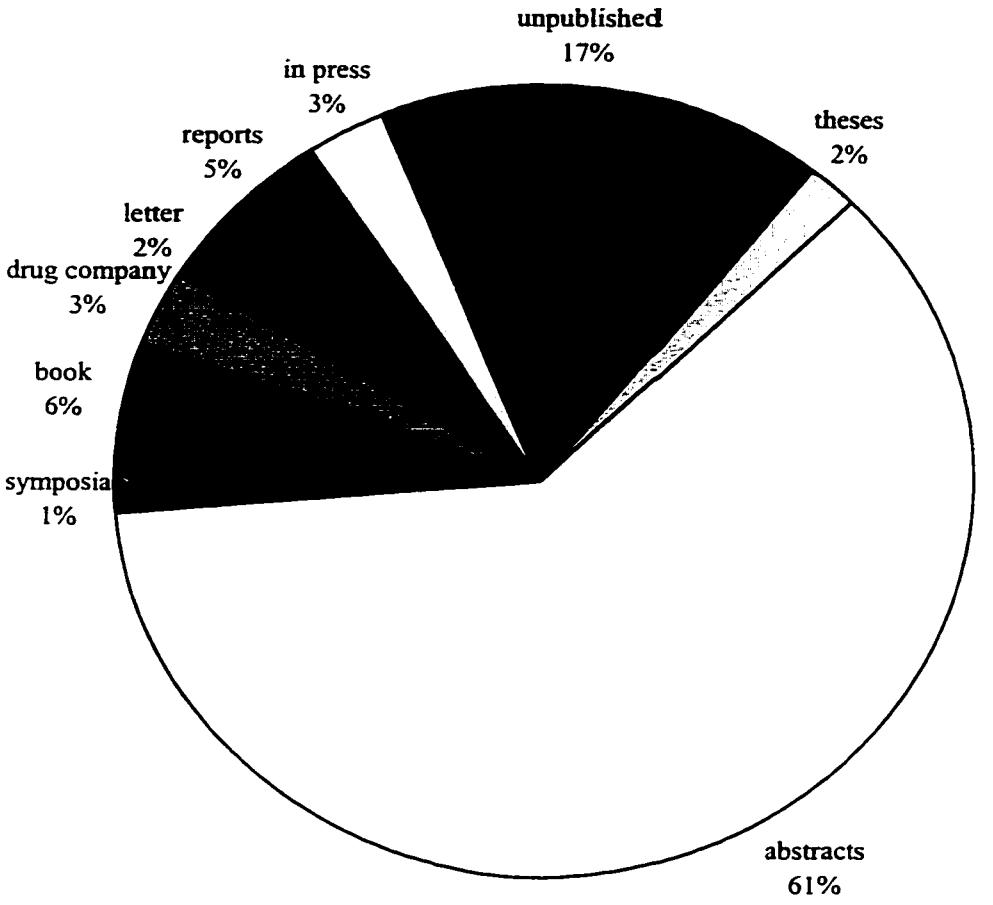


Figure 6: Sources of Grey Literature in the Sample of 33 Published Meta-Analyses. This figure shows the different sources of grey literature in the sample with their relative frequencies.

IV-5 Objective 2

IV-5-1 Objective 2-1 Quality of Reporting of MA

IV-5-1-1 Calibration Set: The quality as scored by each of the two reviewers (LM & DM) for the 5 calibration MA appears in Table 4. There were 14 instances of disagreement in 45 questions, however, the disagreements were mainly (11/14) in level of the presence of an item, for example, one reviewer may have said partially while the other said no. In 3 cases one reviewer said 'no' while the other said 'yes'. Once the discordances were reviewed, agreement was achieved. All questions were clarified, and additional notes were added to the score sheet to facilitate interpretation. For example, question 2: "was the search for evidence reasonable comprehensive?" We agreed that to be scored as reasonable comprehensive, 3 sources had to have been searched, and one had to be something other than electronic (Appendix G-1-2). For item 10, the summary score, the scores differed by 1 point in 4 MA and by 2 points in the remaining MA. It was felt that this was acceptable as the scores were close and the summary item has some room for judgement.

IV-5-1-2 ICC Set: The quality as scored by each of the two reviewers for the 14 ICC MA appears in Table 5. In 12 of the 14 MA scored, there was some disagreement in the scoring of the first nine items of the scale. In only two MA were there times when one reviewer scored yes and the other scored no. Concerning the summary item, in 7 cases there was no difference in the summary scores assigned. In the remaining 7 MA there was never more than 2 points separating our summary scores. The interclass correlation scores for seven of the ten items, including item 10, were above

Table 4: Calibration Exercise for Quality Assessment of Meta analyses using the Oxman / Guyatt index.

MA # 1	**LM**	**DM**
1. Were the search methods used to find evidence stated?	1	1
2. Was the search for evidence reasonably comprehensive?	2	1
3. Were the criteria used for deciding which studies to include reported?	2	1
4. Was bias in the selection of studies avoided?	2	2
5. Were the criteria for assessing the validity of the studies reported?	1	1
6. Was the validity of all studies referred to in the text assessed using appropriate criteria?	2	2
7. Were the methods used to combine the findings of the relevant studies reported?	3	3
8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?	3	3
9. Were the conclusions supported by the data and/or analysis reported?	3	3
10. How wold you rate the scientific quality of this overview?	4	3
MA # 2		
1. Were the search methods used to find evidence stated?	3	3
2. Was the search for evidence reasonably comprehensive?	3	3
3. Were the criteria used for deciding which studies to include reported?	2	3
4. Was bias in the selection of studies avoided?	3	3
5. Were the criteria for assessing the validity of the studies reported?	1	1
6. Was the validity of all studies referred to in the text assessed using appropriate criteria?	3	3
7. Were the methods used to combine the findings of the relevant studies reported?	2	2
8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?	3	1
9. Were the conclusions supported by the data and/or analysis reported?	3	1
10. How wold you rate the scientific quality of this overview?	2	3
MA # 3		
1. Were the search methods used to find evidence stated?	3	3
2. Was the search for evidence reasonably comprehensive?	3	3
3. Were the criteria used for deciding which studies to include reported?	3	3
4. Was bias in the selection of studies avoided?	2	3
5. Were the criteria used for assessing the validity of the included studies reported?	2	1
6. Was the validity of all studies referred to in the text assessed using appropriate criteria?	1	1
7. Were the methods used to combine the findings of the relevant studies reported?	3	3
8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?	3	1
9. Were the conclusions supported by the data and/or analysis reported?	3	2
10. How wold you rate the scientific quality of this overview?	2	1

Table 4: Continued

MA # 4		
1. Were the search methods used to find evidence stated?	2	2
2. Was the search for evidence reasonably comprehensive?	2	1
3. Were the criteria used for deciding which studies to include reported?	1	1
4. Was bias in the selection of studies avoided?	2	3
5. Were the criteria for assessing the validity of the studies reported?	1	1
6. Was the validity of all studies referred to in the text assessed using appropriate criteria?	1	1
7. Were the methods used to combine the findings of the relevant studies reported?	3	3
8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?	3	3
9. Were the conclusions supported by the data and/or analysis reported?	3	3
10. How wold you rate the scientific quality of this overview?	3	2
MA # 5		
1. Were the search methods used to find evidence stated?	2	3
2. Was the search for evidence reasonably comprehensive?	2	1
3. Were the criteria used for deciding which studies to include reported?	3	3
4. Was bias in the selection of studies avoided?	2	2
5. Were the criteria for assessing the validity of the studies reported?	2	1
6. Was the validity of all studies referred to in the text assessed using appropriate criteria?	2	2
7. Were the methods used to combine the findings of the relevant studies reported?	3	3
8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?	3	3
9. Were the conclusions supported by the data and/or analysis reported?	3	3
10. How wold you rate the scientific quality of this overview?	5	3

For items 1 through 9 scoring was done as follows: no = 1, partially = 2, yes = 3
For scoring of item 10 please see appendix G-1-2.
For reasons of space, the questions have been abbreviated, for full questions please see Appendix L 1-2.

Table 5: MA Quality Assessment Calibration Set & ICC Results

MA#	L	D	MA#	L	D	MA#	L	D	MA#	L	D
654	3	2	1014	3	2	1400	1	1	1707	3	3
	3	2		2	2		2	1		3	3
	3	3		3	2		1	1		3	3
	2	3		2	3		2	2		3	3
	1	1		1	1		1	1		3	3
	2	1		2	2		1	2		3	3
	3	3		3	3		3	3		3	3
	3	3		3	3		3	3		3	3
	3	3		3	3		3	3		3	3
	4	3		4	4		2	4		7	6
876	3	1	1059	3	2	1574	3	3	1857	3	3
	3	2		1	1		1	2		3	3
	3	3		3	2		3	3		3	3
	2	1		2	1		2	1		3	3
	1	1		1	1		3	1		3	3
	2	1		2	2		3	1		3	3
	3	3		3	3		3	1		3	3
	3	3		3	3		3	2		3	3
	3	3		3	3		3	2		3	3
	4	2		3	3		3	3		7	6
882	1	1	1153	3	2	1638	3	2	**Q#**	**ICC**	
	2	1		1	2		2	2	**1**	**0.62**	
	1	1		3	2		3	3	**2**	**0.51**	
	2	1		2	1		3	3	**3**	**0.83**	
	1	1		1	1		3	2	**4**	**0.51**	
	2	1		2	1		3	2	**5**	**0.75**	
	3	2		3	3		3	3	**6**	**0.46**	
	3	2		3	3		3	3	**7**	**na***	
	3	2		3	3		3	3	**8**	**na***	
	1	1		3	3		5	5	**9**	**na***	
917	1	1	1350	2	2	1657	1	1	**10**	**0.84**	
	1	1		1	2		2	1			
	1	1		1	1		1	2			
	1	1		2	1		2	1			
	1	1		1	1		1	1			
	1	1		1	1		1	1			
	3	3		3	3		3	3			
	3	2		3	3		3	3			
	3	2		3	3		3	3			
	2	2		2	3		1	2			

Scoring was done as follows: no = 1, partially = 2, yes = 3; for questions please see Appendix L 1-2.
* ICC is high but could not be estimated due to lack of variation (i.e. almost 100% agreement)

0.62, which indicated substantial agreement. In three cases, the ICC coefficients were between 0.46 and 0.51 indicating moderate agreement.

IV-5-1-3 Sample: In reviewing the entire sample (n=73), the overall scientific quality using the Oxman / Guyatt instrument ranged from a low of 1 to a high of 7. The median score for the total sample was 3 (interquartile range 2, 4). The median score for the MA which included grey literature was 3 (interquartile range 2, 5) compared to a median of 2 for both the explicit and implicit groups which did not include grey literature (interquartile range explicit group 2, 3, interquartile range implicit group 1, 4) (Figure 7 and 8). The exclusion of the meta-analyses with non-binary outcomes, from the grey inclusive group, does not seem to have biased the sample, as there was no difference detected in the quality of meta-analyses with binary outcomes and those with continuous outcomes in the grey literature excluded, comparator sample ($\chi^2 = 4.63$, $p = 0.59$).

IV-5-1-3-1 Overall Quality (Q10): Using an ANOVA (Table 6), no significant relationship was found in the summary score for inclusion and exclusion of grey literature (yes, no) ($p = 0.07$). However, MA that include grey literature tend to have a better summary score. No significant relationship was found between the three groups, grey included, grey explicitly excluded, and grey implicitly excluded ($p = 0.196$). When the two groups of no grey literature were separated and compared (Table 7), the lack of a relationship remained ($p = 0.20$).

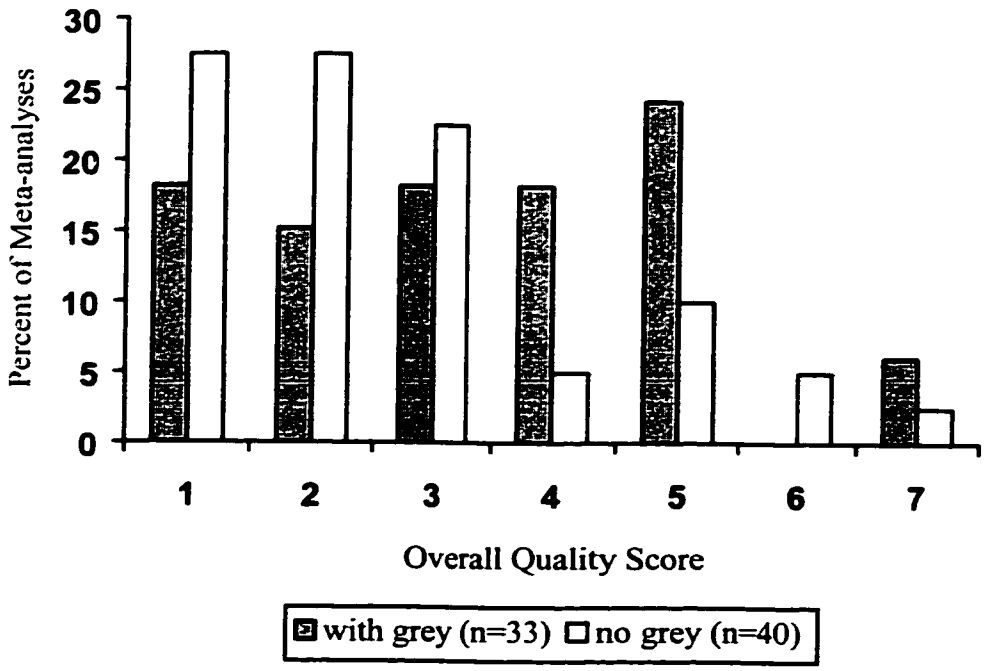


Figure 7: Overall scores for Quality of reporting of Meta-analyses: Summary score of the Oxman & Guyatt Index broken down into two groups; Include Grey Literature and Exclude Grey Literature.

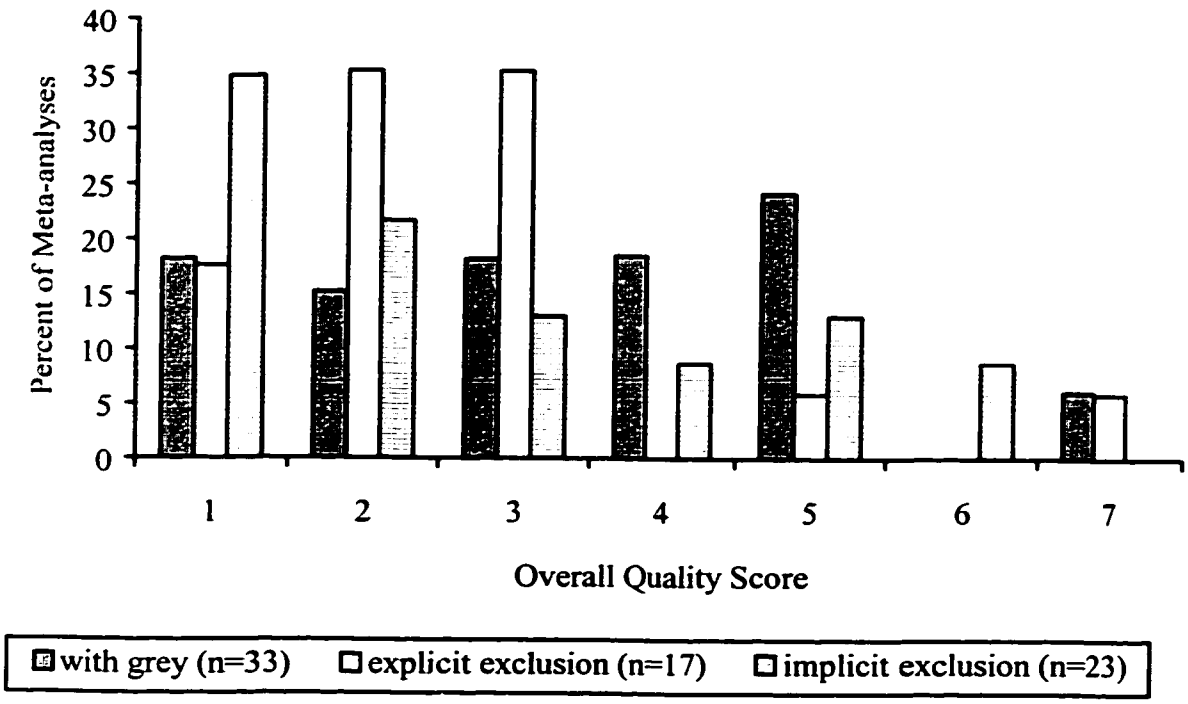


Figure 8: Overall scores for Quality of reporting of Meta-analyses: Summary score of the Oxman & Guyatt Index broken down into three groups; Include Grey Literature, Explicitly Exclude Grey Literature, and Implicitly Exclude Grey Literature.

Table 6: Comparison of Quality of reporting of meta-analyses that include grey literature to a sample that do not.

	MA with Grey n = 33		MA without Grey n = 40	
Oxman & Guyatt Instrument Item:	Yes (%)	partially/ don't know (%)	Yes (%)	partially/ don't know (%)
1. Were the search methods stated?	54.5	33.3	52.5	30.0
2. Was the search for evidence comprehensive?	24.2	54.5	17.5	52.5
3. Were the criteria for deciding which studies to include reported?	69.7	21.2	57.5	17.5
4. Was bias in the selection of studies avoided?	15.2	84.8	10.0	90.0
5. Were the criteria for assessing the validity of studies reported?	69.7	12.1	60.0	7.5
6. Was the validity of studies referred to in the text assessed using appropriate criteria?	51.5	27.3	50.0	47.5
7. Were the methods used to combine the findings reported?	93.9	6.1	77.5	17.5
8. Were the findings combined appropriately relative to the primary question?	97.0	3.0	92.5	7.5
9. Were the conclusions supported by the data and/or analysis reported in the overview?	90.9	9.1	87.5	10.0
10. How would you rate the scientific quality of the overview?				
median score		3		2
interquartile range		2, 5		1, 3

The meta-analyses which scored 'NO' have not been included in this table. They represent the remainder of the distribution ('yes' + 'partially/can't tell' = 100% - 'no').

ANOVA

			Experimental Method				
			Sum of Squares	df	Mean Square	F	Sig.
O10 summary score	main effects	grey*	9.35	1	9.35	3.37	0.070
	residual		196.65	71	2.77		
	total		206.00	72	2.86		

* grey refers to the inclusion of grey literature in the meta-analysis, the two categories are grey literature included, and grey literature excluded.

Table 7: Breakdown of Quality of Reporting of Meta-Analyses without Grey Literature

	Implicit exclusion n = 23		Explicit exclusion n = 17	
Oxman & Guyatt Instrument Item:	Yes (%)	partially/ don't know (%)	Yes (%)	partially/ don't know (%)
1. Were the search methods stated?	47.8	30.5	58.8	29.4
2. Was the search for evidence comprehensive?	21.7	52.2	11.8	52.9
3. Were the criteria for deciding which studies to include reported?	47.8	17.4	70.6	17.6
4. Was bias in the selection of studies avoided?	13.0	87.0	5.9	94.1
5. Were the criteria for assessing the validity of studies reported?	52.2	13.0	70.6	0
6. Was the validity of studies referred to in the text assessed using appropriate criteria?	43.5	52.2	58.8	41.2
7. Were the methods used to combine the findings reported?	73.9	21.8	82.4	11.7
8. Were the findings combined appropriately relative to the primary question?	91.3	8.7	94.1	5.9
9. Were the conclusions supported by the data and/or analysis reported in the overview?	78.3	17.4	100.0	0
10. How would you rate the scientific quality of the overview?				
median score		2		2
interquartile range		1, 4		2, 3

The meta-analyses which scored 'NO' have not been included in this table. They represent the remainder of the distribution ('yes' + 'partially/can't tell' = 100% - 'no').

ANOVA

			Experimental Method				
			Sum of Squares	df	Mean Square	F	Sig.
O10 summary score	main effects	sought*	9.37	2	4.68	1.67	0.196
	residual		196.63	70	2.81		
	total		206.00	72	2.86		

*sought refers to whether grey literature was sought for inclusion, the three categories are grey literature included, grey literature explicitly excluded, and grey literature implicitly excluded.

IV-5-1-3-2 Individual Items (Q1-9): There were no significant differences between the grey literature inclusive group and the two grey literature excluded groups for the individual items of the instrument (Table 8). MA which include grey literature, compared to those that implicitly excluded grey literature tended to score slightly better on all items of the Oxman / Guyatt index. When comparing MA that include grey literature to those which explicitly exclude it, MA with grey literature tended to score slightly higher on all items except those dealing with details of searching and selection of included studies. For these items (items 1,3,5 and 6) the explicit exclusion group scored slightly higher than the grey inclusive group. None of these differences were statistically significant. It was not possible to generate an OR for the explicit group for item 9 (i.e.: 'were the conclusions made by the author(s) supported by the data and / or analysis reported in the overview?') as there were too few MA in either group that scored 'no' or 'partially' on this item.

In the sensitivity analysis, (Table 9) where 'partially' and 'can't tell' were grouped with 'yes' and compared to 'no', there was a significant difference in only one item, item #3, (i.e.:'were the criteria (inclusion/exclusion) used for deciding which studies to include in the overview reported?'). For this item, the MA which implicitly excluded grey literature were more likely to have scored 'yes' or 'partially' than the grey inclusive group (OR: 5.33 (95% CI: 1.23, 23.07)).

Table 8: Effect of Grey Literature on Quality of Reporting (odds ratios).

Outcome Measures: adequately reported vs. not or partially reported.
Methods of analysis: logistic regression
Reference Level = grey literature included.

Scale Item	Grey Excluded OR (95% CI)	
	Implicit (n = 23)	**Explicit (n = 17)**
1. Were the search methods used to find evidence (original research) on the primary question(s) stated?	0.76 (0.26, 2.22)	1.19 (0.36, 3.89)
2. Was the search for evidence reasonably comprehensive?	0.87 (0.24, 3.09)	0.42 (0.78, 2.23)
3. Were the criteria (inclusion/exclusion) used for deciding which studies to include in the overview reported?	0.40 (0.13, 1.20)	1.04 (0.29, 3.75)
4. Was bias in the selection of studies avoided?	0.84 (0.18, 3.93)	0.35 (0.04, 3.27)
5. Were the criteria (methodological quality) used for assessing the validity of the included studies reported?	0.47 (0.16, 1.43)	1.04 (0.29, 3.75)
6. Was the validity of all studies referred to in the text assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)?	0.72 (0.25, 2.11)	1.34 (0.41, 4.39)
7. Were the methods used to combine the findings of the relevant studies (to reach a conclusion) reported?	0.18 (0.03, 1.00)	0.30 (0.05, 2.01)
8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?	0.33 (0.03 3.85)	0.50 (0.03, 8.52)
9. Were the conclusions made by the author(s) supported by the data and/or analysis reported in the overview?	0.36 (0.08, 1.69)	*

*no regression co-efficient was generated here as all of the MA in this group scored "yes" on this item, and most (91%) of the MA in the comparator group (grey literature included) also scored "yes".
OR>1 indicates a better score for the item compared to 'grey literature included'.

Table 9: Effect of Grey Literature on Quality of Reporting (odds ratios).

Sensitivity analysis
Outcome Measures: adequately or partially reported vs. not reported.
Methods of analysis: logistic regression
Reference Level = grey literature included.

Scale Item	Grey Excluded OR (95% CI)	
	Implicit (n = 23)	**Explicit (n = 17)**
1. Were the search methods used to find evidence (original research) on the primary question(s) stated?	2.01 (0.48, 8.50)	0.97 (0.16, 5.90)
2. Was the search for evidence reasonably comprehensive?	1.31 (0.38, 4.58)	2.03 (0.55, 7.42)
3. Were the criteria (inclusion/exclusion) used for deciding which studies to include in the overview reported?	5.33 (1.23, 23.07)	1.33 (0.20, 8.86)
4. Was bias in the selection of studies avoided?	*	*
5. Were the criteria (methodological quality) used for assessing the validity of the included studies reported?	2.40 (0.70, 8.23)	1.87 (0.48, 7.36)
6. Was the validity of all studies referred to in the text assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)?	0.17 (0.19, 1.48)	*
7. Were the methods used to combine the findings of the relevant studies (to reach a conclusion) reported?	*	*
8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?	*	*
9. Were the conclusions made by the author(s) supported by the data and/or analysis reported in the overview?	*	*

*no regression co-efficient was generated here as most of the MA scored "yes" or partially / can't tell.
OR>1 indicates a better score for the item compared to 'grey literature included'.

IV-5-2 Objective 2-2 Quality of Reporting of RCTs

IV-5-2-1 Calibration Set: The quality, as scored by each of the two reviewers, for the 5 calibration RCTs appears in Table 10. There were points of disagreement in scoring 2 of the 5 RCTs. In both there was disagreement on the withdrawals and dropouts item. In one of the RCTs, randomization and study design were also scored differently. We met and reviewed the areas of disagreement and clarified scoring of those items (i.e. withdrawals and drop-outs; to score the point, this item had to be clearly described in the text, if there were no withdrawals or drop outs, this had to be mentioned).

IV-5-2-2 ICC Set: The quality scores, by reviewer, are in Table 11. The main difference remained with the withdrawals and dropouts question. Given that I tended to be more lenient with scoring, my scores may be slightly higher than what has been reported in the literature if the trend holds for the sample. The agreement exercise was done mainly to ensure that the scoring was validated. The overall intraclass correlation coefficient for the Jadad score was = 0.82. This indicates almost perfect agreement <58>. For the allocation concealment item, a kappa score of 0.63 was achieved, which indicates substantial agreement. No level of agreement was generated for the study design item, as we had perfect agreement.

IV-5-2-3 Sample: In looking at the entire sample retrieved (n= 429), the quality scores using Jadad's scale ranged from a high of 5 to a low of 0. The mean score for the total sample was, 2.24 (std dev: 1.17). Using only the grey literature, the scores

Table 10: Calibration Exercise for Quality Assessment of Randomized Controlled Trials Using the Jadad Scale.

TRIAL #	ITEM	DM	LM
1	1. Randomization	1	2
	2. Double blind	2	2
	3. Withdrawals	0	1
	Total	3	5
	Allocation concealment	0	0
	Design	c/o	paral
2	1. Randomization	2	2
	2. Double blind	2	2
	3. Withdrawals	1	1
	Total	5	5
	Allocation concealment	0	0
	Design	paral	paral
3	1. Randomization	1	1
	2. Double blind	1	1
	3. Withdrawals	0	0
	Total	2	2
	Allocation concealment	0	0
	Design	paral	paral
4	1. Randomization	1	1
	2. Double blind	0	0
	3. Withdrawals	0	1
	Total	1	2
	Allocation concealment	0	0
	Design	paral	paral
5	1. Randomization	1	1
	2. Double blind	0	0
	3. Withdrawals	0	0
	Total	1	1
	Allocation concealment	0	0
	Design	paral	paral

see appendix G-2-2 for a description of scoring of this scale.

Table 11: RCT Quality Assessment Scores Using the Jadad Scale: For Calculation of Intra Class Correlation

Trial	ITEM	DM	LM	Trial	ITEM	DM	LM
	1. Randomization	2	2		1. Randomization	2	2
1	2. Double blind	0	0	8	2. Double blind	0	0
	3. Withdrawals	1	1		3. Withdrawals	0	1
	Total	3	3		Total	2	3
	Allocation concealment	0	0		Allocation concealment	0	0
	Design	para	para		Design	para	para
	1. Randomization	2	2		1. Randomization	2	2
2	2. Double blind	0	0	9	2. Double blind	0	0
	3. Withdrawals	1	1		3. Withdrawals	0	1
	Total	3	3		Total	2	3
	Allocation concealment	0	1		Allocation concealment	0	0
	Design	para	para		Design	para	para
	1. Randomization	1	1		1. Randomization	2	2
3	2. Double blind	0	0	10	2. Double blind	0	0
	3. Withdrawals	0	1		3. Withdrawals	1	1
	Total	1	2		Total	3	3
	Allocation concealment	0	0		Allocation concealment	0	0
	Design	para	para		Design	para	para
	1. Randomization	2	2		1. Randomization	2	2
4	2. Double blind	2	1	11	2. Double blind	2	2
	3. Withdrawals	0	0		3. Withdrawals	1	1
	Total	4	3		Total	5	5
	Allocation concealment	0	0		Allocation concealment	0	1
	Design	para	para		Design	para	para
	1. Randomization	2	2		1. Randomization	2	2
5	2. Double blind	1	1	12	2. Double blind	2	2
	3. Withdrawals	1	1		3. Withdrawals	0	0
	Total	4	4		Total	4	4
	Allocation concealment	0	0		Allocation concealment	0	0
	Design	para	para		Design	para	para
	1. Randomization	2	2		1. Randomization	2	2
6	2. Double blind	0	0	13	2. Double blind	2	2
	3. Withdrawals	0	0		3. Withdrawals	0	1
	Total	2	2		Total	4	5
	Allocation concealment	0	0		Allocation concealment	0	0
	Design	para	para		Design	para	para
	1. Randomization	1	1		1. Randomization	2	2
7	2. Double blind	1	1	14	2. Double blind	0	0
	3. Withdrawals	1	1		3. Withdrawals	1	1
	Total	3	3		Total	3	3
	Allocation concealment	0	0		Allocation concealment	0	0
	Design	para	para		Design	para	para

ICC for the Jadad score of trial quality = 0.82.
Kappa for allocation concealment = 0.63

ranged from a high of 4 to a low of 0. In comparing the published literature to the grey literature (Figure 9), the mean score was higher for the published RCTs (2.39, std dev: 1.17) as compared to the grey RCTs (1.54, std dev: 0.84). This difference is statistically significant (mean difference 0.86, 95% CI 0.58, 1.14).

As the interaction there was no interaction between publication status and year of publication each term was considered individually. Publication status was significantly associated with overall quality score. Grey literature was of significantly lower quality than published literature. Year of publication did not significantly contribute to the overall quality score (Table 12).

Initially, I intended to look at allocation concealment and study design to see if they impacted on the quality score. However, there was too little variability with only 6.9% of the studies reporting allocation concealment, and 98% of the studies being parallel group design.

IV-6 Objective 3

IV-6-1 Replication of Meta-Analyses

As demonstrated in Figure 10 and Table 13, the replicated intervention effects were very similar to those that appeared in the publications. Of the 41 MA replicated (in 33 published reports), discrepancies of greater than 10% were only found in 1 case (MA #1451-2). In this case the published report shows a slightly higher incidence difference. However, both the published and the replicated results are significant. It is important to note that the authors report an incidence difference, which was felt to be comparable to the risk difference, as they both represent proportions of patients with the outcome of

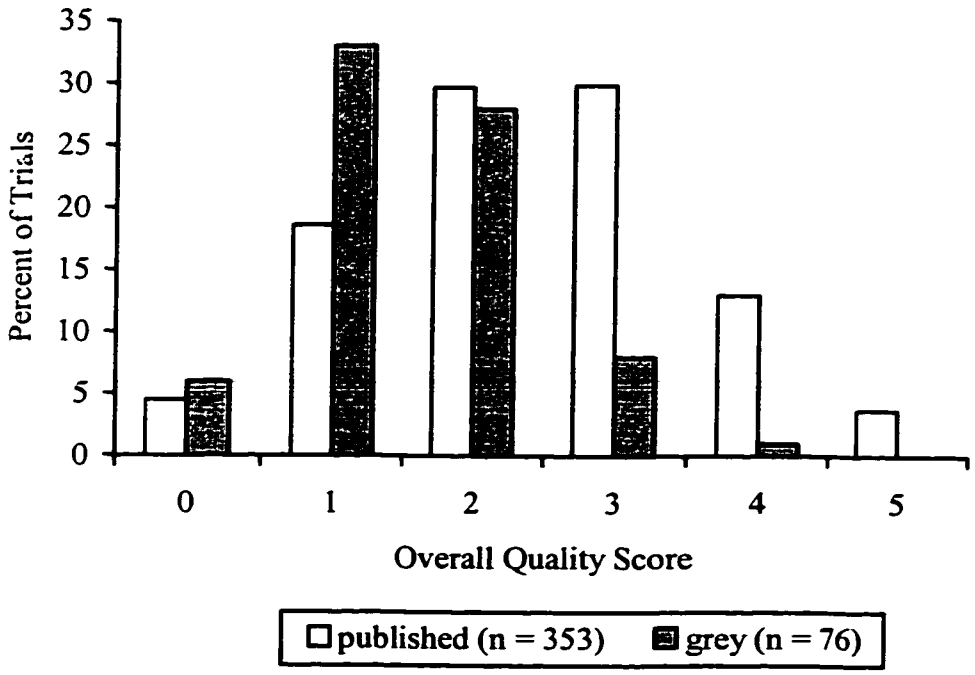


Figure 9: Overall scores for Quality of reporting of the RCTs included in the meta-analyses: Total score of the Jadad Scale broken down by the two groups; 'published' and grey literature.

Table 12: RCT Quality Assessment Scores; presented for the entire sample and then broken down by publication status (i.e. grey literature or 'published' literature)

WHOLE SAMPLE QUALITY

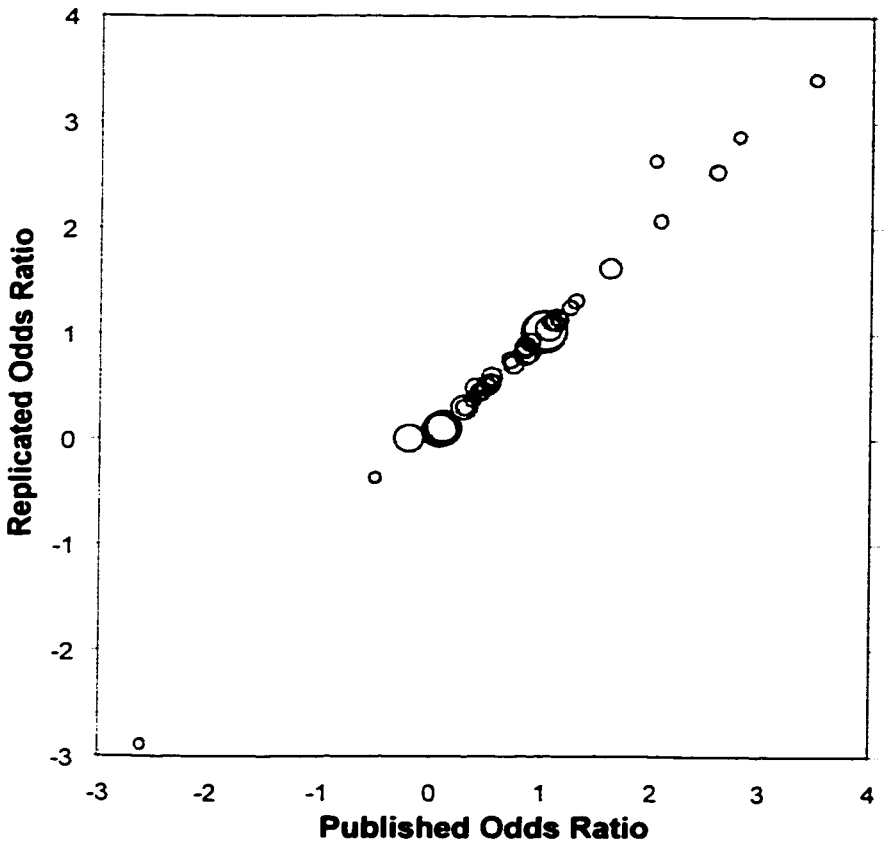
Jadad score item	Mean	Std Dev	Minimum	Maximum	N
Randomization	1.12	0.56	0	2	429
Double - Blinding	0.61	0.71	0	2	429
Withdrawals / Drop-outs	0.51	0.50	0	1	429
Total score	2.25	1.17	0	5	429

	Non Grey n = 353		Grey n = 76	
Jadad Scale Item	Mean	Std Dev	Mean	Std Dev
Randomization	1.17	0.57	0.89	0.42
Double - Blinding	0.65	0.73	0.41	0.52
Withdrawals / Drop-outs	0.57	0.50	0.24	0.43
Total Jadad Score	2.40	1.17	1.54	0.84
Allocation Concealment	7.9%		2.6%	
Parallel group design	98.6%		95.1%	

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Jadad score total	pub status	44.27	1	44.27	35.01	0.000
	year	0.75	3	0.25	0.20	0.897
	pub status * year	2.21	3	0.74	0.58	0.627
	Error	525.97	416	1.26		
	Total corrected	573.20	423			

The number of trials included in this analysis is 424. Five grey items have been excluded as they are unpublished and therefore have no year of publication.



(log scale)

Figure 10: Odds ratios as published versus replicated odds ratios for the 39 analyses included in the 33 meta-analyses which include grey literature. The size of the plotting circles is inversely proportional to the variance of the "published" estimates.

Table 13: Replication of Published Meta-Analyses: Using the Summary Statistic Presented in the Published Meta-analysis.

MA	Statistic	Published			Replicated		
		n	Pt .est.	95% CI	n	Pt. est.	95% CI
406-1	MH RR	7	0.56	0.33, 0.94	7	0.54	0.31, 0.96
406-2	MH RR	6	1.08	0.84, 1.39	6	1.11	0.86, 1.43
406-3	MH RR	2	0.38	0.13, 1.06	2	0.37	0.12, 1.10
450	MH OR	6	40%	10%, 60%	6	0.49	0.32, 0.75
451	MH RR	9	1.13	0.99, 1.29	9	1.13	0.99, 1.29
473*	RR	17	1.03	0.99, 1.06	17	1.03	0.98, 1.09
508	D&L	10	-0.20	-9.20, 8.80	10	0.0084	-0.084, 0.10
627*	Peto OR	8	3.50	2.60, 4.70	8	3.41	2.52, 4.61
634*	MH OR	15	2.60	2.20, 3.00	15	2.54	2.18, 2.97
743*	MH OR	4	2.80	1.69, 4.62	4	2.87	1.70, 4.84
768*	D&L RD	13	0.11	NR	13	0.11	0.03, 0.19
790#*-1	MH OR	6	NR	NR	6	5.51	2.92, 10.39
790#*-2	MH OR	6	NR	NR	6	1.92	0.80, 4.59
818	MH RR	16	0.72	0.46, 1.11	16	0.75	0.50, 1.12
864	OR	24	0.87	0.79, 0.97	24	0.83	0.71, 0.97
935-1*	Peto OR	8	2.07	1.64, 2.62	8	2.07	1.64, 2.62
935-2*	Peto OR	40	1.61	1.46, 1.78	40	1.62	1.46, 1.80
951	RR	14	1.30	0.90, 1.80	14	1.31	0.97, 1.79
970-1*	OR	14	2.02	1.50, 2.72	14	2.64	1.89, 3.70
970-2*	OR	12	1.15	0.83, 1.61	12	1.14	0.84, 1.55
970-3*	OR	7	0.85	0.64, 1.15	7	0.85	0.64, 1.15
1039	Peto OR	6	0.50	0.34, 0.73	6	0.50	0.34, 0.73
1049	MH OR	4	0.46	NR	4	0.45	0.25, 0.79
1210	OR	9	0.75	0.57, 1.06	9	0.71	0.53, 0.95
1216	MH OR	22	1.02	0.97, 1.07	22	1.02	0.97, 1.07
1221-1	D&L OR	9	0.30	0.22, 0.43	9	0.30	0.21, 0.44
1221-2	D&L OR	38	0.85	0.74, 0.97	38	0.81	0.69, 0.97
1251*	risk diff	8	10.5%	4.8%, 16.4%	8	9.95%	3.53%, 16.37%
1325	MH OR	3	0.90	0.70, 1.20	3	0.92	0.72, 1.17
1344	D&L OR	6	0.54	.037, 0.79	6	0.53	0.36, 0.78
1351	risk diff	15	8%	1%, 14%	15	8%	1%, 15%
1353	MH OR	30	0.55	0.40, 0.76	29	0.59	0.43, 0.81
1418	Peto OR	7	0.45	0.28, 0.71	7	0.45	0.28, 0.72
1429*	Peto OR	5	10.24	NR	5	10.14	4.10, 25.10
1433	fishersX^2	18	0.85	NR	18	0.88	0.70, 1.11
1451-1	incid diff	5	-2.60	-6.6, 1.3	5	rd -2.89	-7.80, 2.00
1451-2	incid diff	12	-0.50	-4.2, 3.2	12	rd -0.37	-4.90, 4.24
1626*		23	NR	NR	23	0.28	0.23, 0.34
1637	MH OR	20	1.06	0.96, 1.18	20	1.05	0.94, 1.17
1670	MH OR	4	0.31	0.15, 0.65	4	0.30	0.14, 0.66
1745*	MH OR	4	1.25	0.93, 1.66	4	1.25	0.93, 1.67

no point estimate published in meta-analysis

*outcome is for a positive event, replicated as done, for the regression, we have converted the outcome to a negative event (i.e. survival is converted to mortality)

NR = not reported.

interest. In one case the replication yielded a significant result when the published MA (#1210) had reported a non-significant result.

IV-6-2 Repetition of Meta-Analyses

Table 14 represents the comparison of the meta-analyses (n = 39) as published, with the meta-analyses after having removed the grey literature. When the replicated odds ratio is plotted against the repeated odds ratio (i.e. removal of the grey literature) (Figure 11) a noticeable shift off the diagonal line is observed, indicating a change in the OR with the exclusion of grey literature. A relative change in the point estimate by 10% or more after the removal of the grey literature was observed in 14 of the 41 analyses. In nine of these cases, removal of the grey literature causes the point estimate to shift away from unity. In three MA, the exclusion of the grey literature caused a change in the significance of the results, with the results becoming significant in 2 cases. Figure 12 demonstrates visually the general trend towards more significant results after the removal of grey literature. The paired t-test confirmed the visual interpretation. When the Z score calculated with the grey literature excluded was compared to the Z score of the replicated odds ratios (published and grey inclusive) there was a significant decrease in Z scores ($t = -7.257$, $p < 0.001$). The decrease in Z-scores with the inclusion of grey literature indicates a shift toward accepting the null hypothesis when grey literature is included, which means a decreased chance of getting a significant result. Table 15 presents the results of the sensitivity analysis in which abstracts were removed from the sample.

At the RCT level (n = 467) the logistic regression demonstrates statistically the impact of the grey literature. The regression coefficient for the interaction of intervention and

Table 14: Peto OR for Replicated MA with and without Grey Literature

**outcome in published MA is a positive event, converted here to an unwanted event for this and subsequent analyses.*

MA #	Replicated			Grey Removed		
	N	Pt. est.	95% CI	N	Pt. est.	95% CI
406-1	7	0.56	0.32, 0.96	4	0.72	0.34, 1.52$^{\leftrightarrow}$ ^{XX}
406-2	6	1.11	0.86, 1.43	5	1.07	0.82, 1.40
406-3	2	0.38	0.13, 1.06	1	0.17	0.02, 1.15$^{\leftrightarrow}$
450	6	0.50	0.33, 0.75	5	0.52	0.34, 0.81
451	9	1.22	0.96, 1.56	8	1.23	0.96, 1.58
473*	17	0.77	0.57, 1.05	16	0.80	0.59, 1.10
508	10	1.14	0.91, 1.44	8	1.14	0.90, 1.45
627*	8	0.44	0.33, 0.60	2	1.04	0.59, 1.83$^{\leftrightarrow}$ ^{XX}
634*	15	0.41	0.35, 0.47	13	0.40	0.34, 0.47
743*	4	0.35	0.21, 0.59	3	0.38	0.23, 0.65
768*	13	0.58	0.46, 0.72	10	0.39	0.30, 0.52$^{\leftrightarrow}$
790*-1	6	0.19	0.11, 0.33	5	0.18	0.10, 0.33
790*-2	6	0.49	0.20, 1.20	5	0.37	0.12, 1.12^{+}
818	16	0.67	0.43, 1.05	15	0.67	0.43, 1.06
864	24	0.83	0.71, 0.97	20	0.80	0.67, 0.97
935-1*	8	0.58	0.45, 0.74	7	0.49	0.37, 0.64^{+}
935-2*	40	0.54	0.49, 0.60	36	0.55	0.49, 0.61
951	14	1.56	1.06, 2.29	11	1.54	0.99, 2.38 ^{XX}
970-1*	14	0.40	0.32, 0.50	9	0.46	0.31, 0.69^{+}
970-2*	12	0.85	0.63, 1.14	10	0.88	0.65, 1.20
970-3*	7	1.24	0.92, 1.66	5	1.27	0.93, 1.75
1039	6	0.50	0.34, 0.73	4	0.49	0.30, 0.82
1049	4	0.46	0.27, 0.79	3	0.44	0.25, 0.78
1210	9	0.71	0.53, 0.95	7	0.70	0.52, 0.96
1216	22	1.02	0.97, 1.07	21	1.02	0.97, 1.07
1221-1	9	0.31	0.23, 0.43	4	0.27	0.16, 0.45^{+}
1221-2	38	0.80	0.69, 0.91	34	0.80	0.69, 0.92
1251*	8	2.17	1.62, 2.92	7	1.86	1.36, 2.54^{+}
1325	3	0.92	0.72, 1.17	2	0.88	0.67, 1.17
1344	6	0.54	0.37, 0.77	4	0.50	0.32, 0.76
1351	15	0.74	0.58, 0.94	10	0.53	0.39, 0.73$^{\leftrightarrow}$
1353	29	0.59	0.43, 0.81	21	0.57	0.42, 0.82
1418	7	0.45	0.28, 0.72	6	0.33	0.19, 0.56$^{\leftrightarrow}$
1429*	5	0.10	0.04, 0.24	3	0.13	0.03, 0.49^{+}
1433	18	0.88	0.70, 1.11	15	0.82	0.64, 1.06
1451-1	5	0.82	0.56, 1.20	4	0.76	0.51, 1.14
1451-2	12	0.99	0.75, 1.32	6	0.98	0.70, 1.37
1626*	23	0.28	0.23, 0.34	14	0.28	0.23, 0.35
1637	20	1.05	0.94, 1.17	19	1.05	0.94, 1.17
1670	4	0.34	0.18, 0.66	1	0.27	0.08, 0.91^{+}
1745*	4	0.80	0.60, 1.07	3	0.65	0.41, 1.03^{+}

difference between replicated and grey removed is $\geq 10\%^{+}$; $\geq 25\%^{\leftrightarrow}$; change in significance ^{XX}

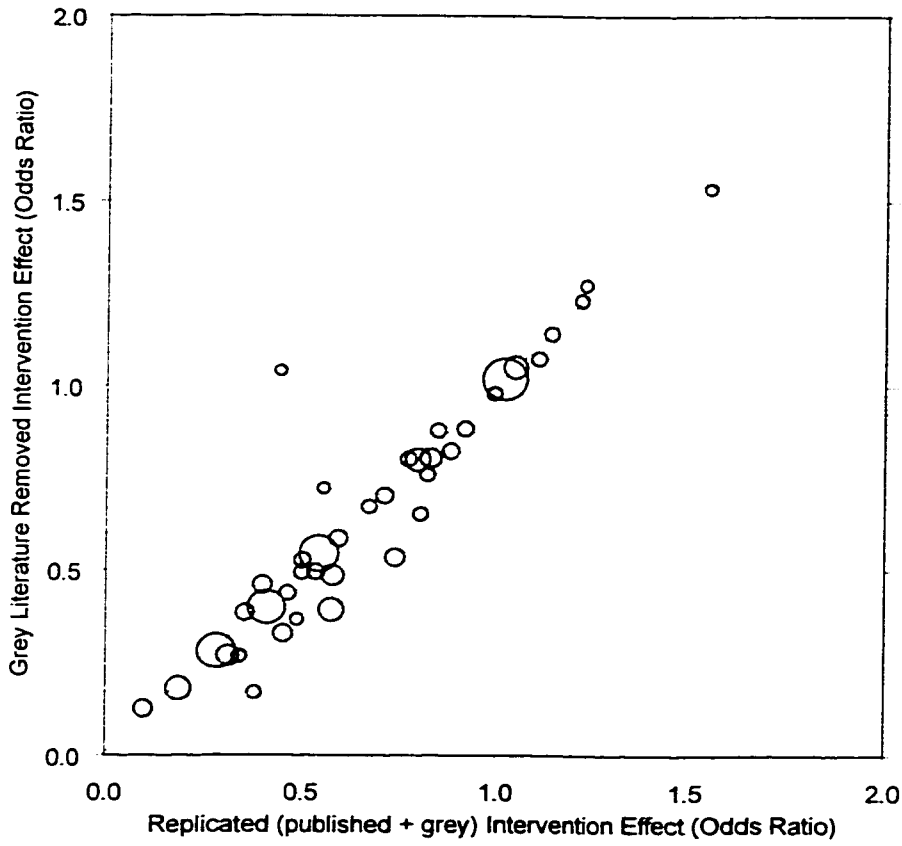


Figure 11: Pooled odds ratios with grey literature included (x-axis) are plotted against the corresponding odds ratios of the same meta-analyses after removal of the grey literature (y-axis). The size of the plotting circles is inversely proportional to the variance of the "grey literature removed" estimates.

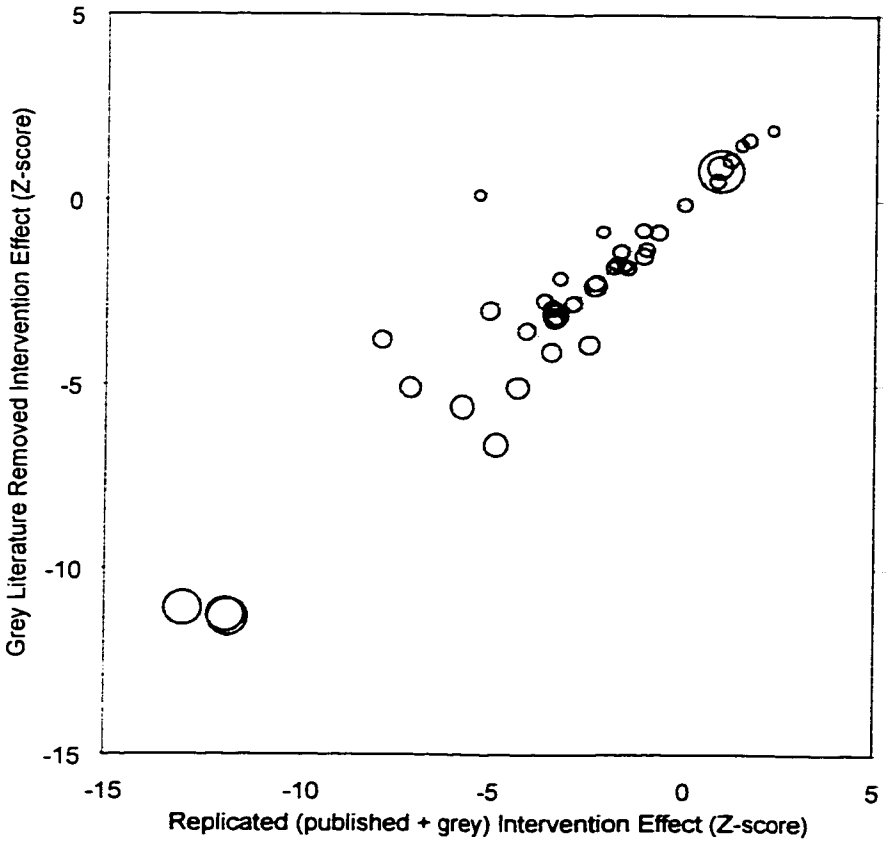


Figure 12: The standardized Z-statistic (log odds ratio divided by its standard error) from the replicated meta-analyses (x-axis) were plotted against the standardized z-statistic from the same meta-analyses with the grey literature removed (y-axis). The size of the plotting circles is inversely proportional to the variance of the "grey literature removed" estimates.

Table 15: Peto OR for Replicated MA with (minus abstracts) and without Grey Literature.

**outcome in published MA is a positive event, converted here to an unwanted event for this and subsequent analyses*

MA #	Replicated			Grey Removed		
	N	Pt. est.	95% CI	N	Pt. est.	95% CI
627*	4	0.75	0.50, 1.13	2	1.04	0.59, 1.83
634*	15	0.41	0.35, 0.47	13	0.40	0.34, 0.47
768*	13	0.58	0.46, 0.72	10	0.39	0.30, 0.52
790*-1	6	0.19	0.11, 0.33	5	0.18	0.10, 0.33
790*-2	6	0.49	0.20, 1.20	5	0.37	0.12, 1.12
864	22	0.83	0.70, 0.99	20	0.80	0.67, 0.97
935-2*	37	0.54	0.49, 0.60	36	0.55	0.49, 0.61
951	14	1.56	1.06, 2.29	11	1.54	0.99, 2.38
970-3*	6	1.28	0.94, 1.74	5	1.27	0.93, 1.75
1039	5	0.52	0.35, 0.76	4	0.49	0.30, 0.82
1049	4	0.46	0.27, 0.79	3	0.44	0.25, 0.78
1221-1	5	0.27	0.16, 0.45	4	0.27	0.16, 0.45
1251*	8	2.17	1.62, 2.92	7	1.86	1.36, 2.54
1325	3	0.92	0.72, 1.17	2	0.88	0.67, 1.17
1344	5	0.48	0.32, 0.71	4	0.50	0.32, 0.76
1353	24	0.60	0.44, 0.84	21	0.57	0.42, 0.82
1451-1	5	0.82	0.56, 1.20	4	0.76	0.51, 1.14
1451-2	12	0.99	0.75, 1.32	6	0.98	0.70, 1.37
1626*	22	0.29	0.24, 0.35	14	0.28	0.23, 0.35
1670	2	0.38	0.16, 0.91	1	0.27	0.08, 0.91

Only Meta-analyses that included abstracts as a source of grey literature are considered here, those that did not contain abstracts would not differ from table 14.

publication status was 1.12 (95% CI 1.01, 1.23). This indicates that on average, the exclusion of grey literature compared to its inclusion (Table 16), results in a greater estimate of the intervention effect. When abstracts were removed from this analysis, this overestimate increased (ROR 1.38, 95% CI 1.15, 1.64). There was no difference between the estimates of intervention effectivness of published trials compared to abstracts (ROR 1.02, 95% CI 0.91, 1.14).

IV-6-2-1 Example: To illustrate the effect grey literature has on the point estimate in an individual meta-analysis, we examined the treatment of chronic venous insufficiency with hydroxyethylrutosides <MA#768> by Poynard and Valterio. This published meta-analysis includes 13 RCTs. Of the included RCTs, 10 were published in journals, 2 were internal reports, and the last one was 'in preparation for publication'. In this MA, we considered only the outcome 'leg pain'. Although the authors reported relief in leg pain, we inferred lack of relief in leg pain from the presented data as we wanted all outcomes reported as negative events. Using all studies included by Poynard and Valterio, the odds ratio was 0.58 (0.46, 0.72) which represents a reduction in the odds of persistent pain by 42%. When the 3 grey items are removed from the analysis, the reduction in persistent pain increased to 63% (OR 0.39, 95% CI: 0.30, 0.52). The published literature alone gives higher estimates of the effectiveness of hydroxyethylrutosides for pain relief in the treatment of chronic venous insufficiency.

Table 16: The Impact of Grey Literature on Meta-Analysis at the Trial Level.

Comparison[1]	# Randomized trials	Grey Effect [exponent of coefficient] ROR (95% CI)[2]
Grey (all) vs Published	102 vs 365	1.12 (1.01, 1.23)
Grey (non-abstract) vs Published	39 vs 365	1.38 (1.15, 1.64)

1. The model: log OR = $\beta_0 + \beta_1$ [trial i^{th}] + β_2 [intervention j^{th}] + β_3 [MA k^{th}] + β_4 [publication status l^{th}] + β_5 [intervention j^{th} * MA k^{th}] + β_6 [intervention j^{th} * publication status l^{th}] + ε (* denotes interaction)

2. Exponent of the interaction term from the logistic regression (Ratio of Odds Ratios) greater than 1 implies that non-grey literature has a larger treatment effect than grey literature.

IV-6-2-2 Quality of Reporting: When quality of reporting is added to the model, the results become non-significant for both the impact of grey literature (ROR 1.04; 95% CI 0.92, 1.16), and for the impact of quality (ROR 0.98; 95% CI 0.91, 1.06) (Table 17). In a sensitivity analysis, as was done for publication alone, abstracts were removed from the analysis. This resulted in a significant effect for publication status (ROR 1.34; 95% CI 1.01, 1.78) but not for quality of reporting (ROR 0.96; 95% CI 0.89, 1.03). Quality on its own also had no effect on the estimate of treatment effectiveness.

Table 17: The Impact of Grey Literature on Meta-Analysis at the Trial Level after taking Quality into Account.

Comparison[1]	# Randomized trials	Grey Effect [exponent of coefficient] ROR (95% CI)[2]
Grey (all) vs Published	102 vs 365	1.04 (0.92, 1.16)
Grey (non-abstracts) vs Published	39 vs 365	1.34 (1.01, 1.78)

1. The basic model: log OR = $\beta_0 + \beta_1$ [trial i^{th}] + β_2 [intervention j^{th}] + β_3 [MA k^{th}] + β_4 [publication status l^{th}] + β_5[jadad score p^{th}] + β_6 [intervention j^{th} * MA k^{th}] + β_7 [intervention j^{th} * publication status l^{th}] + β_8[intervention j^{th} * jadad score p^{th}] + β_9 [publication status l^{th} * jadad score p^{th}] + ε. (* denotes interactions)
Quality was dichotomized at 2 (low quality Jadad score ≤ 2, high quality Jadad score >2.)

2. Exponent of the interaction term from the logistic regression including quality of reporting (Ratio of Odds Ratios)greater than 1 implies that published literature has a larger treatment effect than grey literature.

V Discussion

V-1 Characterization and Prevalence of Grey Literature in Meta Analysis

In the sample of evaluable published meta-analyses, grey literature was found in 33%. These results are slightly higher but consistent with what has previously been found in published meta-analyses catalogued on MEDLINE. Cook et al <40> reported 31% of 150 meta-analyses they found included 'unpublished' literature. Although the term unpublished literature is used, the sources of information are equivalent to those included in my definition of grey literature (abstracts, presentations, theses, dissertations, books, reports, and unpublished). In Cook's sample, using what appears to be a similar sampling method, but for an earlier time period (1989 to 1991), the grey literature accounted for 11% of the trials, while in this sample it accounts for 23%. No reason for this difference could be identified except sample variation and perhaps year of publication. This rate is suggestive that inclusion of grey literature is not uncommon in meta-analytic practice. There remains much debate among researchers and editors over the merit of the inclusion of grey literature in meta-analysis. This is clear when we consider the sample of meta-analyses that explicitly excluded grey literature. The exclusion may be due to a feeling that this literature did not merit inclusion, or it may have simply been excluded because the authors did not have the capacity to identify and retrieve the grey literature. The fact that grey literature is mentioned in the methods section provides evidence of the lack of consensus concerning its inclusion.

To my knowledge there has been little research into the types of grey literature included in meta-analyses, and their prevalence. The findings reported here depend heavily on the

definition of grey literature used. Because of the importance of the definition, I sought input from various people with expertise in the areas of publication bias and library sciences. The prevalence of abstracts in this sample may indicate either, abstracts are not considered grey literature by all authors and therefore less controversial for inclusion, or that abstracts although considered grey literature are not difficult to retrieve, or a combination of the two. Some abstracts are published in conference proceedings and may be catalogued on MEDLINE, which makes them quite accessible, and in some cases peer reviewed.

There is a risk that identifiable grey literature may not be a representative sample of what actually exists <42>. In this sample, this risk is realized in at least one meta-analysis. In the methods section of MA # 951, the following statement indicates a possibly biased sample, and demonstrates the lack of acceptance of the universal inclusion of grey literature. "Except for two unpublished references, which were obtained from the Diarrhoeal Disease Control Programme of the World Health Organization (WHO), only published reports were considered in an attempt to ensure the quality of research. The exceptions were allowed because the quality of the studies were closely monitored by WHO personnel." This biased inclusion is as concerning as a biased inclusion of published literature, especially since there is no empirical evidence to substantiate the authors' concern about quality.

It is likely that in the future the sources of grey literature, if there is such a thing, will change. Efforts are being made to make all research more accessible. These efforts, which include trial registries, databases of grey literature and the open door policies of some drug companies <63>, are being made as there is consensus that publication bias

exists. This thesis provides the empiric evidence that these efforts are justified and needed.

V-2 Quality of Reporting

V-2-1 Meta-Analyses

The overall quality of reporting of the meta-analyses in the sample were slightly lower, median 3 (interquartile range 2,4), than what has been found previously in the published literature. Moher et al found <64>, in their sample of language inclusive meta-analyses, a median quality summary score of 4 (interquartile range 3,4) and Jadad and McQuay <65> found a median quality score of 4 in 80 meta-analyses in the pain literature. Although the median scores reported in both of these papers is higher than the median score reported here, there is agreement that on average the reporting of meta-analyses is of poor quality. Oxman and Guyatt <56> suggest that a score of 3 indicates major flaws and a score of 5 indicates minor flaws. When scores on the individual items are considered, a greater proportion of the reports scored by Jadad and McQuay compared to this sample fulfilled the following criteria: search methods stated, comprehensive search for evidence, criteria for inclusion of studies reported, and avoidance of biased inclusion. This may be due in part to the interpretation of the items; for example, in order for the search to be scored as comprehensive, we insisted on three sources with one being something other than an electronic database. There is no definition in the work by Jadad and McQuay as to what is considered to be a comprehensive search. Sacks et al <66>, using a checklist rather than the Oxman / Guyatt index, also noted deficiencies in the literature search. In their sample of 86 published meta-analyses, they report 35% used more than electronic

databases to identify primary studies. When item 1 and 2 are combined for my sample, the results are similar (37% for the entire sample).

In my sample, the areas of weakness tended to lie with the comprehensiveness of the search (question 2) and the avoidance of bias in the selection of studies (question 4). In the second and fourth questions, 53.4% and 88% of the sample scored 'can't tell' respectively. These two items were areas where lower than substantial agreement was achieved in the calibration set as calculated by the intraclass correlation.

Although these discrepancies may indicate judgement on the part of the evaluator, for the purpose of this research the assessment of quality of reporting was done to make a comparison of the different groups. As one reviewer did all the assessments of quality, interpretation is less likely to impact on the findings.

Quality of reporting of the MA that included grey literature was slightly better than that of those that excluded grey literature (implicitly or explicitly). One might expect that MA that include grey literature would have required a broader and more comprehensive search. Given that item 2 on the Oxman / Guyatt index deals with the inclusiveness of the literature search the possibility of bias exists. I do not feel that quality assessment using the Oxman / Guyatt index has favoured the grey inclusive MA. Among the MA that excluded grey literature (groups combined) 7.5% scored 6 or 7 on the summary item compared to 6.1% of the MA that included grey literature. When item 2 is considered on its own, 24.2% of the grey inclusive MA were scored as 'yes' compared to 21.7% and 11.8% respectively for the implicit and explicit exclusion groups. These differences were not statistically significant (Table 8). In all three groups the majority scored 'can't tell' for item 2. Aside from the comprehensiveness of the search (item 2) the group of meta-

analyses which explicitly excluded grey literature tended to score higher on the searching items (items 1 and 3).

The overall quality of reporting of meta-analysis is an area that is currently receiving a lot of attention. Work on the QUOROM (*QU*ality *O*f *R*eporting *O*f *M*eta-analyses) statement, which aims to improve the quality of reporting of meta-analyses, may help to inform and direct reporting practices.

V-2-2 Randomized Controlled Trials

The overall quality of reporting of the RCTs included in this sample of meta-analyses is low. The mean score for the sample was 2.24 which corresponds to 44.8% of the maximum possible value for the scale. These findings are consistent with other published research. Jadad et al <51> report a mean score of 2.3 in the pain literature. Moher reports 54.8% of maximum in one study of English language RCTs <52> and 51% of maximum for English language, and 46% of maximum for trials published in languages other than English <64>. Egger reports 48.5% of maximum for trials published in German, compared to 46% for those published in English. Using a different instrument, Liberati <16> found trials in breast cancer had an overall score of 50%. The consistent finding is that quality of reporting of RCTs tends to be low. It is striking that the scores tend to be clustered to the lower end of this scale, with 59.2% of the sample scoring 2 or less on the Jadad scale. Reports by Kahn <26,67> suggest that the fertility literature may be of lower quality, as 71% and 78% of the studies they scored were considered low with the Jadad scale.

When the grey literature is compared to the published literature it is of significantly lower quality. Given the make up of the sample, it is not possible to determine if a given source of grey literature is responsible for the difference. The small numbers in some of the sources of grey literature (e.g. theses) prevents the analysis of quality of reporting by source. Not all items were retrieved for quality assessment. Of the items requested from the authors, only 42% of grey items were retrieved, this demonstrates the difficulties encountered by meta-analysts considering the inclusion of grey literature. The identification of grey literature may be the first barrier, but its retrieval can be equally difficult. It is also important to note that, in terms of grey literature retrieved for quality assessment, 80% (61/76) are abstracts. Due to the high proportion of abstracts the noted difference in quality may reflect a difference between abstracts and journal published articles, rather than grey literature and published articles. No other literature is available comparing the quality of reporting of such a diverse group of grey literature to that of published literature. Peer reviewed journal articles have been shown to be of higher quality than non peer reviewed symposium articles in the area of environmental tobacco smoke <68>. The differences in quality of reporting of published versus grey RCTs may be explained by a number of factors:

(1) The sample of grey literature is comprised predominantly of abstracts. When the abstracts were considered separately, the scores ranged from a high of 3 to a low of 0. The abstracts never scored bonus points for describing the methods of randomization, or double blinding. Jadad's scale was not designed for abstracts, and it may therefore not be justified to use this scale, however, there were no other scales available that had been validated for abstracts. To try to account for this deficiency, I looked for the published

literature that had scored 5 on Jadad's scale. I took 7 (54%) of these papers and re-scored only the abstracts. None of the seven maintained their score of 5, in fact, all scored 3 or less. This indicates that the low scores of the grey abstracts may be artificially low, possibly due to space limits. Authors describe dropouts and reasons for dropouts less often in abstracts than in the main text of a paper <69>. Although deficiency exists in abstracts of papers and abstracts for presentation, they may be two separate issues. Authors of the former have included the details in the body of the paper and may not recognize the need to have it in the abstract. Short reports and research letters might be more similar to abstracts submitted to meetings. Deeks and Altman <69> reviewed short reports and research letters in the British Medical Journal and The Lancet to see if they conformed to CONSORT <24>. The CONSORT (the consolidated standards of reporting trials) statement describes items that are recommended for inclusion in the reporting of RCTs. They found that the quality of these short reports and letters was very low. Although randomization was mentioned in 76% of the reports, the generation was only described 17% of the time. Similarly with double blinding, it was mentioned in 52% of reports and described in 9%. The authors suggest that the space limits are likely to be at least in part responsible for this lack of detail. Given that abstracts for meetings are usually restricted to 200 to 300 words, half the limit set for short reports and letters, space is likely an issue for the abstracts in my sample. As many readers use only the abstract for the initial screening of articles, partly due to its availability on electronic databases (i.e. MEDLINE), it is important that it be accurate and complete. This has been demonstrated to be an issue, in a study by Scherer et al <39> 9 of 40 non-RCTs were described explicitly as RCTs in an abstract. The misclassification of study design is

alarming; it casts doubt on the quality of reporting of abstracts and raises questions about their accuracy. One should be able to assess the quality of reporting of abstracts with the Jadad scale. The scale has only three items which are key to the validity of any RCT. Attention should be paid to these three items in the abstract of an RCT.

(2) In the preparation of the grey literature, the authors may have had other motives for preparing the document, and reporting of the methodology may not have been a priority.

V-3 Impact of Grey Literature on the Estimate of Intervention Effect

V-3-1 Replication of MA

Using the data published in the meta-analysis we were able to replicate the published estimates of intervention effectiveness. This step was essential to ensure the data extraction was accurate. The results were expected and are consistent with other reports <49>.

V-3-2 Repetition of MA

The occurrence of publication bias has been well documented, however, the extent to which the exclusion of grey literature impacts on the estimate of effectiveness has not been examined. The results of this research, examining 467 trials, indicate that the exclusion of grey literature can result in a higher estimate of treatment effect by an average of 12%. In the area of quinine for nocturnal leg cramp, Hing et al published a meta-analysis using only published data <71>, they then repeated their meta-analysis with previously unidentified United States Food and Drug Administration documents. They report that the published trials consistently reported larger point estimates for the

efficacy of quinine. The relative risk reduction (RRR) when all trials were pooled (4 published and 3 unpublished) was 21% (95% CI 12%, 30%) compared to a RRR of 43% (95% CI 21%, 65%) when only the published trials were used. Although the inclusion of the unpublished data does not alter the statistical significance of the results, it does demonstrate a decrease in efficacy by about 50%.

When I repeated my analysis, after removing abstracts from the sample, the estimate increased on average by 38%. Abstracts compared to published literature have no impact on the point estimate of treatment effectiveness. Given that abstracts appear to be more like published literature in terms of their impact on the estimates of intervention estimates, and the impact of grey literature excluding abstracts is higher than when they are included, one might suggest that the effect of grey literature is being muted by the inclusion of abstracts. There are strong arguments for not including abstracts with the grey literature; 1) they are frequently catalogued on electronic databases, 2) they are usually published, at least in conference proceedings, 3) they may be peer-reviewed, and 4) they tend to be available. From the existing evidence <72>, it appears as though publication bias exists in abstracts; that is abstracts with positive findings tend to be accepted for presentation at conferences more frequently than those with negative or null findings. In a study <72> of 500 abstracts submitted to the 1991 Society for Academic Emergency Medicine meeting, the best predictors of abstract acceptance were an 'originality score' (OR 2.07 95% CI 1.13, 3.89) and positive findings (OR 1.99 95% CI 1.07, 3.84). Although this may not be indicative of other meetings, it does suggest that abstracts, like papers, are subject to publication bias.

V-3-2-1 Quality: When both quality and publication status are considered together, neither one impacts treatment effectiveness estimates. It is unclear how accurate the estimates of quality of reporting are for the grey literature given the high proportion of abstracts and missing data. With the sample as it is the quality of reporting of grey literature is lower than that of published literature. There could be possible interaction between publication status and quality of reporting.

When abstracts are removed from the sample publication status, in the presence of quality becomes significant. As seen when publication status is considered individually, the exclusion of the grey literature results in a larger estimate of treatment effectiveness when considered in the presence of trial quality (ROR 1.34, 95% CI 1.01, 1.78). Given the previously noted observation that abstracts do not differ from published literature, in terms of impact on treatment effect, this is not unexpected.

V-4 Limitations

The database from which the sample of MA was drawn is based on MA published to 1995. There is nothing in the literature to suggest that the quality of reporting of MA or RCTs have improved since that time. Concerning grey literature, much work has been done in the area of publication bias, but nothing has been done to my knowledge to provide empirical evidence about the merit of including grey literature in MA and it has therefore not become standard practice.

Given my restriction to MA with binary outcomes and no more than 100 RCTs, the results of this study may not be generalizable to large MA (>100RCTs) or those with continuous outcomes.

Many of the limitations of this study are actually limitations or deficiencies with published meta-analyses. Poor referencing of included trials was a problem. In eleven cases MA had to be excluded from the sample because the referencing was so poor that it was impossible to determine which studies were used in the generation of the summary estimates. This has direct implications for the calculated prevalence of the grey literature in MA. If none of these poorly referenced MA include grey literature, then the prevalence would only be 30% (38/126). If they all included grey literature the prevalence would be 39% (49/126). The true prevalence likely lies somewhere between these estimates. In one included MA <1433>, many of the abstracts could not be retrieved due to the poor manner in which they were referenced. As noted earlier in this report, the inclusion of grey literature was biased in at least one of the MA <951> included. The authors of this MA clearly state their inclusion criteria, which although biased in terms of grey literature, allows the reader to judge their decisions. It is impossible to tell if similar conscious or unconscious omissions of grey literature were made from other MA in this sample.

I was unable to retrieve 24 (24%) of the grey items for quality assessment. Without these items, a difference between the quality of published and grey trials was detected. It is unclear if this difference would persist if all the RCTs had been retrieved and assessed. This difference may in fact represent a difference between the quality of reporting of abstracts and published literature, rather than grey and published literature. These unretrieved items would also have provided additional power and may have shed some insight into the interaction between quality and publication status.

When comparing the quality of reporting of the RCTs included in the MA, it is important to remember that what is being compared is not an independent selection of trials, but rather clusters of trials included in each MA.

This work does not address the identification and retrieval of grey literature for those wishing to include it in future meta-analyses. It may be that the time, effort, and cost involved in identifying, locating, and retrieving the grey literature makes its inclusion prohibitive.

The majority of the grey literature items included in this research were abstracts (61%). Although this may be a limitation, it clearly demonstrates what is accessible for inclusion in meta-analyses. The accessibility and other previously mentioned characteristics of abstracts may suggest that they should not be categorized as grey literature. In this study, I felt it was important to classify them as such, as there is much controversy over the merit of their inclusion in meta-analyses. Due to this abundance of abstracts, and lack of other sources, no analyses were possible to investigate differences between types of grey literature. It could not be determined if a given type of grey literature was more responsible for the shift in point estimate than another, except in the case of abstracts.

Future areas for inquiry concern the ability to identify and retrieve grey literature sources, and the time and effort involved in such a venture. Efforts made by various groups, through trial registries, negative trial journals, Internet based grey literature resources, and the 'open books' policies recently adopted by several pharmaceutical companies may make the identification easier. If authors are expected to include grey literature, some guidance must be available on how to identify and retrieve it.

It is also unclear from this research if the impact of grey literature is attributable to one source. Aside from abstracts, no source had a large enough sample to allow specific investigation. This is an area where work is needed, as some types of grey literature may be more difficult to access. It would be beneficial if the sources that impact the intervention effectiveness, if a sub-set, could be identified.

There is also a need to further investigate the interaction between trial quality and publication status with respect to the impact on intervention effectiveness estimates. In this sample of MA, quality in the presence of publication status does not alter intervention estimates. However, work by Moher et al <52> and Schulz et al <57> have shown low quality trials tend to produce higher estimates of the effectiveness of interventions.

V-5 Implications

This work has implications for both meta-analysts and those who use them to direct clinical or policy decisions. The meta-analyst must make every effort to ensure a comprehensive literature search to avoid the effects of publication bias. Given the potential for overestimating the effectiveness of treatments, grey literature should not be systematically excluded. All potentially relevant studies must be sought and considered for inclusion. Studies must only be excluded based on the pre-defined inclusion criteria. For those concerned about the quality of the grey literature, a sensitivity analysis could be used. This could either be done on the basis of the quality of reports, or on the basis of publication status. This approach will help to ensure a comprehensive review by

considering all the available evidence, while allowing the reader to judge the merit of the intervention based on their perception of the quality of the evidence.

VI Conclusion

It has been demonstrated that grey literature makes a significant contribution to meta-analysis. Grey literature should be sought and when appropriate included in systematic reviews and meta-analyses. It is important for meta-analysts, editors and readers to keep in mind the possibility of overestimating effectiveness in the absence of grey literature. Although this work was done on meta-analyses of randomized controlled trials, there are likely similar, if not larger implications for meta-analyses of observational studies, as those studies tend to be more prone to publication bias <30>.

If it is not feasible to identify or include sources of grey literature, efforts should be made to assess the possibility of their existence. Tests for publication bias can help clarify the likelihood that trials were missed.

References

1. Mulrow CD. Rationale for systematic reviews. BMJ 1994;309:507-509.

2. Chalmers TC, Lau J. Meta-analytic stimulus for changes in clinical trials. Statistical Methods in Medical Research 1993;2:161-172.

3. Cooper HM, Hedges LV. The Handbook of research synthesis. New York: Russell Sage Foundation,1994.

4. Pearson K. Report on certain enteric fever inoculation statistics. BMJ. 1904;3:1243-1246.

5. Altman LK. New method of analyzing health data stirs debate. The New York Times. 1990; August 21:C1 -C2.

6. Jenicek M. Meta-analysis in Medicine: Putting experiences together, chapter 9, 269-290. In Epidemiology: The Logic of Modern Medicine. EPIMED, 1995.

7. Greenland S. Quantitative Methods in the Review of Epidemiologic Literature. Epidemiologic Reviews. 1987;9:1-30.

8. Smith ML, Glass GV. Meta-analysis of psychotherapy outcome studies. Am Psychol. 1977;37:752-760.

9. Light RJ, Smith PV. Accumulating evidence: Procedures for resolving contradictions among research studies. Harvard Educational Review 1971;41:429-471.

10. Bailar JC III. The Promise and Problems of Meta-Analysis. N Engl J Med (editorial) 1997;337:559-560

11. Shapiro S. Meta-analysis shmeta-analysis. Am J Epidemiol 1994;140:771-778.

12. Thompson, S. Pocock S. Can meta-analysis be trusted? Lancet 1991;338:1127-1130.

13. Feinstein AR. Meta-analysis: Statistical Alchemy for the 21st century. J Clin Epidemiol 1995;48:71-79.

14. Lau J, Antman EM, Jimenz-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cummulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med 1992;327:248-254.

15. Domanski MJ, Friedman LM. Relative role of meta-analysis and randomized controlled trials in the assessment of medical therapies. editorial. The American Journal of Cardiology. 1994;74:395-396.

16. Liberati A. Meta-analysis: Statistical Alchemy for the 21st century": Discussion. A plea for a more balanced view of meta-analysis and systematic reviews of the effect of health care interventions. J Clin Epidemiol 1995;48:81-86.

17. Strauss S. Lies, damned lies and statistics. The Globe and Mail. 1991; November 2.

18. Pfeifer N. Pooled data of clinical trials not always accurate. Medical Post. 1994; 30(29):12.

19. Sacks HS, Chalmers TC, Blum AL, Berrier J, Pagano D. Endoscopic hemostasis: an effective therapy for bleeding peptic ulcers. JAMA. 1990;262(4):494-499.

20. Crowley P. Chalmers I. Keirse MJ. The effects of corticosteroid administration before preterm delivery: an overview of the evidence from controlled trials. British Journal of Obstetrics & Gynaecology. 1990;97(1):11-25.

21. Lelorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analysis and subsequent large randomized, controlled trials. N Engl J Med. 1997;337:536-542.

22. Ioannidis JPA, Cappelleri JC, Lau J. Comparing results from meta-analysis vs. large trials. JAMA (letter). 1998;280:518-519.

23. Chalmers TC, Berrier J, Sacks HS, Levin H, Reitman D, Nagalingam R. Meta-analysis of clinical trials as scientific discipline. II: Replicate variability and comparisons of studies that agree and disagree. Statistics in Medicine. 1987;6:733-744.

24. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schultz KF, Simel D, Stroup D. Improving the reporting of randomized controlled trials: The CONSORT statement. JAMA 1996;276:637-639.

25. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: Current issues and future directions. International Journal of Technology Assessment in Health Care 1996;12(2):195-208.

26. Khan KS, Daya S, Collins JA, Walter SD. Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. Fertility and Sterility. 1996;65:939-945.

27. Spitzer WO. Meta-meta-analysis: unanswered questions about aggregating data. [editorial; comment]. Journal of Clinical Epidemiology. 44(2):103-7, 1991.

28. Moher D, Klassen TP, Jadad AR. Guides for reading and interpreting systematic reviews: 3. How did the authors synthesize the data and make their conclusions? Archives of Pediatrics & Adolescent Medicine. 1998;152:915-920.

29. Liberati A, Himel HN, Chalmers TC. A Quality Assessment of Randomized Control Trials of Primary Treatment of Breast Cancer. Journal of Clinical Oncology. 1986;4:942-951.

30. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991; 337:867-72.

31. Abby M, Massey MD, Galandiuk S, Polk Jr HC, Peer Review Is an Effective Screening Process to Evaluate Medical Manuscripts. JAMA. 1994; 272(2): 105-107.

32. Last JM. A Dictionary of Epidemiology. Third Edition. Oxford, Oxford University Press. 1995.

33. Squires BP. Peer review under scrutiny. Report on the third international congress in Prague, 1997(editorial). Canadian Family Physician 1998;44:15-16.

34. Egger M. Zellweger-Zahner T. Schneider M. Junker C. Lengeler C. Antes G. Language bias in randomised controlled trials published in English and German. Lancet. 1997;350(9074):326-329.

35. Jefferson T. Redundant publication in biomedical sciences: scientific misconduct or necessity? Science and Engineering Ethics. 1988;4:135-140.

36. Tramer MR, Reynolds DJM, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. BMJ 1997;315:635-640.

37. Huston P, Moher D. Redundancy, disaggregation, and the integrity of medical research. Lancet 1996;347:1024-1026.

38. Petitti N, Duplicate Publication and Correction. N Engl J Med 1997;337(16):1175.

39. Scherer R, Dickerson K, Langenberg P, Full Publication of Results Initially Presented in Abstracts. JAMA. 1994; 272(2): 158-162.

40. Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, McIlroy W, Oxman AD, Should Unpublished Data Be Included in Meta-analyses? JAMA. 1993; 269: 2749-2753.

41. Dickersin K, Min Y-I, Meinert C.L, Factors Influencing Publication of Research Results. JAMA 1992;267(3): 374-378.

42. Dickersin K, The Existence of Publication Bias and Risk Factors for Its Occurrence. JAMA. 1990;263:1385-1389.

43. Coursol A, Wagner EE. Effect of Positive Findings on Submission and Acceptance Rates: A note on meta-analysis bias. Professional Psychology: Research and Practice 1986;17(2):136-137.

44. Halperin EC. EUREKA! It's a negative study! NCMJ 1994;55(2):68-69.

45. Cook D.J, Guyatt G.H, Laupacis A, Sackett D.L, Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1992; 102: 3055-3115.

46. Smith ML. Publication bias and meta-analysis. Evaluation in Education. 1980;4:22-24.

47. Stern JM, Simes RJ. Publication Bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ 1997;315:640-645.

48. Hetherington J, Dickersin K, Chalmers I, Meinert CL. Retrospective and prospective identification of unpublished controlled trials: Lessons from a survey of obstetricians and pediatricians. Pediatrics 1989;84(2):374-380.

49. Medical Editors' Trial Amnesty. Unreported trial registration form. CMAJ 1997;157 (11):1548.

50. Medical Editors' Trial Amnesty. Unreported trial registration form. BMJ. 1997;315:622.

51. Jadad AR, Moore A, Carroll D, Jenkinson C, Reynolds JM, Gavaghan DJ, McQuay HJ. Assessing the Quality of Reports of Randomized Clinical Trials: Is blinding necessary? Controlled Clinical Trials 1996;17:1-12.

52. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609-613.

53. Berlin JA. on behalf of the University of Pennsylvania meta-analysis blinding study group. Does blinding of reader's affect the results of meta-analyses? Lancet 1997;350:185-186.

54. Shea B, Dubé C, Moher D. Assessing the Quality of Reports of Systematic Reviews and Meta-analyses: A Systematic Review of Checklists and Scales. In Egger M, Altman DG. Systematic reviews: 2nd edition. BMJ Publishing group, 1999.

55. Norman DL, Streiner D. Health Measurement scales - A practical guide to their development and use. 2nd Ed. Oxford University Press. Oxford, 1995.

56. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991;44:11:1271-1278.

57. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical Evidence of Bias: Dimensions of Methodological Quality Associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408-412.

58. Donner A, Eliasziw M. Sample size requirements for reliability studies. Statistics in Medicine 1987;6:441-448.

59. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86:420-428.

60. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. British Journal of Cancer. 1977;35:1-39.

61. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. Statistics in Medicine. 1989;8:141-151.

62. Deeks J, Bradburn M, Localio R, Berlin J. Much ado about nothing: statistical methods for meta-analysis with rare events. 2nd Symposium on Systematic Reviews: Beyond the basics Recent Advances, New Challenges, Effective Dissemination Oxford, UK January 1999.

63. Sykes R. Being a modern pharmaceutical company.[editorial] BMJ. 1998;317:1172.

64. Moher D, Pham B, Klassen TP. Schulz KF, Berlin JA, Jadad AR, Liberati A. Does the language of publication of reports of randomized trials influence the estimates of intervention effectiveness reported in meta-analysis? Technical report from the Thomas C. Chalmers Centre for Systematic Reviews.

65. Jadad AR, McQuay HJ. Meta-analyses to evaluate analgesic interventions: A systematic qualitative review of their methodology. J Clin Epidemiol. 1996;49(2):235-243.

66. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med. 1987;316:450-455.

67 Khan KS, Daya S, Jadad AR. The importance of quality of primary studies in producing unbiased systematic reviews. Arch Intern Med 1996;156:661-666.

68. Barnes DE, Bero L. Scientific quality of original research articles on environmental tobacco smoke. Tobacco Control 1997;6:19-26.

69. Froom P, Froom J. Deficiencies in structured medical abstracts. J Clin Epidemiol 1993;46:591-594.

70. Deeks JJ, Altman DG. Inadequate reporting of controlled trials as short reports. Lancet 1998;352:1908.

71. Man-Son-Hing M, Wells G, Lau A. Quinine for nocturnal leg cramps. A meta-analysis including unpublished data. Journal of General Internal Medicine 1998;13:600-6.

72. Callaham ML, Wears RL, Weber EJ, Barton C, Young G. Positive-outcome bias limitations in the outcome of research abstracts submitted to a scientific meeting. JAMA 1998;280(3):254-257.

73. Cook DJ, Mulrow C, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med. 1997;126:376-380.

Appendices

APPENDIX A:

Glossary

Allocation Concealment

the keeping the random numbers, once they have been generated, hidden from all RCT participants (patients, investigators, and outcome assessors) until the time of randomization. <49, 54>

Bias

deviation of results or inferences from the truth, or processes leading to such deviation. Any trend in the collection, analysis, interpretation, publication, or review of data that can lead to conclusions that are systematically different from the truth. <29>

Explicit exclusion of grey literature

grey literature was considered to be explicitly excluded if anywhere in the text the authors mentioned any source of grey literature and stated that they had consciously omitted it.

Grey Literature

reports that are unpublished, have limited distribution, and/ or are not included in bibliographic retrieval systems.

Heterogeneity

is a measure of the between study variance which exists among studies included in a meta-analysis.

Implicit exclusion of grey literature

grey literature was considered to be implicitly excluded if there was no specific mention of any source of grey literature and none was included in the MA.

Meta-Analysis

a systematic review that includes a quantitative measure of the combined results of the included studies.

Point Estimate

refers to the summary statistic in a meta-analysis. It is calculated by combining the outcomes of all included studies

Precision

the amount of error or variance around the point estimate.

'Published'

used here to refer to the RCTs which have been published as full journal articles. Some of the grey items have been published, the term published was selected because there was a need for a title for this group.

Quality (Methodological)

The extent to which the design and conduct of a study are likely to be free of bias.

Ratio of odds ratios [ROR]

ratio of treatment effect odds ratios. A ratio of odds ratios greater than one implies that non-grey literature has a larger treatment effect than grey literature.

Repetition

using the Peto OR the pooled estimates were re-generated, and then regenerated again after all grey items had been removed

Replication

the pooled estimates were replicated using the trial outcome data, and the statistical methods reported in each meta-analysis

Systematic Review

a review in which evidence is systematically identified, appraised and summarized according to a pre-defined explicit methodology <68>

APPENDIX B:

MEDLINE Search Strategy

Set	Search	Results
001	meta-analysis.pt,sh.	2866
002	(meta-anal: or metaanal:) .tw.	2296
003	(quantitativ: review: or quantitativ: overview:) .tw.	79
004	(systematic: review: or systematic: overview:).tw.	287
005	(methodologic: review: or methodologic: overview:).tw.	76
006	review.pt,sh. or review:.tw. or overview:.tw.	634672
007	(intergrative research review: or research integration:).tw.	30
008	7 or quantitativ: synthes:.tw.	61
009	1 or 2 or 3 or 4 or 5 or 8	3920
010	(medline or medlars).ti,sh,ab. or embase.tw.	2371
011	(scisearch or psycinfo or psychinfo).tw.	24
012	(hand search: or manual search:).tw.	169
013	(electronic database: or bibliographic database:).tw.	88
014	(pooling or pooled analys: or mantel heanszel).tw.	1986
015	(peto or der simonian or dersimonian or fixed effect:).tw.	259
016	(psychlit or psyclit).tw.	15
017	10 or 11 or 12 or 13 or 14 or 15 or 16	4663
018	6 and 17	1447
019	18 or 9	5109
020	random:.tw,sh,pt. or placebo:.tw,sh.	175571
021	(clinical trial or controlled clinical trial).pt.	151285
022	randomized controlled trial.pt.	57462
023	double-blind:.tw,sh.	52249
024	20 or 21 or 22 or 23	248821
025	19 and 24	1467

APPENDIX C:

Generation of Random Numbers: SAS PROC PLAN.

```
data a;
        do unit = 1 to &n1;
        output;
        end;
        run;

data a1;
        set a;
        if unit<= 135 then group =1;
        else group =2;
        run;

proc plan seed = 37782;
        factors unit = &n1/noprint;
        output data = a1 out = b;
        run;

data c;
if group = 1 then output c;
run;

proc print data = c;
title "&title";
%mend random;

*%random (423, 227, the first group);
*% random (369, 174, the second group);
% random (455, 135, random numbers);
```

APPENDIX D:

Meta-Analyses that do not include grey literature.

428. Cappelleri JC, Fiore LD, Brophy MT, Deykin D, Lau J. Efficacy and safety of combined anticoagulant and antiplatelet therapy versus anticoagulant monotherapy after mechanical heart- valve replacement: a metaanalysis. American Heart Journal. 1995;130(3 Pt 1):547-552.

429. Marshall JK, Irvine EJ. Rectal aminosalicylate therapy for distal ulcerative colitis: a meta-analysis. Alimentary Pharmacology & Therapeutics. 1995;9(3):293-300.

503. Dexter F, Tinker JH. Comparisons between desflurane and isoflurane or propofol on time to following commands and time to discharge. A metaanalysis. Anesthesiology 1995;83(1):77-82.

529. Meyer TJ, Mark MM. Effects of psychosocial interventions with adult cancer patients: a meta-analysis of randomized experiments. Health Psychology. 1995;14(2):101-108.

564. Conn HO, Poynard T. Corticosteroids and peptic ulcer: meta-analysis of adverse events during steroid therapy. Journal of Internal Medicine 1994;236(6):619-632.

582. Pace F, Maconi G, Molteni P, Minguzzi M, Bianchi Porro G. Meta-analysis of the effect of placebo on the outcome of medically treated reflux esophagitis. Scandinavian Journal of Gastroenterology 1995;30(2):101-105.

603. Anonymous. Laser therapy for retinopathy of prematurity. Laser ROP Study Group [letter]. Archives of Ophthalmology. 1994;112(2):154-156.

609. Messori A, Rampazzo R, Scroccaro G, Olivato R, Bassi C, Falconi M, Pederzoli P, Martini N. Effectiveness of gabexate mesilate in acute pancreatitis. A metaanalysis. Digestive Diseases & Sciences. 1995;40(4):734-738.

610. Hazell P, O'Connell D, Heathcote D, Robertson J, Henry D. Efficacy of tricyclic drugs in treating child and adolescent depression: a meta-analysis. BMJ. 1995;310(6984):897-901.

622. Wong JB, Koff RS, Tine F, Pauker SG. Cost-effectiveness of interferon-alpha 2b treatment for hepatitis B e antigen-positive chronic hepatitis B. Annals of Internal Medicine. 1995;122(9):664-675.

717. Aylward GW, Dunlop IS, Little BC. Meta-analysis of systemic anti-fibrinolytics in traumatic hyphaema. *Eye* 1994;8(Pt 4):440-442.

735. Colditz GA, Brewer TF, Berkey CS, Wilson ME, Burdick E, Fineberg HV, Mosteller F. Efficacy of BCG vaccine in the prevention of tuberculosis. Meta- analysis of the published literature. JAMA. 1994;271(9):698-702.

744. Browman GP. Evidence-based recommendations against neoadjuvant chemotherapy for routine management of patients with squamous cell head and neck cancer. Cancer Investigation. 1994;12(6):662-670.

800. Borris LC, Lassen MR, Jensen HP, Andersen BS, Poulsen KA. Perioperative thrombosis prophylaxis with low molecular weight heparins in elective hip surgery. Clinical and economic considerations. International Journal of Clinical Pharmacology &.Therapeutics. 1994;32(5):262-268.

821. Marino P, Pampallona S, Preatoni A, Cantoni A, Invernizzi F. Chemotherapy vs. supportive care in advanced non-small-cell lung cancer. Results of a meta-analysis of the literature. Chest 1994;106(3):861-865.

930. Berman S, Roark R, Luckey D. Theoretical cost effectiveness of management options for children with persisting middle ear effusions. Pediatrics 1994;93(3):353-363.

990. Simons-Morton DG, Cutler JA, Allender PS. Hypertension treatment trials and stroke occurrence revisited. A quantitative overview. Annals of Epidemiology. 1993;3(5):555-562.

1051. Yurkowski PJ, Plaisance KI. Prevention of auditory sequelae in pediatric bacterial meningitis: a meta-analysis. Pharmacotherapy. 1993;13(5):494-499.

1078. Fraser EJ, Grimes DA, Schulz KF. Immunization as therapy for recurrent spontaneous abortion: a review and meta-analysis. Obstetrics & Gynecology. 1993;82(5):854-859.

1118. Bucher HC, Schmidt JG. Does routine ultrasound scanning improve outcome in pregnancy? Meta-analysis of various outcome measures. BMJ. 1993;307(6895):13-17.

1183. Torri V, Korn EL, Simon R. Dose intensity analysis in advanced ovarian cancer patients. British Journal of Cancer 1993;67(1):190-197.

1185. Gadomski AM. Potential interventions for preventing pneumonia among young children: lack of effect of antibiotic treatment for upper respiratory infections. Pediatric Infectious Disease Journal. 1993;12(2):115-120.

1190. Edmonson JH, Su J, Krook JE. Treatment of ovarian cancer in elderly women. Mayo Clinic-North Central Cancer Treatment Group studies Cancer 1993;71(2 Suppl):615-7.

1284. Knight DB. Patent ductus arteriosus: how important to which babies? Early Human Development 1992;29(1-3):287-292.

1330. Anderson R, Meeker WC, Wirick BE, Mootz RD, Kirk DH, Adams A. A meta-analysis of clinical trials of spinal manipulation. Journal of Manipulative & Physiological Therapeutics. 1992;15(3):181-194.

1332. Salomon P, Kornbluth A, Aisenberg J, Janowitz HD. How effective are current drugs for Crohn's disease? A meta- analysis. Journal of Clinical Gastroenterology 1992;14(3):211-215.

1421. Levin FR, Lehman AF. Meta-analysis of desipramine as an adjunct in the treatment of cocaine addiction. Journal of Clinical Psychopharmacology 1991;11(6):374-378.

1522. Shinton RA, Beevers DG. A meta-analysis of mortality and coronary prevention in hypertensive patients treated with beta-receptor blockers. Journal of Human Hypertension 1990;4 Suppl 2:31-34.

1537. Sacks HS, Berrier J, Nagalingham R, Chalmers TC. Dipyridamole in the treatment of angina pectoris: a meta-analysis. Thrombosis Research.-.Supplement. 1990;12:35-42.

1552. Midgette AS, O'Connor GT, Baron JA, Bell J. Effect of intravenous streptokinase on early mortality in patients with suspected acute myocardial infarction. A meta-analysis by anatomic location of infarction Annals of Internal Medicine 1990;113(12):961-968.

1568. Muldoon MF, Manuck SB, Matthews KA. Lowering cholesterol concentrations and mortality: a quantitative review of primary prevention trials. BMJ. 1990;301(6747):309-314.

1578. Radack K, Wyderski RJ. Conservative management of intermittent claudication. Annals of Internal Medicine 1990;113(2):135-146.

1581. Patten SB. Propranolol and depression: evidence from the antihypertensive trials. Canadian Journal of Psychiatry -.Revue Canadienne de Psychiatrie. 1990;35(3):257-259.

1632. Schmader KE, Studenski S. Are current therapies useful for the prevention of postherpetic neuralgia? A critical analysis of the literature. Journal of General Internal Medicine 1989;4(2):83-89.

1680. Guinan P, Richardson C, Hanna M, Rubenstein M. BCG in the management of superficial bladder cancer. Progress in Clinical & Biological Research. 1989;303:447-453.

1699. Clark P, Tugwell P, Bennett K, Bombardier C. Meta-analysis of injectable gold in rheumatoid arthritis. Journal of Rheumatology. 1989;16(4):442-447.

1724. Infante-Rivard C, Esnaola S, Villeneuve JP. Role of endoscopic variceal sclerotherapy in the long-term management of variceal bleeding: a meta-analysis. Gastroenterology 1989;96(4):1087-1092.

1751. Clagett GP, Reisch JS. Prevention of venous thromboembolism in general surgical patients. Results of meta-analysis. [Review] Annals of Surgery 1988;208(2):227-240.

1757. Sze PC, Reitman D, Pincus MM, Sacks HS, Chalmers TC. Antiplatelet agents in the secondary prevention of stroke: meta- analysis of the randomized control trials. Stroke 1988;19(4):436-442.

1803. Gent M, Roberts RS. A meta-analysis of the studies of dihydroergotamine plus heparin in the prophylaxis of deep vein thrombosis. Chest 1986;89(5 Suppl):396S-400S.

APPENDIX E:

Meta-Analyses which include grey literature.

406. Michels KB, Yusuf S. Does PTCA in acute myocardial infarction affect mortality and reinfarction rates? A Quantitative overview(meta-analysis) of the randomized clinical trials. Circulation 1995;91(2):476-485.

450. Pouleur H, Buyse M. Effects of dipyridamole in combinationwith anticoagulant therapy on survival and thromboembolic events in patients with prosthetic heart valves. A meta-analysis of the randomized trials. J thoracic & ardiovascular Surgery 1995;110(2):463-472.

451. Cronin L, Cook DJ, Carlet J, Heyland DK, King D, Lansang MA, Fisher CJ, Jr. Corticosteroid treatment for sepsis: a critical appraisal and meta-analysis of the literature. Critical Care Medicine 1995;23(8):1430-1439.

473. Galloe AM,Graudal N, Christensen HR, Kampmann JP. Aminoglycosides: single or multiple daily dosing? A meta-analysis on efficacy and safety. European J Clin Pharmacol 1995;48(1):39-43.

508. Lefering R, Neugebauer EA. Steroid controversy in sepsis and septic shock: a meta-analysis. Critical Care Medicine 1995;23(7):1294-1303.

627. Whitehead A, Jones NM. A meta-analysis of clinical trials involving different classifications of response into ordered categories. Statistics in Medicine 1994;13(23-24):2503-2515.

634. Fiore MC, Smith SS, Jorenby DE, Baker TB. The effectiveness of the nicotine patch for smoking cessation. A meta-analysis. JAMA 1994;271(24):1940-1947.

743. Zang WY, Li Wan Po A. The effectiveness of topically applied capsaicin. A meta-analysis. European J Clin Pharmacology 1994;46(6):517-522.

768. Poynard T, Valterio C. Meta-analysis of hydroxyethylrutosides in the treatment of chronic venous insufficiency. Vasa 1994;23(3):244-250.

790. Bressa GM. S-adenosyl-l-methionine (SAMe) as antidepressant: meta-analysis of clinical studies. Acta Neurologica Scandinavica Supplementum 1994;154:7-14.

818. Leizorovicz A, Simonneau G, Decousus H, Boissel JP. Comparison of efficacy and safety of low molecular weight heparin in initial treatment of deep venous thrombosis: a meta-analysis. BMJ 1994;309(6950):299-304.

864. Heyland Dk, Cook DJ, Jaeschke R, Griffith L, Lee HN, Guyatt GH. Selective decontamination of the digestive tract. An overview. Chest 1994;105(4):1221-1229.

935. Silagy C, Mant D, Fowler G, Lodge M. The effectiveness of nicotine replacement therapies in smoking cessation. Online J of Current Clinical Trials 1994;Doc No 113.

951. Brown KH, Peerson JM, Fontaine O. Use of nonhuman milks in the dietary management of young children with acute diarrhea: a meta-analysis of clinical trials. Pediatrics 1994;93(1):17-27.

970. Sutherland LR, May GR, Shaffer EA. Sulfasalazine revisited: a meta-analysis of 5-aminosalicylic acid in the tretemnt of ulcerative colitis. Annals of Internal Medicine 1993;118(7):540-549.

1039. Giles W, Bistis A. Clinical use of Doppler ultrasound in pregnancy: information from six randomized trials. Fetal Diagnosis & Therapy 1993;8(4):247-255.

1049. Zarembski DG, Nolan PE, Jr., Slack MK, Caruso AC. Empiric long-term amiodarone prophlaxis following myocardial infarction. A meta-analysis. Archives of Internal Medicine 1993;153(23):2661-2667.

1210. Pagliaro L, D'Amico G, Sorensen TI, Lebrec D, Burroughs AK, Morabito A, Tine F, Politi F, Traina M. Prevention of first bleed in cirrhosis. A meta-analysis of randomized trials of nonsurgical treatment. Annals of Internal Medicine 1992;117(1):59-70.

1216. Holme I. Cholesterol reduction in single and multifactor randomized trials: relationship to CHD incidence and total mortality as found by meta-analysis of twenty-two trials. Blood Pressure Supplement 1992;4:29-34.

1221. Leizorovics A, Haugh MC, Chapuis FR, Samama MM, Biossel JP. Low molecular weight heparin in prevention of perioperative thrombosis. BMJ 1992;305(6859):913-920.

1251. Deeter RG, Kalman DL, Rogan MP, Chow SC. Thearapy for pharyngitis and tonsillitis caused by group A beta-hemolytic streptococci: a meta-analysis comparing the efficacy and safety of cefadroxil monohydrate versus oral penecillin V. Clinical Therapeutics 1992;14(5):740-754.

1325. Blondel B, Breart G. Home visits for pregnancy complications and management of antenatal care: an overview of three randomized controlled trials. British Journal of Obstetrics &Gynaecology 1992;99(4):283-286.

1344. Fouque D, Laville M, Boissel JP, Chifflet R, Labeeuw M, Zech PY. Controlled low protein diets in chronic renal insufficiency: meta-analysis. BMJ 1992;304(6821):216-220.

1351. Van Ruiswyk J, Byrd JC. Efficacy of prophylactic sclerotherapy for prevention of first variceal hemorrhage. Gastroenterology 1992;102(2):587-597.

1353. Cook DJ, Guyatt GH, Salena BJ, Laine LA. Endoscopic therapy for acute nonvariceal upper gastrointestinal hemorrhage: a meta-analysis. Gastroenterology 1992;102(1):139-148.

1418. Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. BMJ 1991;303(6816):1499-1503.

1429. Hadziyannis SJ. Use of alpha-interpheron in the treatment of chronic delta hepatitis. Journal of Hepatology 1991;13supl 1:S21-26.

1433. Owen J, Winkler CL, Harris BA, Jr., Hauth JC, Smith MC. A randomized, double-blind trial of prostaglandin E2 gel for cervical ripening and meta-analysis. American Journal of Obstetrics & Gynecology 1991;165(4pt1):991-996.

1451. Beasley CM, Jr., Dorseif BE, Bosomworth JC, Sayer ME, Rampey AH, Jr., Heiligenstein JH, Thompson VL, Murphy DJ, Masica DN. Fluoxetine and suicide: a meta-analysis of controlled trials of treatment for depression. BMJ 1991;303(6804):685-692.

1626. Dobrilla G, Comberlato M, Steele A, Vallaperta P. Drug treatment of functional dyspepsia. A meta-analysis of randomized controlled clinical trials. Journal of Clinical Gastroenterology 1989;11(2):169-177.

1637. Held PH, Yusuf S, Furberg CD. Calcium channel blockers in acute myocardial infarction and unstable angina: an overview. BMJ 1989;299(6709):1187-1192.

1670. Boissel JP, Peyrieux JC, Destors JM. Is it possible to reduce the risk of cardiovascular events in subjects suffering from intermittent claudication of the lower limbs? Thrombosis & Haemostasis 1989;62(2):681-685.

1745. Daya S. Efficacy of progesterone support in the luteal phase following in-vitro fertilization and embryo transfer. A meta-analysis of clinical trials. Human Reproduction 1988;3(6):731-734.

APPENDIX F:

Sample Letter Requesting Information from Authors.

Laura McAuley
address, phone & fax numbers
e-mail address

Date

Address

Dear _____,

I am a student in the Masters program in Epidemiology at the University of Ottawa. I have decided to investigate the impact grey literature has on meta-analysis for my thesis. By grey literature I mean reports that are unpublished, have limited distribution, or are not included in bibliographic retrieval systems. I am interested both in terms of impact on quality and on the point estimate.

To consider this question, I have identified a sample of meta-analyses that include grey literature; your meta-analysis _____ is among my sample. I am writing to request copies of the following item(s) that you included in this meta-analysis:

request 1

request 2

I have enclosed a postage paid, self-addressed envelope for your use in accommodating my request. I appreciate your taking the time to consider this request. If you have any questions or concerns please do not hesitate to contact me via one of the means noted above.

Sincerely,

Laura McAuley

Appendix G-1-1

Oxman / Guyatt index:

Oxman and Guyatt's index of the scientific quality of research overviews

The purpose of this index is to evaluate the scientific quality (i.e. adherence to scientific principles) of research overviews (review articles) published in the medical literature. It is not intended to measure literary quality, importance, relevance, originality, or other attributes of overviews.

The index is for assessing overviews of primary ("original") research on pragmatic questions regarding causation, diagnosis, prognosis, therapy or prevention. A research overview is a survey of research. The same principles that apply to epidemiologic survey apply to overviews: a question must be clearly specified, a target population identified and accessed, appropriate information obtained from that population in an unbiased fashion, and conclusions derived, sometimes with the help of formal statistical analysis, as is done in "meta-analyses". The fundamental difference between overviews and epidemiologic surveys is the unit of analysis, not the scientific issues that the questions in this index address.

Since most published overviews do not include a methods section it is difficult to answer some of the questions in the index. Base your answers, as much as possible, on the information provided in the overview. If the methods that were used are reported incompletely relative to a specific item, score that item as "partially". Similarly, if there is no information provided regarding what was done relative to a particular question, score it as "can't tell", unless there is information in the overview to suggest either that the criterion was or was not met.

1. Were the search methods used to find evidence (original research) on the primary question(s) stated?

❑ yes ❑ partially ❑ no

2. Was the search for evidence reasonably comprehensive?

❑ yes ❑ can't tell ❑ no

3. Were the criteria (inclusion/exclusion) used for deciding which studies to include in the overview reported?

❑ yes ❑ partially ❑ no

4. Was bias in the selection of studies avoided?

❑ yes ❑ can't tell ❑ no

5. Were the criteria (methodological quality) used for assessing the validity of the included studies reported?

❑ yes ❑ partially ❑ no

6. Was the validity of all studies referred to in the text assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)?

❑ yes ❑ can't tell ❑ no

7. Were the methods used to combine the findings of the relevant studies (to reach a conclusion) reported?

❑ yes ❑ partially ❑ no

8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?

❑ yes ❑ can't tell ❑ no

For question 8, if no attempt was made to combine findings, and no statement is made regarding the inappropriateness of combining findings, check "no". If a summary (general) estimate is given anywhere in the abstract, the discussion or the summary section of the paper, and it is not reported how the estimate was derived, mark "no" even if there is a statement regarding the limitations of combining the findings of the studies reviewed. If in doubt mark "can't tell".

9. Were the conclusions made by the author(s) supported by the data and/or analysis reported in the overview?

❑ yes ❑ partially ❑ no

For an overview to be scored as "yes" on question 9, data (not just citations) must be reported that supports the main conclusions regarding the primary question(s) that the overview addresses.)

10. How would you rate the scientific quality of the overview?

extensive flaws		major flaws		minor flaws		minimal flaws
❑	❑	❑	❑	❑	❑	❑
1	2	3	4	5	6	7

The score for question 10, the overall scientific quality, should be based on your answers to the first nine questions. The following guidelines can be used to assist with deriving a summary score. If the "can't tell" option is used one or more times on the preceding questions, a review is likely to have minor flaws at best and it is difficult to rule out major flaws (i.e. a score of 4 or lower). It the "no" option is used on question 2, 4, 6 or 8, the review is likely to have major flaws (i.e. a score of 3 or less, depending on the number and degree of the flaws).

APPENDIX G-1-2

Oxman / Guyatt index with points of clarification:

Oxman and Guyatt's index of the scientific quality of research overviews
**Points of clarification have been added to facilitate scoring and decrease interpretation.*

The Introductory paragraph remains unchanged.

1. Were the search methods used to find evidence (original research) on the primary question(s) stated?

❑ yes ❑ partially ❑ no

Yes is given to meta-analysis reporting categories of sources, including years [e.g. databases-medline] used, and whether these categories were named (e.g. medline). Partial points are given for the category of sources [e.g. electronic , hand, register].

2. Was the search for evidence reasonably comprehensive?

❑ yes ❑ can't tell ❑ no

Yes is given if at least three categories, one of which must be electronic with key words stated, and any two others [e.g. hand, register] are reported. Key words and/or MESH terms must be stated.

3. Were the criteria (inclusion/exclusion) used for deciding which studies to include in the overview reported?

❑ yes ❑ partially ❑ no

This item was thought to be reasonably explicit.

4. Was bias in the selection of studies avoided?

❑ yes ❑ can't tell ❑ no

Yes is given if at least two reviewers independently assess for inclusion. A consensus must be reached.

5. Were the criteria (methodological quality) used for assessing the validity of the included studies reported?

❑ yes ❑ partially ❑ no

It was felt that the issues relating to publication bias should not be included in the assessment of this . Yes is given to those meta-analysis reporting "a priori" methods of validity assessment (e.g. if the author(s) chose to include <u>only</u> randomised, double-blind, placebo controlled trials, or allocation concealment as inclusion criteria)

6. Was the validity of all studies referred to in the text assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)?

❑ yes ❑ can't tell ❑ no

This item relates to validity assessment. Yes is given if there is a description of any criteria [either internal or external] used either for inclusion, or for analysis (e.g. sensitivity analysis).

7. Were the methods used to combine the findings of the relevant studies (to reach a conclusion) reported?

❑ yes ❑ partially ❑ no

This item was thought to be reasonably explicit.

8. Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?

❑ yes ❑ can't tell ❑ no

For question 8, if no attempt was made to combine findings, and no statement is made regarding the inappropriateness of combining findings, check "no". If a summary (general) estimate is given anywhere in the abstract, the discussion or the summary section of the paper, and it is not reported how the estimate was derived, mark "no" even if there is a statement regarding the limitations of combining the findings of the studies reviewed. If in doubt mark "can't tell".

9. Were the conclusions made by the author(s) supported by the data and/or analysis reported in the overview?

❑ yes ❑ partially ❑ no

For an overview to be scored as "yes" on question 9, <u>data</u> (not just citations) must be reported that supports the main conclusions regarding the primary question(s) that the overview addresses. If the overview concerns diagnostic/prognostic tests, 'retest is <u>not</u> required' (this ensures that diagnostic/prognostic papers are not scored more rigorously than clinical papers)

10. How would you rate the scientific quality of the overview?

extensive flaws		major flaws		minor flaws		minimal flaws
❑	❑	❑	❑	❑	❑	❑
1	2	3	4	5	6	7

The score for question 10, the overall scientific quality, should be based on your answers to the first nine questions. The following guidelines can be used to assist with deriving a summary score. If the "can't tell" option is used one or more times on the preceding questions, a review is likely to have minor flaws at best and it is difficult to rule out major flaws (i.e. a score of 4 or lower). It the "no" option is used on question 2, 4, 6 or 8, the review is likely to have major flaws (i.e. a score of 3 or less, depending on the number and degree of the flaws).

APPENDIX G-2-1

Quality Assessment of RCTs:

Jadad Scale

			score
1. Was the study described as randomized?	❑ yes	❑ no	
Was the method of randomization described?	❑ yes	❑ no	
If the method of randomization was explained was it appropriate?	❑ yes	❑ no	
Randomization score:			/2
2. Was the study described as double blind?	❑ yes	❑ no	
Was the method of double blinding described?	❑ yes	❑ no	
if the method of blinding was explained was it appropriate?	❑ yes	❑ no	
Double blind score:			/2
3. Was there a description of withdrawals and dropouts?	❑ yes	❑ no	/1
Total			/5

Other items

Was the adequacy of allocation concealment described?	❑ yes	❑ no
Was the design 'cross-over' or 'parallel'?	❑ c. o	❑ para

Scoring of the Jadad Scale:

Either give a score of 1 point for each 'yes' and 0 points for each 'no'. There are no in-between marks.

Give 1 additional point if:	For question 1, the method to generate the sequence of randomization was described <u>and</u> it was **<u>appropriate</u>** (table of random numbers, computer generated, coin tossing etc.)
and / or	If on question 2 the method of double-blinding was described <u>and</u> it was **<u>appropriate</u>** (identical placebo, active placebo, dummy, etc.)
Deduct 1 point if:	For question 1, the method to generate the sequence of randomization was described <u>and</u> it was **<u>inappropriate</u>** (patients were allocated alternately, or according to date of birth, hospital number, etc.)

and / or	For question 1, the study was described as double-blind but the method of blinding was **inappropriate** (e.g. comparison of tablet vs. injection with no double dummy).

APPENDIX G-2-2

Jadad Scale with points of clarification*:

			points
1. Was the study described as randomized?	❑ yes	❑ no	
was the method of randomization described?	❑ yes	❑ no	
if the method of randomization was explained was it appropriate?	❑ yes	❑ no	
Randomization score:			/2
2. Was the study described as double blind?	❑ yes	❑ no	
was the method of double blinding described?	❑ yes	❑ no	
if the method of blinding was explained was it appropriate?	❑ yes	❑ no	
Double blind score:			/2
			/1
3. Was there a description of withdrawals and dropouts?	❑ yes	❑ no	

* *to score yes on this item, the report had to contain a description (either in the text, or in tabular form) of losses in both groups (treatment and control). If there were no losses this had to be stated explicitly.*

Total /5

Other items

Was the adequacy of allocation concealment described? ❑ yes ❑ no

Was the design 'cross-over' or 'parallel'? ❑ c. o ❑ para

Scoring was unchanged.

APPENDIX H:

Derivation of ROR

Model:

$$\log OR = \beta_0 + \beta_1 [\text{trial } i^{th}] + \beta_2 [\text{intervention } j^{th}] + \beta_3 [\text{MA } k^{th}] + \beta_4 [\text{publication status } l^{th}] + \beta_5 [\text{intervention } j^{th} * \text{MA } k^{th}] + \beta_6 [\text{intervention } j^{th} * \text{publication status } l^{th}] + \varepsilon$$

where * denotes an interaction

i: 1, ...,nk	nk is the number of trials within MAk	
j: 0, 1	0 for control	1 for intervention
k: 1, ...m	m is 39	
l: 0, 1	0 for published	1 for grey
ε:	is the random error	

Consider the possible scenarios for intervention and publication status:

Publication status (l)		Intervention (j)		interaction
1	grey	intervention	1	1
1	grey	control	0	0
0	published	intervention	1	0
0	published	control	0	0

Ratio $\frac{\text{OR Grey intervention vs. control}}{\text{OR Published intervention vs. control}}$.

$$ROR = \frac{OR_G}{OR_P}$$

$$= \frac{[(P_1/1-P_1) / (P_o/1-P_o)]_G}{[(P_1/1-P_1) / (P_o/1-P_o)]_P}$$

$$= \frac{\dfrac{e^{\beta 0 + \beta 1\, t + \beta 2\, i + \beta 3\, MA + \beta 4\, ps + \beta 5\, i\, MA + \beta 6\, i\, ps}}{e^{\beta 0 + \beta 1\, t + \beta 3\, MA + \beta 4\, ps}}}{\dfrac{e^{\beta 0 + \beta 1\, t + \beta 2\, i + \beta 3\, MA + \beta 5\, i\, MA}}{e^{\beta 0 + \beta 1\, t + \beta 3\, MA}}};$$

$$= \frac{e^{\beta 2\, i + \beta 5\, i\, MA + \beta 6\, i\, ps}}{e^{\beta 2\, i + \beta 5\, i\, MA}}$$

$$= e^{\beta 6\, i\, ps}$$

$$ROR = e^{\beta 6}$$

APPENDIX I:

Organ system classification

cardiac
gi
depression
infection
reproduction
circulatory
renal
liver
smoking
skin
brain
rhematology
cancer
spinal
eye
addiction
anesthesia