Updating Searches for Systematic Reviews

Margaret Joan Sampson

Department of Information Studies

Aberystwyth University

Submitted in partial fulfillment of the requirements of the

Degree of Doctor of Philosophy

2009

ABERYSTWYTH UNIVERSITY DEPOSIT AGREEMENT FOR ELECTRONIC THESES AND DISSERTATIONS

Details of the Work
I hereby opt to deposit[1] the following item in the digital repository maintained by Aberystwyth University, and/or in any other repository authorized for use by Aberystwyth University:

Author Name: Margaret Sampson

Title: Updating Searches for Systematic Review

Supervisor/Department: Dr. Christine Urquhart, Department of Information Studies
Dr. David Moher, Department of Epidemiology and Community Medicine, University of Ottawa, Canada

Research grant (if any) .....

This item is a product of my own research endeavours and is covered by the agreement below in which the item is referred to as "the Work". It is identical in content to that deposited in the Library, subject to point 4 below.

Non-exclusive Rights
Rights granted to the digital repository through this agreement are entirely non-exclusive. I am free to publish the Work in its present version or future versions elsewhere.

I agree that Aberystwyth University may electronically store, copy or translate the Work to any approved medium or format for the purpose of future preservation and accessibility. Aberystwyth University is not under any obligation to reproduce or display the Work in the same formats or resolutions in which it was originally deposited.

AU Digital Repository
I understand that work deposited in the digital repository will be accessible to a wide variety of people and institutions, including automated agents and search engines via the World Wide Web.

I understand that once the Work is deposited, the item and its metadata may be incorporated into public access catalogues or services, national databases of electronic theses and dissertations such as the British Library's EThOS or any service provided by the National Library of Wales.

I agree as follows:
1. That I am the author or have the authority of the author/s to make this agreement and do hereby give Aberystwyth University the right to make available the Work in the way described above.
2. That the electronic copy of the Work deposited in the digital repository and covered by this agreement, is identical in content to the paper copy of the Work deposited in the Library of Aberystwyth University, subject to point 4 below.

3. That I have exercised reasonable care to ensure that the Work is original and, to the best of my knowledge, does not breach any laws including those relating to defamation, libel and copyright.
4. That I have, in instances where the intellectual property of other authors or copyright holders is included in the Work, gained explicit permission for the inclusion of that material in the Work, and in the electronic form of the Work as accessed through the open access digital repository, or that I have identified and removed that material for which adequate permission has not been obtained and which will be inaccessible via the digital repository.
5. That Aberystwyth University does not hold any obligation to take legal action on behalf of the Depositor, or other rights holders, in the event of a breach of intellectual property rights, or any other right, in the material deposited.
6. The Student's attention is particularly drawn to the next clause
7. That I will indemnify and keep indemnified Aberystwyth University from and against any loss, liability, claim or damage, including without limitation any related legal fees and court costs (on a full indemnity basis), related to any breach by myself of any term of this agreement[2].

Signature ……Date …….

Guidance Notes
Deposit of an electronic version is mandatory, but students are encouraged to opt in to making their theses publicly available via CADAIR. See:

Regulations for Modular Masters' Degrees, paragraphs 38, 39, 41
Regulations for the Submission and Examination of Research Theses, paragraphs 22 & 24
Standing Order 20, paragraphs 41 (MPhil) & 43 (PhD)

Students who previously submitted hard copy versions of theses or dissertations are strongly encouraged to deposit electronic versions and opt in to making them publicly available.

This form may be used by students/former students depositing electronic versions in either of the above scenarios.

The University undertakes to provide appropriate training and guidance with regard to the use of copyright material and its inclusion in theses and dissertations. The supervisor and/or department should assist in checking such inclusions, but the student retains sole and ultimate responsibility for all content so included.

ABSTRACT

Introduction: This thesis examines methods for updating searches for systematic reviews of healthcare interventions. Systematic reviews endeavour to find and synthesize all relevant research as a basis for the practice of evidence-based medicine. They are more useful if they are complete and up-to-date.

Materials and Methods: The sample was 93 meta-analyses in allopathic medicine. Newer randomized controlled trials (RCTs) were sought through MEDLINE searches, and were assessed for relevance by physicians. Two Boolean searches, two similarity searches and one non-database search approach were tested. The Boolean searches were based on a simple subject search paired with a filter selecting only RCTs from Abridged Index Medicus journals or with the balanced Clinical Query. The two similarity searches were Support Vector Machine (SVM) and a Related Article search of PubMed based on the three newest and three largest studies from the original review.

Main Results: Clinical Query provided good recall but with large retrievals. Abridged Index Medicus RCT had smaller retrieval sizes and lower recall, but did detect many large studies. The Related Article search showed the highest recall. Recall with SVM was lower, but retrievals were smaller. RCTs that cited the systematic review being updated were also tested but identified only a small proportion of new evidence.

Relative performance of the test searches was consistent regardless of whether the intervention was a drug, device or procedure. All searches showed variability across clinical areas, but Related Articles RCT showed the most consistency. The pairing of Related Article RCT and Clinical Query gave excellent recall of new relevant material.

Conclusions: Meta-analysts can identify new evidence through a simple structured Boolean search paired with a related articles protocol. By building on the evidence base formed in the original review, related article searching may replace time-consuming non-database methods necessary in conducting original reviews.

DEDICATION

This thesis is dedicated to my father.

Why did we eat the frogs?

TABLE CONTENTS

UPDATING SEARCHES FOR SYSTEMATIC REVIEWS **I**

ABSTRACT **IV**

DEDICATION **V**

TABLE CONTENTS **VI**

LIST OF TABLES **XIII**

LIST OF FIGURES **XV**

ABBREVIATIONS **XVII**

Chapter 1: Introduction 1

1.0 Overview....1

1.1 Statement of the Problem....1

1.2 Rationale 2

1.3 Research Questions....5

1.4 The Research Team 8

1.5 Potential Contributions to the Advancement of Knowledge 11

1.6 Potential Contribution to Practical Information Science Issues....12

Chapter 2: Historical Context of Information Retrieval for Systematic Reviews 13

2.0 Introduction....13

2.1 The Evidence Movement 13

2.12 Progress in Systematic Review Methodology....17

2.2 The Cochrane Library 18

2.3 Information Retrieval....20

2.31 Growth of Online Searching 20

2.4 Information Retrieval in Systematic Reviews 23

2.4 Summary....33

Chapter 3: A Systematic Review of Prior Work on Updating Systematic Reviews 35

3.0 Introduction....35

3.1 Review Methods 35

3.1.1 Search Methods....35

3.1.2 Screening and Selection of Relevant Articles....36

3.2 Results....37

3.2.1 Searching Developments in Updating....37

3.2.2 Step-By-Step Updating Procedures....41
3.2.3 Establishing Priorities for Updating....42
3.2.4 When to Update....45
3.2.5 Hurdles in Updating....47
3.2.6 Developments in Meta-analysis Methods Relevant to Updating....49
3.2.7 Fading of Treatment Efficacy....53
3.2.8 Stability of Observed Effects....55
3.2.9 Harmonization of Updating Efforts....56
3.3 Summary....57

Chapter 4: Methods for the Main Experiment....60
4.0 Introduction....60
4.1 Creating the Cohorts....60
4.1.1 Study Identification....61
4.1.2 Eligibility Criteria....61
4.1.2.1 Main Cohort....61
4.1.2.2 Updated Cochrane Cohort....63
4.1.2.3 AHRQ Cohort....64
4.1.3 Search Strategy....64
4.1.4 Cohort Selection Process....65
4.2 Test Searches: Identification of New Evidence....66
4.2.1 Boolean Searches....67
4.2.1.1 Clinical Query....67
4.2.1.2 AIM RCT....67
4.2.1.3 CENTRAL....68
4.2.2 Citation Searches....68
4.2.2.1 Citing RCTs....68
4.2.3 Similarity Searches....69
4.2.3.1 Related Article RCT....69
4.2.3.2 Support Vector Machine....70
4.2.4 Methods for Executing Each Search....72
4.2.4.1 Clinical Query, AIM RCT, and MA Searches....72
4.2.4.2 Subject Search CENTRAL, inCENTRAL....74
4.2.4.3 Citing RCT Searches....75
4.2.4.4 Related Articles RCT and Related Articles MA Searches....76
4.2.4.5 Support Vector Machine....80
4.2.5 Ranking....83

4.3 Preparation of Material for Screening....84
4.3.1 Building the Updating Spreadsheet....84
4.3.1.1 Screening Worksheet....84
4.3.1.2 Candidate List....85
4.3.1.3 Meta-analytic Calculator....87
4.4 Outcome Measures....88
4.4.1 Assessment of Relevance....89
4.4.1.1 Partial and Full Relevance....90
4.4.2 Establishing the Denominator for Recall....91
4.4.3 Recall of New Studies and Recall of New N....94
4.4.4 Complements to Recall....94
4.4.4.1 Precision....94
4.4.4.2 Specificity....95
4.4.4.3 Fallout or False Positive Rate....95
4.4.5 Summary Measures....96
4.4.5.1 Accuracy....96
4.4.5.2 E....96
4.4.5.3 F....96
4.4.5.4 PosFrac....97
4.4.6 Receiver Operator Characteristic (ROC) Analysis....97
4.4.7 Summary of Outcome Measures....98
4.5 Procedures for Identifying New Relevant Studies....99
4.5.1 Example of Process for Assessing the Need to Update....100
4.5.2 Determination of Need to Update....104
4.5.2.1 Conceptualization....105
4.5.2.2 Classification....106
4.5.3 Quality Control Measures....108
4.5.4 Limitations of the Screening Method....110
4.6 Review-Level Data Extraction....111
4.7 Data Set Creation....112
4.8 Data Analysis....114
Description of the Searches....114
4.9 Search Performance by Intervention Type....115
4.10 Statistical Analysis....116
4.11 Summary of the Main Experiment....117

Chapter 5: Methods - Exploratory Analyses 120
5.0 Introduction 120
5.1 Structural Relationship between Searches 120
5.1.1 Unique Contribution and Overlap 120
5.1.5 Convergence of Multiple Retrieval Methods 121
5.1.3 Related Article Seed Refinement 123
5.1.4 Sufficient Strategies 124
5.1.5 Capture – Recapture 125
5.1.5.1 Data analysis for Capture-Recapture 129
5.1.6 Multidimensional Scaling 132
5.1.7 Correlations 134
5.2 Precision of Searches in a Cross-Sectional Sample 134
5.3 Recall of the Searches in the Original Reviews 137
5.4 Related Article Searching as an Adjunct Search in the Original Review 137
5.5 Performance of the HSSS Revised 138
5.6 Characteristics of the Evidence in Updated Systematic Reviews 141
5.6.1 Where Does the New Evidence Come From? 141
5.6.2 Does Old Evidence Persist? 141
5.7 Is Maturity of the Literature a Predictor of Survival? 142
5.8 Technical Note - Operationalization of Dates 147

Chapter 6: Results 149
6.1 The Cohorts 149
6.1.1 Main Cohort 149
6.1.2 Updated Cochrane Cohort 152
6.1.3 AHRQ Evidence Reviews Cohort 154
6.2 New Evidence 157
6.3 Search Performance 160
6.3.1 Characteristics of the Subject Searches 160
6.3.2 Characteristics of the Authors' Original Boolean Searches 160
6.4 Performance of the Test Searches 161
6.5 Search Performance in the Main Cohort 162
6.5.1 Retrieval Size in the Main Cohort 162
6.5.2 Recall in the Main Cohort 163
6.6 Search Performance in the Updated Cochrane Cohort 165
6.6.1 Retrieval Size in the Updated Cochrane Cohort 165
6.6.2 Recall in the Updated Cochrane Cohort 166

6.7 Search Performance in the AHRQ Evidence Reviews 167
6.7.1 Retrieval Size in the AHRQ Evidence Reviews 168
6.7.2 Recall in the AHRQ Evidence Reviews.....168
6.8 Comparison of Recall of Eligible Studies Pre and Post Signal.....172
6.8.1 Summary of Recall of the Test Searches 173
6.9 Search Performance by Intervention Type 174
6.10 Precision of the Test Searches 175
6.11 Balance of Recall and Precision 176
6.12 Ranking.....178
6.12.1 Ranking in the Updated Cochrane Cohort 178
6.12.2 Ranking in the AHRQ Evidence Reports.....180
6.12.3 Receiver Operating Characteristics (ROC) Analysis 183
6.13 Summary of the Results of the Main Experiment.....185
6.14 CENTRAL.....186
6.15 Central Indexing of New Material for the Main Cohort.....186
6.16 CENTRAL in the Updated Cochrane Cohort.....187
6.17 Searches for Newer Meta-Analyses.....188
6.18 Structural Relationship between Searches.....189
6.18.1 Unique Contribution and Overlap in the Cochrane Cohort 189
6.18.2 Unique Contribution and Overlap in the AHRQ Evidence Reports Cohort 191
6.19 Convergence of Multiple Retrieval Methods.....194
6.20 Population Estimation through Capture – Recapture.....197
6.20.1 Positive Dependence.....198
6.20.2 Negative Dependence 199
6.20.3 Capture – Recapture Population Estimates when Recall of the Ascertainment Methods is Low 200
6.20.4 Capture-Recapture Estimates from Various Ascertainment Methods201
6.21 Multidimensional Scaling.....205
6.22 Correlations of Actual and Test Searches.....208
6.23 Related Article Seed Refinement.....209
6.23.1 Optimizing the Selection Criteria for Seeds 209
6.23.2 MEDLINE Misses as Seed Articles.....212
6.23.3 Performance of Related Article Searching when Retrieval Size is Limited 212
6.24 Sufficient Strategies.....216
6.25 Precision of Searches in a Cross-Sectional Sample.....218
6.26 Recall of the Authors' Searches from the Original Reviews for Updating..221

6.27 Recall of Original Evidence by Searches Used in the Original Reviews 223
6.28 Related Article Searching as an Adjunct Search in the Original Review 223
6.29 Performance of the HSSS Revised 225
6.30 Characteristics of the Evidence in Updated Systematic Reviews....226
6.30.1 Where Does The New Evidence Come From? 226
6.30.2 Journals Contributing New Evidence 227
6.31 Does Old Evidence Persist? 230
6.31.2 Reasons for Exclusion....231
6.32 Is Maturity of the Literature a Predictor of Survival?....233
6.32.1 Survival by Age of Oldest Included Evidence 238
6.32.3 Summary 253

Chapter 7: Discussion and Conclusion....255
7.0 Introduction....255
7.1 Summary of Main Findings 256
7.1.1 Performance of SVM 257
7.1.2 Performance of Citing Reference RCT....258
7.2 Evolutionary Influences....258
7.3 Multiple Database Searching as the Norm....259
7.3.1 Actual *versus* Potential Contribution of Databases....259
7.3.2 Precision Is Reduced In Multiple Database Searches 261
7.4 Supplemental Searches as the Norm....262
7.5 Role of Redundancy....264
7.6 Independent Approaches to the Literature are Necessary....265
7.7 Streamlining Source Selection and Search Limits for the Update....266
7.8 Ways Forward....269
7.9 Surveillance: Setting Targets 271
7.10 Implementation Challenges 272
7.11 Limitations of this Research 273
7.11.1 Screening Method 273
7.11.2 Precision....273
7.11.3 Generalizability 274
7.11.4 Goals Not Achieved 275
7.12 Future Work....276
7.12.1 Related Articles....276
7.12.2 Maturity of the Literature....277
7.12.3 Interagency Collaboration to Automate Support for Updating 278

7.13 Conclusion 279

REFERENCE LIST **281**

APPENDIX 1 – ACKNOWLEDGEMENTS **311**

LIST OF TABLES

Table 1. Change in Terminology of Reports in the MEDLINE Systematic Review Subset 15
Table 2. Citations of Journal-Published Reviews by Indexing Source 75
Table 3. Related Article Pilot Search Results 77
Table 4. Two by Two Table 94
Table 5. Summary of Classification of New Evidence. 106
Table 6. Examples of Authors' Descriptions of Study Flow 135
Table 7. HSSS2006 Counts for Each Element in PubMed and Ovid 139
Table 8. Distribution of Systematic Reviews by Age of First Evidence 143
Table 9. Characteristics of the Cohort of 77 Systematic Reviews Updated by Searching 150
Table 10. Characteristics of the Cochrane Reviews Updated by the Authors 153
Table 11. Characteristics of the Cochrane Update Cohort Outcomes 154
Table 12. Characteristics of the AHRQ Evidence Reports Updated by Searching 156
Table 13. Characteristics of the AHRQ Cohort Outcomes 157
Table 14. Quantity, Sample Size and Age of New Evidence 157
Table 15. How Eligible New Evidence was Identified 159
Table 16. Distribution of New Evidence Amongst the Cohorts Studied 159
Table 17. Comparison of Boolean Search Features 161
Table 18. Number of Records Retrieved Per Systematic Review by Test Searches in the Main Cohort 163
Table 19. On Topic Recall in the Main Cohort 163
Table 20. Recall of Eligible Studies in the Main Cohort 164
Table 21. Recall of Participants from Eligible Studies in the Main Cohort 164
Table 22. Retrieval Size per Systematic Review for Search Methods in the Updated Cochrane Cohort 166
Table 23. Recall of Eligible Studies in the Cochrane Updates 166
Table 24. Recall of N from Eligible Studies in the Cochrane Updates 167
Table 25. Retrieval Size Per Systematic Review for Search Methods in the AHRQ Evidence Reports 168
Table 26. On Topic Recall in the AHRQ Evidence Reports 168
Table 27. Recall of Eligible Studies in the AHRQ Evidence Reports 169
Table 28. Recall of Eligible New Participants in the AHRQ Evidence Reports 169
Table 29. SVM Retrieval of Multiple Publications of Large Trials Eligible for AHRQ Evidence Reports 171
Table 30. Comparison of Recall of Eligible Studies Pre and Post Signal 172
Table 31. Summary: Recall of New Studies, All Cohorts 173
Table 32. Summary: Recall of New Participants, All Cohorts 173
Table 33. Recall of New Participant by Clinical Area, All Cohorts Combined 174
Table 34. Number of Reviews of Drugs, Devices and Procedures 175
Table 35. Recall of the Searches by Intervention Type 175
Table 36. Overall Precision 175
Table 37. Precision and Recall of Ranked Retrieval at Various Cut Points in the Updated Cochrane Cohort 178
Table 38. Recall of the Ranked SVM Retrieval at Various Cut Points – Updated Cochrane Cohort 179
Table 39. Recall of the Ranked Related Article Retrieval at Various Cut Points in the Updated Cochrane Cohort 180
Table 40. Precision and Recall of Ranked Retrieval at Various Cut Points – AHRQ Cohort 181
Table 41. Recall of the Ranked SVM Retrieval at Various Cut Points – AHRQ Cohort 181
Table 42. Recall of the Ranked Related Article Retrieval at Various Cut Points – AHRQ Cohort 182
Table 43. Test of Significance of Area Under the Curve 183

Table 44. Test of Significance of Area Under the Curve, AHRQ Cohort 184
Table 45. Recall in the CENTRAL Subset 187
Table 46. Recall of New Studies Through CENTRAL 187
Table 47. Retrieval Size for Search Methods to Detect New Meta-analyses 188
Table 48. Precision of Meta-analyses 189
Table 49. Unique Component of Each Search from the Cochrane Cohort 190
Table 50. Overlapping Components of Searches from the Cochrane Cohort 190
Table 51. Unique Component of Each Search for AHRQ Evidence Reports 192
Table 52. Relevance by Number of Searches Retrieving a Candidate in the Main Cohort 196
Table 53. Relevance by Number of Types of Searches Retrieving a Candidate in the Main Cohort 196
Table 54. Relevance by Number of Related Article RCT and Clinical Query Searches Retrieving a Candidate in the Main Cohort 197
Table 55. Cross-tabulation of Retrieval Status for Clinical Query and Related Article RCT. 197
Table 56. Comparison of Eligible Records and Capture-Recapture Population Estimates in the Three Cohorts 203
Table 57. Productivity of the Related Article Seeds in the Cochrane Cohort 209
Table 58. Number of Newer Related Randomized Controlled Trials for Seeds that Did or Did not Retrieve New Studies Added in the Cochrane Updates 210
Table 59. Productivity of Largest and Newest Related Article Seeds for the Cochrane Cohort 211
Table 60. Productivity of Related Article Seeds by Position for the Cochrane Cohort 212
Table 61. Productivity of Seeds by Retrieval Status in the Original Cochrane Review 212
Table 62. Retrieval Sizes for Related Article RCT Searches in the Three Cohorts 213
Table 63. Lowest Related Article RCT Rank of Eligible and Ineligible Records 215
Table 64. Search Performance of Related Article RCT at Various Limits to Set Size in the AHRQ Evidence Reports 216
Table 65. Search Strategies Sufficient to Retrieve All Eligible New Studies 217
Table 66. Precision for Different Types of Systematic Reviews 221
Table 67. Recall of New Evidence by the Actual MEDLINE Search Used in the Original Review 222
Table 68. Number of Included Reports from the Original Reviews Indexed in Medline and not Retrieved by the Authors' Search for the Cochrane and AHRQ Cohorts 223
Table 69. Related Article RCT Performance in Retrieving Studies in the Original Reviews of the Updated Cochrane Cohort 225
Table 70. Bibliographic Characteristics of New Evidence 226
Table 71. Journals Publishing the Most Eligible New Evidence 228
Table 72. Journal Impact Factors of New Evidence by Zone 230
Table 73. Distribution of Systematic Reviews by Age of First Evidence 234
Table 74. Size of Systematic Reviews by Age of First Evidence 234
Table 75. Quintile of Age of First Evidence by Specific Qualitative Criteria 236
Table 76. Attribute of the Evidence by Quintile of Age of First Evidence 237
Table 77. Quintile for Age of Oldest Evidence by Signal 239
Table 78. Reviews Including the Oldest Evidence with Major or Invalidating New Evidence 241
Table 79. Titles and Conclusions of Reviews Including the Oldest Evidence but Without Major or Invalidating New Evidence 244
Table 80. Reviews Including the Youngest Evidence with Major or Invalidating New Evidence 247
Table 81. Titles and Conclusions of Reviews Including the Youngest Evidence but Without Major or Invalidating New Evidence 252

LIST OF FIGURES

Figure 1. Hierarchy of strength of evidence to guide treatment decision 16
Figure 2. Online MEDLARS/MEDLINE Searches and Estimated Number of Journal-Published Systematic Reviews1973-1996 22
Figure 3. Example of Subject Search 73
Figure 4. Example of the Search Limits 73
Figure 5. Selecting the Related Article display for the Six Seed Articles 79
Figure 6. Application of Date and Study Design Limits to the Related Article Set 80
Figure 7. Example of XML Output from eLink Related Article Search 82
Figure 8. Example of Replicated Search Strategy 83
Figure 9. Example of Partial SVM Output 83
Figure 10. Example of Screening Worksheet 87
Figure 11. Example of a Receiver Operator Characteristics Plot 97
Figure 12. Overall Process of Determining Updating Status 111
Figure 13. Example of All.xls file 113
Figure 14. Syntax to Convert Retrieval Data from Cases to Variables 114
Figure 15. Excerpt from the Spreadsheet to Identify Sufficient Search Strategies 125
Figure 16. Hypothetical Funnel Plots 146
Figure 17. Flow of Articles in the Formation of the Main Cohort 151
Figure 18. Flow of Articles in the Formation of the Updated Cochrane Cohort 152
Figure 19. Flow of Articles in the Formation of the AHRQ Cohort 155
Figure 20. Recall of On Topic, Eligible and New Participants in the Main Cohort 165
Figure 21. Comparative Performance for the Test Searches 170
Figure 22. Recall and Precision for AHRQ Evidence Reports 176
Figure 23. Recall of New Studies and New Participants and Precision in the Updated Cochrane Reviews 177
Figure 24. Receiver Operator Curve of Eligibility by SVM and Related Article RCT Ranking for the Cochrane Cohort 183
Figure 25. ROC Curve of Eligibility by SVM and Related Article RCT Ranking for the AHRQ Cohort. The positive state is being a target 184
Figure 26. Unique Contribution and Overlap in the Cochrane Cohort 191
Figure 27. Overlap and Unique Component of Test Searches in the AHRQ Cohort 193
Figure 28. All Relevance Categories by Number of Searches and by Number of Methods Retrieving the Record 195
Figure 29. On Topic and Eligible Status by Number of Searches and by Number of Methods Retrieving the Record 195
Figure 30. Capture-Recapture Population Estimates Based on Pairs of Search Methods 201
Figure 31. Derived Stimulus Configuration from Multidimensional Scaling for the Updated Cochrane Reviews and AHRQ Evidence Reports 206
Figure 32. Derived Stimulus Configuration from Multidimensional Scaling for the Main Cohort 206
Figure 33. Correlations of Actual and Test Searches in the Updated Cochrane Review 208
Figure 34. Recall at various Related Article Rankings for the AHRQ and Cochrane Cohorts 213
Figure 35. Precision at Various Related Article Rankings for the AHRQ and Cochrane Cohorts 214
Figure 36. Relationship between Screening Volume and Precision of the Search 220

Figure 37. Relationship between Screening Volume and Number of Included Studies 220
Figure 38. Median Precision by Review Focus 221
Figure 39. Journal Contribution of New Evidence 229
Figure 40. Updated Systematic Reviews that, on Update, Excluded References to Trials Originally Included – Distribution by Age of Oldest Evidence 231
Figure 41. Distribution of Reviews with Signals for Updating by Quintile of Age of First Evidence 235
Figure 42. Distribution of Qualitative Signals by Quintile Based on Age of Evidence where the First Quintile Represents Reviews Including the Newer Trials 236
Figure 43. Survival of 100 Systematic Review 239
Figure 44. Survival of 100 Systematic Reviews by Age of Oldest Evidence 240

ABBREVIATIONS

AHRQ	Agency for Healthcare Quality and Research
AIM RCT	Abridged Index Medicus RCT (one of the searches tested)
AIM	Abridged Index Medicus
ALSCAL	Alternating least squares scaling
AUC	Area under the curve
CADTH	Canadian Agency for Drugs and Technologies in Health
CAS	Chemical Abstract Service
CCT	Controlled clinical trial
CDSS	Clinical Decision Support System
CI	Confidence interval
CINAHL	Cumulative Index to the Nursing and Allied Health Literature
CISTI	Canada Institute for Information Technology
CohortID	Cohort Identification number
COMMA	Comparison of multiple methods of ascertainment
CPCD	Cochrane Pregnancy and Childbirth Database
CQ	Clinical Query (one of the searches tested)
CR	Citing RCT (one of the searches tested)
DERP	Drug Effectiveness Review Project
EBLIG	Evidence Based Librarianship Interest Group
HSSS	Highly Sensitive Search Strategy
$HSSS_{2006}$	Highly Sensitive Search Strategy, 2006 revision
HTA	Health Technology Assessment
HTAi SPIG-IR	Health Technology Assessment international Special Interest Group for Information
InterTASC ISSG	InterTASC Information Specialists' Sub-Group
IRQ	Inter-quartile range
ISI	Institute for Scientific Information
ISRCTN	International Standard Randomized Controlled Trial Number
LIL	Law of Iterated Logarithm
MA	Meta-analysis
MDS	Multidimensional scaling (MDS)
MEDLARS	Medical Literature Analysis and Retrieval System
MEDLINE	Medlars Online
MeSH	Medical Subject Headings

MHDA	MeSH Date
N	Number, usually number of participants in a study
NLM	National Library of Medicine
NNR	Number needed to read
ODPT	Oxford Database of Perinatal Trials
OR	Odds ratio
PICO	Population, Intervention, Comparison, Outcome
PMID	PubMed Identification number
PRESS	Peer Review of Electronic Search Strategies
PRISMA	Preferred Reporting Items for Systematic review and Meta-Analysis
QUOROM	Quality of Reporting of Meta-analyses
RCT	Randomized Controlled Trial
RD	Risk difference
RI MA	Related Article Meta-analysis (one of the searches tested)
RI RCT	Related Article RCT (one of the searches tested)
ROC	Receiver operator curve
RSS	Really Simple Syndication
ScHARR	School of Health and Related Research (University of Sheffield, UK)
SDI	Selective Dissemination of Information
SPSS	Statistical Package for the Social Sciences
SRS	Systematic Review Software
SVM	Support Vector Machine
SYRIAC	Systemic Review Information Automated Collection system
TN	True negative
TP	True positive
TRDG	Trials Registry Development Group
UI	Unique Identifier
UMLS	Unified Medical Language System
WMD	Weighted mean difference

Chapter 1: Introduction

1.0 OVERVIEW

This thesis is about information retrieval issues in updating systematic reviews in healthcare. In this first chapter, I will introduce the problem, describe what a systematic review is, why they must be updated, and state the problems that this research attempts to address. I will present the goals of this thesis, the rationale, the research questions addressed and the potential contribution to the advancement of knowledge as well as its potential contribution to practical issues of updating systematic reviews.

The main work of this thesis is to test the performance of several different search approaches in the search for new evidence for 100 systematic reviews. A sponsored project provided the opportunity to test these searches, so as part of the introduction, I will describe roles and responsibilities, to delineate what my contribution was, and what is the contribution of the research team was.

In the second chapter, I will detail the historical aspects of the development of systematic reviews and the information retrieval methods used in them. The third chapter will provide a literature review of prior work in updating. It also serves as an example to the reader of an updated systematic review, as it was done using the methodology.

The fourth and fifth chapter present the methods used in this research – the fourth chapter details the main experiment in which the experimental search methods are assessed. The fifth chapter covers a variety of topics that help put the main results in context. The discussion will cover the interpretation and application of these results, and demonstrate where advancements in knowledge have been achieved, as well as discussing the additional questions such research inevitably raises.

1.1 STATEMENT OF THE PROBLEM

A systematic review is a state-of-the art literature review designed to address the evidence in a narrowly defined topic, in a manner that is robust against epidemiological

bias. As time passes, a review becomes susceptible to bias as new information may exist that is different from the information synthesized in the review. Of course, until the evidence is actually examined, it is impossible to know if the conclusions of the review remain valid. Therefore, some mechanism is needed to ensure that reviews are up-to-date or at least that they have not been invalidated by newer information.

1.2 RATIONALE

Levine's description of Cochrane Collaboration systematic reviews highlights the elements designed to prevent bias. "The Collaboration's systematic reviews entail exhaustive literature searches and the selection and analysis of studies based on explicit, pre-specified criteria, sometimes, but not always, using meta-analyses, which are formal, mathematical combinations of numerical results."[1] A systematic review should provide the best evidence for clinicians, it should be the starting point for additional clinical trials, and should inform clinical practice guidelines. A systematic review is not able to do any of those things if it is not up to date. An out-of-date review is inherently biased if new studies differ in any systematic way from older studies. The scope note for the MeSH term *bias* is:

> "Any deviation of results or inferences from the truth, or processes leading to such deviation. Bias can result from several sources: one-sided or systematic variations in measurement from the true value (systematic error); flaws in study design; deviation of inferences, interpretations, or analyses based on flawed data or data collection; etc. There is no sense of prejudice or subjectivity implied in the assessment of bias under these conditions."[2]

A ruler with units marked as inches but which were all slightly shorter than an inch would produce biased measurements.

The first wave of systematic reviews was created 30 years old, as will be described in Chapter 2. Even if the findings of those reviews were still valid, few people would have confidence in them unless new work demonstrated their ongoing validity. Thus, some efficient validation process is needed. Systematic reviews have grown substantially in numbers in recent years. There are an estimated 2500 new systematic

reviews published yearly.[3] Few of those are explicit updates, so most may be a novel review that will need updating in the future.

Most developed countries have Health Technology Assessment agencies and many of them face the problem of what to do with an aging fleet of reviews.[4] Many of the issues surrounding maintaining a corpus of reviews relate to surveillance – how to efficiently detect signals that the integrity of the review is being challenged by the appearance of evidence in newly published studies? How to most effectively allocate updating resources across the fleet of reviews to be maintained? Electronic searching and expert opinion have been the principal techniques available for surveillance.[4]

Several rationales for doing an update have been described; 1) increase precision around an effect estimate, 2) monitor changes in the magnitude and/or direction of an effect estimate, 3) minimize impact of time lag bias, 4) timely informing about new developments in a specific field, 5) minimize impact of publication bias, 6) missing or insufficiently detailed data, 7) broaden search comprehensiveness.[5,6] Newly published studies that would be retrievable using the search strategies of the original review would be factors in the first three rationales.

In systematic reviews, the role of the librarian is to develop the search strategies used to assemble the evidence base to be reviewed. They use systematic, reproducible and exhaustive methods to find as much of the research relevant to the topic of the review as possible. In constructing the search strategy, the searcher must attempt to protect the evidence base against publication bias (the preferential publication of positive results), reference bias (the preferential citation of results supporting one's position), language bias (the increased likelihood of certain types of studies being published in English-language sources), and bias from any additional source.[7]

Most systematic reviews conducted today and in the past have relied upon Boolean searches of bibliometric databases to identify the majority of primary studies included in reviews.[8] Expert searchers would normally approach the task by developing a highly sensitive search strategy, using one or more concepts of a carefully constructed

question, and using only limits that have been validated or are clearly relevant to the clinical question. This approach protects the review against missing relevant material, but retrieves a large quantity of irrelevant material that must be reviewed and excluded for cause by expert reviewers. A ratio of 1:9 included versus excluded bibliographic records would be a typical result of screening of a Boolean search result – this will be examined empirically in Chapter 5.

A recent systematic review undertaken by our group shows that there has been some conceptual work on determining when to update, coming mostly from The Cochrane Collaboration and Agency for Health Care Research and Quality (AHRQ). [2] There is one statistical technique that models how the need to update could be predicted by the amount of accumulating new evidence[9] from our research group, a single piece looking at a practical technique in updating the search (updating by using the entry date of database records rather than the publication date), and a number of published techniques in the area of cumulative meta-analysis which help inform the issue of handling multiple testing. Finally, there has been some work looking at the growth of an area and identifying leveling off trends when the need to update may become less frequent.[6] This work will be reviewed in detail in Chapter 3. While there is relatively little published research on updating methods and techniques, an informal listserv survey of a few years ago revealed that information specialists had a variety of techniques for updating the search during the course of the review, to detect evidence as it emerged or just prior to completion of the review.[10] If information retrieval practices in this area are to be evidence-based, then research such as this thesis is needed.

An Expert Working Group on Updating Systematic Review meet in Ottawa Canada in 2006, I was a participant.[11] We developed a conceptual model for the updating of systematic review with three major components; a surveillance function that monitors for the emergence of new evidence, triggering conditions which include the detection of new evidence but also political influences and the gradual erosion of confidence in a review as it ages, and the update itself.[12] This thesis will focus on detecting new evidence.

The ability to conduct electronic searches of bibliographic databases was an enabling technology for systematic reviews. Thirty years ago, the Cochrane Collaboration began publishing systematic reviews, and pledged to update them every year. That objective was unattainable, and was revised to every two years. Even that revised target is under review, as will be discussed in Chapter 3. In the past thirty years, few advances in searching have been made to support updating. While some technical tools such as Selective Dissemination of Information (SDI) exist, and many groups have unpublished working mechanisms for running update searches, this is largely uncharted territory. Anecdotally, the low precision of the searches done in systematic reviews is a barrier to monitoring the literature.

In practice, there are few updated systematic reviews appearing except those undertaken by The Cochrane Collaboration. The literature is replete with systematic reviews on similar topics, although these are rarely explicitly positioned as updates of previous work.[3] It is possible that journal editors are reluctant to publish updates of systematic reviews, and many non-Cochrane systematic reviews are indeed updates of earlier reviews but authors avoid drawing attention to that to enhance their chances of being published.

1.3 RESEARCH QUESTIONS

The major research objective of this thesis is to explore the performance characteristics for various approaches to surveillance, with the aim of identifying approaches that are more efficient for detecting emerging evidence than re-running the high recall search strategy used in reviews.

Updating the evidence base of a systematic review is a different information retrieval problem than forming the original evidence base, and one that has been scarcely addressed. At the time of updating, most (ideally all) relevant studies and many irrelevant articles up to a certain point in timer are known. These studies would have been identified in the course of producing the original systematic review. The information contained in

these true positive (relevant) and true negative (irrelevant) examples has not been harnessed for updating.

The first approach, the one I believe to be most exciting, is to use the results of the screened search for a systematic review to set up an efficient search for surveillance. Surveillance would be to detect the most influential evidence that would render a review in need of update (or at least trigger a consideration of whether an update should be done), based on the accumulation of new evidence. Using support vector machine (SVM), the included studies of this screened set would serve as true positive examples in a training set, and the remaining retrieved records, judged irrelevant, would serve as true negative examples. Updated Cochrane reviews can be used to test the performance of the SVM solution, based on retrievals during the updating period – with the newly included studies being the reference standard against which the performance characteristics, such are recall and precision, are assessed. This mechanism of using updated reviews overcomes the common objection that such machine learning approaches are over fitted to the data, being validated using a "leave one out" approach on the same sample used to develop them. This thesis will advance that aspect of information retrieval and inform more efficient updating.

In the attempt to be exhaustive, systematic review searches have often had very low precision. This makes it challenging to update the literature, as many new irrelevant records will need to be examined, and it makes monitoring for definitive new studies challenging because of the low signal to noise ratio. Anecdotally, the volume of new studies to be assessed has been a barrier to updating. Thus, restricted Boolean searches were tested, to see if higher precision with adequate recall could be achieved. These narrower approaches to subject searching involve MEDLINE searches restricted to *Core Clinical Journals* (previously Abridged Index Medicus),[13] or the *Clinical Queries* of the Hedges team.[14]

Finally, electronic searches for systematic reviews are usually supplemented by a variety of techniques such as checking reference lists, contacting researchers active in the

field and reviewing conference abstracts. The value of these has not been empirically demonstrated, and their role in updating is not known. Therefore, the ability of complementary electronic search approaches, namely Boolean searches and similarity searches, to identify new relevant material is evaluated.

Other approaches I evaluate are *Related Articles* searches in PubMed, citing references, and finally the convergence of multiple retrieval methods. Although many of these approaches have been studied in the past, an examination of their role in updating is novel. The only prior work I have been able to identify is a recent article by Cohen *et al.* 2006,[15] who used a neural network approach to build on the included studies of large systematic reviews. They achieved only modest increases in precision at the very high levels of recall desired for a full update. Yet even these small improvements in precision resulted in meaningful reductions in effort on the part of the review team.

A strength of the research presented here is that the data collection was embedded in another large, well resourced, project. As a result, it was possible to test search performance on a large cohort of high quality and clinically important systematic reviews. The need for updating was determined based on various objective criteria such as the emergence of new studies, accumulation of new research participants in the studies, shift in the point estimate of effect, narrowing of the confidence interval, or emergence of studies either with more rigorous designs or objective clinical end-points instead of surrogate markers or subjective end-points. These criteria were developed by an expert group. The screening of new studies was not done by me, which could potentially have biased results, but by a medical team. Similarly, the decision about whether a systematic review needed to be updated or not was a team decision, made after rigorous examination that often involved intense discussion. Thus, that updating project provided a unique opportunity to demonstrate the performance of various approaches in finding the relevant new evidence.

1.4 THE RESEARCH TEAM

The project that formed the testing ground for these searches was commissioned and funded by Agency for Healthcare Research and Quality, a United States governmental organization. The project was granted to the University of Ottawa Evidence Based Practice Center in January 2006. I was the senior information specialist for that program at the time, and was the "internal leader" for the project. Dr. Kaveh Shojania was the clinical leader. Dr. David Moher, the director of the University of Ottawa Evidence Based Practice Center, was actively involved in guiding and overseeing all phases of the project. The findings of that project have been published as an Evidence Report,[16] and several journal manuscripts derived from the project have been published,[17-19] including one with preliminary results of the searches tested, on which I am the first author.[18]

The protocol for the updating project was developed by this core group of investigators, with input from other team members, and under the guidance of a Technical Expert Panel. The Technical Expert Panel included representatives from several other Evidence Based Practice Centers, the Medical Director of the Agency for Healthcare Quality and Research, journal editors, senior statisticians and epidemiologists. Technical expert panel members and all other team members are formally acknowledged in Appendix 1.

The information science team for the project was lead by myself, and included Jessie McGowan, Tamara Rader, both librarians, Alla Iansavichene, a co-op student with the University of Western Ontario Masters of Library and Information Science program, Raymond Daniel, a library technician, and Dr. Berry de Bruijn, a research scientist. I developed the approach to search strategy evaluation, selected the search approaches to be tested, and developed all procedures for the searches with one exception. I developed the system for tracking which results were found by which methods, and constructing the screening lists that blinded the reviewers to how the records had been identified. I undertook all data analysis related to search performance. Jessie McGowan developed the specific procedures for the Boolean searches and constructed the searches for a pilot of

the method. Jessie also peer reviewed all search strategies that I recreated based on the reports in the original systematic reviews, to ensure their accuracy. Tamara Rader created the Boolean searches for the remaining systematic reviews that were updated through searching. Raymond Daniel downloaded results of the Boolean searches and executed the citation and related article searches according the protocol I developed. He integrated all search results into a unified set for each review to be updated. As well, Raymond produced the spreadsheets and databases of bibliographic records that the reviewers used for screening and recording eligibility decisions. Alla Iansavichene assisted me with data cleaning and did initial coding and classification of the systematic reviews by disease area. She also created the year-by-year search strategies that charted the growth of the literature in these various areas – data that are reported in Evidence Report[16] but not in this thesis.

Dr. Berry de Bruijn ran the Support Vector Machine searches using data sets I prepared from replications of the searches in the original systematic reviews. He was principally responsible for making decision about the settings for the Support Vector Machine tests, but we worked collaboratively through various iterations, assessing performance and deciding on the next strategy. That work arose of discussion between his team at National Research Council and ours over a period of several years as we considered how Support Vector Machine could be harnessed for systematic reviews. These discussions were initiated by ‘Ba Pham, a statistician formerly with our group. Several pilot projects explored the utility of using such techniques to extract basic eligibility criteria from study abstracts, but this was the largest and most fruitful experiment with the technique to date, and the only one related to information retrieval.

The review team consisted of two physicians with a background in research and clinical epidemiology, Dr. Mohamed Ansari and Dr. Jun Ji. They were responsible for screening new records, extracting data from relevant new reports, and integrating that into the systematic reviews being updated. They sought supporting evidence from other published sources and decided how to classify the impact of the new evidence, based on a

worksheet that handled calculations and provided guidance on the criteria for decisions. These were then discussed in a case conference involving the reviewers, Dr. Shojania and myself until consensus was reached on whether the review was in need of update, and what evidence triggered the need for update.

Mary Ocampo extracted various milestone dates from the original and updated reviews, including the date the signaling evidence appeared. Those results are reported in the Evidence Report[16] and one publication in Journal of Clinical Epidemiology[19] but are not a part of this thesis Chantelle Garritty was administrative coordinator for the project and in that capacity handled contract issues, formal correspondence with the sponsoring agency, arranged conference calls with the Technical Expert Panel and undertook numerous other functions that enabled this research. She also completed a pilot survey of agencies that commission or undertake systematic reviews to learn about their updating practices and perceived needs. The results of the pilot were published in the Evidence Report[16] and the results of the full survey are currently under peer review with a journal. That work informs this thesis.

The main analysis from the updating project for the Agency for Healthcare Research and Quality was the survival analysis – how quickly systematic reviews go out of date. That analysis was planned and undertaken by Dr. Shojania. Steve Doucette did the statistical analysis for that part of the project. Those results are reported in the Evidence Report[16] and in the Annals of Internal Medicine.[17]

The analyses described in *Chapter 3 – The Main Experiment* arise out of the data collected as part of the Agency for Healthcare Research and Quality. In this thesis, I have expanded the analysis of considerably beyond what could be completed under a time-limited research contract (the project report was submitted March 15, 2007). As well, the Support Vector Machine material was completed too late for inclusion in the report for that project or the initial journal article. That work described in *Chapter 5 - Exploratory Analyses* represents supplemental analyses that I conducted completely outside the scope of the updating project sponsored by the Agency for Healthcare Research and Quality.

In summary, I was co-investigator on larger project that provided the testing ground for the searches I developed. I planning the search approaches used and the experiment that tested them. I cleaned all search-related data, prepared all files for analysis, and undertook all analysis presented in this thesis, including figures and tables, unless explicitly noted. Of course, many other people assisted with the updating project and with this thesis. I acknowledge them as completely as I am able in Appendix 1.

1.5 POTENTIAL CONTRIBUTIONS TO THE ADVANCEMENT OF KNOWLEDGE

This thesis will contribute two major conceptual advances to the field of information retrieval for systematic reviews. First, it explores the structural relationship between different modes of information retrieval – traditional Boolean searches and similarity-based methods – in the context of updating. It will be shown that these methods are dimensionally distinct and complement each other to provide a level of recall that is difficult to achieve using either method alone. This is an important advance as systematic reviews strive for a very high level of recall.

This thesis also introduces the idea of considering the information density of a document, here measured is the number of patients enrolled in studies that were identified by a search, rather then retrieval of studies, as the unit of measure in assessing the performance of searches in the systematic review context. Systematic reviews attempt to synthesis all available information, and in statistical synthesis, studies with the most patients are given more weight and so influence the results much more than smaller studies. The unit of information retrieval has been the articles that contain the information, not patients participating in studies. This focus may lead to disproportionate effort being expended to find studies that convey little information. Differences in efficiency of searches are more apparent when retrieval is evaluated based on the proportion of new research participants identified rather than the proportion of reports retrieved. Search methods that recall most relevant large studies while retrieving fewer irrelevant studies that need to be screened out permit more efficient surveillance. It then

becomes practical to update based on the appearance of new information, rather than the simple passage of time.

1.6 POTENTIAL CONTRIBUTION TO PRACTICAL INFORMATION SCIENCE ISSUES

This work provides a number of practical contributions to information science. It establishes norms for precision of systematic review searches, making it possible for librarians to bench mark their searches, and yields information for planning and budgeting systematic reviews for accurately.

This thesis contributes a simple and useful procedure for using PubMed related articles to complement the traditional Boolean search strategy. Further, the utility of this simple technique is compared with the more complex Support Vector Machine.

Although there has been a call to report how the search for a systematic review was validated, few reliable techniques exist.[20] This thesis demonstrates a practical method to validate an electronic search by testing it against the included studies identified by any mechanism and indexed in a database. In addition, this work provides the first independent validation of the performance of the recently-revised Highly Sensitive Search Strategy.[21]

Chapter 2: Historical Context of Information Retrieval for Systematic Reviews

2.0 INTRODUCTION

This thesis is about information retrieval issues in updating systematic reviews. Systematic reviews have emerged in the past 30 years as part of the evidence movement within medicine. The history of this movement and of The Cochrane Collaboration have been written elsewhere (*e.g.*,.[1,22-24]) and will be selectively reviewed here. The history of information retrieval is important to the context of this thesis. Again, this will be reviewed only briefly, but the rise of online information retrieval was an important enabling technology for systematic reviews, and it will be shown that the growth of systematic reviews followed the growth of MEDLINE searching closely. Finally, the history of information retrieval in systematic reviews will be examined, looking at milestones and major contributors through the lens of the Cochrane Methods Groups Newsletters, which places them in the context of other developments in systematic review methodology.

This review will also introduce the reader to some of the tools that will be examined later, such as the Cochrane Central Register of Controlled Trials (CENTRAL) and the Randomized Controlled Trial publication type tag available in the MEDLINE database, produced by the US National Library of Medicine.

2.1 THE EVIDENCE MOVEMENT

Archie Cochrane was a physician, born in Scotland in 1909, whose views on the practice of medicine were influenced by his experiences serving in the International Brigade during the Spanish Civil War.[25] Iain Chalmers identified Archie Cochrane's 1972 text "Effectiveness and Efficiency"[26] as a seminal work in the evidence movement in medicine.[27] Cochrane argued that, "because resources would always be limited, they should be used to provide equitably those forms of health care which had been shown in properly designed evaluations to be effective."[28] David Sackett and other leaders in the

field, including Sir Muir Gray and Brian Haynes, published an influential editorial in the journal BMJ in 1996 in which they defined:

> Evidenced Based Medicine: the conscientious, explicit, judicious use of current best evidence in making decisions about the care of individual patients.[29]

Systematic reviews are one tool of evidence-based medicine. Land's Dictionary of Epidemiology[30] defines a systematic review as:

> The application of strategies that limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic. Meta-analysis may be, but is not necessarily, used as part of this process.

It describes meta-analysis as:

> The statistical synthesis of the data from separate but similar, i.e. comparable studies, leading to a quantitative summary of the pooled results.

The term *systematic review* was adopted by those wishing to distinguish such evidence-based reviews from the traditional narrative review. In 1987, Cindy Mulrow published an important empirical study of the quality of 50 review articles published in four prominent medical journals. She assessed them on the basis of eight previously proposed criteria and found that, while most had clearly specified objectives and conclusions, only one had clearly specified methods of identifying, selecting, and validating included information.[31] Huth, the editor of Annals of Internal Medicine who initially rejected but ultimately published the piece, provides a commentary on Mulrow's article, written many years later. He argues that, although review articles were absolutely essential for practitioners, Mulrow demonstrated that these narrative reviews did not meet the standards required of scientific papers.[32]

The first systematic review is thought to be an examination of the literature on scurvy, conducted by James Lind in 1753.[33] The full title of the work is "*A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the subject*". Ian Chalmers traces other early examples of systematic reviews scattered through the next 200 years.[22] The first use of the term "meta-analysis" is credited to Gene

Glass in 1976.[22] Articles titled as systematic reviews began appearing in MEDLINE in 1954, but most of the early literature reviews in the PubMed systematic review subset are described as "critical review of the literature," following Lind's early lead (Table 1, based on MEDLINE results of March 2009).

Table 1. Change in Terminology of Reports in the MEDLINE Systematic Review Subset

Term	1975	1985	1995	2005
Critical review	25	51	111	206
Meta-analysis	0	16	541	2520
Systematic review	0	1	40	1471
Integrative or qualitative review	0	0	8	42
Estimated number of systematic reviews in MEDLINE	25	69	725	3331

Columns do not total to the estimated number of reviews as each review could have used any, all or none of these term in the bibliographic record.

Chalmers traces the roots of the statistical analyses for combining studies back as far as 1860, but he credits Glass with pulling the techniques together.[22] At that time, it was recognized that meta-analysis could provide statistical precision. To ensure validity, it was necessary to set some standard so that the studies included in the meta-analysis were as free from bias as possible. Archie Cochrane set the bar at randomized controlled trials (RCTs). Altman explains that randomized controlled trials have a number of features which, when properly implemented, avoid bias in design or analysis. These are random allocation of participants to the treatment of interest or the comparison treatment, blinding as to which treatment is being given, intention-to-treat analysis in which participants are analyzed even if they failed to complete the treatment to which they were assigned, and a sample size sufficient to reduce the effects of the 'play of chance'.[34]

2.11 EVIDENCE HIERARCHY

Evidence-based medicine is a method in which practice is informed by the best available evidence. Various hierarchies of research designs have been developed, with the

stronger evidence, that is, designs that have more robust protection against bias built into them, placed above weaker designs. The metaphor of a pyramid is used because there tends to be fewer publications for designs higher in the hierarchy. Restricting searches to studies designs with robust designs (those that are at the top of the evidence pyramid) helps narrow down the amount of literature to be appraised. The combination of stronger designs and reduced volume of material to consider is a great benefit for busy clinicians. Systematic reviews and meta-analyses, particularly those restricted to randomized controlled trials, are generally regarded as the highest level of evidence (Figure 1). ScHARR describes a similar hierarchy.[35]

The original hierarchy of strength of evidence for treatment decisions was presented by Guyatt *et al.* in their influential "Users' Guide to the Medical Literature" series in JAMA. This group placed N-of-1 randomized trials above systematic reviews.[36,37] Other hierarchies exist, such as the 5S model developed by Brian Haynes.[38] The five S's are studies, syntheses, synopses, summaries, and systems, recognizing that in many modern healthcare environments, highly formalized point of care systems, called

[*] SUNY Downstate Medical Center.Medical Research Library of Brooklyn. Guide to Research Methods: The Evidence Pyramid. Available at: Evidence-based Medicine

Figure 1. A commonly-used hierarchy of strength of evidence to guide treatment decision.* Used with permission.

Clinical Decision Support Systems (CDSS), guide practice. These too should be evidence-based.

2.12 Progress in Systematic Review Methodology

Methodology advanced quickly for systematic reviews and meta-analyses. Editors of the James Lind Library point out that the first edition of a landmark text titled *Systematic Reviews*[39] was published in 1996 and was less than 100 pages long while the second edition[40] published six years later was nearly 500 pages in length.[24] Of the 26 chapters in the second edition, only one[41] deals with searching.

As methodological complexity has increased, reporting standards have become more stringent, requiring more detailed reporting of the methods used in the systematic review. In 1996, David Moher led a 30-member international study group in the development of the QUOROM statement.[42] QUOROM consolidated the evidence of the day on methodological issues in the conduct of systematic reviews. The group's deliverable was a checklist of 18 items to be reported so that the quality of the review could be assessed. In addition, it called for the use of a flow chart to account for the disposition of all articles identified in the searches used to locate evidence for the review. QUOROM has already had an impact - a recent study found a linear improvement in the number of checklist items reported from 2000 to 2005. For instance, in 2000-2001, 38% of the sample reported adequate details of the search. This rose to 47% in 2002-2003, and 49% in 2004-2005.[43] A revision to the QUORUM statement is in preparation. Called PRISMA, the new version contains a checklist of 27 items, up from 18 in the original version.[44] There were no librarians represented in the original study group, whereas two participated in the updating process.

In summary, the evidence movement is both a philosophy and a set of processes and products. The processes are designed so that both the researcher and the practitioner avoid biased information as much as possible. The products therefore, should be free of bias, and methods have evolved rapidly to help ensure that goal. The evidence pyramid

places research above opinion, synthesized evidence above single studies, and systematic reviews or meta-analyses of randomized controlled trials at or near the pinnacle.

2.2 THE COCHRANE LIBRARY

The Cochrane Collaboration was named in honour of Archie Cochrane, who challenged that, "… a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials."[28] The Cochrane Library was to be the answer to that challenge, developed by the international team who make up The Cochrane Collaboration.

Mark Starr and Ian Chalmers traced the development of The Cochrane Library, which saw its first issue in 1996, and has used an electronic platform from the onset. It had a number of predecessors, the first of which was The Oxford Database of Perinatal Trials (ODPT), which was developed as a source for systematic reviewers. The database was published twice yearly as a complement to several print publications summarizing the evidence. The original medium was 5 1/4 inch computer diskettes. Starr describes the print publications derived from the ODPT as a time slice of the state of the research evidence at the time of publication.

The benefits of this "database plus print" approach were:

> "Unlike more traditional printed articles, the ODPT database system allowed the raw data used in the meta-analyses to be stored with the article, which in turn meant that statistics could be calculated and figures drawn in 'real time', that is, when the figure was displayed. As new studies were added to the database, and new data became available, they were automatically incorporated in the analyses.
>
> One interesting feature when displaying meta-analyses in ODPT was that you could watch the pooled effect estimate and confidence interval change as each trial was added into the analysis, the ordering being based on the assessed quality of each study, its statistical power, or its year of publication."[23]

The format was highly innovative but did not prove to be commercially viable Oxford University Press ceased publishing it in 1992. Update Software redesigned the

product with an emphasis on the systematic reviews. They also enhanced the ease-of-use for the average desktop computer user (a new breed of computer user at the time). The product was re-launched in 1992 as The Cochrane Pregnancy and Childbirth Database (CCPC) – the abbreviation reflects the original name, Cochrane Collaboration Pregnancy and Childbirth database. At the same time, government funding was secured to open the UK Cochrane Centre, with the goal of expanding the work in perinatal medicine to other areas of healthcare. The Centre has employed a librarian from the beginning. The Centre was renamed the UK Cochrane Centre in 1993 with the launch of the international Cochrane Collaboration.

Notable technical features of the reviews at this time were that they were highly structured, enabling users to search and retrieve by population, by intervention or by outcome. There was also a graph generated directly from the stored data, which summarized the outcome data for each review. In 1993, review-authoring software called Review Manager (RevMan) was made available. This enabled authors to develop the review in a structured template that could support varied output formats based on ASCII text, SGML, XML, and various proprietary formats.

In 1994, Update Software demonstrated a prototype of a new CD-ROM publication, The Cochrane Database of Systematic Reviews (CDSR) and it included 615 systematic reviews in the database. By 1996, the Cochrane Controlled Trials Register was added, to help identify primary studies, as was a commenting system that allowed readers to add comments to protocols or finished reviews. The database and register became the Cochrane Library, to which other databases were added as they were developed.

Thus, from the beginning, Cochrane reviews have been designed for an electronic platform, with the ability to incorporate new data as it became available, and to summarize the data in real time. Starr points out that the electronic format is not

constrained by space, allowing more transparent and detailed reporting (including text, graphics and tables) than would generally be possible in print documents.[23] Space was becoming a real issue, given the QUOROM standard's demand for detailed reporting of ever-more complex methods.

2.3 INFORMATION RETRIEVAL

Kagalovsky and Moehr attribute the rise of information retrieval research in the 1950's to two factors. First, there was an influx of documents released after World War II and it was necessary to be able to locate them.[45] Second, the rise of computer technology provided useful tools to organize, index, and retrieve documents.[45]

The United States' National Library of Medicine was a leader, not only in health information retrieval, but generally. It was operating the MEDLARS system (an acronym for Medical Literature Analysis and Retrieval System) and offering on-demand batch searches by 1964.[46] Cheryl Rae Dee has written a recent and comprehensive history of the development of MEDLARS, and explains that the system grew out of efforts to modernize and expedite preparation of the print Index Medicus and other products such as recurring bibliographies.[47] In 1971, MEDLINE (MEDLARS Online) went into operation, allowing off-site, online searches, instead of batch searches. The National Library of Medicine eliminated the in-house batch service in 1973.[46] Other major online databases were developed in this period, often with U.S. government funding and associated with either the National Institutes of Health or the United States Department of Defense. In the case of Chemical Abstract Services (CAS) both were involved.[46]

2.31 Growth of Online Searching

MEDLARS was developed to overcome the limitations of a "card-bound" Index Medicus system. As Adams described in the 1975 Annual Report of the National Library of Medicine, "The fastest card sorter then available could handle 1,000 cards per minute. To search 750,000 cards would take twelve and one-half hours."[48] Adams goes on to

describe the early history of online searching through the National Library of Medicine. An initial development grant was awarded to General Electric in 1960 and the MEDLARS system was operational in 1964.

This system was initially used to develop Index Medicus, but experiments in online information retrieval began soon after, and the 100 journals of Abridged Index Medicus were available to 90 institutional subscribers for online searching in 1970† MEDLINE became available in 55 cities in 1972, with 66 hours of on-line services a week.[48] Searchers were required to undergo two weeks of training, and end-user searching was largely unheard of.[49]

Starr considers the beginning of the perinatal registry, in 1974, to be one of the milestones of The Cochrane Collaboration. This was an effort to identify and index all randomized controlled trials in perinatal medicine.[1] By 1985, 3,500 trials had been identified.[1] Technologies for the registry were a card file of references to perinatal trials, and the development of a MEDLINE search strategy which was run monthly.[23] Today, it is difficult to imagine what it might have been like to try to identify all trials on a topic without access to electronically searchable databases. Indeed, the rise of electronically searchable biomedical databases was an essential enabling technology for systematic reviews.

---

†Here is a description of the previous off-line or batch search process. "Beginning in 1965, searches could be submitted to the National Library of Medicine or to one of the decentralized processing centers that were established in the United States and overseas. Specially trained librarians, who had attended courses that lasted as long as 3 months, then formulated each search and submitted it to a MEDLARS Search Center, where punched cards were fed into a computer, and the resulting printout was shipped back by parcel post. In the United States, turnaround time averaged 4 to 6 weeks." From Coletti 2001,[49] citing Pizer 1969.[50]

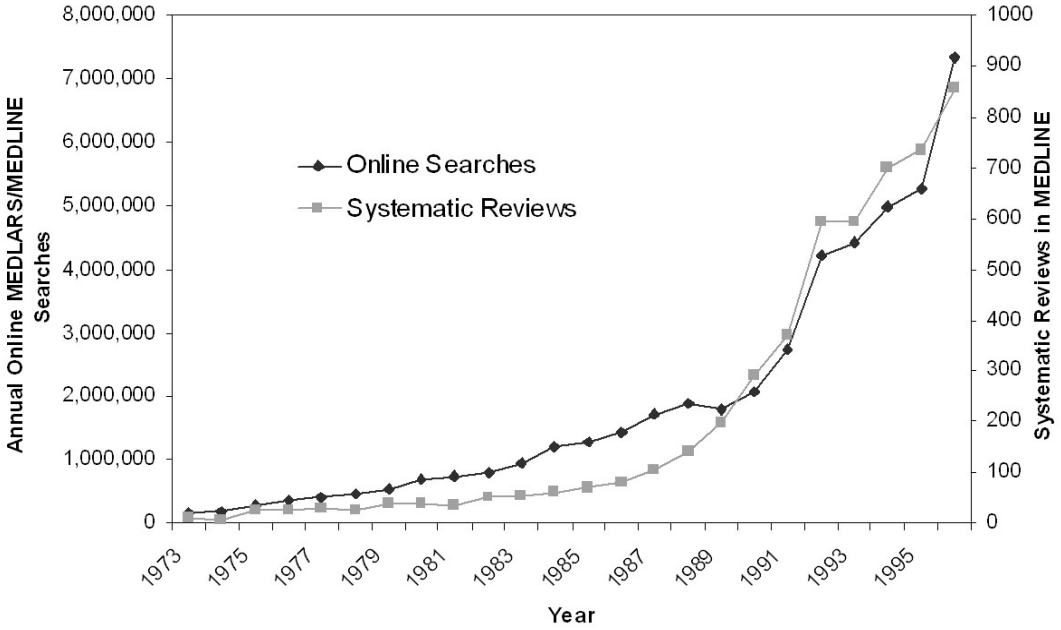


Figure 2. Online MEDLARS/MEDLINE searches and estimated number of journal-published systematic reviews 1973-1996.

Indicators of the growth of both electronic biomedical database searches and of the publication of systematic reviews and meta-analysis are readily accessible. Annual reports of the National Library of Medicine report the number of MEDLARS and later MEDLINE searches conducted through National Library of Medicine,[51-57] and searching MEDLINE itself for the number of records retrieved using the PubMed systematic review subset[58] yield approximations‡ of the number of reviews. These are presented in Figure 2. The years from 1973, the first full year that MEDLARS could be searched online, to 1996 are presented.

As can be seen in Figure 2, 140,000 online MEDLARS searches were being conducted in 1973. By the time the first meta-analysis of perinatal trials was published in 1979, there were half a million MEDLARS searches annually. Meta-analyses and

‡ All records retrieved were examined for 1970-1974 to determine which might be systematic reviews. As the volume rose, numbers retrieved using the subset are divided by 3 to estimate the number that might be assessed as systematic reviews if examined individually. Shojania reported precision of around 50% for the systematic review subset when initially created, but that was with a subject search as well, which would increase precision.[58] Examining a 2004 cohort, we found that 30% of articles retrieved with a modified systematic review search were finally classified as a systematic review.[3] The consistency of these filters across time has not been investigated, although terminology has changed (Table 1).

systematic reviews were not common in the journal literature at the dawn of online searching. Taking the PubMed sys revs subset as a 'surrogate' for the number of sys revs in MEDLINE, the years 1970-1974 yielded only 35 systematic reviews, commonly called critical reviews of the literature. The years 1975–1982, indexed between 25 and 50 publications likely to be systematic reviews. The year 1983 saw 53 such publications. The count of indexed systematic reviews surpassed 100 by 1987, approached 300 by 1990, and 600 by 1992, 700 by 1994, and 874 by 1996. In 1996, there were 7,329,947 online searches of MEDLINE. These figures reflect only those searches done through the National Library of Medicine. By 1982, National Library of Medicine was leasing MEDLARS to private sector provider, of which Bibliographic Retrieval Services (BRS) and DIALOG Information Services Inc. (DIALOG) were the first.[59]

In 1997, PubMed was introduced and anyone could search it through the World Wide Web without any training, any special account, and without any direct cost. The introduction of PubMed resulted in a drastic change in the number of MEDLINE searches.[60] There were 13 million MEDLINE searches during the first ten years shown in Figure 2. Today, there are three million PubMed searches, and approximately 400,000 referrals from Google, every day.[61] Writing in 2001, Donald A. B. Lindberg, Director of the National Library of Medicine explained, "In three years the Library has seen the number of searches on its MEDLINE database rise from seven million searches a year to 250 million. The Library estimates that 30 percent are done by the members of the public for themselves and their families."[62] This may have been the point where ready access to current, synthesized evidence became essential to the practicing clinician.

2.4 INFORMATION RETRIEVAL IN SYSTEMATIC REVIEWS

The original perinatal reviews were supported by a trial registry that was based on index cards. There have been major advances in information retrieval since then. Landmarks were the hiring of an information specialist with the opening of the UK Cochrane Centre in 1992, the development of the CENTRAL,[63] a joint project of The Cochrane Collaboration and National Library of Medicine to identify all controlled trials

in MEDLINE and re-index them with the *Randomized Controlled Trial* or *Controlled Clinical Trial* publication type tags,[41] the publication of the Cochrane Highly Sensitive Search Strategy (HSSS) for identifying randomized controlled trials in MEDLINE in 1994,[64] and its revision in 2006.[21] These major milestones all involve The Cochrane Collaboration, and so the searching-related news from the Cochrane Methods Groups Newsletter, an annual publication beginning in 1997, will be reviewed.

The Cochrane Information Retrieval Methods Group was active by 1997, disbanded in 2001, and resumed operation and was registered as a Cochrane Entity in 2004 with an expanded mandate.

Kay Dickersin and Jean-Pierre Boissel initiated The Trials Registers Development Group (TRDG) at the request of The Cochrane Collaboration steering group (March 1996). The objective of the TRDG was to develop a central register of trials, to support those preparing, maintaining and disseminating systematic reviews within The Cochrane Collaboration.[65] Proposed design elements included fields for design and operational characteristics and results, in addition to the bibliographic record.

At this stage, Cochrane methodological advice through Methods Working Groups was thought necessary for "methods used to prepare and maintain reviews (e.g., statistical methods) and decisions about the methods that are used by the Collaboration to meet its aims (e.g., informatics)." (Mike Clarke, Andy Oxman, Lesley Stewart in a column called "Methods Working Groups: Are They Working?"[65]) There was no mention of literature searching as an essential aspect of review methods.[§] There was however, an Informatics Group convened by David Badger and a Coding and Classification Methods Group, convened by William Hersh. The Coding and Classification group was focused on the indexing of material produced by The Cochrane Collaboration, and on preparing for Cochrane reviews to be indexed in MEDLINE by the National Library of Medicine. The Informatics Group was focused on harnessing new technologies to advance the goals of

[§] I am a co-author on a Cochrane Methodological Review protocol titled "Checking reference lists to find additional studies for systematic reviews" which was published in The Cochrane Library in Issue 1, 2009.

the Collaboration.[65] Original research in searching was being done, regardless of the lack of importance placed on it. In the 1997 Methods Groups Newsletter, the paper by Kay Dickersin, Roberta Scherer and Carol Lefebvre titled "Identifying relevant studies for systematic reviews"[64] was cited as an example of a systematic review of a methodological questin by Andy Oxman in his column "Empirical Methodological Studies".[65] This paper established that conventional search methods might identify only half of the relevant controlled trials from MEDLINE, and introduced the Highly Sensitive Search Strategy.[64]

In the 1998 newsletter, the Dickersin, Scherer and Lefebvre paper was featured in detail. In a commentary on that paper, William Hersh mentions the effort spearheaded by The Cochrane Collaboration to re-index RCTs with the publication type tags.[66] There is no detail of the project in the Methods Groups Newsletters but a full description of the project is found elsewhere.[41] An examination of indexing of studies in MEDLINE had shown that many studies that used the controlled designs were not indexed that way.[41] In 1994, the UK Cochrane Centre and the New England Cochrane Centre began examining records retrieved by the Highly Sensitive Search Strategy to identify additional trials, and the National Library of Medicine re-indexed them as either Controlled Clinical Trials or Randomized Controlled Trials, as appropriate according to the design used. These records were also added to the CENTRAL. Review of studies published from 1965 to 1979 resulted in the indexing of 70,000 additional records as controlled trials in MEDLINE.[41]

Although there are numerous references to the practice of handsearching of journals to identify controlled trials, there is no description of the considerable efforts of the Collaboration to methodically search over 1,700 journals cover-to-cover, as this effort fell under the jurisdiction of the various Cochrane Centres, Fields and Review Groups and Cochrane Fields, rather than to any of the Cochrane Methods Groups.[66] This handsearching effort complemented the electronic re-tagging effort and identified over 17,000 reports of trials not in MEDLINE. These were also added to the trials register.[66]

Also from the 1998 Newsletter, we learn that the Informatics Group was developing an integrated website for the Collaboration. The Coding and Classification

group was considering how to index trials identified from sources other than MEDLINE and how to identify conference abstracts for the trials register.[66]

By 1999, the Coding and Classification Group had become the Information Retrieval Methods Group, convened by William Hersh and Phillipa Middleton. The focus of the group was on improving information retrieval *from* The Cochrane Library. Initiatives included improving documentation, developing an inventory of search engine capabilities, and promoting The Cochrane Library by having a contest for informatics and library students. Students would use the Library in some project that they would present at the following Cochrane Colloquium, with their expenses paid.

Elsewhere in the Methods Groups Newsletter, a report by Barbara Rapp described PubMed and its capabilities. The Related Articles search feature, with relevance-ranked results, was presented, with a concise explanation of how document similarity was determined.[67] As well, in 1999, RevMan software underwent an update to convert it to a 32-bit architecture and some features were added, including a section to describe contributions of each reviewer, and a new format to store references in a structured format.[67]

In 2000, the Information Retrieval Methods Group discussed how to involve more librarians with The Cochrane Collaboration, and how to improve dissemination of methods research, often presented at the Cochrane Colloquium, by including this in the Cochrane Methodology Register. The bulk of the attention, though, was focused on the extensive problems with the searchability of The Cochrane Library, which were documented by a smaller working group drawn from within the Information Retrieval Methods Group. The working group, led by Phillipa Middleton, developed a strategy to communicate issues to the Collaboration and to Update Software, the developers of The Cochrane Library software.[68]

In 2001, the Informatics Group disbanded, as it had been quite inactive and had not fulfilled the key functions of the Methods Groups. This decision was taken by the group when it met at the 2000 Cochrane Colloquium. The Information Retrieval Methods

Group also disbanded, apparently based on a decision by the conveners, who were still William Hersh and Phillipa Middleton. They stated, "We believe there is no longer need for the Group in the Collaboration, and the functions we could perform are actually being performed by others in the Collaboration. Active members of the Group feel we do little more than meet each year at the Colloquium and do not provide any substantive support for the Collaboration. The main functions that the Collaboration needs in terms of information retrieval are; (1) assisting Collaborative Review Group's with searching in the production of systematic reviews, and (2) managing feedback and identifying problems with The Cochrane Library. … Because these functions are maintained by others, and many in the Group are actually interested in methodology research anyway, it is felt that the Group does not serve much purpose for the Collaboration." Members agreed to stay in touch.[69]

In other Cochrane developments, the 2001 newsletter revealed a policy change, communicated through The Cochrane Reviewers' Handbook, that it was now recommended that reviews be updated at least every two years, rather than at least each year. There were plans for major revisions to the part of the Handbook concerned with locating and selecting studies. The International Standard Randomized Controlled Trial Number (ISRCTN) was described, along with the hope that it would simplify identification of multiple reports of a single randomized controlled trial.[69]

Research in progress highlighted two research projects relevant to information retrieval: a study by Sally Hopewell, Mike Clarke, Carol Lefebvre, and Roberta Scherer called "A comparison of handsearching with electronic searching to identify reports of randomized trials"[70] and a 1988 study by Steve McDonald titled "Assessment of the precision of search terms in phases 1 and 2 of the MEDLINE Highly Sensitive Search Strategy for 1994-1997" [which appears not to have been published other than in the Methods newsletter.][69]

Although the Information Retrieval Methods Group had disbanded, there was some attention to issues of searching in the 2002 Methods Groups Newsletter. Karen

Robinson reminded those planning to do methodological reviews of the importance of identifying unpublished methodological studies, noting that 9% of the records in the Cochrane Methodology Register were abstracts from meetings.[71] Sally Hopewell and others had begun a systematic review of research studies examining the effect of grey (semi-published or unpublished) literature on meta-analyses of randomized controlled trials.[71]

In 2003, Carol Lefebvre reported that information science journals were being hand searched to identify relevant methodological research to add to the Cochrane Methodology Register. She also provided examples of evidence-based information science research underway within The Cochrane Collaboration:

> • Assessing which bibliographic databases to search by recording and comparing reports of randomized trials identified in each database and analyzing the overlap.[72]
>
> • Evaluating whether searching MEDLINE is as effective as handsearching MEDLINE-indexed journals to retrieve randomized trials for possible inclusion in systematic reviews.[70]
>
> • Evaluating the comparative effectiveness of handsearching versus electronic searching of a variety of biomedical databases to identify reports of randomized trials for possible inclusion in systematic reviews.[73]
>
> • Comparing cover-to-cover searching of journals by hand with searching the full-text of journal articles electronically on screen and with keyword searching of the full-text of journal articles electronically. (Weir, unpublished)
>
> • Designing objectively-derived highly sensitive search strategies for identifying reports of randomized trials in MEDLINE and EMBASE and reports of systematic reviews / meta-analyses in MEDLINE by identifying terms which occur frequently in 'gold-standards' of known reports but which do not occur frequently in other records in the databases.[74]

Other work relevant to searching was also featured in the newsletter; Egger *et al.*'s work on "How important are comprehensive literature searches and the assessment of trial quality in systematic reviews?"[75] This work reported a complex picture of effect size varying by publication status (published or grey literature), language of publication, trial size and susceptibility to bias. The 2002 Thomas C. Chalmers award went to Pamela Royle, for her work on the importance of published errata to randomized controlled trials.[76,77]

The Information Retrieval Methods Group was attempting to stage a comeback, under the leadership of Carol Lefebvre, Steve Pritchard and Alison Weightman. The draft module was published in the Methods Group Newsletter.[78] The proposed scope was much broader than that of the original group; "The Group will seek to provide advice and support, to conduct research and to facilitate information exchange regarding methods to support the information retrieval activities of The Cochrane Collaboration." Some of the areas identified for empirical research were: "conducting and maintaining systematic reviews of information retrieval methods" and "developing and evaluating retrieval strategies for research evidence to support the systematic review process (systematic reviews, randomized controlled trials and other types of research evidence) for use by The Cochrane Collaboration." The proposed mandate included training searchers in effective information retrieval, the appraisal and evaluation of search strategies, carrying out, supporting and encouraging research, offering policy advice to the Steering Group and other parts of the Collaboration, updating of the searching section of the Cochrane Reviewers' Handbook,[79] advising on good practice in reporting search methods, and formalizing a method for monitoring the quality of searching techniques employed in Cochrane reviews. It was suggested that the Group would complement the work of other Groups, notably the Cochrane Reporting Bias Methods Group, as Egger *et al.*'s work had demonstrated how intertwined the issues of study identification and reporting bias were.[75]

The proposed Information Retrieval Methods Group had identified 100 potential members or supporters by the time the 2004 Methods Group Newsletter was published in June,[80] but the group was not formally registered as a Cochrane entity until November 2004.[81]

By 2005, the group had 150 members and had had a very successful initial meeting, attended by 50 people.[81] Early work of the group was described, and included contributing to the InterTASC web resource of published search filters, and a project in which I was a co-principal investigator, along with Jessie McGowan, to develop a

checklist for assessing the quality of search strategies in Cochrane reviews (Peer Review of Electronic Search Strategies - PRESS).[82,83]

The 2005 Newsletter also featured work by Su Golder *et al.* titled "Developing efficient search strategies to identify papers on adverse events using precision and sensitivity analysis and using statistical analysis", which had been presented at the Colloquium and was published the following year.[84] There was growing interest in harms reflected in other reports in the newsletter.

By the 2006 report to the Methods Groups Newsletter, Jessie McGowan had been recruited to take Steve Pritchard's place as co-convener on his retirement. Major initiatives included the publication of a revision of the Highly Sensitive Search Strategy,[21] revision of the searching chapter of the Cochrane Handbook,[79] and initiatives to address the issue of the descriptions of the searches used in Cochrane reviews, undertaken in collaboration with Health Technology Assessment groups.[85]

Other search-related research that was presented involved the challenges of finding diagnostic studies[86,87] and a new filter for detecting clinically sound treatment studies in EMBASE from the McMaster University group.[88] Meanwhile, work by Lasse Schmidt and Peter Gøtzsche suggested that traditional review articles may not have improved decisively since Mulrow's work in 1987, in the sense that work cited in these non-systematic reviews was biased in favour of studies with positive outcomes.[89]

If harms were a theme of the 2005 newsletter, then updating was the theme in 2006. Rob Scholten described the Updating Working Group and announced a pilot project to set priorities for updating, given that the objectives of updating every two years was not proving to be feasible either for review authors or review groups that needed to edit and approve the reviews.[85] Work by Simon French et al. on "How do conclusions change when Cochrane reviews are updated?" was also reviewed.[90]

The 2007 Methods Groups Newsletter featured the completed work by Wong[88] and Golder.[84] The Thomas C. Chalmers Award for the best presentation at the Cochrane Colloquium was again awarded to a paper on searching; "Comparison of two different

search strategies in identifying literature for a diagnostic test accuracy review of Down's syndrome screening" by Alldred *et al.* unpublished]

News from the Information Retrieval Methods Group included details of representation of information retrieval methods to various Cochrane advisory groups, workshop activity, work on the Handbook revisions, and progress of the PRESS project. In other Methods Group activity, there was a proposal to form an Adverse Events Methods Group, with 30 interested members, but there was no report from the Updating Working Group.[91]

The 2008 Information Retrieval Methods Group column in the newsletter updated the status of projects ongoing from the previous year, and reported on a new initiative to develop a framework and methodology for locating evaluation studies that have previously been identified as hard to access, particularly originating from low- and middle-income countries. The project had been funded by the Cochrane Opportunities Fund and was to be led by one of the co-conveners, Alison Weightman, with participation from three other Cochrane entities.[92,93]

A featured study was Pamela Whiting's report "Systematic reviews of test accuracy should search a range of databases to identify primary studies."[92] Its methodological strengths will be noted below. Relevant to this thesis, a Cochrane methodology review titled: "When and How to Update Systematic Reviews"[94] was reported. The next chapter will be an update of that review.

In summary, the most common topics of research in the Methods Groups Newsletters relevant to systematic review searching were search filter development and testing and studies of the contribution of databases. Searching-related studies focused on Boolean database searching, hand searching of journals, or a comparison of the two. There were no evaluation studies of the contribution of checking reference lists,[**] although the technique is often recommended. There is only one mention and no

[**] I am a co-author on a Cochrane Methodological Review protocol titled "Checking reference lists to find additional studies for systematic reviews" which was published in The Cochrane Library in Issue 1, 2009.

empirical research on similarity searching techniques such as the PubMed Related Articles feature.

The methodological sophistication of the reported information retrieval research has increased over the first 12 volumes of Methods Groups Newsletters. As will be shown later, work on empirical methods for search filter development[74] resulted in the improved recall of the 2006 revision of the Highly Sensitive Search Strategy[21] compared to the original.[64] Progress is also evident in studies of the contribution of databases. The most recent[92] showed much more sophisticated methods for developing the "gold standard". It distinguished whether a study was indexed in a database (coverage) or whether it was retrieved through searching, and finally recognized that the impact of missed studies from a source depends, in part, on whether those missed studies report results that differ consistently from the studies that are identified.

As well as the Cochrane Information Retrieval Methods Group, several groups outside of The Cochrane Collaboration should be mentioned for their contribution to information retrieval in systematic reviews. Health Technology Assessment international Special Interest Group for Information Resources (HTAi SPIG-IR) maintains an active listserv for those interested in information retrieval in the closely allied field of health technology assessment and holds a one-day workshop on information retrieval each year at the HTAi annual meeting. The InterTASC Information Specialists' Sub-Group (InterTASC ISSG) of UK-based HTA information specialists maintains a web site of methodological search filters along with a critical appraisal tool for such filters.

Moving to the broader field of evidence-based librarianship, significant entities include a conference called International Evidence Based Library & Information Practice Conference that has been held every two years, under the leadership of Andrew Booth as International Program Committee Chair. The first conference was held in Sheffield in 2001 and the fifth will be held in Stockholm in July 2009. The Evidence Based Librarianship Interest Group (EBLIG) of the Canadian Library Association sponsors the

open source journal *Evidence Based Library and Information Practice,* now in its fourth year of publication.

2.4 SUMMARY

Systematic reviews of randomized controlled trials are the highest level of evidence to inform medical decisions. The Cochrane Collaboration is the dominant entity creating systematic reviews, with most significant contributors in the field participating. The Cochrane Library was initially designed around a dynamic electronic platform, where updating could occur in real time. Although this real time updating was not feasible in the end, this original goal may have shaped research in statistical techniques of updating, to be seen in the next chapter.

Electronically searchable databases were an important enabling technology preceding the routine production of systematic reviews. Despite this, the proper focus of the Information Retrieval Methods Group was initially framed as supporting the searchability of the Library itself, without a role to play in conducting original research on information retrieval methods for systematic reviews. The Group disbanded. However, original research had been going on from the inception of the Collaboration, with the landmark work being the development of the HSSS. The handsearching and electronic retagging projects were major and highly successful initiatives and they helped develop MEDLINE and the Cochrane Central Register of Controlled Trials (CENTRAL) into vital sources for systematic review work. Strengths of The Cochrane Collaboration include the early development of a trials register, and having a librarian on staff from the founding of the UK Cochrane Centre. These strengths are embodied in the renewed and active Information Retrieval Methods Group.

Much of the subsequent searching research has involved developing filters for additional databases besides MEDLINE, and developing filters for other study designs besides randomized controlled trials. Other ongoing streams of research within the Collaboration (not detailed to any extent here) are work on statistical methods and epidemiological bias. Emphasis on systematic reviews of diagnostic studies and how best

to address harms have emerged as themes within the Collaboration in recent years. These themes have been reflected in information retrieval research. Updating systematic reviews is a most recent theme, and the research on that is reviewed extensively in the next chapter. It will be seen that very little research has been published on how best to update searches for systematic reviews.

Finally, this review has not be a systematic review and is probably subject to all the deficiencies noted by Mulrow.[31] The highlights of the evidence movement, the Cochrane Library and the infrastructure for searching were selected and reported subjectively. The coverage of search-related research and innovations relevant for systematic reviews is restricted to literature featured in The Cochrane Collaboration Methods Groups Newsletter, and the selection process for material featured in the newsletter is bound to have had some biases. However, the intent is only to set the stage, and illustrate the development of the fields and to document for history the evidence presented in these newsletters. In the next chapter, the literature on updating systematic reviews will be reviewed using a more formal review methodology.

Chapter 3: A Systematic Review of Prior Work on Updating Systematic Reviews

3.0 INTRODUCTION

Prior to undertaking the University of Ottawa Evidence-based Practice Center study on updating systematic reviews, from which much of the data in this thesis are drawn,[16] the Ottawa group undertook a methodological systematic review of updating methods. This was published both as a journal article[6] and as a Cochrane methodology review.[94] Although published fairly recently, the search date for the review is 2005, and there has been a great deal of interest and activity related to the issues of updating systematic reviews since that time. Therefore, I updated the search and screened the results for additional relevant material.

3.1 REVIEW METHODS

3.1.1 Search Methods

The following methods were used to identify potentially eligible reports. The electronic search strategy from the original review covered the period from 1966-December 2005 and was re-run in MEDLINE (1950 to January Week 4 2009) using the Ovid interface, and restricted to records entering MEDLINE since the date of the last search (December 2005). The search was also re-run in the Cochrane Methodology Register (Issue 1 2009) using the Wiley interface. In addition, LISTA, a library science database, was searched using the strategy `(update or updating or updated) and (systematic review* or meta-analys*)` in any field, with a publication year of 2005 or later. The other search strategies are available in the appendix to the journal article.[6] The first five PubMed Related Articles were checked for each record from the MEDLINE search that was retained after initial screening, and for all MEDLINE-indexed articles included in the original systematic review.

Citing references were sought for 11 articles of interest[6,17,18,95-102] and for the ten journal-published articles included in the original review, although only citing references published since the search date of the original review were retained for screening. Scopus was used as the source of citing references due to its greater currency.

Abstracts from an international Health Technology Assessment conference called HTAi, held July 6-9. 2008, Montréal, Canada were examined, based on a suggestion of one author contacted (available at http://www.htai2008.org/en_ebook.phtml). Finally, any relevant material already known to me was added to the database.

Full publications of relevant new conference abstracts were sought, and where full publication was not identified, the first or corresponding author was contacted by email with a request to provide the full presentation, or any subsequent publications or other relevant material. Most of those approached responded and either provided material that is more complete or confirmed that there was no additional material available.

3.1.2 Screening and Selection of Relevant Articles

For the original review, a report was eligible if it described the development or use of one or more method, technique, or strategy either for updating or for determining the need to update systematic reviews in health care. Empirical studies, editorial reports or descriptions of such methods that appeared in an updated systematic review were all eligible.[6]

Ideally, selection of articles for a systematic review is based on the consensus of two or more readers that the article meets all inclusion criteria. This helps protect the review against unintended bias or simple misinterpretation on the part of a single reader. This review update was undertaken with only one person (myself) screening and making eligibility decisions, therefore screening was liberal and all records that appeared informative to issues of updating systematic reviews were included.

Only 15 reports, representing seven studies, were eligible for the original review. The evidence base proved to be quite limited, with a small body of literature on cumulative meta-analysis[103-106] and the volume of new evidence needed to overturn

previous meta-analytic results,[9,101] one evaluation of the shelf-life of clinical practice guidelines produced for the Agency for Healthcare Research and Quality,[107] and some suggested approaches to updating from The Cochrane Collaboration.[27,108,109] These findings will be integrated with the new material.

3.2 RESULTS

For this update, 1,411 new records were identified. Of these, 1,188 were from the Boolean database searches, 208 were uniquely from citing references, 13 additional articles were found through Related Articles follow-up searches, and two new articles were provided by authors. All relevant material that I had known of prior to the search was also detected by the search. The title and abstract of these records were examined, and 94 appeared potentially relevant to the systematic review. For these potentially relevant studies, the full article was obtained and examined, and 30 of these were retained.

As well as being more plentiful, the new material was richer in content, reflecting a rapidly developing field. The evidence is summarized below, grouped under the themes of searching developments, developments in meta-analytic methods, fading of treatment efficacy, stability of observed effects, step-by-step updating procedures, determining when to update, establishing priorities for updating, hurdles in updating, and harmonization of updating efforts between agencies.

3.2.1 Searching Developments in Updating

The original systematic review included only one study related to searching, a conference abstract that concluded update searches should be restricted to any record entering the database since the last research, rather than being restricted by publication date.[110]

Turning to the new studies, Cohen has published three papers on document triage. In the 2006 paper, Cohen used a voting perceptron, a type of artificial neural network, to triage documents to be screened for eligibility for systematic reviews. The screened

records for 15 systematic reviews of drug intervention were used as the as training sets.[15] That paper was excluded from the original review, however two new papers[111,112] continue that work and have explicit application to updating. In the second paper,[111] Support Vector Machine is tested, and area under the curve is the evaluation criteria. A significant area under the curve would indicate that the relevance ranking of documents by Support Vector Machine was better than chance. The screened records of the original reviews served as both the training and test sets using split samples. These records had been identified through Boolean searching. The goal was to rank a particular set so that relevant articles would be seen first, and that there would be very high recall, that is, 95% or more of relevant articles would be ranked. This group did not test the ability to detect subsequent publications; instead, they tested the ability of the final algorithm to rank articles from the Boolean searches for the original reviews. Nevertheless, if effective, the methods could be used to find and rank newer articles in order to update the review.

An investigator can control what features of a document are considered by Support Vector Machine to determine relevance. Cohen's group looked both at features that would be specific to a particular review as well as non-topic-specific training. Effective non-topic-specific training has been independently demonstrated by Aphinyanaphongs[113,114] and more recently by they McMaster University group.[115] Non-specific features would be, for example, randomized controlled trials from major medical journals. In these results, Cohen found topic-focused training out-performed non-specific training. Non-specific training still had merit and could be useful in updating where there were very few studies included in the review that could be used as positive training examples, or in contexts other than updating when no topic-specific training set was available.

In terms of document feature sets, Cohen tested title and abstract units, MeSH term variants (MeSH) and UMLS terms identified for each phrase (MMTX) alone and in combinations. MeSH terms are subject headings from the MEDLINE thesaurus. Variants examined were MeSH headings designated as major terms for that article, terms and

subheadings separately, and terms and subheadings together as complete MeSH terms. UMLS terms are terms from the Unified Medical Language System, described by the National Library of Medicine as a meta-thesaurus useful in investigating knowledge representation and retrieval questions.[116] MeSH and text-based features used together provided the best performance. There was no performance advantage to including UMLS features. It has been thought that using UMLS might help compensate for records where MeSH terms were missing, such as new in-process records. However, text-based features performed adequately when MeSH terms were not available, and using UMLS was computationally intensive.

Working backwards, Cohen's initial presentation of this line of inquiry[117] showed that training on a set of relevant documents produced good performance (i.e. high recall) in cross validation using that same set of documents but had much lower performance when applied to subsequently published documents. They noted that, "by its very nature, the field of science changes over time, as does the language used to describe it." Although they undertook some analysis in an attempt to explain the decline in performance, they were unable to do so completely, but concluded that document triage systems may need frequent re-training or even need continuous training. In subsequent papers, they do not bring this out, but train and test on the same sample. This may be methodologically important. MeSH, as any thesaurus, is intended to draw together vartions in terminology to a single concept. This should mitigate regional and temporal variations in usage. MeSH features did not appear to figure prominently in successful ranking.

In a related paper, with Yang as the lead author, this group reports on a system to automatically identify and harvest MEDLINE records for the screening sets used in systematic reviews.[112] The system, called SYRIAC (for the Systemic Review Information Automated Collection system), takes an export of an EndNote screening database, performs some record clean up, submits the records to PubMed, and then adds the EndNote fields and PubMed fields for matched records to a data warehouse, along with reviewer's decisions on the relevance of each item. Eighty-two percent of records from

the screening database were successfully matched with a PubMed record, an improvement from the 30%-50% match rate this group achieved with similar data using PubMed's Batch Citation Matcher. Many of the unmatched records could have been from sources not indexed in MEDLINE, including grey literature.

The point of forming the data warehouse was that records from this system would be used as the true positive and true negative training set for a classifier. The system could automatically update the data at intervals as additional records are screened and classified. Such as system is important as infrastructure for a surveillance system for new material; it would replace many of the manual tasks involved in the preparation of data for this thesis.

The Cochrane Collaboration Updating Steering Group has prepared a draft report which includes a decision tree to determine if a review needs to be updated, and a checklist to determine what changes are needed to the Cochrane Review, section by section.[118] Most of those recommendations will be discussed in *Section 3.2.2*, but the guidance for monitoring for new studies through searches is as follows:

• Auto Alerts to monitor bibliographic databases (i.e., automatic rerun of full search strategy every time new studies are entered into a bibliographic database, with any new hits sent by e-mail)

• Auto Alerts to monitor tables of contents of journals

• PubMed 'Related Articles' feature (citing Sampson 2008[18])

• Citation tracking (via Citation Indexes e.g., Science Citation Index); and

• Checking Cochrane Review Group Specialised Register.

The terms in the search strategy are to be checked periodically for changes and the strategy and associated auto alerts are to be kept up to date. If the search is revised, the details of search strategy are to be reported in the body of the review or in an appendix.

They also advocate reporting the complete date of search, not just month and year, but unfortunately are not clear that the salient information is the date of last update of the database, not the calendar date when the search was run.

3.2.2 Step-By-Step Updating Procedures

The original review identified an early paper by Chalmers *et al.* on preparing and maintaining meta-analyses, which describes the processes used in maintaining the original Oxford Database of Perinatal Reviews described in the previous chapter.[27] The position taken by The Cochrane Collaboration in 1995, on maintaining and updating Cochrane Collaboration Reviews was also reported in the original updating review. Essentially, this involved repeating the search at least every two years, and determining if there were relevant new studies. If none were found, that was noted in the review when it was re-published in the next issue of The Cochrane Library, otherwise, reviewers were to integrate the new material and publish an updated version of the review.[108,109]

The Cochrane Pain, Palliative and Supportive Care Group has a checklist for updating.[119] Designed to ensure standards are met so that the review is as complete as possible prior to submission for editorial and peer review, it covers all aspects of the review. It is equally applicable to an original review as to an update, however for many sections there is guidance on how new material should be integrated and flagged (for instance, changes in some sections warrant mention in the "What's New" section and these are flagged). The Pain, Palliative and Supportive Care Group checklist informed the Cochrane Updating Steering Group's work.[118]

The Cochrane Collaboration Updating Steering Group has considered the Collaboration's updating strategy in some detail. Their report provides section by section guidance on how a Cochrane review should be updated.[118] For some sections, the instructions are as simple as "Contact details: Are they correct? If no: update as necessary." Guidance for other sections, such as the results section, is much more complex, covering the cases of meta-analysis, narrative synthesis in absence of meta-analysis and quantitative analysis other than meta-analysis. Under quantitative analysis they include the quantitative triggers for updating used by Shojania *et al.* in our work on the topic.[16]

Clinical Evidence is a medical reference source, published in both print and electronic format, which is promoted as both evidence-based and up-to-date. The editorial team reviews all topics annually.[120] Their search protocol includes an initial search for high quality systematic reviews. If a systematic review exists, further searches are done only for randomized controlled trials published after the search date of the review, using MEDLINE, Embase, and The Cochrane Library as default sources. Auto Alerts to monitor bibliographic databases are set up and results collated annually. The editorial team now develops a topic plan before each update, allowing them to tailor the approach to the update based on the topic, so some topics will be updated simply based on a repetition of the search, others may have a component added, such as searches for harms data or new treatements.[121]

The Clinical Evidence team appraises any new material using recognized criteria, and good quality evidence is combined into structured summaries by clinicians/specialists. The updated topic is peer reviewed before publication. New studies are highlighted whether they change the conclusion or not. Substantive changes in the evidence are highlighted in the topic summaries. Substantive changes may arise from re-interpretation of existing evidence, appearance of new studies, emergence of harms, either from new studies, clinical experience, or restructuring, such as how the intervention works in different sub-populations, or at different dose levels.[121]

3.2.3 Establishing Priorities for Updating

French examined a cohort of Cochrane Collaboration Reviews that had been published by 1998 and updated by 2002 to see if and how conclusions changed for those reviews that had new studies added in the updated versions.[90] Two reviewers independently classified the conclusions of the pairs of reviews as either unchanged, having minor changes, or having changes that alter the substance or interpretation of the conclusions. The ratio of old and new confidence intervals around the point estimate for the primary outcome was calculated, and any change in significance was noted.

There were 119 eligible reviews. In almost all cases, the width of the confidence interval stayed the same or narrowed. The odds of a change in conclusion increased by 3.3% for every 1% change in the width of the confidence interval. In five cases, statistical significance changed from significant to non-significant, and six reviews with a previously non-significant result for the primary meta-analysis became significant in the update. Conclusions showed substantive change in only 14 of the 119 reviews, and only four of these were accompanied by changes in statistical significance of the primary outcome. In eight of the 14 cases, changes in conclusions concerned outcomes other than the primary outcome. The only factors associated with changed conclusions were change in significance and change in the width of the confidence interval. The authors point out that these are not useful for predicting the need to update.

French *et al.* note that in a number of reviews that had to be excluded because the research question had changed and these new reviews were quite likely to have different conclusions than those of the originals. Although French *et al.* did not find predictors of the need to update, they concluded that the evidence did not support routine time-bound updating, and that the search for indicators of the need to update should continue.[90]

Sutton *et al.*[102] used statistical power as a basis for predicting the amount of new information that is likely to be needed to change the statistical significance of a meta-analysis. The unit of information considered is the number of participants randomized in new trials. Sutton's team has developed software that will calculate the number needed for both random and fixed effects meta-analysis and with relative risk and other measures.[101] The methods were tested on a set of 12 meta-analyses, with studies being removed until the analysis became non-significant. The number of new participants needed was calculated using Sutton's formula and simulations of new studies were developed to test the method's performance.

Sutton also calculated the New Participant Ratio according to the method first proposed in a project in which Barrowman and I were co-principal investigators.[9] This method, which uses the New Participant Ratio, was included in the original review.

Briefly, it focuses on meta-analyses that are not statistically significant. It follows the logic of the "File Drawer" technique for estimating publication bias.[122] That technique asks how many statistically non-significant studies would have to be lingering, unpublished, in a file drawer before they could undermine the statistical significance of a result. The New Participant Ratio instead looks at how many participants would have to emerge in new studies before the non-significant result could become significant, necessitating the update of the review.[9]

Comparing Sutton's estimate of the amount of new information with the New Participant Ratio showed good concordance between the two predicted numbers. Both methods correctly ranked 10 of 12 systematic reviews in terms of likelihood of change. Rankings were the same for both methods for positions 1 through 7 of the 12 positions. Two of the remaining five systematic reviews were ranked in a different sequence by the two methods. Sutton *et al.'s* method produced two anomalies. The 5th ranked study result did not change significantly when the real new studies were re-introduced, and the study ranked least likely to need updating did change significantly. The exceptions were explored to help understand the operation of the method.

Sutton proposes that this method can be used to monitor a suite of reviews. Number of new participants needed for likely change in effect significance would be established for all reviews in a group, such as all reviews by a particular Cochrane Review Group, or all Evidence Reports sponsored by the Agency for Healthcare Research and Quality. Editors would than track the number of new participants from new eligible trials and updating resources would be allocated to those reviews most likely to change. They note that their method could be adapted to clinical limits, to track a certain change in effect size, instead of statistical limits. The software has also implemented the quantitative signals used in the survival analysis of our updating project.[16,17] They suggest that composite signals are a useful avenue for further exploration.

Sutton's method is flexible. Various outcome measures can be included, either fixed or random effects meta-analytic approaches can be accommodated, and signals can

be based on statistical or clinical criteria. Although all 12 meta-analyses tested by Sutton had significant findings that were rolled back to non-significant, the method can be used to predict change from non-significant to significant. While presented as a method to set updating priorities, it could equally be applied to set *a priori* criteria for updating a given meta-analysis.

Sutton *et al.* argues that a strategic approach to updating will reduce the number of updates and reduce the problem of multiple testing (which will be discussed at length in Section 3.2.6). They further suggest that review authors build estimates of the amount of new information needed to warrant updating into the original review.

Both French[90] and Sutton's[102] research found that adding additional information will narrow the confidence interval of a meta-analysis and may result in a change of significance. Both studies also identify cases where the change came in ways that could not be captured in the meta-analysis. Both papers conclude that reviews can require update for reasons other than changes in the primary outcome, and that these may be less predictable. Ultimately, both argue that a system for establishing updating priorities is preferable to updating all reviews on a fixed schedule.

3.2.4 When to Update

The original version of the systematic review of updating cited The Cochrane Collaboration's policy that reviews should be updated every two years. It has become clear that this target is not practical.[95,96,100,123] The Cochrane Collaboration Updating Steering Group started a project to look at updating in September 2007, with support from the Cochrane Opportunities Fund (Kirsty Loudon, personal communication, February 21, 2008). Their draft report titled *Methods to guide decisions of whether, and when, to update Cochrane reviews* takes a new approach, with updating occurring as needed, recognizing that "it is not possible to give a predetermined definitive answer to decide when a Cochrane review should be updated."

Continual monitoring of the literature is recommended to ascertain if the field is moving quickly or slowly, and monitoring strategies are described below. A decision tree

is provided to help authors decide if an update is needed, and the decision to update is the responsibility of the review team. The most likely triggers are the findings of a new study, but other possible triggers, or combination of triggers, might include; new evidence (e.g., information about new treatment regimes, harms, economic data, or outcome measures), new methodology (e.g., new statistical techniques, or changes in the Cochrane Handbook for Systematic Reviews of Interventions or RevMan), or factors such as age of the review, imminent use of the review in policy or guidelines, and response to feedback from users of the review.[118]

The Drug Effectiveness Review Project (DERP) of the Agency for Healthcare Quality and Research produces systematic reviews examining classes of drugs, such as newer antihistamines, Alzheimer's drugs and drugs to treat attention deficit hyperactivity disorder. The DERP team have presented a system for deciding when a review done as part of the program needs to be updated.[124,125]

In their 2007 presentation at the Cochrane Colloquium, they outlined a surveillance process for identifying reviews in need of updating. This involves annual literature reviews in MEDLINE and searches of the web sites of the American and Canadian regulatory agencies for prescription drugs. Identified material is screened by the author of the original review according to the inclusion criteria of the original review. New information is presented to a decision making panel with representatives of those who fund, commission and use the reviews. Applying the approach resulted in a decision to update in 43% of 23 topics in the first report[124] and 53% of 30 topics a year later.[125] Decisions were analyzed in terms of the number of new trials, new drugs, new indications for the drugs and new safety alerts, resulting in nine decision rules. All reviews with more than 26 new controlled clinical trials had been chosen for update (n=5). Beyond that, the decision rules were much more complex and anomalies occurred. For example, when the decision tree was pruned to force fewer rules, the result was that all reviews with more than 26 new trials or fewer than 11 new trials would be updated, but if the number of new trials fell between 11 and 26, no update would take place.[124]

Survival analyses showed median survival times of 9 to 23 months, with all psychiatric topics going out of date within 12 months. However, none of the parameters considered were predictive of the need to update.[124] The team concluded that surveillance did work well for setting updating priorities but that no simple formula could codify the judgments of the decision making teams.[124] The DERP team is also involved with the work of Cohen *et al.*[15]

3.2.5 Hurdles in Updating

Several authors describe logistical and motivational challenges involved in updating systematic reviews. Linde describes the situation of several Cochrane Collaboration Reviews in complementary and alternative medicine for which he is an author.[96] He notes that complementary and alternative medicine has seen a growth of higher quality and larger trials. These are positive developments, however, they increase the work involved in updating a systematic review. In some cases, inclusion criteria needed to be changed to reflect the availability of trials with stronger study designs that therefore were less susceptible to bias. In these cases, the review has to be re-worked extensively. In another review, the early trials were mostly placebo controlled and did not use standard diagnostic criteria. Newer trials used a variety of controls and the standardized diagnostic criteria, meaning that they cannot simply be combined with the trials in the original review – differences in results have to be interpreted in light of the difference in how the trials were conducted. This means that the review needs considerable restructuring. Because of the extent of changes, the revision has to undergo complete peer review, adding time and complexity to the process.

In another case, Linde reported that there was little new material, but the review had been done in the early days of The Cochrane Collaboration and it would take a great deal of work, including re-doing much of the data extraction and analysis, to meet current reporting standards.[96] In this same case, the topic of the review was no longer a research focus for the review authors. In another example, a safety review was undertaken as well as an efficacy review but only the efficacy review was submitted as a Cochrane review

since the prospective of updating the safety review on a regular basis was too daunting. Linde concludes that successful updating is critical to the long-term survival of The Cochrane Collaboration but that regular updates cannot be done without dedicated financial support.

Ervin[95] cites previous reports of limited success in meeting updating targets.[90,126] Henderson[127] and Koch[123] provide similar results. Ervin notes that authors' motivations for undertaking an original review vary, and may influence their motivation to update. For instance, a review may be done as part of a course requirement or as a necessity to obtain funding for primary research, such as a randomized controlled trial. The necessary motivation may not be there to maintain the review through future updates. Ervin reviews the theory of planned behaviour and suggests how it could be applied to assessing motivation for updating. Further, such work sheds light on the practical assistance needed to overcome identified barriers, such as financial support and access to librarians and methodologists.

Jacquerioz *et al.* report on discussions in March 2007 to develop strategies to increase the impact of systematic reviews on maternal and childbirth health issues.[128] The meeting was sponsored by the Agency for Healthcare Research and Quality but had a strong Cochrane influence and looked at creating, updating and disseminating systematic reviews. In terms of updating, participants considered it more appropriate to have the frequency of updates depend upon the topic, as well as the number and quality of new data, rather than at a fixed time, with some rapidly moving disciplines requiring more frequent updating than other slower moving fields. Streamlining of information flow through centralized searching and notification when a certain quantity of new information had accumulated was thought to be helpful. The New Participant Ratio[9] was suggested as a potentially fruitful approach to identifying the need to update null reviews. Identifying willing workers for updating was a second theme. Some suggestions were a matching system to pair interested volunteers and authors seeking assistance as well as promoting updates as 'training' for learning how to prepare a review *de novo*.

French *et al.* hypothesize that the existence of new evidence could be a barrier to updating - i.e. that reviews with little new evidence may be preferentially updated due to workload issues.[90] They do not provide any evidence to support or refute the existence of such a phenomenon, they merely point out that studying only reviews that have been updated may give a biased picture of the true number of reviews in which conclusions would change with updating.

Garritty *et al.* conducted a survey of updating issues and practices of agencies that conduct or sponsor systematic reviews.[4] The survey asked about barriers to updating and the leading barriers, and the percent of respondents who endorsed them somewhat or strongly were: limited funding and resources (72%), reviewer motivation (55%), lack of academic recognition for updating (49%), the need to re-do data extraction (45%), and need to re-assess study quality (38%).

3.2.6 Developments in Meta-analysis Methods Relevant to Updating

The original review on updating included a number of studies involving cumulative meta-analysis. Cumulative meta-analysis is a statistical procedure in which the effect estimate based on all studies combined is sequentially updated by incorporating each newly available study. It should be capable of identifying the earliest time that a definitive conclusion can be reached. A significant limitation is that the chance of falsely rejecting the null hypothesis (making a type 1 error) increases due to repeated hypothesis testing. The original review examined the work of Pogue and Yusuf on sequential monitoring boundaries.[104] They adapt a system that was initially developed for use in clinical trials where the significance level of each testing of the null hypothesis is made more stringent.

Turning to the new research, a number of recent closely related papers present and explore the concept of trial sequential analysis based on information size.[129-132] This is the sample size that would be needed to detect a certain effect size with statistically acceptable risks of falsely accepting or rejecting the null hypothesis.[130] Information size is analogous to sample size calculations for clinical trials. However, unlike the situation of

repeat testing within a single clinical trial, sample size for meta-analysis must take into account heterogeneity between studies.[131]

Trial sequential analysis establishes boundaries where the effect size needed to achieve statistical significance is adjusted upward when the optimal information size has not yet been achieved.[131] In one study, a quarter of meta-analyses that had statistically significant findings and had achieved the necessary heterogeneity-adjusted information size would have had false positive results in earlier versions, had the meta-analysis been updated as each new trial appeared.[131]

Thorlund's paper examines meta-analyses with both the required information size and with significant findings.[131] Meta-analysis was done retrospectively with uncorrected cumulative meta-analysis and with monitoring boundaries established using trial sequential analysis. They claim that in the sample studies, trial sequential analysis eliminated false positive results, which would have been common in cumulative meta-analysis.

Borm and Donders argue that trial sequential analysis can only partially adjust for repeated updating of meta-analyses because of the many unknown factors and unplanned pattern of studies and few, if any, meta-analyses have stopping rules.[133] They suggest that the assumptions underlying the notion of "optimum information size" that would produce a convincing result are often violated as new trials continue to be conducted, even after the optimum information size is attained. These authors propose a "failsafe *k*", which they describe as comparable to the "failsafe N" approach for evaluating the impact of publication bias. Publication bias is the tendency to preferentially publish studies that support a particular point of view – usually preference is given to publishing studies or outcomes with studies that show positive result. Failsafe N represents the number of missing studies with null effects that could be added to a meta-analysis without undermining the statistical significance of the results. Borm and Donders' new measure, failsafe *k*, is the "maximum number of times a meta-analysis can be updated before the type I error exceeds the threshold of statistical significance." Through simulations, they

evaluate how periodic updates alter the type 1 error rate. The experimental variables they used were publication bias, frequency of updates, the stopping rule for drop in p, the power of the trials, and heterogeneity.

Hu *et al.* also look at controlling type 1 error in cumulative meta-analysis, using a method they call LIL, for law of iterated logarithm.[134] Their 2003 paper[103] demonstrated a method for use with meta-analyses of continuous outcomes (such as amount of weight loss), while the 2007 paper looks at binary outcomes (such as mortality). Like Borm and Donders' failsafe *k* method, the LIL method does not require pre-specification of the maximum information, that is, it is not necessary to assume that new research will stop once a definitive result has been achieved through meta-analysis.

Hu *et al.* re-examine Lau's famous example of 33 trials comparing intravenous streptokinase with placebo.[135] Lau demonstrated that by 1973, it would have been apparent, through cumulative meta-analysis, that more patients hospitalized for acute myocardial infarction were alive after 30 days if treated with streptokinase than with placebo. In fact, the full benefit of this life-saving therapy was not recognized until 1988. This paper has been highly influential and has been cited over 500 times.†† Re-analyzing Lau's data and applying the LIL correction for repeat testing, Hu argues that compelling evidence of the superiority of this treatment was only available in 1988, or, using slightly less conservative methods, by 1977.

In this research, the investigators only update the meta-analyses once each year, rather than with the publication of each new study. Annual updating was the frequency originally intended by The Cochrane Collaboration. Hu *et al.* argue that the decision to update a meta-analysis should be based on increments of new information, where information is the sample size, or number of participants in studies. Updating should not be done after each study or after a certain amount of time – even with statistical correction, it is best to minimize the number of inspections.

†† As of March 21, 2009, Web of Science recorded 535 citing articles.

While much of the attention has focused on inflated type 1 error due to repeat testing in cumulative meta-analysis, Bender *et al.* point out that multiplicity exists in many forms in meta-analysis; multiple outcomes, multiple groups, multiple time points, multiple effect measures, subgroup analyses, and multiple looks at accumulating data.[136] They suggest that there is no completely satisfactory solution now, and they advise consumers of systematic reviews to be aware of the problem when evaluating the evidence. In terms of multiplicity due to multiple looks at accumulating data in updating of meta-analysis, they summarize the work of Pogue and Yusuf,[104] Wetterslev *et al.*,[132] and Hu *et al.*[134] in monitoring boundaries and other corrections. Bender *et al.* point out that if the accumulating trials are interpreted within a Bayesian framework, there is no need to adjust for multiplicity. They introduce an important distinction between cumulative meta-analysis completed retrospectively, such Lau *et al.'s* demonstration that evidence supported intravenous streptokinase as thrombolytic therapy for acute myocardial infarction years before that effect was widely recognized,[135] and cumulative meta-analysis prepared as the evidence accumulates and for the purpose of supporting decisions. Bender *et al.* argue that the former does not need to address multiple testing, but the latter does. In summary, Bender recommends that reviewers, "incorporate sample size considerations and the expected degree of heterogeneity for the primary effect measure(s) in prospectively planned cumulative meta-analyses, and that they adjust the threshold for statistical significance or the inflation of the test statistic to account for future multiple looks as data accumulates, if the required sample size has not yet been reached and if future review updates are planned."[136]

Although there are a number of papers focusing on controlling type 1 error rate in cumulative meta-analysis, in practice, most systematic reviews appear to be updated in batch mode, rather than as each new study appears. A search for "update" and "cumulative", limited to systematic review or meta-analysis in PubMed, yielded only 14 records, two of which were the updating systematic review being updated here.[6,94] Many were irrelevant due to the use of "cumulative" in other contexts, such as cumulative

toxicity, cumulative indices, cumulative incidence, lifetime cumulative and average exposures, and Cumulative Index to the Nursing and Allied Health Literature - the CINAHL database. On examination, four of the remaining records did not appear to refer to an updated systematic review, and two were updated guidelines but were not based on cumulative meta-analysis. Finally, only three readily discernible examples of updated systematic reviews using cumulative meta-analysis could be identified as of February 28, 2009.[137-139]

3.2.7 Fading of Treatment Efficacy

Ioannidis studied highly cited treatment studies and found that they generally presented positive results and often showed large effects of treatment.[140] While many of these findings were supported by the results of subsequent studies, 16% were refuted and another 16% had stronger effects than subsequent studies – significantly more than a control sample of less frequently cited trials. Ioannidis points out numerous influences could account for this finding. Since then, an number of interesting results have appeared.[140]

Gehr *et al.* discussed the fading of reported effectiveness.[141] They state that the real effect of a medical therapy should be constant over time, but many factors may influence reports of effectiveness. They studied four drugs; Pravastatin and Atorvastatin to lower cholesterol and the anti-glaucoma drugs Timolol and Latanoprost. They looked at randomized controlled trials that either tested these drugs or used them as a control to test another newer therapy. Three of these drugs showed statistically significant decreases in reported efficacy over time, the other drug (Atorvastatin) showed constant effects. They tested a number of potential explanatory variables, including publication bias, and found that the only consistent factor was that sicker patients were treated in the early studies. Study size, and whether the drug of interest was the test or control for the trial, made little or no difference. They speculate that other factors may be that positive trials are published more quickly than negative trials (a form of publication bias), and that improvements in the conduct of studies over the years make exaggerated positive results

less likely – in this case the apparent fading would be an artifact. They caution systematic reviewers that such time effects may need to be considered when interpreting results.

McAlister and Mohamed looks at the evolution of evidence[142] and described "regression toward the truth". Much of the McAlister and Mohamed paper is a commentary on Vaitkus and Brar,[143] who studied a series of 27 randomized trials and 17 meta-analyses examining N-acetylcysteine for the prevention of contrast induced nephropathy. They described Vaitkus & Brar's findings that earlier studies reported larger treatment effects than later studies as "regression to the truth effect". Vaitkus and Brar restricted their observations to randomized controlled trials, thus allowing them to make this observation with confounding publication date with study design.[143] McAlister attempted to replicate this finding using meta-analysis included in a comparison of paired Cochrane Collaboration Reviews and industry-supported meta-analysis on the same topic. They found that the majority of the meta-analyses showed smaller treatment effects than the first trial, and smaller treatment effects than the largest trial. They argued that this demonstrates the strength of meta-analysis, that meta-analysis is inherently conservative, and that all relevant studies should be included, rather then than excluding small or early studies as unreliable.[142]

The editorial team of Clinical Evidence reviewed all substantive changes in two print issues (December 2005 and June 206). In that period, covering about one year of updates, 1,807 treatment topics were updated, 23% had substantive changes and 17% of those entailed changes in the characterization of how effective the treatment was. Results were evenly split, with 51% the treatments reclassified as more beneficial, and 49% reclassified as less beneficial based on the updated evidence or revised interpretation.[121] In subsequent data from the August 2007 edition[144] and from the February 2008 edition,[145] 52% and 55% of topics with substantive changes were reclassified as less beneficial or more harmful than previously categorized.

3.2.8 Stability of Observed Effects

Stability of observed effects is related to the ideas of optimum information size. It may be possible to look at a body of research results and declare them stable and unlikely to change in the future. Presumably, further updating would be unproductive unless there was some abrupt shift in how the illness was manifested or a new treatment or prevention strategy emerged.

In the original updating review, the work of Mullen, Muellerleile and Bryant was presented.[106] They used cumulative meta-analytic techniques and gauged the "cumulative slope" or rate of change in effect size with the addition of each new study. The smaller the magnitude of the slope of the regression line, the greater the confidence that the pooled effect size was becoming stable. When the rate of change slows sufficiently, it may be appropriate to stop updating a meta-analysis.

In this update, Muellerleile and Mullen used simulation to develop the indicators of sufficiency and stability in cumulative meta-analysis used in updating public health interventions, and then applied them to existing, published datasets.[146] Stability refers to the leveling of the slope described above. Sufficiency is a new element in their work and refers to statistical significance – whether there is sufficient evidence that the phenomenon exists. This is examined through another variant of the failsafe number used by Barrowman *et al.*[9] and by Borm and Donders[133] – the number of unpublished, irretrievable studies with null effects that would need to exist to undermine the statistical significance of a finding.[122] Muellerleile and Mullen created a failsafe ratio that is calculated based on the failsafe number of studies and the number of studies already in the collection. From the point in time when that ratio exceeds 1.0, they argued that there would be no need for additional research to establish that the phenomenon under study did exist.

Applying this work to two examples, they found that cumulative meta-analysis of evaluations of heart-healthy eating programs reached sufficiency and stability quickly, and subsequent evaluations have represented excessive time and effort. On the other

hand, a similar analysis of drug abuse prevention programs showed that the meta-analysis achieved stability some time ago – the size of the treatment effect has remained more or less constant, but sufficiency has never been established, implying that future evaluations of such programs are not worthwhile.[146]

3.2.9 Harmonization of Updating Efforts

Harmonization refers to common standards and coordination of efforts between agencies. The idea of harmonizing updating efforts between agencies arose during a consensus development workshop with the international Expert Working Group on Updating Systematic Reviews that pre-dated the Ottawa Evidence-Based Practice Center's primary research on updating.[11] Harmonization was included as a theme by Garritty *et al.*[4] in a survey of agencies that sponsor systematic reviews and Health Technology Assessments. Seventy percent of respondents somewhat or strongly agreed that centralizing efforts across institutions or agencies that produce systematic reviews would be a worthwhile efficiency, although numerous barriers to such inter-agency cooperation were perceived. Several articles address this, if somewhat peripherally.

Whitlock's group[147] outlined a series of steps that can help reviewers reach reasoned decisions about the incorporation of existing systematic reviews into a new systematic review and enumerate potential hazards to consider in doing so. They discussed the challenges and suggest that a reporting standard be developed for review of reviews. They suggested using previous reviews at least to serve as a crosscheck on the completeness of the search for primary studies and as a source for primary studies. They suggested updating the searches even for very recent reviews. They noted exceptions, particularly when the systematic review being undertaken links several bodies of knowledge, one or more of which may be firmly established, such as the health benefits of smoking cessation. In these cases, no update of the search was warranted. Other than as a means of identifying primary studies, they urged caution in the re-use of reviews, suggesting that only high quality reviews be retained, and that some verification of the accuracy of screening decisions and data extraction be done. Whether or not previous

reviews are incorporated, Whitlock *et al.* recommended that reviewers address any inconsistencies between previous reviews and the current review. Methods used to locate, screen, and appraise reviews and to resolve discrepancies should be reported in a reproducible and transparent manner. They noted that reporting of the handling of previous reviews is not now a part of the QUOROM guidelines.[42]

McAlister[142] does not suggest a method for harmonizing, but addresses the issue of redundancy in the literature, with particular reference to the large number of reviews on N-acetylcysteine. In terms of research priorities, "17 meta-analyses examining a narrow clinical question which contributes relatively little to the global burden of disease represents a disappointing duplication of effort, particularly in light of the fact that the therapeutic options for many diseases which cause substantial disability adjusted life years lost worldwide have not been systematically examined at all." This is a sobering observation in light of Chalmers' initial challenge that, "… a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials."[28]

The Cochrane Collaboration recently conducted a strategic review of its structures, processes and governance. Members of the Collaboration ranked the recommendations, and the recommendation, "to develop a partnership strategy to engage other systematic review producers and knowledge packagers" was ranked sixth out of 26 recommendations.[148]

3.3 SUMMARY

The systematic review updating work reported in this thesis began in 2005, and coincided with an upsurge of interest in the topic. While The Cochrane Collaboration's initial objective of updating all reviews annually was discarded in favour of bi-annual updating, even that target now appears neither achievable nor necessary. Most work has focused on updating on an "as needed" basis; however, it is not at all clear how the need can best be ascertained. The work on controlling type 1 error in cumulative meta-analysis that was reviewed in the initial systematic review on updating continues, although there

are relatively few examples of cumulative meta-analysis being used in actual updates of systematic reviews. Still, the work on controlling type 1 error is relevant to the extent that concepts such as optimum information size can inform when to update.

Apparent changes in efficacy between an original and updated meta-analysis need to be interpreted in light of a small body of work that suggests that the true, and often lower, level of effectiveness of medical treatments may only be apparent over time. Some research explores how best to determine when the stability of observed effects has occurred.

The Cochrane Collaboration and Drug Effectiveness Review Program (DERP) have both outlined protocols for deciding when to update, and they are generally based on expert opinion informed to some extent by new evidence. The Cochrane Collaboration, DERP, and Clinical Evidence have drafted search protocols to use for monitoring, but none of these has been evaluated, and the Cochrane protocol is cumbersome.

Sutton proposes prospective determination of the number of new participants in trials that will be needed to change a significance level of a meta-analysis, and using the accumulation of new studies to inform priorities for updating. The issue remains how best to identify those new participants. Sutton and Cohen's groups have both developed information systems for automating some of the tasks of surveillance. Sutton has developed a system to assess power and determine the number of new participants to be watched for. Cohen's group has automated the process of building training sets to use with classifiers.

Inter-agency cooperation in the refining and in the practical application of methods such as those developed by Sutton and Cohen's groups, could help overcome some of the perceived barriers to harmonization of updating efforts across the corpus of systematic reviews.

Themes run through several of these topics. Some findings are stable, and need no further examination, such as the effects of smoking cessation or heart-healthy eating programs. Updating should be based on the appearance of new information, rather than

the passage of time. New information may either take the form of 1) the number of new study participants (rather than the number of new studies) or 2) the emergence of non-linear information. Examples of non-linear information include the emergence of previously unrecognized harms, introductions of new competing treatments, or some new understanding of sub-populations whose disease condition or response to treatment may differ in some important way, meaning that the application of the treatment should be refined.

Chapter 4: Methods for the Main Experiment

4.0 INTRODUCTION

The main experiment tests several search strategies to determine their performance in the task of identifying important new evidence to updating systematic reviews. These searches are tested in three cohorts of clinically important systematic reviews of reasonable quality. The test searches were used to identify new evidence to see if those reviews became out of date during the observation period. Accuracy of the conclusions based on the new evidence found was verified against confirmatory sources. The performances of the test searches were determined using recall as the primary outcome and, in some cases, precision as a secondary outcome.

This chapter presents the methods used to select the cohort, construct the searches, and prepare the search results for screening, the screening procedures themselves, and the confirmation of findings. In addition, methods and procedures for the creation of the dataset used for analysis, the outcome measures chosen, along with the rationale for selection, and the statistical procedures for analyzing the data for the main experiment are presented.

4.1 CREATING THE COHORTS

A cohort of 109 systematic reviews served as the basis for evaluating the performance of test searches, and for examining the nature of changes in evidence over time. The methods used to identify, select and describe this cohort are reported in this section. These methods are those of the Agency for Healthcare Research and Quality (AHRQ) updating project, published as an AHRQ Technical Report.[16] The investigators, principally Dr. Kaveh Shojania, determined the eligibility criteria, a signal for the need to update a systematic review, and definitions of major and invalidating new evidence. A Technical Expert Panel advised on and approved the protocol.

4.1.1 Study Identification

Systematic reviews were used as the sampling frame. They were identifed from commentaries in ACP Journal Club, a bimonthly publication of the American College of Physicians that aims "to select from the biomedical literature articles that report original studies and systematic reviews that warrant immediate attention by physicians attempting to keep pace with important advances in internal medicine."[149] The process used for ACP Journal Club selection involves "reliable application of explicit criteria for scientific merit, followed by assessment of relevance to medical practice by clinical specialists." Moreover, systematic reviews indexed in ACP Journal Club must meet specific standards. They must report the search methods used and the inclusion and exclusion criteria. Eligible systematic reviews must include at least one study that would, on its own, meet the criteria for write-up in ACP Journal club, according to the standards set for the relevant study type. Therefore, a systematic review of a treatment would need to include as least one study with random allocation of participants to comparison groups, endpoint assessment of at least 80% of those entering the investigation, and an outcome measure of known or probable clinical importance.[149] Thus, choosing this sampling frame allowed us to identify systematic reviews of reasonable quality that are directly relevant to clinical practice.[16]

4.1.2 Eligibility Criteria

4.1.2.1 Main Cohort

In addition to selection by ACP Journal Club, the following eligibility criteria were imposed. Eligible systematic reviews must have evaluated the clinical benefit or harm of a specific drug, device, or procedure (or of a class of drugs, devices or procedures). There must have been a quantitative synthesis (meta-analysis) that included a point estimate and 95% confidence interval for at least one clinical outcome (e.g. disease endpoint, functional status or mortality) or established intermediate outcome (e.g., blood pressure, glycemic control, standard instrument for measuring disease activity, such

as a depression scale). Evaluations of alternative and complementary medicines, as well as educational and behavioural interventions were not eligible.

Furthermore, the systematic review must have been published between 1995 and 2005, but with the date of the final search conducted no later than December 31 2004. This was to ensure at least one full year for new evidence to appear.

At least one conventional meta-analytic estimate of treatment benefit or harm must have been reported in the form of a relative risk, odds ratio, or absolute risk difference for binary outcomes and weighted mean differences for continuous outcomes. Up to four eligible benefits and up to two eligible harms were studied. We excluded individual patient data meta-analyses, meta-regressions, and indirect meta-analyses because of the difficulty of determining whether or not data from new trials would alter previous quantitative results. Studies reporting standardized effect sizes were excluded to avoid the complexity of assessing new data reported using various different outcome scales, because it would be problematic to determine which, if any, outcome would have been one that would have been incorporated into the standardized effect measure.

The systematic review must have included at least one randomized controlled trial, in keeping with ACP Journal Club criteria, but other included studies in the qualifying meta-analyses were restricted to quasi-randomized or controlled clinical trials.

The cohort was limited to no more than 30 Cochrane Collaboration Reviews, as these have been found to differ in many ways from other systematic reviews in the peer reviewed literature on the basis of style and possibly on topic coverage.[3,126] A sample size of 100 systematic reviews was established for the survival analysis of systematic review. This was thought to be large enough to evaluate up to five predictors of survival, but also reflect the practical resource limitation. A subset of the explicit updates was established to test the experimental search methods against the search conducted for the review. This subset and the AHRQ cohort described below were used to test the performance of Support Vector Machine (SVM). Several additional inclusion criteria were imposed for

that purpose, and the reviews with explicit updates were screened against those additional criteria.

4.1.2.2 Updated Cochrane Cohort

All Cochrane Collaboration Reviews found eligible for the main cohort (n=27) were assessed for eligibility against the following additional criteria. 1) The text of a review and an updated version was available. 2) The search for the original review was comprehensive (defined, following the Oxman and Guyatt quality assessment scale for meta-analyses, as a search of MEDLINE and at least one other electronic bibliographic database, and one or more non-database method such as hand searching or checking reference lists[150]) and therefore likely to identify most relevant studies. This was important for three reasons. First, if the updated systematic review missed important new evidence, than the denominator for recall would be underestimated and this could inflate the apparent recall of the test searches. Second, precision estimates for the new searches could be deflated because the numerator would be underestimated if the test searches identified relevant new material that was missed by the original reviewers. Both recall and precision inaccuracies might differentially disadvantage the test search methods least like those used by the systematic reviewers in performing The Cochrane Collaboration updates or the AHRQ searches. Third, SVM requires a training set of true positive examples, and a more complete search should provide a better training set. 3) The MEDLINE search of the original review and the update were reported in enough detail that an information specialist experienced in systematic review searching could replicate it. This permitted re-creation of the true negative set, and determination of the retrieval size of the author's search. 4) The search dates for MEDLINE were reported so the update interval covered by the review could be established. 5) A list of included studies was presented in the original review and the update so that true positives could be identified. 6) At least ten RCTs or quasi-RCTs must have been included in the original review. This requirement was to give an adequate true positive training set for SVM. Finally, 7) at least 67% of papers included in the original review had to be retained in the

update. It was also designed to exclude reviews where the inclusion criteria had changed substantially between the original and the update. Cochrane Collaboration Reviews were selected as they were the majority of reviews that had updates, and they were most likely to report all necessary information.

4.1.2.3 AHRQ Cohort

A cohort of ten Evidence Reviews published by the AHRQ was selected. AHRQ sponsored the updating research, and agency officials were interested to see how well the methods developed for determining need of update for journal and Cochrane Collaboration Reviews would work when applied to the type of systematic review published by the AHRQ. Based on an examination of characteristics of AHRQ Evidence Reports prepared for a previous proposal for methodological research,[151] AHRQ reports more closely resemble health technology assessment reports than conventional systematic reviews of treatment interventions in that they tend to be multi-faceted, including issues of diagnosis and prognosis, not just treatment issues. The same eligibility criteria were applied to the selection of AHRQ reports that were used to select the main cohort except that coverage in ACP Journal Club was not required. We also applied the additional eligibility criteria associated with the updated Cochrane cohort except that the AHRQ reports did not require an update.

4.1.3 Search Strategy

Reviews were identified through a search of the ACP Journal Club database on Ovid,[152] undertaken January 31, 2006. The database contained articles from ACP Journal Club and Evidence Based Medicine published from 1991 to November/December 2005. We sought systematic review and meta-analyses using a search targeted at the standardized titles and abstracts used by ACP Journal Club.[153] The search to identify candidates screened for inclusion in the cohort was:

1. review$.ti.
2. meta-analy$.mp.
3. data sources.ab.

4. (search$ or medline).ab.
5. or/1-4
6. limit 5 to articles with commentary

4.1.4 Cohort Selection Process

Records retrieved by the search were downloaded into Reference Manager citation management software,[154] then uploaded to SRS™[155] a web-based platform for systematic reviews. Eligibility assessments were conducted by physicians with experience in clinical epidemiology. Records were screened in alphabetical order by first author until 100 eligible reviews (with a maximum of 30 Cochrane Collaboration Reviews) were identified.

Two reviewers screened each record for eligibility based on title, abstract and indexing terms. Records with consensus in favour of eligibility were promoted and a second assessment was made based on the full report, to confirm eligibility. Once screening was completed, a review of the cohort was made to ensure that only one systematic review of a particular topic was included, to avoid double counting the same changes in evidence (or lack thereof). Reviews were considered to overlap when they had a common population and intervention. When an eligible review was identified as an explicit update of an earlier review (e.g., in the case of Cochrane Collaboration Reviews, which are updated and reissued periodically as a matter of policy), we used the earliest version in the time frame of 1995-2005, even if this was not the version reviewed by ACP Journal Club. We used the version of that review in The Cochrane Library, Issue 1 2006 as the most current version. If the review had been withdrawn, we used the final version prior to withdrawal. The Effective Practice and Organization of Care review group kindly provided access to previous issues of The Cochrane Library.

When more than one review on the same topic was identified among journal published reviews, only the earliest was included. Subsequent reviews on the same topic were retained as an aid to identifying newer evidence, but were excluded from the cohort.

4.2 TEST SEARCHES: IDENTIFICATION OF NEW EVIDENCE

A necessary part of the process of determining if a cohort review was out of date was to find new relevant evidence. It was determined *a priori* that this would be used as an opportunity to evaluate a number of alternate approaches to searching. This section describes the test searches which from the core of this research.

Three broad approaches to information retrieval are Boolean searches, similarity searches and citation searches. Boolean searches are typically used in systematic reviews. These are searches constructed of search strings and use operators (i.e., AND, OR, NOT) to express the search logic. I have recently reported on several variants of this approach.(3) The second type is a similarity search which works from known examples to find new relevant material. Citation-based methods identify material citing or cited by relevant examples. Checking reference lists is an example of citation searching used in systematic reviews.[156]

Currently, the approach to searching when updating systematic reviews has been to repeat the search methods of the original review. This using includes electronic Boolean searches of multiple bibliographic databases plus measures such as contacting experts and checking reference lists. Monitoring has largely been a matter of re-running electronic searches or considering nominations or studies found incidentally by reviewers.

Similarity searches have not been described frequently in systematic review methods. When updating a systematic review, unlike the case when performing a *de novo* review, a comprehensive collection of relevant studies is available as a basis for similarity searching. Here, I develop and test two variants of similarity searching, PubMed Related articles restricted to Randomized Controlled Trials (RI RCT) and Support Vector Machine.

4.2.1 Boolean Searches

4.2.1.1 Clinical Query

The first subject search variant involves the *Clinical Query,* for high quality clinical studies. This query was developed by the Hedges team, based at McMaster University, Hamilton, Canada, who have developed and validated a number of search strategies to find high quality studies to answer clinical questions related to treatment, prognosis, harms and aetiology and numerous other topics for MEDLINE and other databases.[157] This was chosen because it is developed specifically to identify high quality studies, and these may represent the best evidence for inclusion in systematic reviews. The filters are designed to find methodologically sound primary studies quickly,[14] although one recent case report found it had the same recall but lower precision than the original Highly Sensitive Search Strategy.[158] This study included only systematic reviews of treatment interventions, but filters created and validated using the same methodology exist for other studies types, such as diagnostic studies[159] and health services research,[160] thus making the methods adaptable for types of systematic reviews other than intervention effectiveness.

4.2.1.2 AIM RCT

The second subject search focused on randomized controlled trials published in journals included in MEDLINE's Core Clinical Journals subset.[13] This journal subset was formerly known as Abridged Index Medicus, and this approach will be referred to as AIM RCT. Tsay and Yang examined the bibliometric characteristics of randomized controlled trials and found that of the 42 journals comprising the first zone of the Bradford distribution (i.e. the small number of journals in which contain the nucleus of the literature) 25 of these were NLM "core clinical journals"[161] Searching this journal subset potentially provides an efficient way to find the papers most influential in overturning the findings of a systematic review.

4.2.1.3 CENTRAL

The final subject searching approach was to use the Cochrane Central Register of Controlled Trials (CENTRAL) to identify new studies of potential relevance. CENTRAL was created and is maintained with the object of being as complete a source for identifying new trials as possible.[41]

4.2.2 Citation Searches

The above Boolean approaches based on subject searches require expertise to create the search strategy for each review. Knowledge of searching techniques is required, and the person undertaking the search must form a good understanding of the topic of the review. The additional methods tested here are algorithmic and not dependent on knowledge of the field or interpretation of the question, an observation made by Pao and Lee in regards to citation searching.[162]

4.2.2.1 Citing RCTs

Citing reference searching, where papers citing the systematic review itself are retrieved, could be used it identify relevant new material. Patsopoulos *et al.*[163,163] have recently demonstrated that meta-analyses received more citations than any other study design both in the first two years and in the longer term. Using citing references to identify related material may have greater utility in monitoring for newly emerging relevant literature related to a particular meta-analysis given the propensity to cite them. It remains to be seen whether it is a useful mechanism to identify subsequent RCTs. Citation of systematic reviews may become more common as journals and some granting agencies introduce formal requirements that prior literature be systematically considered.[164] If citations of systematic reviews follow the same citation patterns as RCTs,[140] highly cited systematic reviews may be more likely to be refuted through subsequent research, thus more likely to be in need of update than less highly cited systematic reviews. Thus, even if the performance metrics of citation tracking indicate

poor performance as a surveillance tool, a high citation rate may indicate that closer surveillance through other methods is warranted.

Bernstram *et al.*, in experimental work, found citation-based algorithms useful in improving MEDLINE recall, finding them superior to non-citation-based methods.[165] An advantage over Clinical Queries remained, even after citation lag was introduced experimentally. The authors noted however, that this is not the optimal approach for those seeking the most recent information available, as is the case in systematic reviews. They also note that not all important relevant items are cited by related works, so such an approach is necessarily incomplete.[165]

In other work, Pao and Lee found that recent seed articles will not have had time to be cited often, while those published more than 12 years prior seemed to produce little, so there may be a window of ten years or so where the method is productive.[162] Algorithmic approaches may also be useful in areas that are difficult to search, such as cross-disciplinary or emerging topics.[166]

4.2.3 Similarity Searches

4.2.3.1 Related Article RCT

PubMed's *Related Articles* feature provides an avenue to identify potentially relevant material through query-by-example. It may be possible to identify the core literature by starting with known relevant items and searching for related items (query expansion). Some advantages of such a method include that it is relatively quick, can be replicated, is not dependent on searcher skill, and relies only on access to the published review and PubMed. Some disadvantages are that its performance characteristics (precision and recall) have not been formally demonstrated and it may not be possible to re-run the searches automatically at intervals, a characteristic that is desirable in a surveillance method.

The PubMed *Related Article* function retrieves pre-computed nearest neighbours. Terms, including weighted free text and index terms, are given a local weight based on

frequency in the document, adjusted for document length, and a global weight based on frequency in the corpus. Near neighbours are bibliographic records with high scores when all the terms common to both records are considered.[167] Lin *et al.* have described the inner workings of the related article algorithm, PMRA, describing it as a topic-based content similarity model.[168,169] They found that the related article search was a well-used feature of PubMed. Based on query logs gathered during a one-week period in June 2007 roughly a fifth of all non-trivial PubMed user sessions contain at least one related article search.[168]

In previous work, Lui and Altman used the *Related Articles* feature to incrementally update a bibliography, achieving recall of 75%, a strict precision of 32% and a partial precision of 42%.[170] Bernstram show preliminary evidence that better results can be obtained by using multiple seed articles and using a combination of nearest neighbour sets.[165]

The approach used here, which will be to select a few of the included studies as seed articles for Related Articles searching, has an analogue counterpart in pearl-growing where the indexing of relevant articles informs the search strategy. White explored taking a number of citations to a relevant article that jointly express the topic of interest, and using these as seed to generate descriptor terms for searching.[171] Schlosser *et al.*[172] describe the use of pearl-growing in developing a Boolean search strategy for systematic reviews. They advocate a comprehensive approach where a set of relevant articles is assembled, and these are searched in each database under consideration to determining relevant indexing terms.

4.2.3.2 Support Vector Machine

The final approach used here and potentially useful for surveillance of emerging evidence is support vector machine (SVM), a machine learning technique or classifier that operates by finding an algorithm that defines a "vector space" distinguishing True Positives from True Negatives in a training set. That algorithm was applied to classify new examples.[173] Here, I used the search result from the original review as the training

set, with True Positives being the studies included in the review (the relevant retrievals) and True Negatives being the studies found by the authors' search but excluded from the review (irrelevant retrievals). The new examples to be classified are articles published since the first search. Those classified as Positives by the SVM are retrieved.

Pavlidis and Wapinski argued that the strength and widespread applicability of the SVM are due to a number of advantages it holds compared to other machine learning techniques. These include a strong theoretical foundation, scalability, making it suitable for use with large databases (such as MEDLINE), flexibility, including the ability to incorporate prior knowledge, and finally its accuracy.[174]

Machine learning approaches have been explored in systematic reviews. Cohen tried to explain variations in performance across the topics by reference to the number of features without great success. They may have overlooked one important correlation present in their results; the correlation between precision of the initial query and the precision of the neural network was 0.911. That is, the searches that were easier for the librarian were also easier for the neural network. They do acknowledge that there would have to be something non-obvious about the retrieval task for the classifier to work from in order for it to out-perform the human query builder. SVM may have an advantage over the neural network approach used by Cohen in that it can incorporate a much larger number of features. Cohen also used a very high threshold for recall. Performance may be better in a context where the objective is literature surveillance to detect a signal to update, rather than when the objective is exhaustive identification of evidence. Still, at the high recall threshold used, they concluded that machine learning did bring efficiencies to the updating process. It may be that using machine learning for surveillance can introduce efficiencies by focusing updating efforts on those reviews most likely to be out of date, a fundamentally different approach.

O'Blenis *et al.*[175] reported exploratory work using Naïve Bayes Classifier algorithm to take over the classification of results once reviewers have screened enough records to provide a training set. This classifier also returns a ranked result set that allows

the documents to be considered in order of likely relevance. Considering documents in order of relevance is the most efficient way to monitor emerging literature for an accumulation of new evidence that would render a SR in need of update. Cohen identified that in the actual update, such relevance ranking is not of great interest.[15] They considered the possibility of using "Really Simple Syndication" (RSS) feeds to notify investigators of new material, stating "these notifications would be useful both for focusing the reviewer's time on the most likely articles to include high quality topic-specific evidence, as well as to alert the review team when a sufficient amount of new evidence had accumulated to warrant an update of a given drug class review."[15]

4.2.4 Methods for Executing Each Search

4.2.4.1 Clinical Query, AIM RCT, and MA Searches

The Boolean approaches require knowledge of search techniques. A structured approach was taken to developing these queries in order to reduce operator dependence and maximize the generalizability of results. Jessie McGowan, a senior librarian with extensive experience in systematic review searches completed the first ten systematic reviews needing update searches. Then, working with Tamara Rader who executed the remaining searches, the following step-by-step instructions for developing and executing these searches in the Ovid MEDLINE interface were developed. Part one involves analyzing the topic and searching the subject area (Figure 3).

Part 1

1. Select article to be updated and read the abstract
2. Identify key words for the condition, population (if it's children or elderly for example) and for the intervention
3. Read the methods section to determine when the search was performed. If no date is given, search from the year before the date of publication.
4. Open OVID MEDLINE
5. Enter key words for the condition and determine the most appropriate MeSH heading. If a suitable MeSH heading is found, then there is no need to combine with natural language keywords.
6. Enter key words for the intervention and determine the most appropriate MeSH heading. If a suitable MeSH heading is found, there is no need to combine keywords.
7. If necessary combine using "OR" the keywords & MeSH headings you have selected for the intervention & condition.

8. Then combine the 2 concepts (Intervention AND Condition)

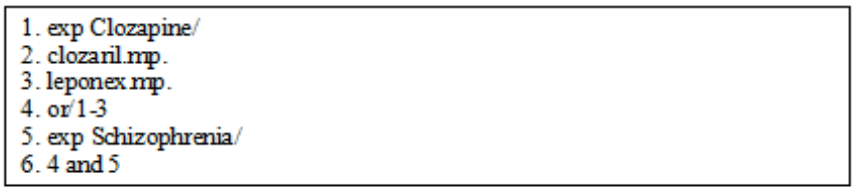


Figure 3. Example of Subject Search for CohortID 150

Part 2

Part 2 involves limiting the Subject Search to type of study design using pre-determined OVID "limits" – here is where the differences for CQ, AIM RCT and MA searches are specified. Naming conventions and download instructions specified in Part 2 reflect the record keeping steps needed for this project.

9. Limit the search to the year you noted earlier. Either the year the search was done, or the year before the paper was published.
10. To limit your results to RCT-AIM results, click "More Limits"
11. Select the line of your subject search limited by year
12. Check the box marked AIM Abridged Index Medicus
13. Scroll down until you get to Publication Types
14. Select Randomized Controlled Trials
15. Click "Limit Search"

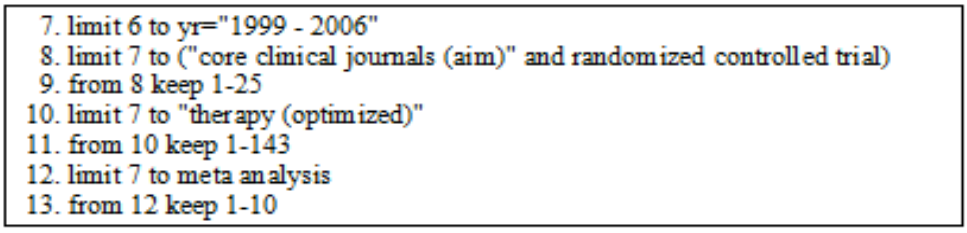


Figure 4. Example of the Limits for CohortID 150

16. Save the search results by selecting options at the bottom of the screen called Results Manager.
17. In the first column, select the option: "All in this set"
18. In the second column, click the red box "Select Fields"
19. DE-select all the fields EXCEPT for UI – unique identifier
20. In the third column select: Brief (Titles) Display and Include Search History
21. In the Sort Keys section for Primary, choose "year of publication" and "descending" and in Secondary, choose "Entry Date" and "descending"
22. Click Save
23. The file name should reflect the cohort number, the method of search and the number of results for example: 151 RCT AIM (25).txt
24. Return to Main Search Page
25. To limit your results to CQ, click "More Limits"

26. Select the line of your subject search limited by year
27. Scroll down until you get to Clinical Queries
28. Select Therapy (Optimized)
29. Click “Limit Search”
30. Perform steps 16-22
31. Again, the file name should reflect the cohort number, the method of search and the number of results for example: 151 CQ (143).txt
32. Return to Main Search Page
33. To limit your results to MA (Meta-analysis), click “More Limits”
34. Select the line of your subject search limited by year
35. Scroll down until you get to Publication Types
36. Select Meta-analysis
37. Click “Limit Search”
38. Perform steps 16-22
39. Again, the file name should reflect the cohort number, the method of search and the number of results for example: 151 MA (10).txt
40. Return to the main search page
41. Now you are done the subject searching. TO SAVE the search click “Save Search/alert”
42. You will be prompted for your user name: “epcupdating” and password “xxxxxx”
43. Name the search SS 151 and select “permanent”
44. Select save
45. Return to main page and delete the set to begin a new search.

4.2.4.2 Subject Search CENTRAL, inCENTRAL

The CENTRAL database (The Cochrane Library, Issue 3, 2006) was searched through the Ovid interface. The subject search developed for each review was used (Figure 3), with the following adjustments to compensate for difference in operation of MeSH term searches between Ovid MEDLINE and Ovid CENTRAL, which were apparent from the pilot. Starred MeSH headings (the MeSH headings representing the major focus of the article) were converted to unstarred terms, as starring provided inconsistent results particularly for exploded subject headings. ‘‘Exp Antibacterial agents/’’ was replaced with ‘‘exp Antibiotics/.’’ CENTRAL is a secondary database, where trials are identified from other databases or from handsearch (as examples) thus there could be a lag of years between publication and identification and indexing in CENTRAL. CENTRAL does not allow limiting by record entry date. Not all records have MEDLINE indexing terms. Therefore, we tested CENTRAL only for cohort reviews where a signal for updating occurred to provide as basis for comparison of retrieval size

when search methods are held constant, and to get an indication of how complete the indexing of the newer relevant studies was, without correcting for indexing lag. CENTRAL records were not added to the screening pool, but were merged into the big database so that extent of coverage of the relevant material could be determined.

4.2.4.3 Citing RCT Searches

Several sources index citing references, including ISI Science Citation Index, Scopus and Ovid. Scopus and Ovid are relatively new providers of this service. ISI Science Citation Index is well established. One research report indicated that Scopus yielded more citing references than ISI in health science,[176] our experience matched this. We piloted the search for citing references in Scopus by using the first 50 systematic reviews that were found eligible for the cohort and then selected four for testing in the other databases, selecting two highly cited systematic reviews and two cited less often. In most cases, Scopus provided more cited references than ISI for journal published systematic reviews (Table 2). Ovid's indexing of citing references appeared quite incomplete. Scopus was therefore used as the citing reference source for journal-published systematic reviews.

Table 2. Citations of Journal-Published Reviews by Indexing Source

		Scopus		Ovid		ISI	
CohortID	Y	Year	N	Year	N	Year	N
6	2000	2000	7	2000	11		
		2005	1	2005	11	2005	0
13	1999	2000	31	2000	26	Not found	
		2001	1	2001	26		
		2002	1	2002	26		
120	2001	2000	9	2000	23	Not found	
		2001	5	2001	23		
		2004	3	2004	23		
264	1999	2000	1	-	-	Not found	
		2001	9	2001	11		
		2002	1	2002	11		
		2004	2	2004	11		

The pilot identified an issue with the indexing of citations in Scopus for Cochrane Collaboration Reviews. Cochrane Collaboration Reviews are re-issued with each quarterly issue of The Cochrane Library. Scopus appeared to count citations to issues from different years as though they were distinct systematic reviews, with the number cumulating only annually, but not across years. Subsequent testing showed that Ovid indexes cumulatively. ISI did not routinely index Cochrane reviews at the time this work was undertaken (Table 2). Since the reason for identifying citing references was to identify subsequent randomized controlled trials and systematic reviews on the same topic, a citation to any issue of the review was of interest. The Ovid interface was therefore selected as the citing source for Cochrane Collaboration Reviews, as it gave more integrated coverage.

Having selected the interface for Cochrane and Journal-published reviews, the actual identification of citing references was done by looking up the index review in the citation database (Scopus or Ovid) and then searching the citing references in PubMed to determine which of the citing reference was an RCT. I undertook the pilot searches and Raymond Daniel, a library technician, completed the remaining citing RCT searches. The PubMed IDs for all citing RCTs for that review were downloaded and later integrated into the corresponding All.xls file.

4.2.4.4 Related Articles RCT and Related Articles MA Searches

Originally, I intended that the seed articles should be those identified by one of the systematic reviewers knowledgeable about field as the "pivotal studies" from the review to be updated. As the method was otherwise quite mechanical, requiring no specialized knowledge, I decided instead to select seed articles using quantitative criteria that could be easily applied. The following criteria were standardized during a pilot with ten systematic reviews. The three largest and three newest included studies were selected. Where these overlapped, they were not replaced. So if two of the three newest studies were also among the largest, only four studies would be used. Where multiple reports of

the same study were included in a systematic review, only one was used, and the one that appeared to be the main report was selected as the seed. The selected seed articles were identified in PubMed. If one of the seed articles was not indexed in PubMed, it was dropped and not replaced. Relevance was not assessed in the pilot, only the viability of the method, in the sense that the seeds could be readily identified, and that the number of seeds generated a reasonable number of results (Table 3). As the method did appear viable, it was adopted for this project. I undertook the ten Related Article pilot searches. Raymond Daniel, a library technician, completed the remaining Related Article searches using an established protocol, initially under supervision and then independently.

Table 3. Related Article Pilot Search Results

Cohort	Search start date	N of productive seeds	N of RCTs	N of MAs
31	1997/01/01	5	177	10
38	1995/01/01	6	200	10
40	1998/01/01	5	183	7
43	1998/02/01	3	26	8
50	2003/02/01	4	84	10
56	1998/01/01	3	86	7
70	1996/10/01	5	222	3
94	2004/01/01	5	47	1
96	1999/12/01	4	48	7
104	1994/01/01	5	132	8

Seed items were identified from the screening worksheet (Figure 3). In the example shown, the three newest studies are PubMed IDs 9857355, 10782589 and 11427285. The three largest studies are 6755247, 6233921 and 9651632. The PubMed IDs for seeds were entered into the PubMed search box, and retrieved. The Related Article Display was selected. (Figure 5).

Figure 5. Selecting the Related Article display for the six seed articles

The 1156 related articles are returned, sorted by relevance, with the seed articles appearing highest in the list. A null search is set up for records with a publication type of Randomized Controlled Trial and a PubMed entry date after the search date of the original review, which is December 15, 1994 in this case. A null search consisting only of limits, with no entry in the search box, is needed as a related article search cannot be limited directly. In the History tab, the related article search and limit are combined (Figure 6). The number of hits was noted and the records were downloaded. The limits were then change to the search date used plus one year, and publication type of Meta-Analysis. The number of hits was noted and the PubMed IDs were copied from the UI List display, and saved to a text file for later integration into the All.xls files.

Figure 6. Application of Date and Study Design Limits to the Related Article Set

As this method did not provide access to the relevance scores, the searches were repeated later in the project using the eLink utility, a different search interface than that provided at the main PubMed search screen. Elink is one of a number of Entrez programming utilities that are designed to retrieve search results for use in subsequent applications. The PubMed eLink utility allows searching of nearest neighbours to PubMed IDs submitted as a URL query. The query for CohortID 497 would be:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11881837,12205648,12409683,12480794,10881251,10325453&term=Randomized Controlled Trial[ptyp]&cmd=neighbor_score

Search Parameters, many of which are optional, include Database (db), Record Identifier (id, here PubMed ID), Relative Date or Date Range, Date Type (edat=database entry date, mdat=date of last modification or pdat=publication date) and one Search Limit (term). A single search limit can be applied, in this case Randomized Controlled Trial[ptyp] was applied to restrict the retrieval to nearest neighbours that were RCTs. As MHDA was not available as a date type in eLink, no date limit was used. The results are returned in XML (Figure 7).

Figure 7. XML Output from eLink Related Article Search

The XML was pasted into a Microsoft Word document where the global search-and-replace function was used to convert field separators to tabs. The tab-separated result was then pasted into a spreadsheet, which resulted in each field becoming a spreadsheet column. Irrelevant fields were removed by deleting those columns, until only the PubMed ID and similarity score remained. A column representing the CohortID was added. These were then pasted into an SPSS file, and the SPSS function Merge Files / Add Variables was used to add the similarity scores to the big database. Because MHDA was not available through the eLink interface, records from this related article query that did not match the PubMed ID of an existing Related Article RCT retrieval were ignored.

4.2.4.5 Support Vector Machine

For each systematic review in the set of six updated Cochrane Collaboration

Reviews and ten AHRQ Evidence Reports selected for study, I replicated the reviewer's original MEDLINE search strategy if it was in Ovid, or translated it to Ovid in order to derive the true positive and true negative training sets (Figure 8). Jessie McGowan verified the replications. True positives were those MEDLINE indexed records that represented the included studies of the original review, regardless of whether they were found by the authors' search of MEDLINE. The true negative training set included those MEDLINE records that would have been retrieved by the authors' search at the time they executed it, but which were not included in the reviews. These were records that would have been screened and found ineligible by the reviewers. For the updated Cochrane Collaboration Reviews, targets were predetermined to be those studies included in the updated review that were not in the original, and which entered MEDLINE after the search for the original review. The full retrieval set, at least to the point where the review was found to be in need of update was reviewed for the AHRQ cohort.

I prepared all material for the SVM searches, which were executed by Berry de Bruijn, a Research Officer with the Interactive Information program of the Canada Institute for Information Technology (CISTI), part of the National Research Council.

1. exp Pharyngitis/
2. pharyngit$.mp.
3. exp Tonsillitis/
4. tonsillit$.mp.
5. (Sore adj throat$).mp.
6. or/1-5
7. exp antibiotics/
8. antibiot$.mp.
9. or/7-8
10. and/6,9
11. RANDOMIZED CONTROLLED TRIAL.pt.
12. CONTROLLED CLINICAL TRIAL.pt.
13. RANDOMIZED CONTROLLED TRIALS.sh.
14. RANDOM ALLOCATION.sh.
15. DOUBLE BLIND METHOD.sh.
16. SINGLE-BLIND METHOD.sh.
17. or/11-16
18. (ANIMALS not HUMAN).sh.
19. 17 not 18
20. CLINICAL TRIAL.pt.
21. exp CLINICAL TRIALS/
22. (clin$ adj25 trial$).ti,ab.
23. ((singl$ or doubl$ or trebl$ or tripl$) adj25 (blind$ or mask$)).ti,ab.
24. PLACEBOS.sh.

25. placebo$.ti,ab.
26. random$.ti,ab.
27. RESEARCH DESIGN.sh.
28. or/20-27
29. 28 not 18
30. 29 not 19
31. 19 or 30
32. 10 and 31
33. remove duplicates from 32
34. ("3057159" or "1905799" or "8945796" or "9219402" or "9270458" or "9116551" or "9051558" or "10634735").an.
35. 34 not 33
36. (1996$ or 1997$ or 1998$ or 1999$ or 200$ or 2001$ or 2002$ or 200301$ or 200302$ or 200303$ or 200304$ or 2003050$).ed.
37. 36 and 33
38. 37 not 34

Figure 8. Replicated Search Strategy.

Lines 1-32 is the replicated search. Line 34 gathers the MEDLINE-indexed included studies and is the true positive set. Line 35 identified MEDLINE misses. Line 37 limits replication of the authors' search to the dates covered by the authors' search. Line 38 isolates the material the authors' would have screened and excluded – it is the true negative set. The study is Del Mar 1997.[177]

SVM was run on MEDLINE hosted locally at the National Research Council NRC. MEDLINE was refreshed prior to running the searches. The approach was piloted with one AHRQ report, CohortID 91 and various configurations were tried, observing the placement of targets within the retrieval. In initial testing, a fine-grained filter consisting of TP vs. TN was then applied, both with and within MeSH headings. This reduced the set size, resulting in a few targets being missed, but targets appeared to rank well within the larger set. Including MeSH at that stage influenced rankings only, rankings presented here were performed without MeSH. Partial output from AHRQ CohortID 91 is shown in Figure 9 .

12426518	0.9966345513440674
12543892	0.9966021216946064
11952508	0.9938684975082074
11816513	0.9936448066127455
15051778	0.9920208569565627
12546524	0.9919643299087394
12480972	0.9917385507389284
11273875	0.9913514482614061
12087568	0.9905956782953245
12460041	0.9894039400806446
12782431	0.9892496149487965
12566906	0.9888238472359361
14993489	0.9884684747577197
12372943	0.9880165229049765

11287776	0.9876216701378340
12427142	0.9872590452040522
15083958	0.9866843725189881
11918762	0.9864536532034286
11814098	0.9860111210055162
12091844	0.9857550204709897
12942577	0.9850231752050475
11676354	0.9839380986250016
14733416	0.9828451302897062
12612947	**0.9820890963490960**
12455724	0.9820751418793477
14514745	0.982030710854638
12631092	0.9813900934974515

Figure 9. Partial SVM Output For CohortID 91.
A target shown in bold in position 24 of 54449 ranked records

First, the MEDLINE set was filtered with the revised HSSS,[21] yielding 105,951 records. Preliminary testing indicated that this filter would retrieve 99.1% of MEDLINE indexed relevant new evidence from the cohort of 100. This filter was added primarily to improve processing speed, relative to running the classifier against all of MEDLINE. With the filter, processing for one systematic review took about ten minutes. Next, a coarse grained filter was applied; TP and TN were combined in a single set labeled POS. Adjacent PubMed IDs (the next higher PubMed ID to each member of POS, as long as that higher number was not itself a member of POS) were combined in a set to represent WORLD. The model was trained on POS vs. WORLD, using words and phrases from title, abstract, author names, and journal name. This model was run against the records passing the HSSS filter (n=105,951). PubMed ID and relevance scores were recorded.

4.2.5 Ranking

I have previously reported ranking of large Boolean search results with statistical search engines that work from document similarity ratings.[178] Any successful ranking approach has application in screening the update set as reviewers could screen a small set of likely candidates. O'Blenis *et al.* also report efforts at using machine learning to relevance rank large sets interactively so reviewers can screen those records most likely to be relevant first.[175]

The similarity searches, Related Article and SVM, return relevance scores. This provides searchers with the option of establishing thresholds for retrieval (e.g. a pre-

determined retrieval size, or a minimum relevance score). The utility of these rankings were analyzed two ways. First, the proportion of targets placed above certain cut-points was tested against the proportion expected if ranking was random. Simple Interactive Statistical Analysis (available at: http://www.quantitativeskills.com/sisa/distributions/binomial.htm) was used for this analysis after testing it by replicating the results of previous research.[178] Second, for RI RCT and SVM searches, the ROC curve of recall at various cut points was constructed.

4.3 PREPARATION OF MATERIAL FOR SCREENING

4.3.1 Building the Updating Spreadsheet

An updating spreadsheet was created for each systematic review in the cohort. It contained a worksheet showing studies included in the original review and new candidate studies and new meta-analyses. Meta-analytic calculator worksheets enabled the reviewers to enter meta-analytic results from the original review and add data from newer studies. Two summary sheets allowed reviewers to summarize major findings and classify results in terms of need to update, type of new evidence, and survival times.

4.3.1.1 Screening Worksheet

The worksheet of the spreadsheet that showed the included studies and candidate studies was prepared by the IS team. This worksheet was used by reviewers to record their relevance assessment of the candidate studies here, and for eligible studies, recorded the number of participants included in the trial.

Once a cohort review was confirmed to be eligible, all included studies were identified, each was looked up in Ovid MEDLINE, and if it was found, the bibliographic record was downloaded into a Reference Manager database created for that systematic review. Place holding records were added to the database for studies included in the review but not found in MEDLINE. Two bibliographies were created using specially created Reference Manager output formats.

The first bibliography was a tab-delimited file which was used as the basis of the screening worksheet used to track existing and new evidence. The file contained the unique Reference ID assigned by Reference Manager, the surname of the first author, year of publication, PubMed ID and date of MEDLINE indexing (MHDA). Columns were added to the spreadsheet to record variables not available as MEDLINE fields. The number of participants included in each study was extracted from the abstract, or if necessary, the systematic review, or if necessary, from the original article and these data were added to the spreadsheet. Additional columns in the spreadsheet were added for the CohortID number used to identify the systematic review within the project, the type of record (*Original*, except when an explicit update of the systematic review subsequently excluded the record, in which case the type was coded as *Original Only*), whether the record was on topic, whether it was eligible (both were coded as "Y" for all studies included in the original review), and a column in which the reviewers could add notes.

The second bibliography was a simple list of the PubMed IDs. These were saved as a text file which was edited to create an Ovid search string in the form:

(18788016 or 17348209 or 17209469 or 17201639 or 17195692 or 17194559 or 17187296 or 17185740).ui.

This string was used in two ways. First, I used it in the procedure that established which of the included and MEDLINE-indexed studies would have been found by the author's MEDLINE search, for those cohort reviews where that was tested. Second, I used it as the true positive set of records for SVM searches, for those systematic reviews where SVM was used.

4.3.1.2 Candidate List

New meta-analyses and primary studies were listed after the *Original* and *Original Only* studies on the screening worksheet. The construction of these lists is as follows.

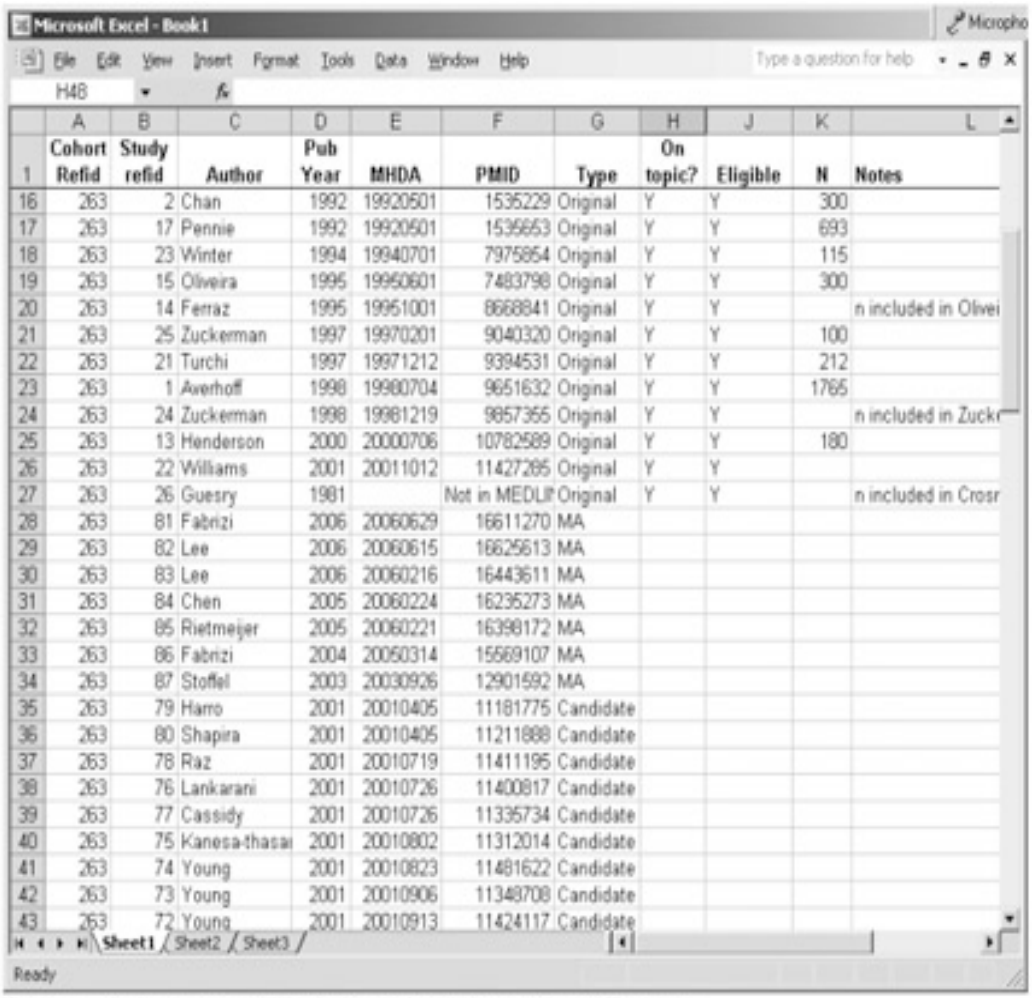
Each test search was run and results were downloaded and saved. PubMed ID numbers were isolated from the retrievals. The procedure for this varied – in PubMed, the

PubMed IDs can be displayed as a simple list. In Ovid and SVM retrievals, download included some extraneous material which had to be edited out to isolate the PubMed IDs. Once isolated, the PMIDs retrieved by each search were pasted into a spreadsheet called the All.xls file. Columns were added for the CohortID, the unique identifier for the systematic review used in the project, and a code indicating the search method that found that item.

An integration search was used to resolve PubMed IDs into a set with no duplicates. An Ovid search string was prepared from the PubMed IDs in the All.xls file, with not attempt to remove duplicates. This string was run in Ovid MEDLINE, and the search strategy was saved in the project account. The retrieval set was first limited to records with the publication type "meta-analysis", and these were downloaded in reverse chronological order. The remaining records were then downloaded, in chronological order.

A separate candidates database was created for each systematic review in the cohort and the records downloaded from the integration search were imported into it, first the newer meta-analyses, followed by the primary studies which were the candidates. From this file, a tab delimited text file was created using a custom bibliographic output style, and this file was imported into the spreadsheet that contained the basic information for the studies included in the original. Type was coded as either *MA* (meaning meta-analyses) or *Candidate* for these new records.

Neither the Reference Manager database of candidates nor the candidate list generated from it contained any information about which search methods had retrieved the record, effectively blinding the reviewers who assessed the candidates for eligibility. An example of a screening spreadsheet is shown Figure 10.



Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Help

H48

	A	B	C	D	E	F	G	H	J	K	L
1	Cohort Refid	Study refid	Author	Pub Year	MHDA	PMID	Type	On topic?	Eligible	N	Notes
16	263	2	Chan	1992	19920501	1535229	Original	Y	Y	300	
17	263	17	Pennie	1992	19920501	1535653	Original	Y	Y	693	
18	263	23	Winter	1994	19940701	7975854	Original	Y	Y	115	
19	263	15	Oliveira	1995	19950601	7483798	Original	Y	Y	300	
20	263	14	Ferraz	1995	19951001	8668841	Original	Y	Y		n included in Olivei
21	263	25	Zuckerman	1997	19970201	9040320	Original	Y	Y	100	
22	263	21	Turchi	1997	19971212	9394531	Original	Y	Y	212	
23	263	1	Averhoff	1998	19980704	9651632	Original	Y	Y	1765	
24	263	24	Zuckerman	1998	19981219	9857355	Original	Y	Y		n included in Zuckr
25	263	13	Henderson	2000	20000706	10782589	Original	Y	Y	180	
26	263	22	Williams	2001	20011012	11427285	Original	Y	Y		
27	263	26	Guesry	1981		Not in MEDLI	Original	Y	Y		n included in Crosr
28	263	81	Fabrizi	2006	20060629	16611270	MA				
29	263	82	Lee	2006	20060615	16625613	MA				
30	263	83	Lee	2006	20060216	16443611	MA				
31	263	84	Chen	2005	20060224	16235273	MA				
32	263	85	Rietmejer	2005	20060221	16398172	MA				
33	263	86	Fabrizi	2004	20050314	15569107	MA				
34	263	87	Stoffel	2003	20030926	12901592	MA				
35	263	79	Harro	2001	20010405	11181775	Candidate				
36	263	80	Shapira	2001	20010405	11211888	Candidate				
37	263	78	Raz	2001	20010719	11411195	Candidate				
38	263	76	Lankarani	2001	20010726	11400817	Candidate				
39	263	77	Cassidy	2001	20010726	11335734	Candidate				
40	263	75	Kanesa-thasai	2001	20010802	11312014	Candidate				
41	263	74	Young	2001	20010823	11481622	Candidate				
42	263	73	Young	2001	20010906	11348708	Candidate				
43	263	72	Young	2001	20010913	11424117	Candidate				

Sheet1 Sheet2 Sheet3

Ready

Figure 10. Screening Worksheet for a Systematic Review

Studies included in the original review are presented first, arranged oldest to newest and are assessed as eligible and on topic. The number of study participants (N) is extracted. For the second and subsequent report of the same study, no new N is attributed. Study refids 14, 24 and 26 are examples of this. Next, newer meta-analyses (MAs) are shown, in reverse chronological order. Finally, newer primary studies (Candidates) are shown. These are assessed for eligibility by the review, so have no initial value. N is extracted and recoded by the reviewer for the first instance of candidate studies assessed as eligible.

4.3.1.3 Meta-analytic Calculator

Once the *Include Studies* panel of the updating spreadsheet was prepared by the IS team, the meta-analytic worksheet was given to the reviewer. The reviewer also received

the Reference Manager database containing the bibliographic record with abstract for the newer MAs and the Candidates found by the test searches.

A spreadsheet template was created by Alison Jennings that allowed the reviewer to enter a meta-analytic value from a systematic review, and add new data sequentially. Separate template worksheets were designed for each of the eligible meta-analytic measures: odds ratio, relative risk, risk difference and weighted mean difference. The reviewer assigned to the systematic review selected between one and six eligible outcomes (up to four benefits and up to two harms) for assessment, and duplicated the worksheet for the appropriate summary measure as needed. The reviewer then transcribed the basic meta-analytic result for each selected outcome from the systematic review to the appropriate calculator worksheet. Data from eligible candidate studies were added, and the meta-analysis was updated after each entry. Conditional formatting alerted the reviewer when a stopping rule was met. Once the stopping rule was met for an outcome, the reviewer transcribed data to the summary page of the spreadsheet, and classified the type of new evidence and calculated the survival time. A more detailed example of the function and use of the meta-analytic calculator is provided below (Section 3.5.1).

4.4 OUTCOME MEASURES

Following Tague-Sutcliffe's pragmatics of information retrieval experiments,[179] the experimental units of this experiment are the searches of the query, where the query is the topic of the systematic review, and the different strategies for the same query are assessed on the basis of recall and precision. Recall is the proportion of relevant documents correctly retrieved, and precision is the proportion of retrieved records that are correctly retrieved. Recall and precision are used as the principle basis for evaluation of search performance in these experiments. Saracevic attributes the introduction of these measures to Kent, Berry, Leuhrs, & Perry, 1955 ([180] Citing Kent, Berry, Leuhrs, & Perry, 1955). Two persistent problems in information retrieval evaluation research are establishing relevance and establishing the denominator for calculating recall. A third issue, one that has received less discussion, is the statistical measure used to summarize

outcomes. Assessment of relevance and establishment of the denominator will be discussed here. Summary measure will be discussed with the description of statistical analysis. Finally, alternative outcome measures will be considered briefly in this section.

4.4.1 Assessment of Relevance

The main criteria that will be used to assess search performance are recall and precision. Both of these are relevance-based measures,[181] and indeed "relevance is a, if not even the, key notion in information science in general and information retrieval in particular"[180] and so the definition of relevance used in these studies needs to be discussed.

Borlund argues that relevance in multifaceted, with the consequence that there is poor consensus among information scientists as to its meaning.[182] Blair argues that it is utility that matters, and that utility can be assessed only by the originator of the query.[183] Borlund argues that situational relevance many be the most realistic type of relevance to users in most applications.[182] Situational relevance is defined by Saracevic as "usefulness in decision making" and so is nearly synonymous with utility.[180]

Topicality or "aboutness" is distinguished from relevance by utility (see Tague Sutcliffe [179]). Aboutness can be expressed through indexing and is independent from the need of any user.[180,184] Overall, there seems to be some consensus that relevance from a user perspective, rather than from a system perspective, (or what Kemp calls pertinence[185]) is subjective, is in the eye of the beholder, and is the more important outcome in information retrieval. The theory of relevance should inform system and interface design.[186]

However relevance is operationalized, it should measure what is important in the search, and the relevance judgments should be made consistently.[183] A challenge to this is the observation that, "real-life IR experiments demonstrate that a user's relevance judgment can change during a search through interacting with a search engine, viewing retrieved documents, and so forth."[187] An example in the systematic review context would be a second report, even a covert double publication[188] of the same randomized controlled

trial. The first and second report may be equally eligible, but having found the first report, the utility of the second report is low. van der Weide and van Bommel explore this through the concept of the incremental information value of documents, which declines when previous similar documents have be obtained.[184] This topicality can be absolute, but relevance is not. This leads to Ellis' assertion of a paradox of relevance‡‡ "The more one uses the "real" relevance, the less one can measure it."[185]

4.4.1.1 Partial and Full Relevance

We assessed both topicality and utility. Topicality was the assessment by the reviewer that the retrieval was "*On Topic*" – that is, should probably have been retrieved by a search for that systematic review. *On Topic* is more of a system-side measure, the extent to which records that were retrieved ought to have been retrieved and are not "false drops" even thought they may not be fully relevant. Utility was operationalised as "*Eligible*" – the retrieved item was both *On Topic* and could be included in the analysis of one of the six chosen outcomes. *Eligible* was a narrower standard than *On Topic*, and all *Eligible* records were assumed to also be *On Topic. Eligible* also reflects a more user-side approach. This relevance standard of *Eligible* could also be narrower than the inclusion criteria for the original systematic review. A report on an outcome not eligible or not chosen for the updating exercise or presenting data in a form that could not be included in the meta-analysis of one of the selected outcomes could be fully *Eligible* for the review yet rated as *On Topic* but not *Eligible* in this study. Finally, to address the issue of second and subsequent reports of the same trial where the later reports are relevance but not useful, reviewers assessed such reports as *On Topic*, and if they reported on the same selected outcomes, as *Eligible*, but no new participants were attributed to them.

A still narrower definition of relevance would be Hersh's outcome-oriented approach where only the material that ultimately has an impact on outcomes is

‡‡ Also attributed to Saracevic 1975 by Hersh[189]

relevant.[189] In this research, this would be operationalized as those new studies that actually rendered the systematic review in need of update.

The designation of *On Topic* corresponds well to aboutness, at least conceptually. The designation of *Eligible* is a subset of the useful articles; they are useful in a particular sense of bringing new information to the meta-analysis. The variable "N" quantifies the new information, and could be zero where that new information is not novel.

There can be articles within the candidate list that are useful but not *Eligible*. These articles could provide important background information, inform on methods or attune the reviewer to controversies or doubts about the treatment under study, or as Kagolovsky describes it, provide important prompts to the user.[45] Saracevic notes that "people may and do derive relevance from ideas and clues in articles that no system could readily recognize, at least as yet. But, that depends also on domain expertise."[180] White reviews the sorts of hidden associations that document users perceive that make seemingly disparate works relevant and lead to knowledge discovery, and suggests that co-citation is more apt than indexing terms to capture this sort of relevance.[186]

4.4.2 Establishing the Denominator for Recall

Relevance is a challenge for both precision and recall, but recall faces the additional challenge of establishing the denominator. The denominator for the calculation of recall is the total number of relevant studies in the dataset. Tague-Sutcliffe[179] suggests five possible approaches to this task; 1) predetermine the relevant set in some way, 2) use a small test set, 3) assess a random sample of unretrieved items, 4) use relative recall, 5) use a very broad search. Each of these will be considered in slightly more detail and from the perspective of systematic review work.

1. The relevant set is predetermined in some way. Tague-Sutcliffe gives the example of using the title of a paper as the basis for a query, and the cited documents as the relevant set. This is very similar to the approach used with updated Cochrane reviews in this experiment – it was assumed that the reviewers found and included (and so cited) the newer relevant documents.

Tague-Sutcliffe notes that this approach is arbitrary, and there may be other documents in the database that are also relevant. To guard against this, additional inclusion criteria were imposed to – the searches must have been comprehensive and a reproducible MEDLINE search strategy must have been presented. In addition, the possibility that additional relevant documents went undetected will be explored with the updated Cochrane Collaboration Rreviews through capture-recapture methods described later.

2. A small test set is used, and all documents are assessed for relevance. Tague-Sutcliffe notes that this may not be a reliable guide to results with a larger file. This approach is not relevant for these investigations.

3. A random sample is taken of non-retrieved set and all documents in the sample are assessed. Tague-Sutcliffe notes that the low base rate of relevant documents in most databases renders the necessary sample size impractical. An approach used to overcome this problem in the epidemiological literature, sequential screening, could be applied to an information retrieval problem, and has parallels with systematic review screening process.[190] In sequential screening, a preliminary screening instrument with high sensitivity but low specificity identifies a high-risk subgroup, which is evaluated more rigorously. By raising the base rate of the second sample, screening efficiency is enhanced. This is similar to the two stage screening common in systematic reviews, where the retrieval is biased toward recall (sensitivity), with low precision (positive predictive value), and a rapid screening is done using readily available titles and abstracts. In the second phase, the complete document is assessed, a more expensive test, but precision of this second phase is greatly enhanced by increasing the base rate in the second sample. Derogatis provides a clear and concise description of the rationale for sequential testing.[190] An informative example can be seen in Benn's simulations of various sequential testing strategies for prenatal Down syndrome screening, where the false positive and

false negative consequences are marked, the base rate is low, and tests vary in their performance characteristics, risks and invasiveness.[191] While this approach has some parallels with the systematic review process and is potentially useful in establishing a denominator, it was not used in this research.

4. In comparative tests, relative recall is calculated, where the denominator includes all relevant documents found by any method. Tague-Sutcliffe notes that this method work well in comparative tests, but results may not generalize to other databases or other query sets. It has previously been argued that this method is a useful solution in the case of systematic reviews with comprehensive searches, and approximates the recall seen with a gold standard based on hand searching.[192] It is the main method used with the 77 systematic reviews that were updated with additional searching. The recall of each test search has as the denominator the relevant records identified by any of the test searches or included in newer systematic reviews on the same or similar topics.

5. Finally, in a large database, use a very broad search, or carry out a complete review of subsets known to be rich in relevant items, and use the total number of relevant items found as the denominator. This is the approach used in much search validation work, where the results of a hand search are used as the "gold standard' and to form the denominator.[193]

In summary, the main approaches used in these experiments to ensuring a complete denominator for recall is the predetermined relevant set, used for the updated Cochrane reviews, and relative recall for those reviews updated by screening newer systematic reviews and candidates identified by the test searches. Achieving perfect recall is a trivial task, achieved by simply retrieving every record in the database. Useful recall trades off against something, and complementary measures to recall are precision, fallout and specificity.

4.4.3 Recall of New Studies and Recall of New N

As well as assessing recall of new studies, recall of new N was examined. In meta-analysis, larger studies can carry more weight and thus may be more likely to render a review in need of update.

4.4.4 Complements to Recall

Some measure of efficiency of retrieval is often used as a complement to recall. Here precision, specificity and fallout will be described. These measures, as well as recall, can be derived from a 2x2 table **(Table 4).**

Table 4. Two by Two Table

	Relevant	Not Relevant	
Retrieved	A	B	A+B
Not Retrieved	C	D	C+D
	A+C	B+D	A+B+C+D

	Relevant	Not Relevant
Retrieved	True Positive	False Positive
Not Retrieved	False Negative	True Negative

Recall is calculated as

Recall = A/(A+C)

4.4.4.1 Precision

Precision is the proportion of relevant documents retrieved by the search and is the equivalent of the positive predictive value of a diagnostic test.[181] It is calculated as

Precision = A/(A+B) (Table 4)

Number needed to read (NNR) is the reciprocal of precision and represents screening efficiency or screening burden for the user. Lendgrebe suggests that when class imbalance is present (a very small proportion of the database is relevant), precision is the better measure than accuracy or sensitivity. Both precision and sensitivity are influenced by such a skewed distribution of relevant and irrelevant items, but precision remains much more sensitive to performance in retrieval of the small number of relevant items.[194]

Precision is a classic indicator of search quality, and there are no particular challenges to its use. There is a trade-off between recall and precision, however it is less marked than the trade-off between recall and specificity.[181]

4.4.4.2 Specificity

Specificity is used in diagnostic tests, thus is a familiar measure within the epidemiological community. It is the proportion of irrelevant records that are not retrieved by the search and is calculated as

Specificity = D/(B+D) (Table 4)

Specificity has been used extensively in filter development, in particular, it is part of the standard methodology of the Hedges team.[157] Values for specificity are almost always very large in information retrieval when a very large database is used, as most references are not retrieved.[181] It is used by the Hedges group to compare the relative efficiency of candidate search terms within the same database.[157]

4.4.4.3 Fallout or False Positive Rate

Fallout is the proportion of non-relevant documents that are retrieved (false positives) out of all non-relevant documents (true negatives plus false positives).[195] It is calculated as

Fallout = B / (B+D) (Table 4)

As with recall, all relevant documents need to be identified in order to establish the denominator, which is in this case one minus the number of relevant documents. As with specificity, fallout deals only with the non-relevant items. Van Rijsbergen (p. 149) argues that there is a functional relationship between recall, precision and fallout determined by the proportion (density) of relevant documents in the collection which he calls G or generality, calculated as

G = Number of relevant documents/ Total number of documents.

G = (A+C) / (A+B+C+D) (Table 4)

and

Precision = (Recall * G) / ((Recall * G) + Fallout * (1-G))

Precision = (A/(A+C)) * G / ((A/(A+C) * G) + ((B/(B+D) * 1-G)))

4.4.5 Summary Measures

The appeal of summary measures is that they provide a single indicator of search performance, but the disadvantage is that the interpretation of the measure is not obvious. Here three summary measures are briefly considered as alternatives to examining both precision and recall as the primary and secondary outcome measures.

4.4.5.1 Accuracy

Accuracy looks at the overall percent of correct classification across the whole database, it is; 1-(false positives + false negatives) or 1-(A+D). It is highly influenced by class priors, or the distribution of true negatives and true positives,[194] and so is of little use in searching where most records are not relevant. More formally by Lendgrebe *et al.*,[194] citing Weiss[196];"In an imbalanced setting, where the prior probability of the positive class is significantly less than the negative class (the ratio of these being defined as the skew or λ), accuracy is inadequate as a performance measure since it becomes biased towards the majority class."

4.4.5.2 E

E was proposed by van Rijsbergen as a general measure of effectiveness [195] (p. 174) which can be weighted toward recall or precision depending on which of those attributes is more important in the situation being considered by altering the value of α, where α is the relative value of recall.[181]

4.4.5.3 F

F measure[195] is a combined measure used in the precision-recall case. It is the geometric mean of precision and recall, in which the two measures are weighted equally, defined as

F= 2* Recall * Precision / (Recall + Precision) .

4.4.5.4 PosFrac

Posfrac = (TP+FP)/N or (A + B) / A+B+C+D

. "This measure is useful in applications requiring second-stage manual processing of the positive outcomes of the classifier (such as medical screening tests), and estimates the reduction in manual effort provided by the classification model."[194] While this is potentially interesting given the sequential screening approach used in systematic reviews, it has not previously been used in that context, and there seems no compelling reason to adopt it here.

4.4.6 Receiver Operator Characteristic (ROC) Analysis

Receiver Operator Characteristic (ROC) analysis is commonly used to evaluate classifiers where subjects are classified into two categories[197] and can be used for SVM

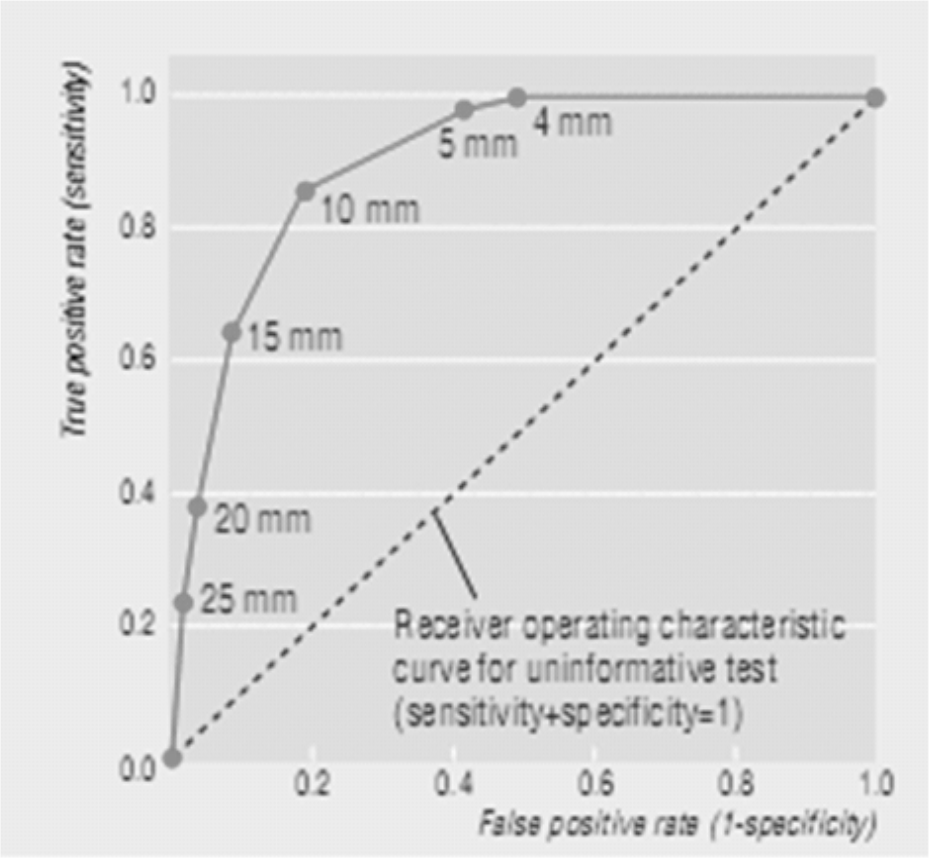


Figure 11. Example of Receiver Operator Characteristics plot. Reprinted with permision of BMJ.

and Related Article RCTs, where threshold of retrieval can be varied according to relevance score and records are classified as relevant or irrelevant. Area under the ROC (AUC) can be viewed as a performance measure that is integrated over a region of possible operating points (decision thresholds). Curves close to the rising diagonal are indicative of poor performance, while larger areas under the curve, which occur when the line rises sharply, are indicative of good performance .[198] Figure 11 illustrates the ROC curve for an effective screening test, of endovaginal ultrasonography for detecting endometrial cancer, originally published by Deeks.[198]

Receiver Operator Characteristic (ROC) analysis is commonly used to evaluate classifiers where subjects are classified into two categories[197] and can be used for SVM and Related Article RCTs, where threshold of retrieval can be varied according to relevance score and records are classified as relevant or irrelevant. Area under the ROC (AUC) can be viewed as a performance measure that is integrated over a region of possible operating points (decision thresholds). Curves close to the rising diagonal are indicative of poor performance, while larger areas under the curve, which occur when the line rises sharply, are indicative of good performance.[198] The figure below illustrates the ROC curve for an effective screening test, of endovaginal ultrasonography for detecting endometrial cancer, originally published by Deeks.[198]

Search methods with dichotomous relevance values of 0 or 1 do not permit altered retrieval thresholds and so ROC analysis is not possible – this is the case with the Boolean searches and citing reference method.[194]

4.4.7 Summary of Outcome Measures

"Precision and recall remain standard measures of effectiveness to this day, with some variations on the theme. They measure the probability of agreement between what the system retrieved or constructed as relevant (systems relevance), and what a user or user surrogate assessed or derived as relevant (user relevance)."[180] Sampson *et al.* have previously defended the use of these parameters in the context of systematic review

searches; “the information retrieval paradigm used in systematic reviews is classically suited to evaluation using the measures of recall and precision. Retrieval occurs in batch mode, although preliminary work may be exploratory and interactive. High recall and high precision are sought, with large retrieval sets being the norm. Retrieved documents are classified into a binary relevance scheme as eligible or ineligible for inclusion in the review. Finally, measures are taken to minimize the subjectivity or idiosyncrasy of the relevance assessment: the search result is evaluated against explicit criteria, often by two reviewers who must reach consensus, so that the work could be independently replicated.”[192]

4.5 PROCEDURES FOR IDENTIFYING NEW RELEVANT STUDIES

After eligibility was confirmed and the screening spreadsheet was developed, the reviewer received a package with a copy of the original systematic review (and the update, if one was available), the spreadsheet with the included studies, newer meta-analyses, and candidates as well as the template for all data extraction and calculation, and the Reference Manager database containing the bibliographic record for the newer MAs and the candidates.

Each reviewer read the systematic review, and then stated the question of the review in PICO format, recording this on the updating spreadsheet, also noting major inclusion criteria. PICO is a structured approach to posing questions commonly used in evidence-based medicine, where P is the population of interest (often the disease), I is the intervention of interest, C is the comparator, and O is the outcome.[199] Relevant new evidence would have to match all elements, as well as be a suitable study design.

The reviewer then selected up to four efficacy outcomes and two harms following the eligibility criteria of a clinical outcome (e.g. disease endpoint, functional status or mortality) or established intermediate outcome (e.g., blood pressure, glycemic control, standard instrument for measuring disease activity, such as a depression scale) with a meta-analysis with point estimate and 95% confidence interval.

The selected outcomes were recorded on the updating spreadsheet. For each outcome, the meta-analytic measure used (such as odds ratio or weighted mean difference), the number of trials in the original, the number of participants in total and number of participants in the largest included trial was recorded. For each selected outcome, the reviewer added a new worksheet using the template associated with the particular meta-analytic measure. The reviewer next recorded the pooled estimate and upper and lower 95 percentile confidence intervals. As new studies were identified, the data from the new trial was added. The spreadsheet calculator computed a new pooled estimate and confidence interval using a fixed effects model that included the additional evidence.

4.5.1 Example of Process for Assessing the Need to Update

An example may add clarity to the screening and updating assessment. A meta-analysis published in 2002 examined angiotensin receptor blockers in heart failure.[200] The population was patients with symptomatic heart failure and the intervention was angiotensin receptor blockade (ARB). The comparison was either placebo or angiotensin-converting enzyme (ACE) inhibitors. The primary outcome of the original review was all cause mortality and the secondary outcome was hospitalization for heart failure. Important inclusion criteria were random allocation, parallel group design, blinded and treatment of at least four weeks duration. Sixteen randomized controlled trials published between 1995 and 2001 were included in the original review. Study size ranged from 33 patients in a 1999 study to 5014 patients in a 2001 study.

Three meta-analyses from that review were selected as outcomes for updating and were transcribed to the updating spreadsheet. The three were mortality in trials comparing ARB *versus* eitherplacebo or ACE inhibitors, hospitalizations in trials of ARB *versus* either placebo or ACE inhibitors and mortality in studies of ARB *versus* placebo. All three meta-analyses had used odds ratios, none were significant in the original systematic review. In all cases the confidence interval spanned the line of effect. If the line of effect

was crossed, it indicates that the odds of outcome occurring did not differ statistically between groups.

Three new trials were found for the first outcome. The event rate (number of deaths in the course of the study), was recorded for the intervention and control group, as was the number of patients studied. As each of the three new studies was added, the spreadsheet calculated the odds of the event occurring in each group, and the odds ratio between groups for that study. The pooled odds ratio and upper and lower bounds of the 95% confidence interval were also computed. The pooled result was tested to see if it had become significant, indicated by a Z score of 1.96 or greater, corresponding to a p value of 0.05 or less. The robustness of this finding was tested, since a change from just non-significant to just significant (or *vice versa*) could easily be overturned with the addition of the next new study and so was not a compelling reason to perform an update. Marginal significance was operationalized as a Z score between 1.881 and 2.054. Next, the old and new upper and lower confidence interval were compared to see if there had been a change in the width of 95% confidence interval by at least 50% (i.e., CI new / CI old ≤ 0.5 or CI new / CI old ≥ 1.5). Robustness of this change was also examined.

Figures pertaining to size criteria, namely number of new studies, number of new participants, ratio of number of new to old studies and ratio of number of new and old N were computed. An increase in number of patients by 50% or more (either overall for a specific outcome) or an increase in number of trials by 50% or more (either overall for a specific outcome) or a new study more than three times the size of the largest study in the original systematic review were all flagged.

In our example, the first three new eligible studies all compared ARB with placebo and examined mortality, one of the selected outcomes. All three trials were published in 2003. The odds ratio in the original review was 0.68 with a 95% confidence interval from 0.38 to 1.22, so there was no clear advantage to either the treatment or the control. After the first study was added, the odds ratio was virtually unchanged at 0.673 (0.388 - 1.168). The addition of the second and third new studies changed the pooled

estimate to 0.837 (0.697 - 1.006) and 0.835 (0.696 - 1.003) respectively. This brings the odds ratio to borderline significance, which would have been an issue if the question were whether the crossing of the line was robust, but indeed the line of effect was not crossed, so this result is ignored. The width of the confidence interval was reduced by more than 50%, and this is noted, but this finding is not in itself sufficient to signal that the review needs updating, as nothing about this would be likely to change clinical practice. For this outcome, we do see that there has been a greater than 50% change in the number of patients studied. The largest trial in the original review comparing ARB and placebo and reporting on mortality involved 844 participants. These three new studies had 218, 2028 and 292 new participants respectively. Had the largest new study been more than three times the size of the largest study in the original, this would have been flagged.

In summary, considering the material published between May 15, 2001, and the end of 2003, there is a large body of new evidence on mortality, which increased the precision of the point estimate, but which did not shift the point estimate much and did not change it from being non-significant to significant. Since the conclusions and treatment implications would remain unchanged, the update can wait.

For the outcome of hospitalization for heart failure, the original pooled odds ratio was 0.86 with a 95% confidence interval of 0.69 - 1.06, close to significant. A meta-analysis by Lee, published in 2004, reported on this outcome. One large new study included in the Lee meta-analysis was added to the original estimate. The odds ratio of the new study was 0.640 with a confidence interval from 0.530 - 0.780, indicating fewer hospitalization in the intervention group (ARB) than in the control group (placebo). Pooling this new result with the previous results, the updated odds ratio became 0.730 with a 95% confidence interval of 0.633 - 0.843. Because 1.00 (equality) is outside the confidence interval, we can say that the pooled estimate of the effect now supports that ARB is superior to placebo. Based on this outcome, this therapy, which was previously characterized as promising or likely beneficial, is now more clearly favourable and

uncertainty has been reduced. A signal for updating has occurred; as this finding should be communicate so that clinicians can change their practice.

The reviewer would now transcribe the flagged evidence and the stopping criteria to summary page of the updating spreadsheet and seek an expert opinion source that would affirm or challenge the conclusion that the intervention significantly reduces hospitalizations for heart failure relative to placebo, and should be adopted in practice. As this updating project was retrospective, it had the benefit of hindsight, and could examine current sources of expert opinion such as Clinical Evidence (BMJ Publishing Group) or even more recent meta-analyses to confirm the findings of the updating exercise.

The reviewer's work in this is to assess the new studies for eligibility, extract the relevant data (here the number of events and number of participants in each group), and enter these in the spreadsheet, which computes the new variables and flags changes that exceed the thresholds set. Confirmatory sources were then consulted, and the case presented and discussed at a case conference where the classification could be accepted, or the evidence re-interpreted. On some occasions, if expert sources indicated that the finding of the original meta-analysis was no longer considered best practice, additional searching could be requested or spot searching could be done to identify the trials or meta-analysis that appeared pivotal in informing current best practice. Any new evidence found this way would be added to the updating spreadsheet as *Reviewer Nominated* to capture that the search had missed important evidence.

In this example, a total of eight eligible new studies were identified, all by examining a single new systematic review published in 2004 (Lee) The original review was considered current as of May 15, 2001, the last search date reported. The first eligible new evidence appeared in February 6, 2002, and the threshold for updating was crossed October 27, 2003. The survival time of the original systematic review was calculated by subtracting the original publication date from the indexing date of the new evidence resulting in the signal. It was 623 days (about one year and eight months). In the systematic reviews where no signal for updating occurred, the survival was calculated as

the number of days between the original publication data and the end date of our update searches.

As in this example, reviewers first assessed newer meta-analyses, starting with the newest. The logic of this screening order was the newest review would be most likely to include all the relevant evidence. If the newer systematic review was *On Topic,* the primary studies included in it were assessed for relevance (both *On Topic* of the systematic review and *Eligible* for the outcomes selected for updating). These primary studies were then sought in the candidate list, and their relevance recorded. If assessed as *Eligible*, the total number of study participants (N), both in the intervention and control arms, was recorded, not the number included in the specific meta-analyses used in the updating. If a previous report of an eligible study had been assessed as eligible, than no N was attributed to subsequent instances. If a primary study was not present in the candidate list, but was assessed as *Eligible*, it was added to the end of the list, and N recorded. These were treated as false negatives and were included in the denominator for recall for each systematic review.

If all newer systematic reviews were examined without any stopping rules being met, the reviewer would then begin screening the candidate list, from oldest to newest. The logic for this screening order was that we wanted to identify the earliest point where the review was in need of update, and for efficiency, we wanted to stop screening just beyond this point. A margin was included so that if a statistical threshold was met with one study, but the estimate reverted back to its original state with the inclusion of the next eligible study, than the review was not considered in need of update. The same assessment and recording procedure was used with the candidates as was used with the primary studies in newer meta-analyses.

4.5.2 Determination of Need to Update

The need to update was determined by the appearance of major or potentially invalidating new evidence. Quoting from the survival analysis, potentially invalidating new evidence "would make one no longer want clinicians or policymakers to base

decisions on the original findings (such as a pivotal trial that characterized treatment effectiveness in terms opposite of those in the original systematic review)." Major changes in evidence "would affect clinical decision making in important ways without invalidating the previous results (such as the identification of patient populations for whom treatment is more or less beneficial)."[17]

4.5.2.1 Conceptualization

Prior to this project, an international workshop was held in Ottawa, Canada in March 2006 to consider various aspects of updating systematic reviews.[11] One goal of the workshop was to create a framework. An issue explored in the workshop was the notion of triggers for updating. Triggers could arise from the evidence, or from external factors including the passage of time (as uncertainty builds with the passage of time), but in general, a trigger occurs when a state of uncertainty exists about the continued validity of the review and the resolution of that uncertainty is sufficiently important from a policy or scientific standpoint that resources are committed and an update is initiated. Alternatively, if update is not practical, a review could be withdrawn.

The Updating project attempted to operationalize triggers arising from new evidence. In this phase, the new evidence is screened and added to the meta-analysis. Certain rules determined whether the result has changed in important ways, or otherwise the weight of new evidence would indicate that the review should be updated. The "stopping rules" or "signals for updating" were probably the most difficult aspect of the protocol to develop and were mainly the work of Kaveh Shojania and David Moher. Ideally, stopping rules would operate in such a way that could be independently replicated by others and would produce meaningful results – that is, in cases where the update was triggered, the finding of the updated review would be different in some important way from the findings of the original review. Ultimately, as the systematic review is a tool for evidence-base medicine, the differences should be great enough to warrant a change in practice.[189] Both qualitative and a quantitative signals were explored through length

debate within the AHRQ updating project team and between the team and the Technical Expert Panel.

4.5.2.2 Classification

In the end, the triggers, or definitions of signals that a review needed to be updated, where chosen to be consistent with previous work comparing whether two sets of results relating to the same question were concordant or discordant.[17] A number of these were presented in the example of the actual assessment of the original meta-analytic result in the face of new evidence. The complete set is presented in Table 5.

Table 5. Summary of Classification of New Evidence.
Reviewers first assessed amount of new evidence (C criteria) and statistical impact (B criteria), then sought expert opinion that would support or refute the change or lack of change in outcome (A criteria). The direction of change implied by the evidence (increased or decreased certainty in efficacy, and direction of change (increased or decreased) in therapeutic use implied by the findings, were then evaluated. Final classification was then made.

Expert Criteria: type in the source of expert opinions below, and quote the expert opinion	
A1	Pivotal trial*, new meta-analysis, practice guideline (from major specialty organization or published in peer-reviewed journal), or recent textbook (e.g., UpToDate) characterizes the treatment in opposite terms to those in the cohort review: definitely or probably effective → definitely or probably ineffective or vice versa (i.e., ineffective → effective).
A2	Pivotal trial, new meta-analysis, practice guideline, or recent textbook calls into question the use of the treatment on the basis of harm (i.e., the treatment would no longer be recommended because risks outweigh benefits). A new result for harm that does not undermine use, but has clear potential to affect treatment decisions would count as a notable change (specifically, A4).
A3	Pivotal trial, new meta-analysis, practice guideline, or recent textbook characterized another treatment as significantly superior to the one evaluated in the cohort review (based on efficacy or harm).
A4	Pivotal trial, new meta-analysis, "discordant" meta-analysis, trial indexed in ACP J Club, more recent practice guideline, or recent textbook does not contradict the previous review, but adds an important caveat, about the patient populations who benefit, way in which treatment has to be delivered in order to derive benefit, increases in harm that are not sufficient to undermine use altogether, but would clearly temper affect clinical decision making.

A5	Instead of a caveat, there has been expansion of the role of the treatment (e.g., the treatment has now been shown to be of benefit in primary prevention, not just secondary; or now shown to be of benefit in children or aged population etc).
A6	Therapy previously characterized as “promising”, “likely beneficial” or similar description and now characterized as definitely beneficial. Conversely, if original review characterized therapy as “probably not effective” and now characterized as "definitely not effective". This criterion can be summarized as a movement of one category on the 5-point scale in Direction of Change.
A7	Discordant meta-analysis or trial indexed in ACP Journal Club characterized the treatment in sufficiently different terms to the cohort review that disagreement would have met criteria for ‘Major change’ (A1) except the source was not a pivotal trial, new-meta-analysis, or more recent practice guideline, or recent textbook.
Statistical Criteria	
B1	Line of no effect crossed.
B2	The new result indicates a relative change in effect size of at least 50% (e.g., RRnew / RRold <=0.5 or RRnew / RRold >=1.5 – same applies for OR, RD, WMD).
B3	New and old point estimates differ significantly
Size Criteria	
C1	Increase in number of patients ≥ 50% (either overall for a specific outcome) †
C2	Increase in number of trials ≥50% (either overall for a specific outcome)
C3	The biggest New Study ≥ 3 times larger than the previous largest trial
C4	Width of 95% confidence interval changed by at least 50% (i.e., CI new / CI old <=0.5 or CI new / CI old >=1.5)
Direction of change: Certainty	
	Consider the five position scale in which a study conclusion (‘X is beneficial’ or ‘Y is harmful’) is characterized as: Definitely true | Possibly true | Unclear | Possibly not true | Definitely not true Change in certainty is defined in terms of the absolute distance from the center position ("unclear"). When the new result lies further from the center than the result in the cohort review, then certainty has increased; if the new result is closer to the center, then certainty has decreased. And, if the distance is the same to the center (even if on opposite sides) then certainty has not changed.

Direction of change: Therapeutic Use
Effective to ineffective (A1, A6, or A7) = Decreased utility
Harm (A2 or A4) = Decreased utility
A new treatment is superior (A3) = Decreased utility
Caveat (A4) = Decreased utility
Ineffective à Effective (A1, A6, or A7) = Increased utility
Expansion (A5) = Increased utility
Final Classification
‡ Potentially invalidating
§ Major (indicate though if based on expert or stats by itself involving primary or harder outcome)
Minor (at least one size criteria without any of the other criteria)
None of the above (meaning at least one eligible new study found, but not meeting any of the criteria)
No new studies found

Source: Summary page, updating spreadsheet template
*A pivotal trials was defined as one having a sample size of at least 3 times the largest previous trial, published in one the top 5 medical journals (New England Journal of Medicine, The Lancet, JAMA, Annals of Internal Medicine or BMJ) based on journal impact factor.
†In other words, either the total number of patients in all of the included trials has increased by at least 50% or the number of patients who contributed to the analysis of one of the included outcomes has increased by at least 50%.
‡ Assign potentially invalidating if expert criteria A1-A3 are met.
§ Assign major if expert criteria A4-A7 are met, regardless of statistical criteria; or ii) statistical criteria on the primary outcome or a harder outcome. Harder outcome is defined as either a harder outcome than any of the outcomes in the cohort meta-analysis or an outcome harder than the ones that had statistically significant results in the cohort meta-analysis. E.g., if the cohort review showed significant reduction in myocardial infarctions and then a significant result for mortality would count if the cohort review's result for mortality was not significant or if mortality had not been analyzed at all.

4.5.3 Quality Control Measures

A number of quality control measures were taken to ensure objective and repeatable assessment of eligibility of cohort systematic review and new evidence. Initial reviewer calibration was undertaken to ensure reviewers assessing material consistently, and double screening was used for selecting the systematic reviews making up the cohort,

with discrepancies discussed and resolved by consensus. When assessing the need to update each review, reviewers were first to achieve consensus on the PICO elements. The PICO elements informed eligibility of newer material. Resources did not permit dual screening of candidates.

One challenge in information retrieval experiments is that assessment of relevance may be influenced by material previously found and thus shift over time.[187] Tague-Sutcliffe identifies sequence effects due to either learning or fatigue as a challenge in repeated measures designs where the results of various search approaches are assessed.[179] Here, integrating all query results for an update into a single candidate list eliminated the need for multiple passes and blinded the reviewers to which search identified the candidates. Focusing first on the results of newer meta-analyses and assessing only to the point where a signal occurred limited the screening burden and hopefully also screening fatigue. Providing an explicit method to record second and subsequent publications of the same date was intended to minimize sequence effects.[187] Automating the calculation of all quantitative signals minimized human error in assessments.

Case conference (KS, JJ, MA, MS) was used to review and finalize the determination of updating status and final classification of the new evidence as potentially invalidating, major, minor, new studies only, or no new evidence. This was an opportunity to assess the expert opinion, and explore reasons for discordance. Additional searching was requested if results seemed incomplete, based on current opinion or the knowledge of any members of the review team.

One member of the review team (JJ) performed quality checks when extracting data from the updating spreadsheets to transcribe into the main survival data set. I reviewed all screening worksheets when extracting data for the main search performance data set. In either case, incomplete or incongruent data were referred back to the reviewer for completion or clarification.

4.5.4 Limitations of the Screening Method

This screening approach has a number of implications for the recall and precision estimates. First, it will be apparent that not all candidates were assessed. Second, there are systematic differences between the amount of evidence assessed for reviews that did go out of date and those that did not go out of date. For those systematic reviews where the results were overturned with new meta-analyses, it was possible that only a few candidates were assessed, and only relevant candidates were classified – all these would be *On Topic* and most would be *Eligible.* On the other hand, for systematic reviews that were not in need of update, the entire candidate list would be evaluated, including retrieved records that were not eligible. Recall of signaling evidence can be calculated but precision cannot be accurately established from these data. Precision of the searches requires that the whole retrieval, or at least a systematic portion of it, be assessed. Complete screening of the AHRQ cohort was done to permit precision estimates.

An additional, but less problematic, situation is that some evidence that was not added into the calculator could be assessed as relevant. This occurs when new meta-analysis includes studies published subsequent to the study that the signal resulted from. In our ARB example, there was one smaller study that appeared several months after the signal date, and it is recorded as *On Topic* and *Eligible*, with N attributed. Thus, recall is recall of evidence identified by our methods as relevant – it may not correspond exactly with recall of evidence included in one or more of the updated outcomes prior to the stopping rule. The assessment process is presented graphically in Figure 12.

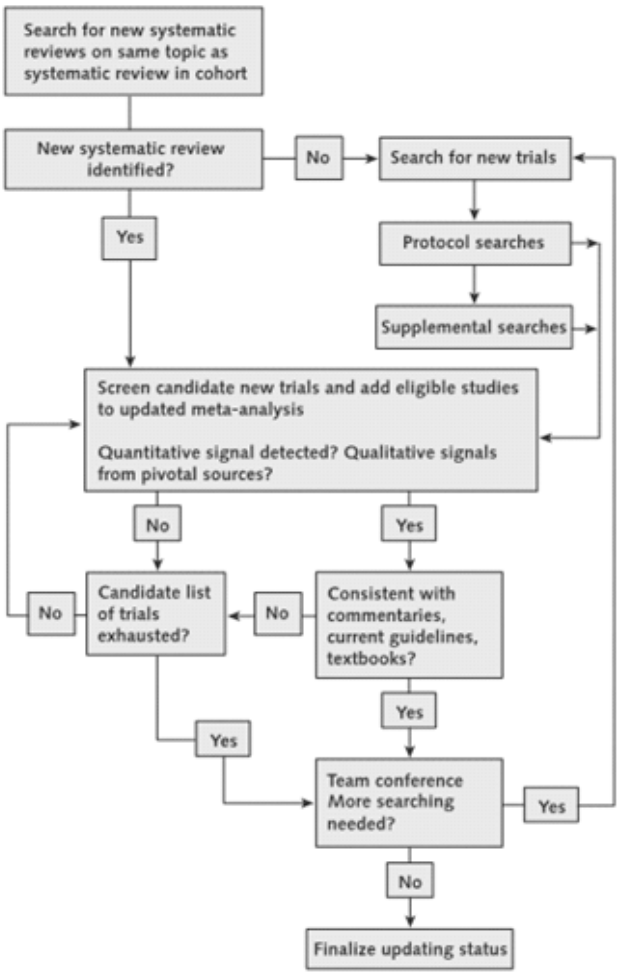


Figure 12. Overall Process of Determining Updating Status
Reprinted with permission of Annals of Internal Medicine

4.6 REVIEW-LEVEL DATA EXTRACTION

Characteristics of reviews that were extracted included first author, journal, journal impact factor, PubMed ID, type (Cochrane Collaboration Review, journal published review or health technology assessment report), date of publication, MHDA, date of the most recent search, clinical area, whether updated by searches or by a newer systematic review, the number of studies included in the original review, the total number of participants represented in the original review, the intervention type (drug, device or procedure), control type (standard care, placebo, alternative drug/ device/ procedure or multiple controls) and study designs included (RCT only or RCTs plus quasi-RCTs or controlled clinical trials (CCTs)).

Where possible, these variables were extracted from the MEDLINE record. Otherwise, they were extracted by a member of the team – mostly by the IS team, but the intervention type, control type, and study design were determined by the review team.

All systematic reviews in the cohort were classified according to clinical area. This variable was used to describe the sample, and forms the basis for the analysis of growth of the literature. Several factors were considered in the classification: ISI journal classification, the Cochrane CollaborationReview Group where the topic might be placed, the high level MEDLINE subject heading (MeSH) that the population (condition) would be indexed under and finally, which heading on the AHRQ Evidence Report web page might best suit the review. Where a review seemed to fit two or more categories, we recorded all applicable. We then sought to reduce the number of classifications to around ten, with most reviews falling into one of the ten categories. I undertook review with another reviewer (AI) and our consensus was reviewed and confirmed by a medically trained reviewer (MA). We arrived at 13 categories and adopted the ISI journal classification to name them.

4.7 DATA SET CREATION

Creation of the master dataset containing all studies included in the original reviews, in the updates, the newer meta-analysis, candidates, and reviewer nominated records, was a major undertaking.

The procudures I developed and used were:

1. All.xls files were pasted into an SPSS[197] file and pivoted
2. The screening worksheets of the updating spreadsheets were cleaned and pasted into another SPSS file
3. The two SPSS files were merged
4. Another SPSS file had characteristics of the reviews, survival, and final classification. This was created for the survival analysis.[17]
5. Necessary review-level data was added to the main dataset by a series of recodes.

The all.xls files recorded which searches had retrieved each PubMed ID. These were organized so that eachretrieval was a case with three variables; PubMed ID, CohortID for which it had been retrieved, and the search method that had retrieved it (Figure 13).

Highlighting in Figure 13 shows that 3 search methods retrieved PubMed ID 10667492.

It was necessary to arrange the data so that each PubMed ID was a case, and each search approach was a dichotomous variable where 1 signified that the search had retrieved that PMID and where 0 signified that it had not retrieved it. The SPSS Restructure facility was used to restructure cases into variables (Figure 14).

	A	B	C	D
1	**13**	**10667492**	**CENTRAL**	
2	13	10745332	CENTRAL	
3	13	10890228	CENTRAL	
4	13	11060526	CENTRAL	
5	13	11705075	CENTRAL	
6	13	12873543	CENTRAL	
7	13	12884156	CENTRAL	
8	13	15803051	CENTRAL	
9	13	16220082	CENTRAL	
10	**13**	**10667492**	**CQ**	
11	13	10745332	CQ	
12	13	10796589	CQ	
13	13	10890225	CQ	
14	13	10890228	CQ	
15	13	11060526	CQ	
16	13	11405973	CQ	
17	13	15798866	CQ	
18	13	15803051	CQ	
19	13	16189088	CQ	
20	13	16220082	CQ	
21	13	16499288	CQ	
22	13	16540956	CQ	
23	13	16614620	CQ	
24	13	15803051	CR	
25	13	16540956	CR	
26	13	0	MA	
27	**13**	**10667492**	**RCT AIM**	
28	13	11060526	RCT AIM	
29	13	16540956	RCT AIM	
30	13	16094039	RI MA	

Sheet1 Sheet2

Figure 13. All.xls file for CohortID 13

After restructuring, I integrated the retrieval data with relevance data using the SPSS

```
RECODE
 Method
 ('RCT AIM'=1) (ELSE=0) INTO AIM_RCT .
RECODE
 Method
 ('CQ'=1) (ELSE=0) INTO CQ .
RECODE
 Method
 ('RI RCT'=1) (ELSE=0) INTO RI_RCT .
RECODE
 Method
 ('CR'=1) (ELSE=0) INTO CR .
EXECUTE .
SORT CASES BY CohortID PMID.
CASESTOVARS
 /ID = CohortID PMID
 /GROUPBY = VARIABLE .
EXECUTE.
COMPUTE RI_RCT = max(RI_RCT.1,RI_RCT.2,RI_RCT.3,RI_RCT.4,RI_RCT.5,
RI_RCT.6,RI_RCT.7,RI_RCT.8,RI_RCT.9,RI_RCT.10,RI_RCT.11,RI_RCT.12,RI_RCT.13) .
COMPUTE AIM_RCT = max(AIM_RCT.1,AIM_RCT.2,AIM_RCT.3,AIM_RCT.4,AIM_RCT.5,
AIM_RCT.6,AIM_RCT.7,AIM_RCT.8,AIM_RCT.9,AIM_RCT.10,AIM_RCT.11,AIM_RCT.12,AI
M_RCT.13) .
Compute CR= max(CR.1,CR.2,CR.3,CR.4,CR.5,CR.6,CR.7,CR.8,CR.9,CR.10,
CR.11,CR.12,CR.13) .
Compute CQ= max(CQ.1,CQ.2,CQ.3,CQ.4,CQ.5,CQ.6,CQ.7,CQ.8,CQ.9,CQ.10,
CQ.11,CQ.12,CQ.13) .
Execute.
```

Figure 14. Syntax to convert retrieval data from cases to variables

merge files procedure. A primary key of CohortID and PubMed ID was used to match cases from the two datasets. The compound primary key was necessary as some PubMed IDs had been retrieved for more than one systematic review in the cohort. After the two datasets were merged, the review-level data was added to the dataset using a series of recode commands (Figure 14).

4.8 DATA ANALYSIS

Description of the Searches

Boolean searches, both those used in the test searches and those used by the original review authors in the AHRQ and Updated Cochrane cohorts, were characterized on a number of dimension. The number of the following features employed was

considered; free text, explode MeSH terms, use of starring or major MeSH terms, subheading (either attached or floating), and filters. Filters were any query statements designed to include or exclude specific study designs or publication types. This included the clinical queries, AIM RCT, highly sensitive search strategy, but others as well, including the query "(prevention or preventive or therapy or therapeutic or treatment).ti,sh."[201] and, in another Evidence Report,[202] the construction:

16. 15 or 14

17. 16 not letter.pt.

18. 17 not editorial.pt.

19. 18 not case report/

20. case report/ and clinical trial.pt.

21. 18 and 20

22. 19 or 21

Language limits and manoeuvres to limit to human studies or exclude animal-only studies were counted.

The number of search terms was determined by making a copy of the search in Word and deleting any lines that contained limits, filters or were Boolean combinations of other line such as "(1 or 2) and 3". Boolean ORs, ANDs and adjacency operators were replaced with carriage returns to isolate each term on a separate line. The lines were then numbered, and the number of lines was taken as the number of search terms.

4.9 SEARCH PERFORMANCE BY INTERVENTION TYPE

Day *et al.*[203] found that simplified search strategies achieved 100% recall in pharmacological interventions but as low as 40% recall in non-pharmacological interventions in musculoskeletal and pain systematic reviews. Recall of the searches will be examined in a subgroup analysis of systematic reviews in which the intervention studies was classed as a drug (n=89, 86%), compared with those where the intervention was a device or a procedure (N=16, 14%).

4.10 STATISTICAL ANALYSIS

Data were analyzed quantitatively through SPSS version 16.[197] For the main analyses, the independent variable is search method. Recall was the primary outcome measure (dependent variable), and precision and its variant, number needed to read, were secondary outcome measures. Most analysis was descriptive rather than inferential, and visual presentation was used where practical. Recall and precision, which are ratios (proportions), are interval level data. Recall and precision were aggregated across queries, both with overall mean (macro-averaging) and average of measures for individual systematic reviews (micro-averaging). The former gives more weight to the document and the later gives more weight to the query in the averaging process[179] (see p. 483). Classifications, as nominal level descriptive data, were summarized as counts. Rankings and relevance scores, as ordinal data, were summarized using median and inter-quartile range.

Most graphical presentations were prepared using Plot for Mac OS X.[204] Survival plots and ROC curves were prepared using SPSS.[197] Growth of the literature plots were prepared using CurveExpert 1.3.[205] Venn diagrams were prepared using Adobe Photoshop Elements.[206]

Calculation of recall was performed in SPSS. Each search method was coded as either 1 (retrieved) or 0 (not retrieved) for each record considered. Recall could be calculated directly by selecting records that were assessed as eligible, and taking the mean for each search type for those records. For instance, if the review with CohortID of 109 had 10 new records assessed as *Eligible* by the reviewers, if 9 of those 10 records had SVM coded as 1 and the remaining record has SVM coded as 0, the mean of SVM would be 0.90, and this, as the proportion retrieved, equalled recall.

Precision could not be calculated directly in SPSS. For each search type, all records retrieved by those searches were retrieved in SPSS, and a cross tabulation was made of Target by Set. Target was 1 if the record had been assessed as eligible. The

number retrieved and number eligible were transcribed to Excel, where the proportion of retrieved that were eligible was calculated.

ROC was used for those test searches that provide relevance scores and so permit adjustment to the threshold for retrieval – the SVM and Related Article searches. Area under the curve was calculated with the SPSS ROC procedure which plots an ROC curve, and computes the area under the ROC curve with confidence intervals.[197]

4.11 SUMMARY OF THE MAIN EXPERIMENT

The design and methods can be summarized following Tague-Sutcliffe's "Pragmatics of Information Retrieval Revisited" [179]

Decision 1. To test or not to test?

Test – there is a need and the work is novel (see Chapters 2 and 3).

Decision 2. What kind of test?

A laboratory test with standardized users (reviewers see Section 4.5), a single database, standardized searchers, and specified search constraints (see Section 4.2).

Decision 3. How to operationalise the variables?

The independent variables are the test searches; a broader Boolean search (Clinical Query) and narrower version (AIM RCT), two similarity searches – RI RCT and SVM, and finally a citation search – citing RCTs (see Section 4.2). Dependent measures are principally recall and, where possible, precision (see Section 4.4.7).

Decision 4. What database to use?

All search approaches were tested in MEDLINE, although CENTRAL was examined for comparison (see Section 4.2.4.2).

Decision 5. Where to get queries?

The queries are the populations and interventions of systematic reviews of drugs, devices or procedures in allopathic medicine from ACP Journal Club (See 4.1.2).

Decision 6. How to process the queries?

Queries were processed through publicly available interfaces (Ovid MEDLINE or PubMed) for all but SVM where NRC's SVM interface was used (see Section 4.2).

Retrievals were assembled into single sets and presented to reviewers in spreadsheets (see Section 4.3). Assessed retrievals were restructured in SPSS into repeated measures formats (see Section 4.7).

Decision 7. How will factors under investigation be assigned to experimental units?

The factors under investigation are the search strategies and the experimental unit corresponds to the systematic review to be updated. For example, each systematic review to be updated could have been randomly assigned to one search approach. Instead, all searches were used for all systematic reviews in a repeated measures or within subject design, maximizing data use and reducing inter-query variation. The exception of support vector machine which was not assigned to all systematic reviews in the cohort but to two subsets – AHRQ reviews and updated Cochrane reviews that met certain objective criteria (see Section 4.8).

Decision 8. How to collect the data?

Data collected to describe the sample and the retrieved records was derived whenever possible from bibliographic records, or where that was not possible, was abstracted by one or more members of the research team (see Section 4.6). Data on relevance of retrievals was determined by the decision of authors of updated reviews, where available, but otherwise was collected by reviewers who assessed records, blind to retrieval method, for topicality and relevance to the systematic review to be updated (see Section 4.3.1.3). Classification of the impact of the new material was based on group discussion and consultation with published sources of expert opinion (see Section 4.5.2).

Decision 9. How to analyze the data?

Data were analyzed quantitatively, using recall as the primary outcome measure, and precision and its variant, number needed to read, as the secondary outcome measures (see Section 4.4.7). Most analysis was descriptive rather than inferential, and visual presentation was used where practical. Recall and precision were aggregated across

queries, both with overall and average means. Rankings and relevance scores were also considered (see Section 4.8).

Chapter 5: Methods - Exploratory Analyses

5.0 INTRODUCTION

The previous chapter laid out the methods for identifying a cohort of clinically important systematic review of reasonably quality, identifying new evidence using test searches to see if those reviews became out of date during the observation period, and determining the performance of those searches. This chapter explores other information retrieval issues in updating systematic reviews. Aspects of both the searches and of the resulting evidence will be explored. First, the structural relationship between searches will be examined to explore why those searches perform as they do. Understanding the performance of the original searches informs updating practices and helps place the performance of the tested searches in context. Thus, precision of the searches in a cross-sectional sample of systematic reviews from 2004 is examined, and recall of the systematic review searches for the updated Cochrane cohort and the AHRQ cohort is examined.

Characteristics of the new evidence will be examined, as will the question of whether the survival of the review depends on how old the evidence in it is. Not all evidence is necessarily retained in an updated systematic review. Comparison of the nature of material retained and not retained may provide insights into how evidence shifts over time, which may have implications for updating systematic reviews.

5.1 STRUCTURAL RELATIONSHIP BETWEEN SEARCHES

5.1.1 Unique Contribution and Overlap

Unique retrievals are those records retrieved by only one of the searches tested. Frequencies of unique retrievals, unique *On Topic* retrievals, unique *Eligible* retrievals, unique *Eligible studies with N attributed* were calculated for the various test searches for all three cohorts. Unique retrievals were identified through a series of SPSS Select statements that selected records retrieved by the test search and not by any of the others.

New variables were created to represent the unique component, named in the manner; RI_RCT_Only, CQ_Only *et cetera*. Crosstabs then identified the components of those variable that were *On Topic*, *Eligible*, and had *New N Attributed*. Searches examined were Related Article RCT, Citing RCT, SVM95, SVM200point5, Clinical Query, and Subject Search in CENTRAL. Abridged Index Medicus RCT was not tested since it is a narrower version of Clinical Query, and so yields no unique material. The records found by the reviewer authors *Actual* searches but not by any of the test searches was also identified and reported, but since the authors' searches were highly variable in recall, unique retrievals where examined for the test searches relative to each other, regardless of whether the review authors' actual search would or did also find the record.

Overlap was also calculated between all pairs in the set Related Article RCT, Citing RCT, SVM95, SVM200point5, Clinical Query, Subject Search in CENTRAL and Abridged Index Medicus RCT. The size of the unique component of each search and the degree of overlap between searches was presented as a proportional Venn diagram, were the diameter of the circle representing each search retrieval was calculated as:

(1) $$D = 2\sqrt{\frac{n}{\Pi}}$$

And n is the size of the retrieval.

5.1.5 Convergence of Multiple Retrieval Methods

Prior research in contexts other than systematic review searches suggests that retrieval of the same citation by multiple methods signals its relevance. Saracevic and Kantor found a striking increase in precision when items were retrieved by more than one searcher, and they reviewed previous work looking at overlapping retrieval in a variety of contexts, concluding that the degree of agreement in representation of concepts in search terms was consistently low, but the overlapping area was of great interest. They proposed a "super-strategy" in which searches of a question independently by several searchers, and the overlap in the retrieved sets is examined.[166,170] It has also been shown that articles indexed in more than one database form a core literature that is more likely to be

relevant.[207] More directly relevant to the current question, Pao and Lee found a 5% overlap between MEDLINE subject searching and citation searching, but 86% of those identified by both methods were judged at least partially relevant and 68% were judged relevant.[208] Wu and McClean reviewed the prior work in data fusion, a technique derived from multisensor processing and applied to multiple document lists for the same information need, results from different information retrieval systems, or as in this thesis, different retrieval strategies in the same system. They tested three fusion techniques in four datasets and found, in the univariate analyses, that the effectiveness of combining strategies (fusion) was associated with low overlap between queries and high precision for the individual queries. In multivariate analysis, high precision became the least important factor, while a high standard deviation of mean average precision across the query methods was predictive of good performance of the fused results. They explained their results; "if some results are better than some others, then these good results are more likely to share some common opinion, and their common opinion will dominate the whole group; while those poor results share less common opinion, and their effect on fusion is limited. On the other hand, if all the results are close in performance, then no one result or several results can dominate the whole group, and less improvement can be made by data fusion." They also note that overlap is most informative when the systems are relatively different, and that the number of quite different systems cannot be very large.[209] Beitzel also found that improvements were only seen when the different query methods produced a fairly large number of unique relevant results.[210]

Lee explored the phenomenon using TREC data and concluded that different query methods retrieve similar sets of relevant documents but retrieve different sets of non-relevant documents. He found that ranking the combined results based on rank within the retrieved sets was more effective than weighting based on similarities scores.[211]

The association between the relevance of records and the number of methods retrieving the record was examined to see whether the association would hold true in systematic review searches. This analysis is an extension of the work to find sufficient

search strategies in the cohort of systematic reviews updated through searching. In the main cohort, the records found through the test searches and assessed for eligibility were selected. The number of different search methods that retrieved each was calculated by summing the scores for Related Article RCT, Clinical Query, Abridged Index Medicus RCT and Citing RCT. Each search was scored as 1 if it retrieved a given record and 0 if it did not retrieve that record. Thus, the sum of the scores for these variables indicated how many of the searches retrieved the record in question. SVM was not included in this analysis as results were not available for the entire cohort. The number of searches identifying each of the Eligible (n=423) and ineligible records (n=8410) was assessed. Differences between the number of methods retrieving ineligible records and the number of methods retrieving eligible records were calculated and tested using χ^2. A significant difference, in favour of relevant records being retrieved by more of the searches, would suggest that convergence was signalling relevance, or maybe more likely, that many of those records picked up by a minority of searches are being found by those searches for some idiosyncratic reason other than relevance.

The subject searches are not independent – for example, Abridged Index Medicus RCT is narrower to Clinical Query – so the three subject search methods were collapsed and retrieval convergence of Boolean, similarity and citing searches were examined – using fewer systems that are less alike, following Wu and McLean.[209] The number of types retrieving each record was computed, and χ^2 calculated between the number of search types retrieving (1-3) and eligibly. Finally, as so few records were identified by the citing reference search, the eligibility of records found by both Clinical Query and Related Article RCT was tested against retrieval by only one or the other of these searches, representing the subject search and similarity search approaches.

5.1.3 Related Article Seed Refinement

The performance of the Related Articles was explored in detail in the cohort of updated Cochrane reviews. The three newest and three largest studies included in the original review were selected as seeds for the related article searches. There was no

empirical basis for chosing this criteria. The rationale was simply that large studies would carry the most weight in a meta-analysis and the new studies would represent any secular trend. It is possible that other seed selection criteria could provide a better proformance. To explore this possibility, seeds are classified based on qualifications as seeds (independent variable) and performance (dependent variable). The strength of the association between these variables was assessed. The objective was to determine whether only one or the other type of seed (large or new) is needed. Reducing the number of unproductive seeds could improve precision. In the updated systematic reviews, each new included study was tested to determine which seed it was related to. Seed yielding one or more new included studies were considered productive. Seeds not yielding any new studies were considered unproductive. Seeds that were related to all new included studies were considered super seeds. The number of productive seeds and the number of articles related to each productive seed was reported. Productivity of MEDLINE Misses used as seeds was compared to non-misses used as seeds.

The feasibility of improving precision of the Related Article searches by limiting the retrieval size was explored. Rankings were examined for the two cohorts where support vector machine was also tested, the six updated Cochrane reviews and the ten AHRQ Evidence Reports. Position of eligible record in the ranking of related article retrieval was recorded, and recall and precision of the truncated searches was computed.

5.1.4 Sufficient Strategies

Given some overlap exists between searches, and that more overlap is expected in relevant than irrelevant articles, each additional search method beyond the minimum set needed to identify all or nearly all eligible records would reduce precision without improving recall. Sufficient search strategies are combinations of searches that would retrieve all relevant studies. These were determined for the 72 systematic reviews in the main cohort where one or more new relevant studies were identified. In a spreadsheet, each Eligible record is listed, along with its type, and the searches that retrieved the record (Figure 15). The combination of search methods needed to retrieve all candidate

studies assessed as Eligible was determined by inspection. This combination was considered a sufficient strategy for that systematic review. Where two or more combinations would retrieval all new *Eligible* records, the retrieval size of the different combinations was calculate to identify which combination had the highest precision.

Cohort ID	PMID	Type	RI RCT	AIM RCT	CR	CQ	CENTRAL	
13	12884156	Candidate	0	0	0	1	1	No sufficient combination
	12225613	Candidate	1	0	0	0	0	
	12512504	Nomination	0	0	0	0	0	
22	12324552	Candidate	1	0	1	0	0	RI_RCT
	12126819	Candidate	1	0	1	0	0	
	11425770	Candidate	1	0	0	0	0	
	11216966	Candidate	1	0	1	0	0	
26	16714187	Nomination	0	0	0	0	1	CENTRAL
31	10023943	Candidate	1	0	0	0	0	AIM_RCT
	10376614	Candidate	1	0	0	0	0	
	10714728	Candidate	1	1	0	1	1	
	11386263	Candidate	1	0	0	0	1	
	11386264	Candidate	1	0	0	0	1	
33	16236688	Candidate	0	0	0	1	0	CQ
	15643104	Candidate	1	0	0	1	0	
	15605617	Candidate	1	0	0	1	0	
37	16144890	Candidate	1	0	0	1	0	RI_RCT or CQ

Figure 15. Excerpt from the Spreadsheet Used to Identify Sufficient Search Strategies

5.1.5 Capture – Recapture

The Capture-Recapture technique is used by biologists to estimate the size of animal populations. A sample is captured, tagged and released. A subsequent sample is captured, and the proportion of previously captured animals is noted.[212] Total population size is estimated from the numbers of new animals, recaptured animals, and animals not appearing in the subsequent sample. This technique has also been called "Comparison of multiple methods of ascertainment" (COMMA).[213]

Capture-recapture has been used in epidemiology to estimate ascertainment rates for diseases.[214] Capture-recapture has been used to detect missing studies from systematic reviews[213] and to detect publication bias.[215] Most recently, the technique has been applied

to the task of deciding when a search is finished by estimating the total number of on topic and eligible studies likely to be in the literature and determining if additional sources might yield a worthwhile harvest.[216] Qualitative research, and qualitative systematic reviews also may use the idea of 'saturation' – that sufficient data have been collected.[217]

The technique is used here to test the hypothesis that there were missing studies in the updated Cochrane reviews. As will be elaborated below, positive dependence of sources will tend to produce an underestimate of the true population size, and negative dependence will tend to produce an over-estimate.[218] Therefore, the behaviour of capture-recapture estimates based on pairs of searches can also help illustrate the relationship between searches when compared to the total number of relevant studies found by any of the searches.

Capture-recapture analysis requires data in a format that permits calculation of the intersection of data from different sources. In this case, the PubMed unique identifier provides an ideal mechanism. Diagnostic quality of the data is important.[218] This thesis uses two different methods to assess relevance, equivalent to the diagnostic methods in epidemiological studies. In two of the samples, a reviewer from our team assessed relevance. In the third sample, relevance was determined by the decision of the original author to include a study in the update of a systematic review.

Source dependence is a main limitation of the capture-recapture method when the objective is to estimate the population size. "Any two sources *A* and *B* are dependent if the overall, i.e., average, probability of members of a population who appear in their intersection or 'overlap' is equal to the product of the average probabilities of appearing in *A,* and *B."*[218] If, as is the case with AIM RCT, one is a nested set of the other, the probability of being in the smaller source (say 0.20) will be larger than the product of the individual probabilities (say 0.20 * 0.30 or 0.06), and the population will be overestimated. Where there is no overlap, the dependence will be negative as the

probably of being in either is larger than the probability of being in both, which is 0.00, and so the population size will be underestimated.

The situation is more complex when more than two sources exist but in the examples of extreme overlap and no overlap, the same errors of estimation will occur. In fact, Bennett *et al.* argue, usually there is variation in overlap with some pairs of sources being more dependent than others are, making the overall dependence difficult to quantify.[215] Hook and Regal note that even when not independent, if the investigator can ascertain the likely direction of dependence (positive or negative) than the capture-recapture method is still useful as it can be used to set plausible upper and lower boundaries on the true population size.[4] Multiple pairs, with different directions of dependence could conceivably by employed to make such estimates.

An assumption of the method is that each with a source, each member of the population has the same "ascertainability" or probability of capture.[218] Superficially, this seems unlikely to occur in the case of searching. It is the variability in indexing and in the text the authors use in the title and abstract that makes searching difficult. Still, within this sample, some restrictions have been imposed that reduce variability between studies. These include that the intervention must be a single or class of drug, medical device or procedure, and not a CAM therapy. In fact most (87% in the main cohort - see Section 6.9) were drug interventions. Further, the study designs must have been randomized controlled trials or controlled clinical trials. Searching and indexing is well developed for these intervention types and studies designs and reporting guidelines have been in place for the duration of the update interval that should ease indexing and retrieval.[219]

The final assumption, as outlined by Hook and Regal, is "that the population under study is 'closed,' i.e., that there are no entries or losses during the study period."[218] This assumption is met in the case of the test searches, in that all samples were drawn very nearly simultaneously, and limited by entry date into the database.

Hook and Regal outline the computational aspects. Conceptually, it is a case of multiple 2x2 tables, where the cell containing population members missed by both

ascertainment methods are being estimated. These estimates can be solved using Bayesian or non-Bayesian models.[218]

Funnel plot methods (described below) are another approach to estimating missing studies. It is assumed that the non-identification results from some bias. These biases, such as those described by Song[220] and Tricco,[7] influence identification and selection and so violate the assumption of equal ascertainability.[215] Bennett *et al*. used the Capture-Recapture method to estimate publication bias, or the number of studies missing altogether from the literature.[215] Ascertainment methods were electronic database searching, contacting experts, and hand searching of journals. Bennett *et al.* compared estimates from capture-recapture methods to funnel plots in a case study involving resistance training for older adults. The objective of the exercise was to estimate the robustness of the conclusions of the review in the face of a potentially incomplete evidence base. Eight databases were searched, reference lists of identified trials and relevant reviews were examined, conference proceedings were searched by hand and the first author of each relevant study was contacted in an effort to identify additional studies.

The funnel plot indicated that studies were missing, and using the most plausible model of dependencies between sources, Capture-Recapture resulted in an estimate that that two studies were missed overall (95% CI 0, 8).[215] The capture-recapture model that seemed most plausible showed a positive dependency (interaction) between databases and experts as well as experts and hand searching.[215]

Spoor *et al.* applied the capture-recapture technique to evaluate the completeness of a systematic review search in randomized controlled trials published in the journal Diabetic Medicine. They compared a MEDLINE search using the original Highly Sensitive Search Strategy (HSSS) for detecting randomized clinical trials[64] to the results of hand searching, and estimated that the population of trials was 160 (95% confidence interval 158-164) and the number of studies still unidentified after the two searches at 2 (0-6).[3] They obtained identical results with maximum likelihood estimator and with Chapman's method for small samples.[213]

Capture-Recapture will be used here to compare the estimates of various combinations of searches against the number found through all test searches. Dependencies between methods, which should be positive in the case of searches of the same type (i.e., between similarity methods) while less dependency is expected between different type of searches (i.e., dependency between similarity searches will be more that between a similarity search and a Boolean search). The expectation is that using two methods of ascertainment that are relatively independent will provide more complete retrieval than using two methods of ascertainment that have greater dependency.

5.1.5.1 Data analysis for Capture-Recapture

Spoor *et al.* estimated the total number of randomized controlled trials using the maximum likelihood estimator N=M(n/m) where:

M = number of publications identified in MEDLINE

n = number identified by hand searching

m = number identified by both

The estimate of the total population size is:

N=(M+1) (n+1)/(m+1) -1, rounded to the nearest whole number.[213]

Their results were easily reproduced using this equation, however Spoor *et al.* provide no information on the calculation of the confidence interval around this estimate, so other measures were sought.

Jensen reviewed unbiased estimators and presented a method for calculated confidence intervals for Bailey's nearly unbiased estimator for sampling with replacement.[221]

Under Bailey's method, the derivation of the point estimate for the total population size is slightly different from Spoor's method:

(2)
$$N = \frac{M(n+1)}{m+1}$$

The confidence interval around N is bounded by the reciprocals of

$$\frac{Mn}{m + 1.96\sqrt{(nM/N)(1-M/N)}}$$

Using these methods on Spoor's data, the resulting estimate of the total number of RCTs and the 95% confidence interval are 160 (157, 164), very similar to Spoor's result of 160 (158-164).

This method was extended to the current data. Since the best estimate should be obtained by the two methods with highest recall and most likely to be independent, Related Article RCT and Clinical Query were used. Targets retrieved by either Related Article RCT or Clinical Query were selected from all sets and a cross tabulation of Related Article RCT and Clinical Query was performed to determine

M = number of publications identified in Related Article RCT

n = number identified by Clinical Query

m = number identified by both.

While this approach, of selecting the two methods that appear independent and have the highest recall, may be reasonable, no precedents were identified in the literature. Another approach is to collapse methods. Bennett *et al.* used this approach in comparing sources searched for evidence that Progressive Resistance Training reduces disability in the elderly. In this analysis all ascertainment methods were collapsed into three approaches; 1) databases, 2) hand searching of conference abstracts or reference lists and 3) nomination by experts. Bennett does not fully explain the rationale for collapsing methods, but presumably, this was done to reduce the complexity of the model and deals with positive dependencies between sources within an approach.

In our sample, the true number of relevant articles is most closely approximated in the AHRQ Evidence Report set. In that set, all records retrieved by the searches were assessed directly. In the updated Cochrane cohort, new material was identified by the author searches for the updated systematic reviews. In the main cohort, newer systematic

reviews on the same or similar topics were first examined, and in many cases is was not necessary to examine all candidates to determine that the review was out of date. Therefore, in the Cochrane and Main cohorts, it is possible or even likely that there were additional records in the retrievals that were relevant.

In this experiment, the test searches can be grouped into subject searches (represented by Clinical Queries and Abridged Index Medicus RCTs although Abridged Index Medicus RCTs is redundant) and similarity searches (represented by Related Article RCT and SVM200point5). Citing References was a third method, but it is not included as its yield was poor and computing a maximum likelihood estimator based on three samples exceeds my computational abilities.

In order to determine the yield by approach, a variable named RI RCTorSVM200point5 was computed as the maximum of Related Article RCT and SVM200point5, and so equals 1 if a target had been retrieved by either of these search methods. Then all targets that were detected by either Clinical Query or Related Article RCT or SVM200point5 were retrieved, and a cross tabulation of Clinical Query and RI_RCTorSVM200point5 was performed. The total number of targets identified by each approach and their intersection was derived from the tables. This was entered into the equation for a two-source capture-recapture analysis, again using Bailey's estimate and Jensen's method for confidence intervals.

The above analysis provides an estimate of the total number of studies that might exist, some of which might not have been found. However, this and other analyses also provide an opportunity to explore the behaviour of the Capture-Recapture method when searches with demonstrated poorer recall are used to estimate the population size. The tendency to over- and under-estimate can be exploited to explore the dependencies between searches by comparing estimated with actual number, where underestimates indicate dependence between searches. In particular, the estimate obtained by using the Authors' MEDLINE search (Actual) and the Update search in CENTRAL, and Clinical Queries and the subject search in CENTRAL, and Clinical Queries and the subject search

in CENTRAL, will be tested to see if these can be considered as independent approaches to trial identification.

5.1.6 Multidimensional Scaling

Multidimensional scaling (MDS) is an exploratory technique used to visualize proximities. As few assumptions need to be made about data distributions, it is less restrictive than Factor Analysis. Factor analysis works from correlations and requires that the underlying data are distributed as multivariate normal, and that the relationships are linear, therefore it is easily distorted with small samples.[[222] p. 603] MDS works from similarity measures.[223] Thus, multidimensional scaling can situate the searches in two or more dimensions. Examining where different search methods lie on the dimensions may be useful for interpreting factors that influence the overlapping and unique components of the various searches. That understanding may inform optimum combinations of searches.

The Jaccard coefficient is a similarity measure that has been used in information science for co-citation analysis[224] and to examine the overlap between databases,[225] between retrieval methods, and between library collections,[226,227] so will be used as an example of the computational method used. The Jaccard coefficient (Sj) measures similarity of sets by dividing the size of the intersection by the size of the union of the sets. What is being measured is the similarity of the retrieved sets. In this case, the sets are the retrieved records of the searches under study.

(4) $Sj = a/(a+b+c)$, where

a = number of retrievals common to both searches

b=number of retrievals unique to first search method

c=number of retrievals unique to second search method. (Adapted from [225])

The Jaccard coefficient varies between 0 and 1 with larger values indicating greater similarity. The Jaccard coefficient has two features appropriate for describing overlap: it is self-normalizing and as it considers only records retrieved by one or the other of the pair of searches ignoring records retrieved by neither, (i.e., mutual lack of

ownership or joint absences is not a basis for similarity[227]) bias is avoided in small sample sizes.[224,228] Equal weight is given to matches and nonmatches.[197]

The proximity matrix is a similarity matrix created in SPSS using the variables (called stimuli) and then analyzed using the SPSS procedure ALSCAL (for alternating least squares scaling). The proximity matrix used as input for the MDS procedure was required to have at least five variables or stimuli, which were the searches. As the main cohort had only four searches for primary studies (Related Article RCT, Clinical Query, Abridged Index Medicus RCT, and Citing RCT), the two searches used to identify newer meta-analyses were added to the modelling. The first search used the subject search developed for Clinical Query but had the MEDLINE Meta-analysis publication type added as a limit. The second search to find newer meta-analyses was the Related Articles search result, limited to publication type meta-analysis. In both meta-analyses searches, the earliest date for the search was one year after the search data for the systematic review being updated. A separate multidimensional scaling analysis was run combining the two cohorts in which SVM was tested (the AHRQ Evidence Reports and the six updated Cochrane reviews), adding the SVM200point5 search. The searches for newer meta-analyses were not run for these cohorts, so were not included in the modelling.

The default two-dimensional Euclidean distance model was used for the analysis.[197] Euclidean distance is the 'ordinary' distance between two points.[229] In this model, input is assumed to be square symmetric matrices with data elements that are at the ordinal level of measurement (which is suitable for binary outcome data). The default output includes the improvement in Young's S-stress for successive iterations, two measures of fit for each input matrix (Kruskal's stress and the squared correlation, RSQ), and the derived configurations for each of the dimensions.[197]

Goodness of fit was assessed by examining the stress values. High stress values reflect that the input proximities cannot be well represented with a small number of dimensions. While MDS, as its name implies, can scale on multiple dimension, with so few input variables two or, at most, three meaningful dimensions would be expected.

Resulting dimensions of the derived stimulus configuration plots were visually inspected for interpretability and clustering.[230] If necessary, correlations between the scores of the searches on the resulting dimensions and aspects such as retrieval size or recall will be examined *post hoc*.[231]

5.1.7 Correlations

In work by Cohen *et al.* studying the potential for reducing workload in systematic review preparation using automated citation classification,[15] there was a very high correlation (0.91) between precision of the original query developed by the librarian for use in performing the systematic review, and the precision of the classifier tested. This result is not reported directly, but can be derived from Table 7 of the paper. This raises the concern that some topics are simply easier to search than others – when the librarian who constructed the search used in the review was able to achieve good precision, the classifier was also able to achieve good precision. The utility of a similarity search or ranking scheme is limited if it works well only in those cases where human searchers can achieve relatively high precision. Therefore, the correlation between precision of the *Actual* search and the precision of SVM and Related Article RCT will be examined in the updated Cochrane reviews

5.2 PRECISION OF SEARCHES IN A CROSS-SECTIONAL SAMPLE

There are no published norms for recall and precision of systematic review searches against which the performance of the test search can be compared (although most sources state that searches should be of high recall and so will usually sacrifice precision.[79,232] In 1999, the QUORUM standard for reporting systematic reviews was published, mandating an accounting of the flow of bibliographic references through the screening process.[42] Precision can be calculated from these data.

A cross sectional sample of systematic reviews first indexed in MEDLINE in December of 2004 provides a recent cohort for such an examination.[3] Three hundred systematic reviews were included, of which 125 (41.7%) were Cochrane reviews.

Investigators recorded whether the published review had a description of review flow in text, tables or in a QUOROM flow diagram. Of 300 included studies, 109 reported on this aspect, and 20 of these provided a flow diagram.

In new work, I examined the full text of these 109 systematic reviews to determine, where possible, the total retrieval of all searches, the number of unique references retrieved from all searches (whether electronic or manual), the number of records screened after duplicates were removed, the number of passing initial screening on the basis of the bibliographic record and obtained in full-text version for more detailed evaluation by reviewers and finally, the number of studies included in the systematic review. Data were taken from the QUOROM flow diagram, where available, or were extracted from the text of the review or determined by counting references to included studies.

There were many ambiguities in the authors' descriptions. Where it was unclear whether the number being reported reflected that total number of unique records retrieved or the number examined in full text, efforts were made to roughly replicate the MEDLINE search to determine which was more probable. Wherever possible, the N eligible was taken as the number of papers found eligible, but sometimes only the number of eligible studies were reported, and this number was used. Often it was not clear if papers or studies were reported (the revised PRISMA encourages more precise reporting in this regard[44]). The descriptions of article flow and the way these were coded for the final four systematic reviews examined is shown in Table 6.

Table 6. Examples of Authors' Descriptions of Study Flow

Text
"The search of the Cochrane Library … failed to yield any studies related to the subject of the review. The final search in MEDLINE produced 1889 papers. The abstracts were assessed and the relevant papers were narrowed down to 85. … Copies of the papers were obtained.… Before reading the papers, their references were screened for relevant non-retrieved papers. These search produced another 50 papers. Their abstracts were read and it was decided to retrieve only 31 of them.… The handsearch of … did not produce any studies that were not already identified as

relevant or not relevant by the electronic search. Out of total of 116 papers, 23…"[233]

N retrieved before duplicate removal	N of records screened	N articles retrieved and screened	N eligible
n/a	1946	116	23

"Results: The literature search revealed 1,984 abstracts. Included were 244, 42, and 12 articles in the first, second and third selections steps, respectively."[234] Steps 1, 2 and 4 corresponded to the abstract, full text and final inclusion.

N retrieved before duplicate removal	N of records screened	N articles retrieved and screened	N eligible
n/a	1984	244	12

"Thirty-six relevant randomized controlled trials have been presently identified, from an original pool of 1365 studies generated by the search (which identified 27 trials) and from updates to the search conducted over 2000 to August 2002 for the CCDAN Controlled Trials Register."[235] The exact total screened is not known but 1365 would be the minimum, and this figure is used. Trials, not eligible publications are reported but the number of eligible articles was determined by counting references in the Included Study list.

N retrieved before duplicate removal	N of records screened	N articles retrieved and screened	N eligible
n/a	1365	n/a	54

"We screened 645 abstracts and articles from which…. The results of this literature review are summarized in tables 1-5."[236] The number of references in those tables was counted to determine the number of included studies.

N retrieved before duplicate removal	N of records screened	N articles retrieved and screened	N eligible
n/a	645	n/a	14

"We identified over 3900 papers using the keyword and reference list search. From perusal of the abstracts, over 70 papers were retrieved and of these 19 papers met the study inclusion criteria."[237]

N retrieved before duplicate removal	N of records screened	N articles retrieved and screened	N eligible
n/a	3900	70	19

Precision was calculated when the number of included studies and the number of unique retrievals could both be determined, and was simply the ratio of these two numbers. Overall and median precision were calculated and data are presented graphically.

Each systematic review in the epidemiology study was classified as to primary focus; treatment or prevention, diagnosis, epidemiology or other. Other included systematic reviews focusing on harms, education, associational studies, instrumentation and research methods. Median precision was calculated for the reviews of each type.

5.3 RECALL OF THE SEARCHES IN THE ORIGINAL REVIEWS

Relative recall is the recall of one component of a multifaceted search compared to the recall of all methods.[192,238] Relative recall of the MEDLINE search can be calculated for the two cohorts were the searches used by the authors of the original reviews were replicated, the updated Cochrane reviews and the AHRQ cohort. For these two cohorts, the MEDLINE indexing status of all included studies was determined as part of the identification of the SVM true positive training set (Figure 6, Chapter 3). As has been previously described, the MEDLINE search result of the authors' search, less the records for the included studies, formed the true negative training set. Those included studies indexed in MEDLINE but not retrieved by the authors' search were identified by taking the true positives set NOT the true negative set.

The recall of the authors' searches was calculated as follows:

$$\text{Recall} = \frac{\textit{Studies retrieved by the replicated Original search}}{\textit{All MEDLINE}-\textit{indexed included studies}} \quad (5)$$

5.4 RELATED ARTICLE SEARCHING AS AN ADJUNCT SEARCH IN THE ORIGINAL REVIEW

The utility of Related Article searching as an adjunct to the Boolean search in original systematic review was explored. Once an original review is at the stage where screening has been complete, the searcher is in a position to identify the three largest and three newest eligible trials. It would be possible to run the Related Article search at that stage, and identify material not already captured by the other searches. Kastner *et al.* provide an example of such as search for missed material, although studies found through

the database searches that has passed initial screening and were retrieved in full text were used as seeds, and the procedure was run before the non-database searches (cited references, grey literature sources, trial registries and contacting experts) were undertaken. They note that all of the relevant new studies found were published after the date of the database searches, i.e., no trials missed by the Boolean database searches were found by the related article searches.[216]

In the updated Cochrane reviews, the Related Article RCT procedure was run and date limited to the period covered by the searches of the original review. Retrieval size, recall of included studies, recall of MEDLINE misses from the original and position of MEDLINE misses within the retrieved set was calculated to determine if the technique could identify MEDLINE-indexed material not found by the authors' MEDLINE search.

5.5 PERFORMANCE OF THE HSSS REVISED

The 2006 revision of Cochrane Collaboration's revised Highly Sensitive Search Strategy ($HSSS_{2006}$) was used as a pre-filter for the SVM tests, however, at the time there had been no independent validation of that revised filter. In addition, although analyzing the reason for retrieval failure in all MEDLINE misses is beyond the scope of this thesis, these data do provide an opportunity to see how often the methodological filter would have been a cause of retrieval failure. It is possible that the subject search would also have failed to identify the record; however, that would not be possible to test without replicating all of the searches.

The filter tested was derived from the paper and erratum[21] that presented the various strategies for identifying randomized controlled trials from MEDLINE. The strategy with the highest recall was assessed as the $HSSS_{2006}$. In Glanville *et al.*'s sample, the highest recall was achieved with *Strategy A* (recall of 0.993 and precision of 0.213) and this strategy was tested.

Glanville *et al.* presented the strategy in PubMed format. For the purposes of this work, the search was translated into OVID MEDLINE syntax. This was done in as a four-stage process. First, the string, as presented in the erratum, was submitted to PubMed, and

the search was limited to retrieve MEDLINE records only. The number of retrieved records was noted. Second, the "Details" box of PubMed was examined. PubMed translates the users' input, and the actual search run is presented in the Details box. That search string was submitted to PubMed as a series of single search commands (Table 7), which were then re-combined using Boolean expressions to verify that the single statements produced the same number of records as the original submission. The number of record retrieved by each search command was noted, so that they could be compared with the OVID counts. Third, these single term queries were combined, and the total was verified as being the same as the original PubMed search. Finally, each single term was submitted to OVID MEDLINE and counts were verified against the PubMed counts to confirm that the translation was accurate.

Table 7. HSSS2006 Counts for Each Element in PubMed and Ovid

Query*	PubMed	PubMed limited to MEDLINE	Query	Ovid
	N	N		N
#21 not #24	1941487	1882219	9 not 12	1859937
#22 not #23	3142301	3129042	10 not 11	3077794
humans [mh]	9624337	9624336	humans.sh.	9694282
animals [mh]	12766638	12753378	animals.sh.[†]	4055990
#14 or #15 or #16 or #17 or #18 or #19 or #20	2268708	2209409	or/1-8	2188346
Groups [tiab]	834178	790190	groups.ti,ab.	789810
trial [tiab]	199739	190416	Trial.ti,ab.	190850
Randomly [tiab]	116285	108677	randomly.ti,ab.	109125
dt [sh]	1155260	1155260	dt.fs.[‡]	1163190
placebo [tiab]	103776	100300	placebo.ti,ab.	100919
randomized [tiab]	164877	157307	Randomized.ti,ab.	158443
clinical trial [pt]	494988	494359	clinical trial.pt.	434900

*Search strings are presented in the reverse order of entry.
[†]exp Animals/ yields 12865972 hits and the final number when exp animals/ is used is 1859169.
[‡]dt.sh. results in zero hits in Ovid, so dt.fs. was used.

Two anomalies were noted in the translation. First, PubMed explodes all subject headings unless a no-explode command is specified. In Ovid, exploding the animal term results in a very large retrieval. It was determined that searching Animal.sh. yielded the same final numbers, once all search terms were combined, as were obtained by searching

animal[mh] in PubMed, although the counts for the single query statement are quite different in the two interfaces. The second adjustment needed was the subheading DT (for drug therapy) is searched as a subject heading in PubMed (dt[sh]). A literal translation of this to dt.sh. yielded no hits in Ovid, but searching DT as a floating subheading (dt.fs.) yielded a comparable result to the PubMed search string.

Searches were conducted April 30, 2007. At that time, the Ovid database coverage dates were 1950 to April Week 3 2007. The retrieval size of the PubMed version of the HSSS was 1882219 when limited to MEDLINE, and the OVID translation yielded 1859937 (a ratio of 1.012).

The final OVID version of the filter was:

(clinical trial.pt. or randomized.ti,ab. or placebo.ti,ab. or dt.fs. or randomly.ti,ab. or trial.ti,ab. or groups.ti,ab.) not (animals/ not humans/)

The translated search was tested against relevant evidence from three sources; the MEDLINE-indexed signalling evidence (n=58), targets from Cochrane SVM collection (originally n=27, before "old" studies added to the review were excluded), and all *Eligible* records from the screened material (n=687). After overlap was removed, 695 pieces of MEDLINE-indexed relevant evidence remained.

PubMed IDs for these 695 records were submitted as a query, the $HSSS_{2006}$ was run, and the PubMed IDs missed by the $HSSS_{2006}$ were determined through a NOT statement. Missed records were downloaded for examination. Recall of the $HSSS_{2006}$ was calculated:

(6)
$$\mathrm{Re}\,call = \frac{PubMed\ IDs\ that\ passed\ the\ HSSS_{20006}\ filter}{All\ tested\ PubMed\ IDs}$$

A similar procedure was used to test the original version of the highly sensitive search strategy. Retrieval size of the two versions of the HSSS was compared, without duplicate removal as set size exceeded the limit for duplicate removal in OVID MEDLINE 1950 to January Week 2 2009.

5.6 CHARACTERISTICS OF THE EVIDENCE IN UPDATED SYSTEMATIC REVIEWS

5.6.1 Where Does the New Evidence Come From?

The PubMed IDs for all *Eligible* new evidence were extracted from the big dataset and were submitted to PubMed. Records were downloaded and imported into Reference Manager where duplicates between cohorts were removed. After duplicate removal, the Reference Manager indexes were purged, and frequencies of authors and journals were prepared using the Reference Manager feature of "print to file, include reference count". Numbers of authors and journals were determined by examining the database properties.

The distribution of journals was examined according to Bradford's law which states that journals in a single field can be divided into three parts, each containing the same number of articles: 1) a core of journals on the subject, relatively few in number, that produces approximately one-third of all the articles, 2) a second zone, containing the same number of articles as the first, but a greater number of journals, and 3) a third zone, where the ratios of journals in the three zones are $1:n:n^2$, where n is some constant.[239] In Bradford's study of physics journals, 9 journals contributed one-third of the articles, 5 times 9 (45), produced the next third, and 5 times 5 times 9 journals (225), produced the final third.[240]

Journal impact factor of the journals contributing new evidence were examined, based on the 2007 Journal Citation Report.[241] The proportion of evidence from Abridged Index Medicus journals was calculated. The Abridged Index Medicus title list of February 2008 was used.[13]

5.6.2 Does Old Evidence Persist?

The phenomenon of trials included in the original review but excluded from updates was noticed. These are the trials coded as *Original Only*. Two possible explanations for such exclusions from subsequent updates are, first, that the inclusion criteria of the systematic reviews changed in subsequent versions. For instance, in the update, only placebo controlled trials are eligible, whereas trials using placebo or active

controls were included in the original. A second reason for subsequent exclusion would be that an error had been made and a trial considered eligible in the original could be found to be ineligible later on.

Frequency and characteristics of *Original Only* evidence, including publication date relative to other studies included in the review, indexing status (indexed in MEDLINE or not), and sample size are examined. Cochrane reviews list study exclusion reasons, and these were recorded for each of the *Original Only* studies. Finally, the distribution of signals for updating was compared for those updated Cochrane reviews with and without *Original Only* studies, through cross tabulation and χ^2.

5.7 IS MATURITY OF THE LITERATURE A PREDICTOR OF SURVIVAL?

The need for update is a function of the stability of the evidence. If little or no new evidence is emerging, or the evidence is largely confirmatory, representing more of the same, than updating a review is unlikely to result in treatment implications. For the main survival analysis,[17] all aspects of survival were studied relative to either the date of the search or the date of publication of the systematic reviews in the cohort. The maturity of the evidence base in the original review may also warrant consideration as a factor in the survival of the findings of the review. Reviews done very close to the first evidence may be less stable (likely to go out of date more quickly or more easily) than those done long after the first evidence emerged.

Expecting that factors may not change in linear fashion as the time since the first studies increases, age of the oldest evidence was examined categorically, with the observed range divided into quintiles. Those reviews used in the survival analysis were studied (n=100). A variable representing the span between the year of publication of the earliest evidence included in the review and the year of the search used in the review was computed in two steps. First, in the study-level dataset, all included studies that were either *Original* or *Original Only* were selected and cases summaries were created that reported the minimum publication year for these studies for each CohortID. This output was transcribed to the database containing the review-level data and age of the oldest trial

was computed as the year of the oldest trial minus the year of the search done for the systematic review, in whole years. These were sorted from shortest to longest time span and the rank of each observation was calculated. The array was divided into fifths, as close to the 20th, 40th, 60th, 80th and 100th percentiles of this distribution as possible. As whole years were used there were many tied values, consequently the number of systematic reviews in each quintile varies (Table 8). The first quintile includes the youngest evidence – those reviews where all included studies were published within the six years prior to the search. The fifth quintile contains the oldest evidence – the oldest trials included in these reviews were published at least 23 years prior to the search. For the review with the longest evidence base, the oldest trial preceded the search by 47 years.

Table 8. Distribution of Systematic Reviews by Age of First Evidence

		Maximum Age of First Evidence at Search	
Quintile		Years	N of Systematic Reviews in the Quintile
Youngest	1	6	25
	2	10	17
	3	16.4	18
	4	22	23
Oldest	5	47	17

Characteristics of these quintiles in terms of the mean and standard deviation of the number of studies included in the review and of the number of patients in the reviews were calculated. Survival analysis was conducted using years from the final search date to primary survival by quintile of oldest evidence. If there is a major or potentially invalidating signal, primary survival is calculated as the publication date of the original systematic review to the earliest of the date of new evidence resulting in final classification, the date of the qualitative signal, or the date of the quantitative signal involving the primary outcome. Otherwise, survival is the time from the final search date of the original systematic review to September 1, 2006

Distribution of signals (reviews with potentially invalidating or major new evidence (n=59)) across the cohorts was examined and tested using χ^2. The nature of new evidence emerging in the different quintiles was examined. The three qualitative signals that could occur in reviews with potentially invalidating new evidence were; opposing findings; substantial harm; or a superior new treatment. The four qualitative signals for major changes in evidence were; important changes in effectiveness short of 'opposing findings'; clinically important expansion of treatment; clinically important caveat and finally; opposing findings from discordant meta-analysis or non-pivotal trial. This is presented graphically.

Heterogeneity and publication bias are two issues that could be present in the evidence in the original review. Both introduce challenges in the interpretation of evidence. Heterogeneity is simply variation seen among the results of different trials in the meta-analysis. Some variation is expected due to chance, but additional variation may be introduced through differences in patient selection, aspects of study design, or other factors that complicate interpretation of a pooled result.[198] The Cochrane Handbook now distinguishes between clinical diversity, methodological diversity and statistical heterogeneity.[242] Diversity can contribute to statistical heterogeneity, and the main question is whether that diversity provides meaningful clinical information, or whether it obscures the true result.[242] For example, DerSimonian and Levine showed how differences in patient groups and in the way studies were conducted accounted for the strong positive effect of calcium supplementation for prevention of preeclampsia seen in a meta-analysis which was not substantiated in a subsequent large trial that controlled for clinical and methodological diversity.[243]

Statistical heterogeneity is a question of whether the estimates of intervention effectiveness from different sources vary more than would be expected by chance.[244] It can be tested multiple ways, but one method that is easily interpreted is a χ^2 value. If the χ^2 value is equal to the number of studies less one (the degrees of freedom of the test) than there is no measurable heterogeneity present. Larger values of χ^2indicate that more

heterogeneity is present.[244] Although some measures exist for managing heterogeneity statistically, namely, random-effects meta-analysis, it is also necessary to consider if studies should be combined.[242,243] Moher *et al.* found that heterogeneity was investigated in 68% of the 2004 cohort of systematic reviews but this figure rose to 91% in those systematic reviews where a meta-analysis was performed.[3]

Heterogeneity was assessed in 93% of the systematic reviews in the sample used here for the survival analysis. Presence or suspicion of heterogeneity was a factor in 61% of the reviews, and was a significant predictor of survival in the multivariate analysis.[17]

Publication bias is the greater likelihood that trials with certain characteristics (usually positive findings) will be published than trials without those characteristics.[7,220,245-247] Publication bias can also be seen in the delayed publication of trials with negative or ambiguous findings, relative to those with positive results.[248] Published trials are more easily located for inclusion in systematic reviews, so when signs of publication bias are observable, it may be that the results of meta-analysis do not represent the underlying truth, making these meta-analyses more susceptible to being overturned as subsequent results emerge.

Publication bias is usually assessed graphically, through funnel plots in which sample size or standard error (or some other measure of study precision) is plotted against treatment effect size. More scatter is expected by chance in smaller studies than in larger studies however, chance would scatter the estimates symmetrically.[75] Lack of symmetry in the lower part of the plot can indicate publication bias, or overestimation of effect in small studies due to methodological weakness (Figure 16).[249] Copas and Shi characterize this lack of symmetry as non-ignorable non-random missing data.[250] [Kjaerdard explains the statistical analysis of funnel plot asymmetry nicely[251]] Moher *et al.* found that 31% of the 2004 cohort of systematic reviews considered publication bias.[3]

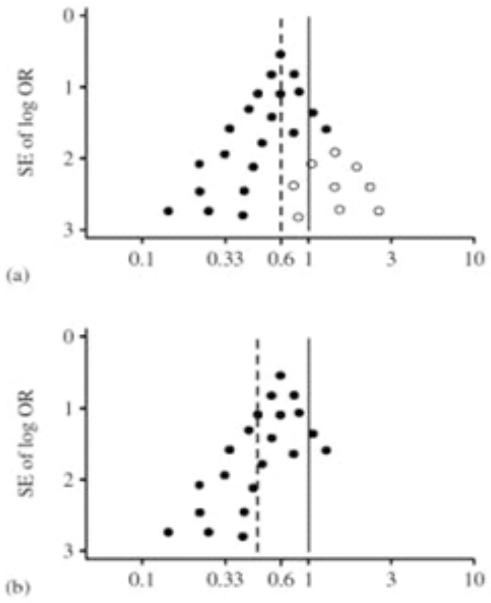


Figure 16. Hypothetical Funnel Plots. (A) asymmetrical plot in the absence of bias (open circles indicate smaller studies showing no beneficial effects; (b) asymmetrical plot in the presence of publication bias (smaller studies showing no beneficial effects are missing)

Publication bias was assessed as an explanatory variable in the survival analysis. The text of the systematic reviews in the main cohort was examined for any mention that the authors' assessed publication bias, and whether publication bias was described as present or suspected. Some assert that publication bias is present in almost all research, and can be suspected even if the funnel plot is symmetrical.[252] Sixty percent of the main cohort used in the survival analysis reported on publication bias, and it was present or suspected in 18% of the sample. Publication bias was not found to be associated with survival,[17] however as it is likely to be indicative of missing studies, it is relevant as an information retrieval issue.

Finally, the nature of systematic reviews surviving and those with signals for updating was examined. Conclusions from the abstract of the original reviews were presented alongside the qualitative signal for those systematic reviews with major or potentially invalidating new evidence. The nature of that evidence was taken from summaries prepared for each systematic review in the course of the updating process. The

summaries were written by a member of our project team, Jun Ji or Mohammed Ansari, and were scrutinized and finalized in case conference.

5.8 TECHNICAL NOTE - OPERATIONALIZATION OF DATES

Two aspects of dates are important for this thesis, as updating involves moving forward from a particular point of time. Calculating survival and quantifying the amount of new information both depend on a starting date and it is important that the starting dates be calculated consistently across reviews in the cohort. Secondly, this research took place over a span of time, and is retrospective. Therefore, it is necessary to determine what material would have been present in a database and would have been retrievable contemporaneously with the reviewers' search.

Date variables extracted from the systematic reviews include most recent search date, date of submission for publication, date of acceptance for publication, date of publication and date of indexing, date of first new evidence and final survival date. Some of these variables were used only for a study of time lags in the preparation and publication of systematic reviews.[19] Priority for date sources was the MEDLINE record, the publication itself, or an estimate. If partial dates were reported, the missing components were imputed as follows: if month is missing, the 6th month was assumed, if day of month was missing, the 15th was assumed. Dates were extracted and then verified by a second person. Data cleaning runs were undertaken in SPSS to test for improbable values such as an acceptance date earlier than the submission date. All such cases were due to date imputation and the imputed value was changed to the earliest logical value.

Search date was taken as the date reported in the methods section of the systematic review - the most recent date was used if more than one was given. Cochrane reviews have a number of date fields as part of the review template and the most recent among these three dates was recorded; the date reported in search strategy section in the body of the report or the "Date new studies sought but none found", or the "Date new studies found and included/excluded". The field "Date new studies found but not yet included/excluded" was ignored. For database dates, the end date reported for MEDLINE

searching was used (i.e., 1966-June Week 4, 2003) if available. If the MEDLINE date was not reported, any other database end date was used. If no end date was reported, the variable was treated as missing.

Submission date was the submission date or the date the manuscript was received as stated in the article. These data were available only for journal published reviews - AHRQ and Cochrane reviews had no equivalent. Acceptance date was the date stated in the article.

For all types of reviews, the publication date was taken from the Ovid MEDLINE records, for indexing date, the MHDA from the Ovid MEDLINE record was used. The publication date for original Cochrane reviews that no longer appeared in MEDLINE was taken as the date of most recent amendment. When no "most recent search date" could be determined, we used publication date minus one year. In the survival analysis, survival was calculated with publication date as the start date.[17]

For dating records of new material, it was necessary to determine not when the material was published, but rather when it was in the database, and when it was fully indexed. Several of the searches rely on the publication type *Randomized Controlled Trial*, which is assigned during indexing, and is not available for PubMed *in process* records. Related Article nearest neighbour scores are pre-computed for fast processing[167] however, even records as supplied by the publisher have related articles, and are retrieved as related articles. Still, we were interested in Related Articles that were also randomized controlled trials, so relevant related articles not yet fully indexed might be missed. Thus, for MEDLINE retrievals, MHDA was used instead of publication date or database entry date to establish the sequence and age of records.

Chapter 6: Results

6.1 THE COHORTS

6.1.1 Main Cohort

The search of ACP Journal Club yielded 651 records for potentially eligible systematic reviews. Records were downloaded sorted by author surname, to provide a quasi-random screening order. Three hundred and twenty five reviews were screened to achieving the target sample size of 100 reviews. 165 records were excluded on the basis of the ACP Journal Club record, and 160 articles were excluded after assessment of the full-text article. Exclusion reasons are shown in (Figure 17).

Seventy-two of the systematic reviews were journal-published reviews. Twenty-seven were Cochrane Collaboration Reviews, one was a health technology assessment published by the Canadian Agency for Drugs and Technologies in Health.

Survival data for this cohort have been published.[17] Fifty-seven (57%, 95% confidence interval, 47% to 67%) of the reviews had a signal for updating during the observation period, and median survival of these was 5.5 years (CI, 4.6 to 7.6 years) from publication. For the 57 reviews with signals for updating, median time to event was 3.0 years (inter-quartile range, 0.9 to 5.1 years) from publication.

Seventy-four of the systematic reviews were assessed entirely against new evidence found through four test searches; Clinical Query, Abridged Index Medicus RCTs, Citing RCTs, and Related Article RCTs. Three systematic reviews were first assessed through a known update, but then further update searching was done using the test searches to bridge from the end of the update to September 2006, the end of our search period.

These 77 systematic reviews are the basis for the main evaluation of these four test searches. Each of the 77 systematic reviews included a median of 13 studies (inter-quartile range, 8 to 21) and 2666 participants (inter-quartile range, 1284 to 8345) (Table

9). Most reviews evaluated drug therapies; the most common clinical content areas were cardiovascular medicine, gastroenterology and neurology. We were able to identify at least one new eligible trial for 70 systematic reviews, with a median of 5 new trials (inter-quartile range, 1 to 6) and 1,185 patients (inter-quartile range, 173 to 5,054) per review. These findings are very similar to the findings for the cohort of 100.

Table 9. Characteristics of the Cohort of 77 Systematic Reviews Updated by Searching

Characteristic	77 Updated By Search	100 Studied For Main Survival Analysis
	N	N
Publication type		
Peer-reviewed journal article	64	72
Cochrane review	12	27
Health technology assessment*	1	1
Therapy evaluated		
Medications	67	85
Medical devices	3	8
Procedures	7	7
Clinical topic area		
Cardiovascular	19	20
Gastroenterology	11	13
Neurology	10	11
Other 10 categories	<8 each	<10 each
Publication period		
January 1995–February 1997	13	16
March 1997–April 1999	15	22
May 1999–June 2001	18	25
July 2001–August 2003	16	20
September 2003–December 2005	15	16
Median included trials	13 (IQR, 8-21)	13 (IQR, 8–21)
Median included participants	2666 (IQR 1284-8354)	2663 (IQR, 1281–8371)

IQR is inter-quartile range

Twenty three systematic reviews were assessed exclusively against an updated systematic review, most of these were Cochrane reviews. All of these were included in the cohort survival study, but here only those meeting the additional selection criteria necessary for study with SVM were retained for detailed study.

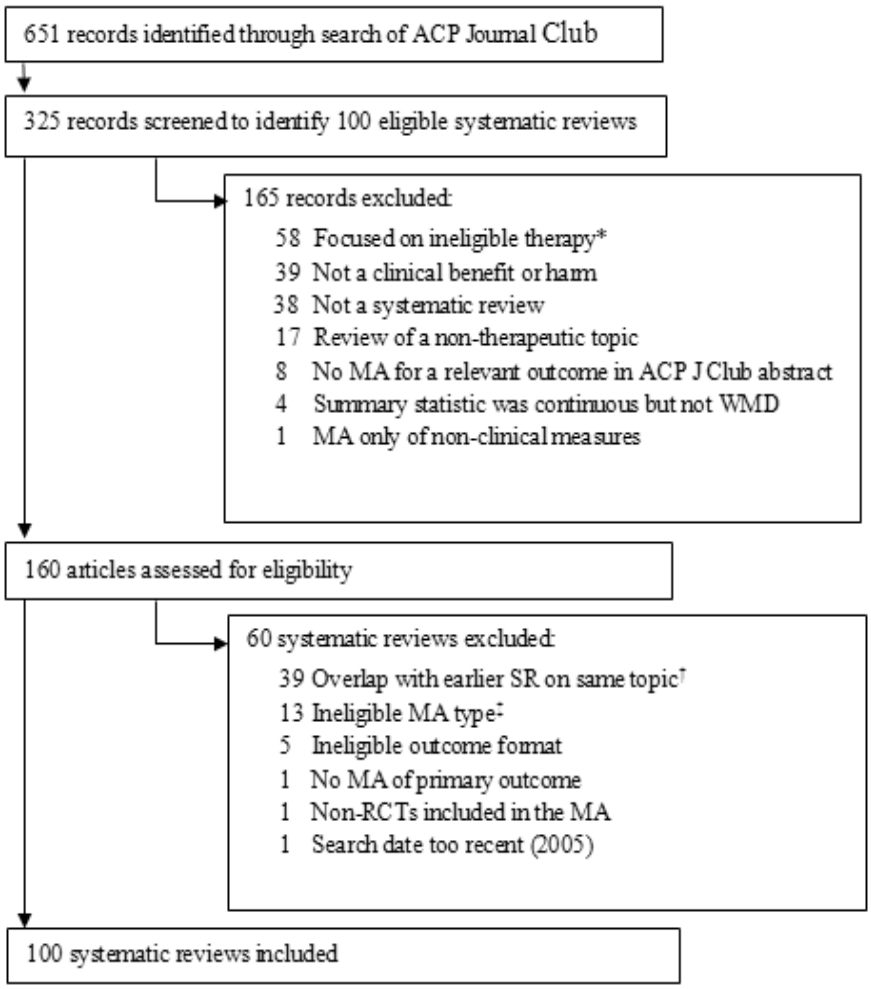


Figure 17. Flow of Articles in the Formation of the Main Cohort

*This category includes reviews not focused on a specific class of drug, device or procedure, as well as ones focused on educational or behavioral interventions, or complementary therapies.

†This category includes updates of systematic reviews already in cohort, topics similar to that of a systematic review already included, or the journal version of an included Cochrane review

‡ This category includes meta-analysis using individual patient data without regular meta-analysis, meta-regression, or indirect meta-analysis.

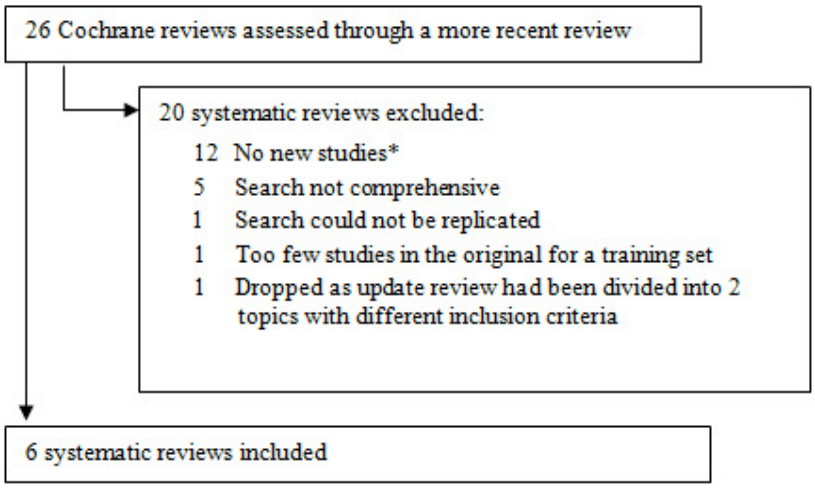


Figure 18. Flow of Articles in the Formation of the Updated Cochrane Cohort

Cochrane reviews for which the most recent update had the same studies as the original review were updated through test searches and are included in the cohort of 77.

6.1.2 Updated Cochrane Cohort

Twenty-six Cochrane reviews were assessed through an existing updated systematic review. These were screened against further eligibility criteria to assemble a cohort in which to study SVM. Six met all eligibility criteria and SVM searching was done. Exclusions reasons are shown in Figure 18.

Clinical areas of the individual reviews, year of publication of the original and updated reviews, number of newer studies included in the update, and final classification are shown in Table 10. These six systematic reviews included a median of 17 studies (inter-quartile range, 14 to 20) and 8679 participants (inter-quartile range, 4085 to 50,109) making them much larger than was typical for the main cohort. Again, most reviews evaluated drug therapies; the most common clinical content areas differed from the main cohort, with a clustering in infectious diseases (Table 11). All had at least one

new eligible trial, indeed this was an inclusion criteria for this cohort. There was a median of five new trials (inter-quartile range, 2 to 6) and a median of 2469 participants (inter-quartile range, 730 to 8731) per review. These are larger studies than were typical of the new studies in the main cohort.

Table 10. Characteristics of the Cochrane Reviews Updated by the Authors

Characteristic	6 Updated By Authors N	100 Studied For Main Survival Analysis N
Publication type		
Peer-reviewed journal article	-	72
Cochrane review	6	27
Health technology assessment	-	1
Therapy evaluated		
Medications	5	85
Medical devices	1	8
Procedures	-	7
Clinical topic area		
Cardiovascular	-	20
Gastroenterology	-	13
Neurology	-	11
Infectious Disease	3	9
Critical Care	1	8
Respiratory Systems	1	9
Urology and Nephrology	1	5
Other 6 categories	-	<10 each
Publication period		
January 1995–February 1997	-	16
March 1997–April 1999	3	22
May 1999–June 2001	1	25
July 2001–August 2003	2	20
September 2003–December 2005	-	16
Median included trials	17 (IQR, 14-20)	13 (IQR, 8–21)
Median included participants	8679 (IQR 4085-50,109)	2663 (IQR, 1281–8371)

IQR is inter-quartile range

Table 11. Characteristics of the Cochrane Update Cohort Outcomes

Topic	Year of Original	Year of Update	Interval (Years)	New Studies in Update Interval	Final Classification	Specific Criteria Met
Critical care	2001	2005	2.75	2	Potentially invalidated	A1
Respiratory	1999	2005	7.73	1	Minor	-
Urology & Nephrology	2001	2005	3.04	1	Major	A4
Infectious diseases	1997	2004	4.39	6	Minor	-
Infectious diseases	2001	2004	5.96	5	Minor	-
Infectious diseases	1999	2004	6.34	5	Minor	-

6.1.3 AHRQ Evidence Reviews Cohort

One hundred and twenty four AHRQ Evidence Reports were identified from the AHRQ web site[253] and screened against the inclusion criteria of the main cohort, except that they there was no requirement that they have been reviewed by ACP Journal Club.

Fourteen were assessed for signals for updating and ten of these met the additional inclusion criteria for the SVM study cohort. Exclusion reasons are shown in Figure 19. Clinical area of the individual reviews, year of publication of the original and final classification are shown in Table 12.

These ten Evidence Reports included a median of 96 studies (inter-quartile range, 31.75 to 121.5) and 22,830 participants (inter-quartile range, 14172 to 49687) making them much larger than was typical for the main cohort (Table 12) or the updated Cochrane reviews (Table 10). Again, most reviews evaluated drug therapies, however in this cohort there were reviews that examined multiple treatment modalities, such as medication or surgery. The most common clinical content areas differed from the main cohort, with the specialties of obstetrics/gynecology and psychiatry having greater representation. These Evidence Reports were somewhat newer than the range of ages

seen in the Cochrane and Main cohorts. The Evidence Practice Center program was established in 1997, and the first reports appeared in 1998 and 1999.[253] At least one new eligible study was identified for each Evidence Report. There was a median of 20 new trials (inter-quartile range, 6 to 41) per evidence report, and these new trials involving a median of 18,471 participants each (inter-quartile range, 1283 to 26,853). These are the largest studies found for any of the cohorts.

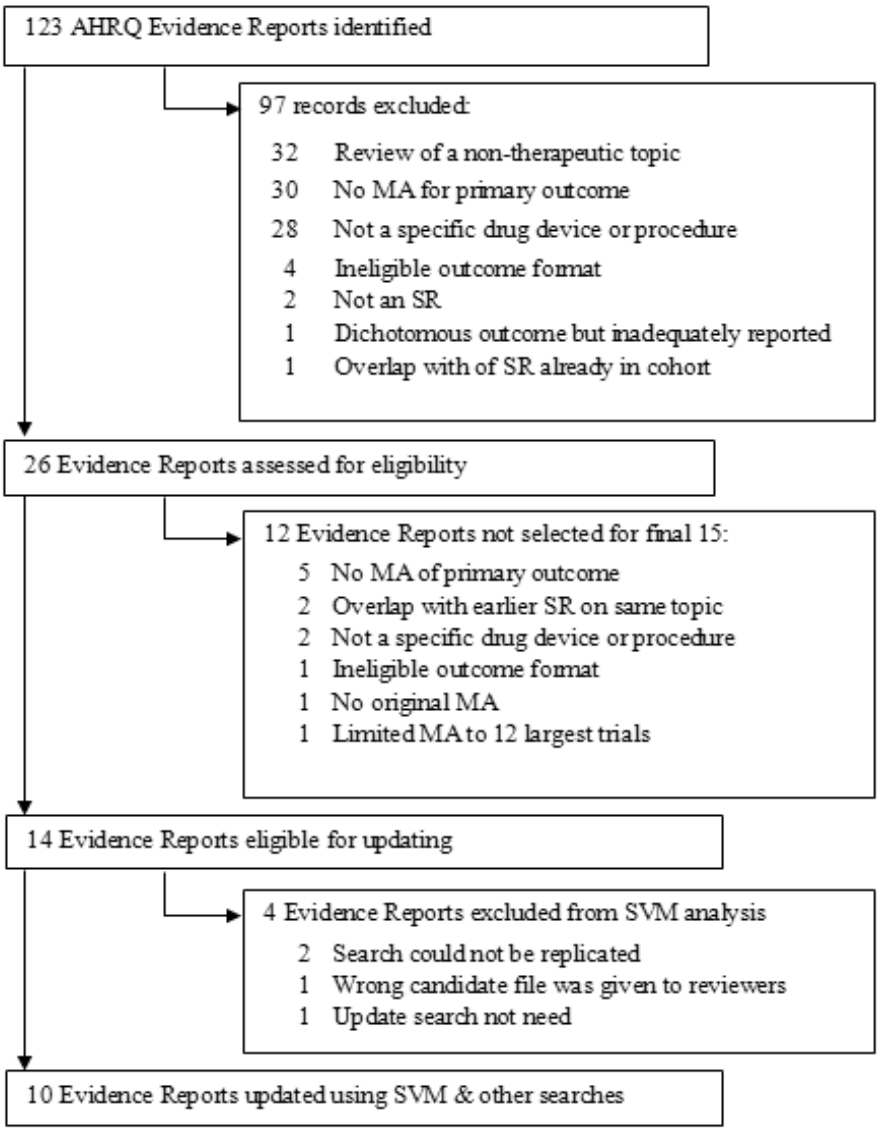


Figure 19. Flow of Articles in the Formation of the AHRQ Cohort

Table 12. Characteristics of the AHRQ Evidence Reports Updated by Searching

Characteristic	10 Evidence Reports	100 Studied for Main Survival Analysis
	N	N
Publication type		
Peer-reviewed journal article	-	72
Cochrane review	-	27
Health technology assessment	10	1
Therapy evaluated		
Medications	9*	85
Medical devices	1	8
Procedures	3	7
Clinical topic area		
Cardiovascular	2	20
Gastroenterology	-	13
Neurology	1	11
Obstetrics and Gynecology	3	5
Psychiatry	2	5
Other 8 categories	1 each	<10 each
Publication period		
January 1995–February 1997	-	16
March 1997–April 1999	-	22
May 1999–June 2001	3	25
July 2001–August 2003	3	20
September 2003–December 2005	4	16
Median included trials	96 (IQR, 31.75-121.5)	13 (IQR, 8–21)
Median included participants	22830 (IQR 14172-49687)	2663 (IQR, 1281–8371)

IQR is inter-quartile range

*One Evidence Report included both drugs and devices, another Evidence Report included both drugs and procedures.

Table 13. Characteristics of the AHRQ Cohort Outcomes

Classification	Year of Original	Final Classification	Specific Criteria Met
Neurology	2004	Major	A4
Cardiac & Cardiovascular Systems	2003	Major	A4 And B2
Obstetrics and Gynecology	2000	Potentially invalidating	A2
Psychiatry	2004	Minor	-
Psychiatry	2003	Major	A6
Obstetrics and Gynecology	2003	Potentially invalidating	A1
Cardiac & Cardiovascular Systems	2000	Major	A4 and A6
Oncology	2003	Major	A4
Endocrinology	2004	Major	A5
Obstetrics and Gynecology	2000	Potentially invalidating	A1

6.2 NEW EVIDENCE

Across the three cohorts (Main77, Cochane6, AHRQ10) and all methods of identification (searches, reviewer nomination and updates) 18,960 new records were identified. Of these, 11,586 new records were evaluated and 860 eligible new records, representing 601 different studies, were identified. Median publication date of new eligible studies was a decade newer than the median date of studies in the original reviews. Median sample size increased 78% from 100 participants in studies included in the original reviews to 178 participants in the newer studies identified in updating (Table 14).

Table 14. Quantity, Sample Size and Age of New Evidence

	Original Evidence		New Evidence	
	N or Median	1st and 3rd Quartiles	N or Median	1st and 3rd Quartiles
Eligible Records	3117		860	
Eligible Studies	2758		601	
Publication Date	1993	1988, 1998	2003	2000, 2004
Sample size	100	40, 313	178	71.25, 538.25

Most eligible new evidence was identified through the test searches (Table 15), with important exceptions. In the main cohort, 8% of new studies were identified by reviewer nomination – these were mostly newer primary studies identified by examining the included studies of newer meta-analyses on a similar topic, but could also be studies known to our review team but not identified by the test searches. These reviewer nominated studies represented incomplete identification of new evidence by the test searches and were included in the demoninator for calculations of recall.

In the main cohort, some new evidence was identified from updates. Nine systematic reviews were assessed exclusively on the basis of newer systematic reviews on the same topic. When two systematic reviews on the same topic passed screening and were eligible for inclusion, the older was kept in the cohort and the newer was excluded but retained as a source of new studies – three such journal-published reviews were updated exclusively on the basis of these newer reviews on the same topic, and six Cochrane reviews had updates sufficient to determine their survival, but were not eligible for the updating cohort. These updates identified 95 new studies across the nine systematic reviews. Three systematic reviews had updates, but required additional searching to extend the period of coverage, although the updates identified 13 new studies. Two of these were Cochrane reviews and one was a journal-published systematic review.

The Cochrane Update cohort only included studies found in updates, by definition. For the AHRQ cohort, all candidate studies were screened, and new meta-analyses were not provided as a short cut to identifying signals for update, and as a result, there were no reviewer nominations in this cohort, and no studies identified by updates (Table 15).

Table 15. How Eligible New Evidence was Identified

	Candidates from Searching N (%)	Reviewer Nomination N (%)	Updates N (%)	Total Total (%)
Main Cohort	416 (90.6)	43 (9.4)	0 (0.0)	459 (100)
Cochrane Updates	0 (0)	0 (0)	20 (100.0)	20 (100)
AHRQ Evidence Reviews	270 (100.0)	0 (0)	0 (0)	270 (100)
Total	686 (91.6)	43 (5.7)	20 (2.7)	

Distribution of the new evidence across the three cohorts is shown in Table 16. Although the Main Cohort includes 83% of the reviews, only 61% of the new eligible studies and 69% of new N are associated with that cohort. The Cochrane cohort of systematic reviews updated by the original review teams represents 7% of reviews examined, yielded only 3% of the eligible new studies (targets for retrieval), but 6% of new participants – suggesting fewer studies but with larger sample sizes were included. The AHRQ10 cohort, in which all new candidate studies were screened, made up 11% of the reviews, but 37% of the new evidence and 25% of the new participants. The disproportionate amount of new evidence in the AHRQ cohort relative to the main cohort likely reflect the complete screening of the candidate list in the AHRQ cohort – identification of new evidence for the main cohort is likely incomplete as it relied somewhat on rapid identification of newer evidence through newer systematic reviews (see Section 4.5.1).

Table 16. Distribution of New Evidence Amongst the Cohorts Studied

Cohort	% of Cohort	% of Targets	% of New N
Main Cohort (N=77)	82.8%	60.5%	68.7%
Cochrane Updates (N=6)	6.5%	2.5%	6.3%
AHRQ Evidence Reports (N=10)	10.8%	37.0%	25.0%
Total	100.00%	100.0%	100.00%

6.3 SEARCH PERFORMANCE

6.3.1 Characteristics of the Subject Searches

Characteristics of the 77 subject searches for main cohort have been previously reported.[18] The mean number of terms used was 3.6 (standard deviation 1.59). Thirty-six of the 77 searches (46%) used free text terms, 58 (74%) used exploded MeSH terms, 21 (27%) used starring, and 2 (3%) used subheadings. The mean (standard deviation) of features used was 1.51 (0.58). The specificity clinical query was used for 13 (17%) of the searches (Table 17). Examples of subject searches are available electronically as Appendix C of the technical report,[16]

For the AHRQ cohort, the mean (standard deviation) number of terms used in the Boolean test searches was 12 (7.5). All of the searches (100%) used both free text terms and exploded MeSH terms, none (0%) used starring or subheadings. The mean (standard deviation) of features used was 2 (0.0). The specificity clinical query was used for 6 (60%) of the searches. These searches were quite different from those used in the main cohort, relying on devices that increase recall (such as explosion and free text terms), and then using a hedge designed for higher precision. This may reflect the increased number of interventions examined in the AHRQ reports, or it could reflect a shift in style over the course of the project, as the searches were done over eight months and the AHRQ searches were done last. The correlation between number of terms and numeric sequence in the main cohort is -0.04, suggesting that complexity of the Evidence Reports is the more likely explanation.

6.3.2 Characteristics of the Authors' Original Boolean Searches

Original searches were replicated for the Cochrane and AHRQ cohorts. More terms were used in the author searches than in our test searches – in the case of the AHRQ Evidence Reports, the only direct comparison possible, the authors of the original reviews used more than twice as many terms as were used in our searches. Similar numbers of search features were used in the author searches as in our test searches. Free

text search strings were used in all of the AHRQ and half of the Cochrane searches. Exploded MeSH terms were employed universally. Limiting to major MeSH terms was never used and subheadings were used sparingly – in one (10%) AHRQ search and two (33%) of the Cochrane updating searches prepared by the review authors.

Table 17. Comparison of Boolean Search Features

	Test Searches		Author Searches	
	Main 77	AHRQ	AHRQ Original	Cochrane Update
Mean N of Terms (SD)	3.6 (1.59)	12 (7.5)	27.4 (15.7)	18.3 (22.1)
Mean N of Features (SD)	1.51 (0.58)	2 (0.0)	2.1 (0.32)	1.7 (0.52)
Methodological Filter Used	100%	100%	70%	100%
Language Limit Applied	0%	0%	50%	0%
Human Limit Applied	0%	0%	90%	66.6%

6.4 PERFORMANCE OF THE TEST SEARCHES

Several aspects of performance were considered. Recall of new *On Topic* and *Eligible* studies, recall of new participants, and overall retrieval size all inform slightly different aspect of search performance (see Section 4.4.1.1). Somewhat different information is available for the different cohorts, for example, the search strategies of the original reviews could not be replicated for many of the systematic reviews in the main cohort, so, while it gives us a good overall picture of the performance of the various strategies, the test strategies cannot be compared directly to the performance of the authors' original search. Therefore, the performance of the various searches is first presented cohort by cohort, and finally, all are consolidated in large tables for the main outcome measures of recall of new studies (Table 31) and recall of new study participants (Table 32).

Precision cannot be calculated for all cohorts, as not all records were assessed. Retrieval size will be examined for comparative purposes. As will be seen later, there is a positive association between retrieval size and recall, which is to be expected if precision

and recall are inversely related. Recall can be calculated for all cohorts and both recall of new studies and recall of new participants can be considered. Results of the main cohort will be shown first, followed by the two cohorts with included SVM tests. Finally, the results from the three cohorts will be presented together (Table 31).

6.5 Search Performance in the Main Cohort

The searches tested in the cohort of 77 systematic reviews updated by searching are Clinical Queries, AIM RCTs, Citing RCTs, and Related Articles RCTs. New studies were assessed until a stopping point was reached, but not all candidates up to the stopping point were assessed, as reviewers focused on newer meta-analyses to identify relevant studies. Indexing status of relevant articles was tested retroactively in CENTRAL only for those systematic reviews in the main cohort that had a potentially invalidating or major signal for updating.

6.5.1 Retrieval Size in the Main Cohort

Size of the retrievals for the test searches are shown in Table 18. There are great discrepancies between retrieval size across the search methods, for example, the maximum retrieval size for Clinical Queries is more than ten times that of the Citing RCT search. On the other hand, the number of null retrievals, systematic reviews in which the search method found no new articles at all, was ten-fold higher for Citing RCT searches than for Clinical Queries. Number of records retrieved is not reported for CENTRAL, as only relevant items were looked up, no subject searching was done.

Table 18. Number of Records Retrieved Per Systematic Review by Test Searches in the Main Cohort

Retrieval Method	Median	1st Quartile	3rd Quartile	Maximum	Null Retrievals N	(%)
Clinical Queries	53	17	155	861	1	(1.3)
AIM RCTs	11	2	40	250	10	(13.0)
Citing RCTs	4	0	11	124	23	(29.9)
Related Articles RCTs	100	62	142	376	1	(1.3)

6.5.2 Recall in the Main Cohort

Recall of the searches was taken as the proportion of all *Eligible* new evidence identified by each search. Recall of new N was taken as the proportion of new participants from the *Eligible* new studies that were identified by each search. For the main cohort of 77 systematic reviews, recall of *On Topic* studies, *Eligible* studies, and recall of new participants from eligible studies are shown in Table 19 through Table 21.

Table 19. On Topic Recall in the Main Cohort

Retrieval Method	N of Studies Identified	Recall
Clinical Query	1133	.68
Abridged Index Medicus RCT	288	.17
Citing RCT	119	.07
Related Article RCT	1075	.64
Total N	1673	

Recall of *On Topic* records varied greatly between the searches. Only 7% of the *On Topic* records were randomized controlled trials that cited the original review – this was the lowest recall of any of the search approaches, while 68% of the *On Topic* material was identified by our standardized subject search limited by the Clinical Query.

Table 20. Recall of Eligible Studies in the Main Cohort

Retrieval Method	N of Studies Identified	Recall
Clinical Query	233	.51
Abridged Index Medicus RCT	95	.21
Citing RCT	61	.13
Related Article RCT	354	.78
Total N	456	

Search performance in recall of new *Eligible* studies followed a similar pattern as recall of *On Topic* evidence but two strategies, Abridged Index Medics RCTs and Citing RCT, made gains in recall while Clinical Queries dropped back both in level of recall and rank, falling to second place behind Related Article RCTs.

Recall of new N is shown in Table 21. A total of 578,126 participants were enrolled in the eligible new studies identified for the systematic reviews in the main cohort. Eighty-three percent of this total was enrolled in the studies detected by the Related Article RCT search, for example.

Table 21. Recall of Participants from Eligible Studies in the Main Cohort

Retrieval method	Retrieved N from Eligible Studies	Recall
Clinical Query	285,491	0.49
Abridged Index Medicus RCT	249,627	0.43
Citing RCT	159,525	0.28
Related Article RCT	480,245	0.83
Total	578,126	

Comparing across these three measures, we find that for three of the searches, recall improved as the criteria tightened. For example, the AIM RCT search identified only 288 of the 1673 records judged to be On Topic (recall = 0.17). When only Eligible records were considered, the AIM RCT search showed slightly better performance, finding 95 of 459 Eligible records (recall = 0.21), however, those 95 records represented studies that

contained 249,627 of the 578,126 new participants from these studies. Recall of new participants is thus 0.43, approaching the performance of the Clinical Queries that had identified four times as many *On Topic* records.

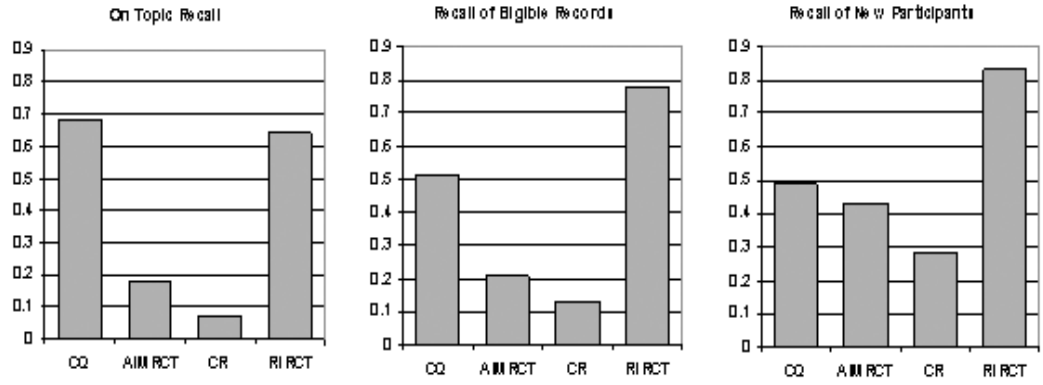


Figure 20. Recall of On Topic, Eligible and New Participants in the Main Cohort

6.6 SEARCH PERFORMANCE IN THE UPDATED COCHRANE COHORT

The six reviews in this cohort were Cochrane reviews that were updated by the authors of the reviews. As we did not do any screening, there is no assessment of the search performances for *On Topic* studies. *Eligible* new studies were those included in the updated review and which were newer than the search date of the original review. All new included studies were tested to see if they were indexed in CENTRAL, and the performance of the author's search was tested in CENTRAL, as was the performance of our subject search used in the Clinical Query and Abridged Index Medicus RCT test searches, with the modifications described in Section 4.2.4.2.

Support Vector Machine was tested in the cohort using four different standards – one using all records with a relevance score of 0.5 or more, up to a maximum of 200 retrievals (SVM200point5) and sets consisting of all records with relevance scores of 0.95 or more (SVM95), 0.90 or more (SVM90) and 0.80 or more (SVM80).

6.6.1 Retrieval Size in the Updated Cochrane Cohort

As with the main cohort, not all candidates retrieved by the searches were assessed, and so precision cannot be established, but the retrieval sizes of the various searches can be compared (Table 22).

Table 22. Retrieval Size per Systematic Review for Search Methods in the Updated Cochrane Cohort

Retrieval Method	Median	1st Quartile	3rd Quartile	Maximum	Null Retrievals	
					N	%
Clinical Queries	70	24	96	145	0	0
Abridged Index Medicus RCTs	17	6	20	22	1	16.7
Citing RCTs	2	1	2	3	0	0
Related Articles RCTs	67	56	88	174	0	0
SVM*	32	27	32	39	0	0

*Total number ranked by SVM

6.6.2 Recall in the Updated Cochrane Cohort

Recall of included new studies is shown in Table 22 and recall of new participants is shown in Table 24.

Table 23. Recall of Eligible Studies in the Cochrane Updates

Retrieval Method	N of Studies Identified	Recall
Clinical Query	16	.80
Abridged Index Medicus RCT	9	.45
Citing RCT	1	.05
Related Article RCT	20	1.00
SVM200point5	19	.95
SVM80	20	1.00
SVM90	20	1.00
SVM95	18	.90
Total N	20	

Table 24. Recall of N from Eligible Studies in the Cochrane Updates

Retrieval method	Retrieved N from Eligible Studies	Recall
Clinical Query	33,638	0.95
Abridged Index Medicus RCT	25,395	0.72
Citing RCT	561	0.02
Related Article RCT	35,504	1.00
SVM200point5	34,287	0.97
SVM95	35,504	1.00
SVM90	35,504	1.00
SVM95	34,167	0.96
Total N	35,504	

This picture differs from the picture in the main cohort in that recall of studies and new participants is much higher, with the exception of Citing RCTs, where only one small study was identified. It should be noted that there is one very large trial of 17, 966 which accounted for a third of the total N. All search methods but Citing RCTs identified this study, and this accounts for the extremely low recall of Citing RCTs relative to the other test searches.

6.7 SEARCH PERFORMANCE IN THE AHRQ EVIDENCE REVIEWS

Ten AHRQ Evidence Reports comprise the final cohort. What differs about this set is that all retrieved candidates were assessed, at least up to the stopping point, which was the point where major or potentially invalidating new evidence was detected. This complete screening enables calculation of precision for the various searches and provides a better assessment of the SVM, as all SVM retrievals were assessed. In comparison, in the Cochrane Cohort, SVM assessment was limited to determining which of the studies found by the authors through conventional search methods were also detected by SVM, there was not opportunity to assess the yield of material uniquely identified by SVM or by another the other test searches. As well, in this cohort, recall of both *On Topic* and

Eligible studies can be determined. The results for this cohort are therefore the most comprehensive of the three.

6.7.1 Retrieval Size in the AHRQ Evidence Reviews

Table 25. Retrieval Size Per Systematic Review for Search Methods in the AHRQ Evidence Reports

Retrieval Method	Median	1st Quartile	3rd Quartile	Maximum	Null Retrievals	
					N	%
Clinical Queries	147	78	212	418	0	0.0
Abridged Index Medicus RCTs	31	23	44	128	0	0.0
Citing RCTs	0	0	0	28	8	80.0
Related Articles RCTs	79	45	106	172	0	0.0
SVM200point5	60	37	85	104	0	0.0

Comparing these results to the retrieval sizes in the main cohort, we see a reversal in position between Clinical Queries and Related Articles RCT – while Related Article RCTs were the largest result sets in the main cohort, Clinical Queries are the largest here. Support Vector Machine returned small retrievals than either clinical Queries or Related Article RCTs in both this cohort and in the Cochrane updates, and in both cases, returned larger sets than the strategy limited to RCTs from Abridged Index Medicus journals. While all search methods yielded some null retrievals in the main cohort, here, as in the Cochrane Updates, the only null retrievals were for Citing RCTs.

6.7.2 Recall in the AHRQ Evidence Reviews

Table 26. On Topic Recall in the AHRQ Evidence Reports

Retrieval Method	N of Studies Identified	Recall
Clinical Query	640	0.74
Abridged Index Medicus RCT	208	0.24
Citing RCT	17	0.02
Related Article RCT	373	0.43
SVM200point5	242	0.28
Total N	862	

Table 27. Recall of Eligible Studies in the AHRQ Evidence Reports

Retrieval Method	N of Studies Identified	Recall
Clinical Query	181	0.67
Abridged Index Medicus RCT	66	0.24
Citing RCT	0	0.00
Related Article RCT	154	0.57
SVM200point5	125	0.46
Total N	270	

Table 28. Recall of Eligible New Participants in the AHRQ Evidence Reports

Retrieval Method	Retrieved N from Eligible Studies	Recall
Clinical Query	163,352	0.79
Abridged Index Medicus RCT	151,065	0.73
Citing RCT	0	0.00
Related Article RCT	150,585	0.73
SVM200point5	89,347	0.43
Total N	205,844	

Comparing the three measures of recall (Figure 21), we see a familiar decline in the retrieval performance of Clinical Queries between *On Topic* and *Eligible* records, but unlike the pattern in the other two cohorts, there is a rebound when recall of new participants is considered. AIM RCT remains strong in recall of new participants probably driving the strong performance of the Clinical Queries here, as Clinical Queries would have little unique contribution of new participants beyond those contributed by the Abriged Index Medicus subset. Citing RCTs was not a useful approach in this cohort. It did not identify a single relevant piece of evidence. Related Article RCT repeated the pattern seen in the main cohort of gains with each tightening of the relevance criteria, although only matching AIM RCTs for recall of new N. This method identified all new evidence included in the Cochrane review.

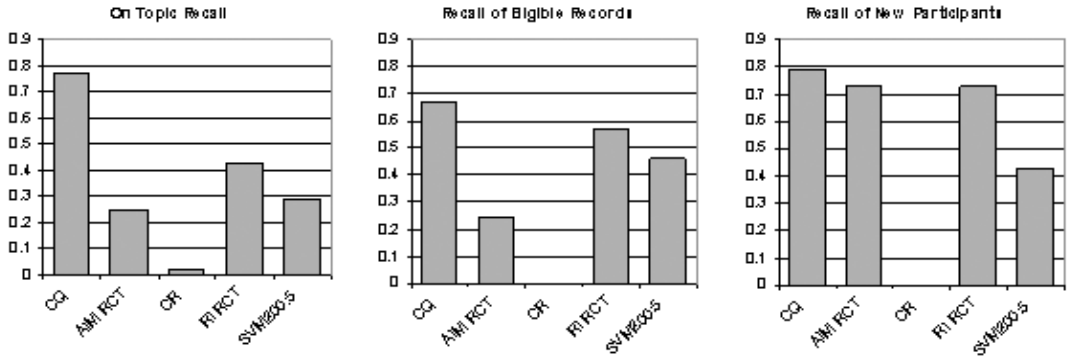


Figure 21. Comparative Performance for the Test Searches

Support Vector Machine, despite strong performance in detecting the new evidence in the Cochrane reviews, did not show as strongly here, with recall below that of the other viable methods.

In this cohort, there were four eligible new studies that had more than 10,000 participants. Two of these four were included in two AHRQ Evidence Reports in this cohort, and so carry considerable weight in the recall of new N. Those reviews examined antioxidants for the prevention and treatment of cardiovascular disease[201] and for the prevention and treatment of cancer.[254] The first of these two trials was the Women's Health Study.[255] It was eligible for one systematic review in the main cohort as well, and was the second largest eligible new trial in any of the cohorts, with 39,876 participants. The other, the fourth largest in the AHRQ cohort, with 13,017 participants, was the SU.VI.MAX study,[256] which was also eligible for one review in the main cohort. The related article RCT and Clinical Query searches retrieved the main report for both these studies for both AHRQ Evidence Report for which they were eligible, while SVM retrieved the main report of Women's Health Study only for the cardiovascular Evidence Report, and did not retrieve the SU.VI.MAX study for either review. These misses accounted for most of the deficit in recall of new participants seen with SVM.

As might be expected with such large and complex trials, both of these studies had had multiple reports, with N attributed to a single publication. How did SVM do at retrieving other relevant reports of the same study? The Reference Manager databases and

reviewer spreadsheets for these two Evidence Reports were examined for other reports of these studies that would be eligible or where the title or abstract would clearly indicate the existence of one of the eligible reports (Table 29).

Table 29. SVM Retrieval of Multiple Publications of Large Trials Eligible for AHRQ Evidence Reports

	Antioxidants for Cardiovascular Disease[201]		Antioxidants for Cancer[254]	
	Relevant?	Retrieved by SVM?	Relevant?	Retrieved by SVM?
Womens Health Study				
*Lee, 2005[257]	Yes	Yes	Yes	No
Buring, 2006[258]	Yes	No	Yes	No
Liu, 2006[259]	Yes	Yes	…	..
Song, 2004a[260]	No	No	No	No
Schaumberg, 2003[261]	No	No	No	No
Christen, 2004[262]	…	…	Yes	No
Cook, 2005[263]	…	…	No	No
Song, 2004b[264]	…	…	No	No
SU.VI.MAX				
*Hercberg, 2004[256]	Yes	No	Yes	No
Galan, 2005[265]	Yes	Yes	Yes	No
Hercberg, 2002[266]	Yes	No	Yes	No
Galan, 2003 [267]	No	No	No	No
Czernichow, 2005[268]	Yes	No	…	…
Zureik, 2004 [269]	Yes	No	…	…
Meyer, 2005 [270]	…	…	Yes	No
Malvy, 2001[271]	…	…	Yes	No
Hercberg, 2005 [272]	…	…	No	No

*Indicates primary report for that review.
… indicates that the report was not retrieved for that Evidence Report by any search, and so was not assessed.

For Women's Health Study, there were seven additional reports other than the report to which N had been attributed that were in the candidate list for one or the other of these reviews. SVM found the main report and one additional report for the cardiovascular Evidence Review but none of the reports assessed as *Eligible* for the cancer Evidence report. For the SU.VI.MAX study, there were eight reports in addition to

the report to which N had been attributed. For cardiovascular Evidence Review, SVM identified one of these four additional reports assessed as *Eligible*. For the cancer Evidence Review, four reports in addition to the index report were assessed as *Eligible* and SVM retrieved one of these.

If SVM was credited with identifying the SU.VI.MAX trial for the cardiovascular Evidence Report, recall of new participants would increase from 0.43 to 0.49 (total new n retrieved would go from 89,347 to 102,364). This is still low relative to the Clinical Queries and Related Article searches (Table 28). As well, it is possible, had this method of determining if any reports of a large trials, not just the first eligible, be computed for all trials, the other search methods might also have shown higher recall of new N.

6.8 COMPARISON OF RECALL OF ELIGIBLE STUDIES PRE AND POST SIGNAL

Although the intention was to completely assess all candidates, in two of the ten Evidence Reports the reviewers stopped assessments at the signal. To assess the potential for bias based on this early stopping, recall was considered for studies up to the point of the signal, and studies appearing after the signal, performance was very similar for the two periods, thus the incomplete data for two Evidence Reports would not appear to complicate interpretation of the results.

Table 30. Comparison of Recall of Eligible Studies Pre and Post Signal

		Related Article RCT	SVM200 point5	Clinical Query	Abridged Index Medicus RCT	Citing RCT
Pre-signal	N=151	0.61	0.46	0.66	0.23	0.00
Post-signal	N=135	0.52	0.47	0.68	0.26	0.00
Total	N=286	0.57	0.46	0.67	0.24	0.00

6.8.1 Summary of Recall of the Test Searches

Results have been presented for three recall measures in up to eight test searches in three cohorts. Table 31 and Table 32 compile these for two of the recall measure, recall of *Eligible* studies, which is the most conventional measure, and recall or identification relevant new participants, which has not commonly been reported but may be of greater interest in the updating context.

Table 31. Summary: Recall of New Studies, All Cohorts

	Main77	Cochrane Updates	AHRQ	All Combined	Cochrane & AHRQ Combined
Clinical Query	0.51	0.81	0.67	0.56	0.65
Abridged Index Medicus RCT	0.21	0.45	0.24	0.22	0.25
Citing RCTs	0.13	0.05	0.00	0.08	0.00
Related Article RCTs	0.78	1.00	0.55	0.69	0.57
SVM .5 or 200	NA	0.95	0.46	NA	0.48
SVM .8	NA	1.00	NA	NA	NA
SVM .9	NA	1.00	NA	NA	NA
SVM .95	NA	0.90	NA	NA	NA

NA indicates not available for this cohort.

Table 32. Summary: Recall of New Participants, All Cohorts

	Main77	Cochrane Updates	AHRQ	Overall, All Cohorts	Overall, Cochrane & AHRQ Cohorts
Clinical Query	0.49	0.95	0.77	0.59	0.80
Abridged Index Medicus RCT	0.43	0.72	0.73	0.52	0.72
Citing RCTs	0.28	0.02	0.00	0.20	0.002
Related Article RCTs	0.83	1.00	0.73	0.81	0.76
SVM .5 or 200	NA	0.97	0.46	NA	0.50
SVM 8	NA	1.00	NA	NA	NA
SVM .9	NA	1.00	NA	NA	NA
SVM .95	NA	0.96	NA	NA	NA

NA indicates not available for this cohort.

Table 33. Recall of New Participant by Clinical Area, All Cohorts Combined

	N of Reviews	Clinical Query	Abridged Index Medicus RCT	Citing RCT	Related Article RCT	SVM
Cardiac & Cardiovascular System	21	0.707	0.682	0.382	0.400	0.650
Clinical Neurology	11	0.299	0.251	0.009	0.642	0.494
Critical Care Medicine	11	0.144	0.119	0.333	0.943	1.000
Endocrinology & Metabolism	6	0.976	0.783	0.000	0.766	0.446
Gastroenterology & Hepatology	8	0.406	0.107	0.066	0.949	NA
Infectious Diseases	8	0.938	0.790	0.033	0.905	0.977
Obstetrics and Gynecology	7	0.244	0.030	0.149	1.000	0.988
Oncology	5	0.992	0.965	0.013	0.493	0.009
Peripheral Vascular Diseases	5	0.435	0.396	0.050	0.126	0.199
Psychiatry	4	0.515	0.196	0.000	0.990	0.492
Respiratory System	3	0.200	0.000	0.053	1.000	1.000
Rheumatology	3	1.000	1.000	0.000	0.793	NA
Urology & Nephrology	1	0.922	0.216	0.055	1.000	1.000
Total	97	0.587	0.488	0.155	0.744	0.520

NA – recall could not be computed as there were no instances of Gastroenterology & Hepatology or Rheumatology in the AHRQ or Cochrane update sets

6.9 SEARCH PERFORMANCE BY INTERVENTION TYPE

All systematic reviews focused on interventions, either treatment or prevention. Considering search performance according to the type of intervention, drug, device or procedure, we see that almost all reviews focused primarily on drugs (Table 34) and most new evidence was for these reviews (Table 35). While the number of eligible articles for the few systematic reviews of devices and procedures are small, there do not seem to be great differences in search performance across intervention type.

Table 34. Number of Reviews of Drugs, Devices and Procedures

	Main Cohort	Updated Cochrane Reviews	AHRQ Evidence Reports	Total	%
Drug	67	5	9	81	87.1%
Device	3	1	0	4	4.3%
Procedure	7	0	1	8	8.6%

Table 35. Recall of the Searches by Intervention Type

	Drug N=702	Device N=20	Procedure N=40	Overall N=762
Clinical Query	0.56	0.50	0.65	0.56
Abridged Index Medicus RCT	0.22	0.20	0.35	0.22
Citing RCT	0.08	0.00	0.20	0.08
Related Article RCT	0.68	1.00	0.73	0.69
SVM200point5	0.46	1.00	0.67	0.48

6.10 PRECISION OF THE TEST SEARCHES

Precision can only be assessed with complete confidence in the cohort of AHRQ Evidence Reviews, as it was only in this cohort that all candidates were screened.

Table 36. Overall Precision

	N Eligible	N of Candidates Retrieved	Precision	Rank
Clinical Query	181	1637	0.1106	5
Abridged Index Medicus RCT	66	441	0.1497	4
Citing RCT	17	33	0.5152	1
Related Article RCT	156	814	0.1916	2
SVM200point5	125	659	0.1897	3

Precision ranged from 0.515 to a low of 0.111. There is a clear associated with retrieval size – rank of precision (high to low) is almost the same as rank of N of candidate retrieved (low to high) with the exception of Abridged Index Medicus RCT, which has lower precision than would be expected based on it fairly small retrieval size. It

should be noted that for SVM, fixing the retrieval size may results in good precision with large numbers of relevant records and poor precision with few records – in either case the denominator is fixed.

6.11 BALANCE OF RECALL AND PRECISION

Similarity search methods had some advantage over Boolean searches in terms of improved precision, but at some cost to recall. The correlation between overall precision and recall precision and recall of new studies (r= -0.823) (Figure 22) and between of new N remains negative (r= - 0.942).

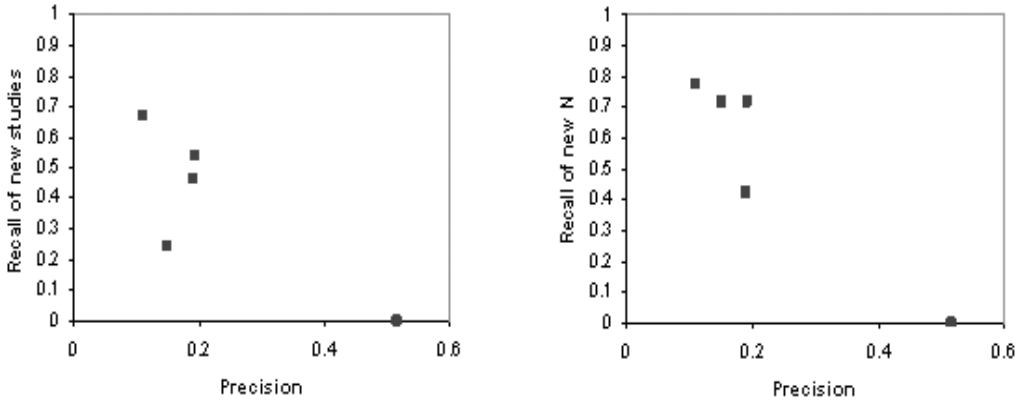


Figure 22. Recall and Precision for AHRQ Evidence Reports.
Recall of new studies (left) and new participants (right) by precision of searches for new evidence for AHRQ Evidence Reports. From top to bottom, new N; Clinical Query, Related Article RCT, SVM, Abridged Index Medicus RCT, and Citing RCT. From top to bottom for new N ; Clinical Query, Abridged Index Medicus RCT, Related Article RCT, SVM, Citing Reference. Circles indicate relevance ranked similarity search methods, squares indicate skill-based Boolean search methods, triangle is citing reference method.

Turning to the cohort of updated Cochrane reviews, precision can be considered if it is assumed that the authors of the updated Cochrane searches successfully identified all relevant new evidence shows the performance of all searches tested against this cohort,

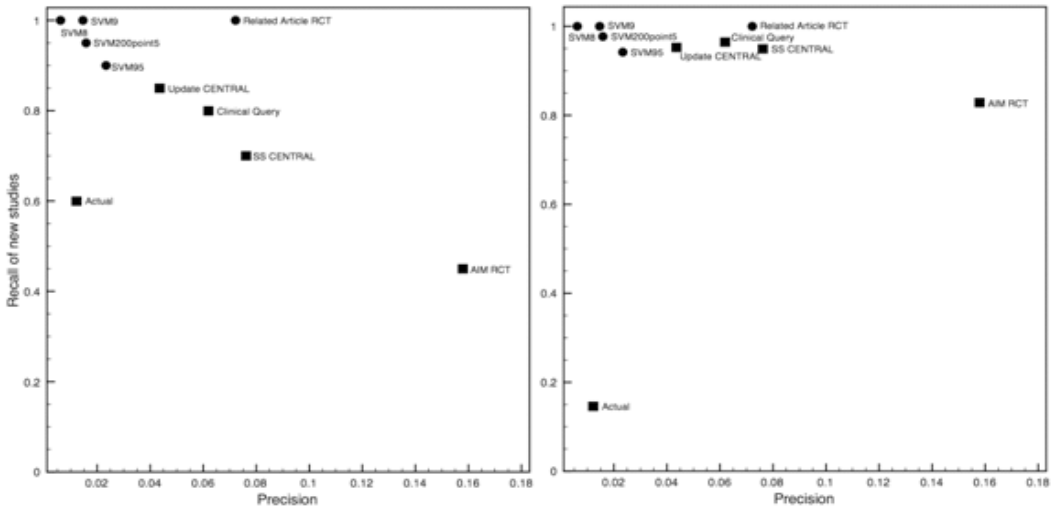


Figure 23. Recall of New Studies (Left) and New Participants (Right) and Precision in the Updated Cochrane Reviews

including several that will be discussed below; the authors' actual update search (see Section 6.26 for complete results) and the update search and the subject search used for Clinical Queries and AIM RCT searches when run in the CENTRAL registry of controlled trials, part of the Cochrane Library (see Section 6.14 for complete results). When the recall of new eligible studies is considered, with the exception of the authors' actual update search (Actual) and the Related Article RCT search, there is a negative and near- perfect linear association between recall and precision. The correlation between overall precision and recall of new studies for the test searches – that is, excluding the author searches, is r= -0.876 (Figure 23, left panel). The picture changes somewhat when recall of new participants is balanced against precision (Figure 23, right panel). Boolean searches, at a disadvantage in the recall of new studies, now approach the similarity-based searches in terms of recall, and precision becomes the more salient distinguishing factor. The authors' own searches (Actual) had inferior precision compared to the more formulaic Boolean searches. The correlation between overall precision and recall of new

participants remains negative for the test searches – that is, excluding the author searches (r= -0.808).

6.12 RANKING

6.12.1 Ranking in the Updated Cochrane Cohort

In the Cochrane set, both the SVM and Related Article RCT searches placed more than half of the targets (the studies added in the updates that were newer than the original review) in the top twency when the retrievals were ranked by relevance scores (Table 37). Thirteen of twenty targets were ranked in the top ten by either SVM or the Related Article RCT search (Table 38 and Table 39). While both ranking systems placed a high proportion of targets in the top ten, only two of the included new studies were ranked in the top ten by both methods (15.4%). For all systematic reviews in this cohort, any target that was retrieved by the search method would have been identified by the time the first fifty records were examined, if records were screened in order by relevance. This is the case for both search methods.

Table 37. Precision and Recall of Ranked Retrieval at Various Cut Points in the Updated Cochrane Cohort

	SVM		Related Article RCT	
	Recall	Precision	Recall	Precision
Top 10	0.35	0.117	0.40	0.133
Top 20	0.55	0.092	0.70	0.117
Top 30	0.75	0.078	0.95	0.094
Top 40	0.80	0.063	1.00	0.079
Top 50	0.95	0.053	1.00	0.067
Top 100	0.95	0.032	…	…

The SVM runs for the six Cochrane reviews examined here identified 546,014 MEDLINE records with a relevance score greater than zero. However, for the binomial test, the retrieval size was taken as the count of all retrievals with a relevance score of 0.8

or greater (the SVM.8 search), making the test more conservative. Seven of twenty targets were ranked within the top ten. A binomial test revealed that this significantly exceeded chance ($p < 0.001$). All targets were ranked within the top 250, also exceeding chance, ($p<.001$). Complete results are shown in Table 38.

Table 38. Recall of the Ranked SVM Retrieval at Various Cut Points – Updated Cochrane Cohort

Cut point	Records ranked above the cut point (*a*)	Records retrieved from MEDLINE (*b*)	Included studies ranked above the cut point (*c*)	Targets (*d*)	Proportion records above the cut point (p=*a*/*b*)	Recall above the cut point (q=*c*/*d*)	P value 2 sided
10	60	3207	7	20	0.019	0.35	<0.000
20	120	3207	11	20	0.037	0.55	<0.000
30	180	3207	14	20	0.056	0.70	<0.000
40	240	3207	15	20	0.075	0.75	<0.000
50	300	3207	16	20	0.094	0.80	<0.000
100	600	3207	19	20	0.187	0.95	<0.000
150	900	3207	19	20	0.281	0.95	<0.000
200	1200	3207	19	20	0.374	0.95	<0.000
250	1500	3207	20	20	0.468	1.00	<0.000

Suppose there are *b* records in the initial retrieval. Suppose that *a* records (here we consider a range of values for a from 10 to 250) are considered, yielding a proportion p = *a*/*b*. Note that *a* is reported for each systematic review while *b* is reported for the entire sample – this is accounted for in the calculation of p (p=6*a*/*b*). Suppose further that this subset includes *c* of the *d* relevant records from the initial retrieval, i.e. a proportion q = *c*/*d*. If the ranking performs no better than would be expected by chance, then we would expect q = p. The single sided binomial test returns the exact probability of getting the value observed or any larger value, considering the valued expected by chance. [adapted from Sampson et al 2006[178]]

*The cut point exceeds the retrieval size in many cases, so a is less than 6 times the cut point.

Related Article RCT rankings put slightly more targets above each cut point, relative to SVM, however, results were similar overall and all targets were ranked in the top fifty. Testing within the set, ranking performed better than chance through the top twenty only, due to small set sizes. Eight of twenty targets were ranked within the top ten.

A binomial test revealed that this significantly exceeded chance, $p < 0.040$. Complete results are shown in Table 39.

Table 39. Recall of the Ranked Related Article Retrieval at Various Cut Points in the Updated Cochrane Cohort

Cut point	Records ranked above the cut point (*a*)	Records retrieved from MEDLINE (b)	Included studies ranked above the cut point (c)	Targets (d)	Proportion records above the cut point (P=a/b)	Recall above the cut point (Q=c/d)	P value 2 sided
10	60	265	8	20	0.226	0.40	0.040
20	120	265	14	20	0.453	0.70	0.016
30	180	265	17	20	0.679	0.85	0.052
40	240	265	19	20	0.906	0.95	0.287
50	300	265	20	20	1.000	1.00	-

6.12.2 Ranking in the AHRQ Evidence Reports

Of 286 *Eligible* new records for the Evidence Reports in the cohort, 125 were retrieved by SVM (43.7%) and 154 by Related Article RCT searches (53.8%). Looking at the retrieved portion, both the SVM and Related Article RCT ranking placed half of the *Eligible* retrieved records in the top thirty in terms of relevance scores (Table 40).

Sixty-one of 286 (20.3%) *Eligible* new studies were placed in the top ten by either SVM or Related Article RCT. Of those, ten (16.4%) were in the top ten by both the SVM and Related Article RCT rankings. This is similar to the overlap found in the updated Cochrane reviews where two of the 13 targets (15.4%) placed in the top ten were placed there by both ranking systems. All relevant records identified by the search method were within the first 150 records for SVM (Table 41) and within the first 200 records for Related Article RCT, and all but one were in the top 150 (Table 42).

Table 40. Precision and Recall of Ranked Retrieval at Various Cut Points – AHRQ Cohort

	SVM			Related Article RCT		
Cut point	Relative Recall*	Recall†	Precision	Relative Recall*	Recall †	Precision
10	0.10	0.22	0.270	0.12	0.22	0.340
20	0.17	0.39	0.245	0.21	0.39	0.316
30	0.23	0.52	0.226	0.28	0.51	0.272
40	0.28	0.62	0.216	0.32	0.59	0.260
50	0.31	0.70	0.204	0.36	0.66	0.240
100	0.49	0.94	0.186	0.50	0.92	0.204

*Recall into that rank, using all Eligible studies as the denominator. The denominator is 286 for both SVM and Related Article RCT.
†Of those recalled by that search method, proportion placed at or above that rank. The denominator is 125 for SVM and 154 for Related Article RCT.

Testing the obtained versus expected proportion above each threshold with the binomial test, both rankings performed significantly better than chance, and provided similar performance (Table 41 and Table 42). In the cohort of updated Cochrane reviews ranking performed better than chance through the top twenty only, due to small set retrieval sizes. With this larger set, ranking was significantly better than chance throughout the range and up until the point where all eligible studies had been retrieved, and this is the case for both SVM and Related Article RCT.

Table 41. Recall of the Ranked SVM Retrieval at Various Cut Points – AHRQ Cohort

Cut point	Records ranked above the cut point (*a*)	Records retrieved from MEDLINE (*b*)	Included studies ranked above the cut point (*c*)	Targets (*d*)	Proportion of records above the cut point (p=*a*/*b*)	Recall above the cut point (q=*c*/*d*)	P value, 2 sided
10	100	612	27	125	0.016	0.22	<0.000
20	200	612	49	125	0.033	0.39	<0.000
30	287	612	65	125	0.049	0.52	<0.000
40	359	612	78	125	0.065	0.62	<0.000
50	429	612	88	125	0.082	0.70	<0.000
100	608	612	118	125	0.163	0.94	<0.000

150	612	612	125	125	0.245	1.00	<0.000
200	612	612	125	125	0.327	1.00	<0.000
250	612	612	125	125	0.409	1.00	<0.000

*The cut point exceeds the retrieval size in many cases, so a is less than 10 times the cut point.

Table 42. Recall of the Ranked Related Article Retrieval at Various Cut Points – AHRQ Cohort

Cut point	Records ranked above the cut point (*a*)	Records retrieved from MEDLINE (*b*)	Included studies ranked above the cut point (*c*)	Targets (*d*)	Proportion of records above the cut point (p=*a*/*b*)	Recall above the cut point (q=*c*/*d*)	P value, 2 sided
10	100	814	34	154	0.123	0.22	<0.000
20	190	814	60	154	0.233	0.39	<0.000
30	270	814	79	154	0.332	0.51	<0.000
40	350	814	91	154	0.430	0.59	<0.000
50	421	814	101	154	0.517	0.66	<0.000
100	675	814	141	154	0.829	0.92	<0.000
150	787	814	153	154	0.967	0.99	<0.000
200	814	814	154	154	1.000	1.00	<0.000
250	814	814	154	154	1.000	1.00	<0.000

*The cut point exceeds the retrieval size in many cases, so *a* is less than 10 times the cut point.

Comparing the ranking performance in SVM and Related Article RCT (Table 40), the Related Article RCT search put a slightly higher proportion of all *Eligible* records into the top portion of the retrieval at cut points 10 through 50, but by the top 100, where almost all relevant retrieved records were ranked by both searches, the difference in relative recall was negligible. When only *Eligible* records retrieved by that search method were considered, recall above a certain cut point was very similar thorough out the range for both SVM and Related Article RCT, thus the two ranking systems are equally effective. Precision declines steadily for both searches as the cut points increase, but the drop off is more marked for Related Article RCT than for SVM. The correlation between relative recall and precision and recall and precision at the different cut points is -0.99 for

SVM and -0.97 for Related Article RCT. Differences in ranked performance are due to differences in retrieval effectiveness rather than ranking effectiveness.

6.12.3 Receiver Operating Characteristics (ROC) Analysis

Receiver Operating Characteristics can be calculated for the search methods that rank results, SVM and Related Article RCT.

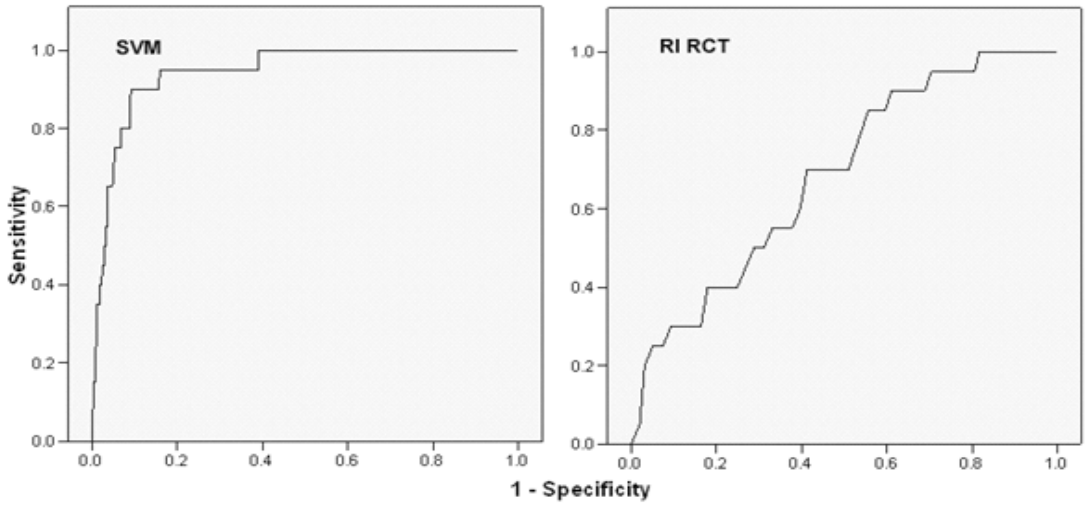


Figure 24. Receiver Operator Curve of Eligibility by SVM and Related Article RCT Ranking for the Cochrane Cohort. The positive state is being a target.

Table 43. Test of Significance of Area Under the Curve

				Asymptotic 95% Confidence Interval	
	Area	Std. Error*	Asymptotic Sig.†	Upper Bound	Lower Bound
SVM Rank	0.945	0.019	0.000	0.907	0.982
Related Article RCT Rank	0.684	0.057	0.006	0.573	0.795

* Under the nonparametric assumption

† Null hypothesis: true area = 0.5

The area under the curve is tested against the null hypothesis that the true area = 0.5, the value that would be obtained if the true shape of the curve were a diagonal. That occurs when, for all values of sensitivity, sensitivity=1-specificity, and the variable tested offers no improvement over chance. Here, although ranking was useful in determining

relevance, the gain was much greater for SVM, with area under the curve was 0.945, than for Related Article RCT, where the area under the curve was 0.684. The more impressive receiver operating characteristics curve for SVM is due to effectively distinguishing a small number of targets from a large set of irrelevant records (Figure 24).

Turning to the AHRQ cohort, the receiver operating characteristics curves are less impressive (Figure 25) being much closer to the diagonal and here the area under the curve is not different from chance for SVM (Table 44).

Table 44. Test of Significance of Area Under the Curve, AHRQ Cohort

				Asymptotic 95% Confidence Interval	
	Area	Std. Error*	Asymptotic Sig.†	Upper Bound	Lower Bound
SVM Rank	0.553	0.031	0.067	0.493	0.613
Related Article RCT Rank	0.660	0.024	0.000	0.613	0.708

* Under the nonparametric assumption

† Null hypothesis: true area = 0.5

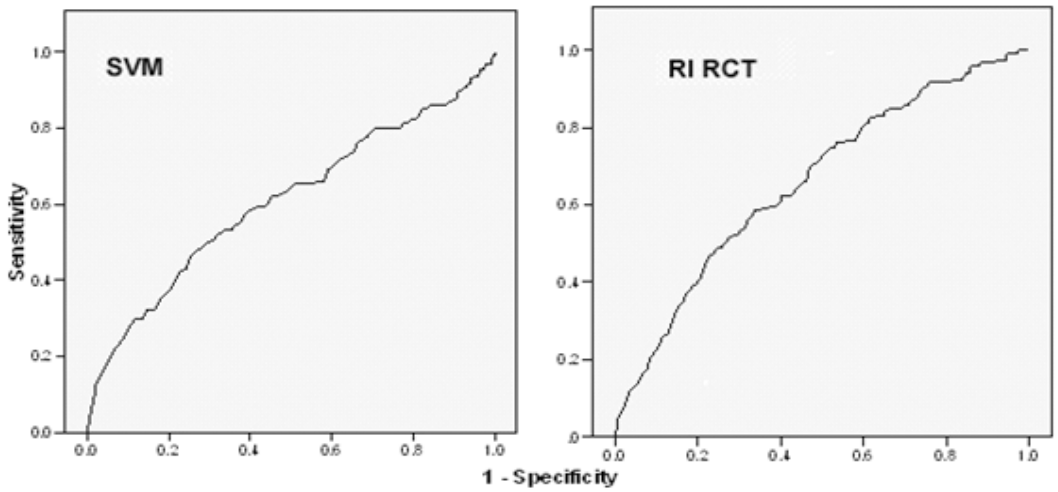


Figure 25. ROC Curve of Eligibility by SVM and Related Article RCT Ranking for the AHRQ Cohort. The positive state is being a target.

6.13 SUMMARY OF THE RESULTS OF THE MAIN EXPERIMENT

The main experiment examined the performance of searches in three cohorts. Two Boolean searches were based on a subject search. The subject searches were simple, using a mean of 3.6 terms and 1.5 search features. The subject search was paired with a filter selecting only RCTs from Abridged Index Medicus journals and with the Clinical Query.

Clinical Query provided good recall but with large retrievals. Abridged Index Medicus RCT had smaller retrieval sizes and identified fewer new *Eligible* studies, but did detect many large studies, so performed well when recall of new participants was considered. The two similarity searches were SVM, where various cut points were tested in the Cochrane and AHRQ cohorts, and an algorithm based on related article searching in PubMed. The similarity searches were similar to the Clinical Query approach, but the Related Article search showed the highest recall of new *Eligible* studies overall, with SVM lower, and trailing the Clinical Query. SVM's advantage was smaller retrieval sizes. A high proportion of Citing RCTs were *Eligible* for inclusion in updated reviews, but this method identified only a small proportion of all relevant new studies.

Relative performance of the test searches was fairly stable across the three cohorts, although there was a ceiling effect in the updated Cochrane reviews, where most searches performed strongly. Relative performance was also fairly stable regardless whether the intervention was a drug, device or procedure, and whether the evidence appeared before or after the signal that an update was necessary.

Precision, tested in the AHQQ cohort and updated Cochrane Cohort, ranged from 0.111 to 0.192 for the more productive test searches but reached 0.515 for the Citing RCT method. Similarity methods showed higher precision than did the two Boolean approaches. Precision and recall showed a strong negative correlation whether recall of new studies or recall of new participants was considered.

Ranking was examined for the similarity searches. Placement of *Eligible* studies near the top exceeded chance for both the SVM and Related Article RCT searches. When receiver operating characteristic curves were examined for these searches, SVM

outperformed Related Article RCT in the Cochrane set, although the area under the curve was significantly greater than chance for both. Area under the curve was less impressive in the AHRQ Evidence Report searches, with only the Related Article RCT search improving over chance.

6.14 CENTRAL

6.15 Central Indexing of New Material for the Main Cohort

CENTRAL was not one of the methods used to build the candidate study list screened by the reviewers, it was added in after the fact to allow us to determine how complete the coverage of relevant new evidence is in CENTRAL. It is possible that additional relevant studies would have been found had we searched it to build the candidate lists. We tested CENTRAL coverage of *Eligible* new studies for those systematic reviews in the main cohort that had a signal for updating (N=48). Several systematic reviews were reclassified after the CENTRAL searching was done, thus data is missing for five systematic reviews subsequently assessed as having major or potentially invalidating new evidence.

Recall of each test search for this subset is presented in Table 45 so CENTRAL coverage can be compared with the recall of the other searches. The performance of the test searches in this subset is, however, very similar to the performance for the entire main cohort for all three measures of recall (Table 20 - Table 22). This subset contained 27 reviewer nominated record - records identified because they were included in another systematic review, were known to our team, or were found through *ad hoc* searching. CENTRAL indexed only one (0.04%) of these.

Table 45. Recall in the CENTRAL Subset

	N of Records Identified	On Topic Recall	N of Records Identified	Eligible Recall	N of New Participants Identified	Recall of New N
Clinical Query	699	0.65	143	0.45	244,612	0.51
Abridged Index Medicus RCT	168	0.16	64	0.20	227,281	0.47
Citing RCT	76	0.07	38	0.12	154,472	0.32
Related Article RCT	671	0.62	246	0.77	390,383	0.81
CENTRAL	469	0.44	121	0.38	246,055	0.51
Total N	1077		320		481,770	

6.16 CENTRAL in the Updated Cochrane Cohort

The twenty targets, representing newer studies added to the six Cochrane reviews in the update were examined to determine if they were indexed in CENTRAL, and if the subject search developed by our team would have retrieved them, and if the search used by the authors of the updated review would have retrieved them. Recall of included studies and recall of new participants are presented in **Table 46**.

Table 46. Recall of New Studies Through CENTRAL

Search	Records Identified	Proportion Identified	New Participants Identified*	Proportion Identified
Eligible studies indexed in CENTRAL	20	1.00	35,504	1.00
Update search in CENTRAL	14	0.70	26,125	0.78
Subject Search in CENTRAL	14	0.70	32,825	0.98
Total	20		35,504	

*19 of the new included records had N attributes

Although the authors' search used for the update and our subject search recalled the same number of the included studies, the structured subject search had the advantage of retrieving all but the smallest studies.

6.17 SEARCHES FOR NEWER META-ANALYSES

The retrospective nature of the AHRQ Updating project made it practical to identify newer meta-analyses which would aid in identifying new primary studies. Two searches were used to do this – the subject search limited by the MEDLINE publication type Meta-analysis, and the result of the Related Articles algorithm limited to the MEDLINE publication type Meta-analysis. As efficient identification of newer meta-analyses is potentially useful in a variety of contexts, the performance of these searches is examined here. These searches are for the main cohort of 77 systematic reviews. 1193 purported new meta-analyses identified, 793 by the subject search and 539 by the related articles approach (139 were identified by both).

Table 47. Retrieval Size for Search Methods to Detect New Meta-analyses

Search Method	Median	1st Quartile	3rd Quartile	Maximum	Null Retrievals No.	Null Retrievals (%)
Subject Search – MA	6	3	13	63	4	5.2
Related Articles – MA	5	2	9	68	8	10.4

Median retrieval size was approximately equal across the two searches, with the main distinguishing feature being a slightly higher number of null retrievals – searches yielding no new material – with the related article algorithm (Table 47). Null retrievals are more common here than with the searches for primary studies for two reasons. First, there are fewer meta-analyses conducted than there are primary studies.[§§] Second, the observation period is shorter for meta-analyses. Primary studies were searched starting from the end date of the search of the original review, while the searches for newer meta-

[§§] There are examples where the ratio of meta-analyses to primary studies is high. Biondi-Zoccai et al studies 10 meta-analyses of the role of acetylcysteine in the prevention of contrast associated nephropathy, for which there were 28 trials[273] In the case of tissue plasminogen activator for stroke, a PubMed search of those MeSH headings finds 63 records with the publication type of randomized controlled trial and 68 systematic reviews, using the systematic review topic limit. Ten of these 68 are classified as meta-analyses.

analyses commenced one year after the search of the original review to allow some time for new primary studies to accumulate (see Methods 3.16).

Table 48. Precision of Meta-analyses

	N Retrieved	N assessed	N Eligible	Precision
Subject Search – Meta-analyses	793	171	74	0.433
Related Articles – Meta-analyses	539	119	49	0.412

Of 230 assessed records from the meta-analyses search, 171 were retrieved by the subject search and 119 were retrieved by Related Article search (Table 48). These retreivals are the denominator for calculation of precision of the search methods. Seventy-four of meta-analyses identified through the subject search were *Eligible* (precision=0.433) while 49 of the 119 assessed records found through related article searching were *Eligible* (precision =0.421). Sixty-six purported meta-analyses were retrieved by both searches and 33 (50%) of these were *Eligible.*

6.18 STRUCTURAL RELATIONSHIP BETWEEN SEARCHES

6.18.1 Unique Contribution and Overlap in the Cochrane Cohort

The overlap and unique component of the material retrieved by the test searches for the updated Cochrane cohort was examined (see Methods 4.2.1) and is shown in Figure 26. The size the of the circle is proportional to the size of the retrieval. In this figure, the overlap and unique portions are approximations. Exact figures for the unique components are presented in Table 49. Accurate figures for overlap are presented in Table 50. This figure is simplified to show only test searches (not the actual searches used in the review) and only one SVM search is shown.

Table 49. Unique Component of Each Search from the Cochrane Cohort

	Records Retrieved Uniquely by that Search Method	Total Retrieval Size	% of Retrieval Unique to that Search Method
Clinical Query	66	258	23.2
Abridged Index Medicus RCT	0	57	Nil
Citing RCT	0	1	Nil
Related Article RCT	81	259	31.3
SVM200point5*	984	1200	82.0
SVM95*†	591	771	76.7
SSCENTRAL†	15	184	7.9
Actual†	507	985	51.5

* SVM95 and SVM200point5unique were unique relative to the other search types, but not necessarily unique relative to any SVM variant, including SVM95 vs. SVM200point5. Other searches are unique relative to SVM if the records were not found by any of the SVM methods.
†Not shown in Figure 26. Actual is tested as unique against all others. Others as tested as unique against the other searches listed, but not against Actual.

Table 50. Overlapping Components of Searches from the Cochrane Cohort

	Clinical Query n=258	Abridged Index Medicus RCT n=57	Citing RCT n=1	Related Article RCT n=259	SVM200 point5 n=1200
Clinical Query	258				
Abridged Index Medicus RCT	57	57			
Citing RCT	1	1	1		
Related Article RCT	99	38	1	259	
SVM200point5	146	52	1	149	1200
SVM95	125	44	1	124	1031
Subject Search in CENTRAL	158	52	1	78	129
Authors' Actual Search	198	34	1	94	313

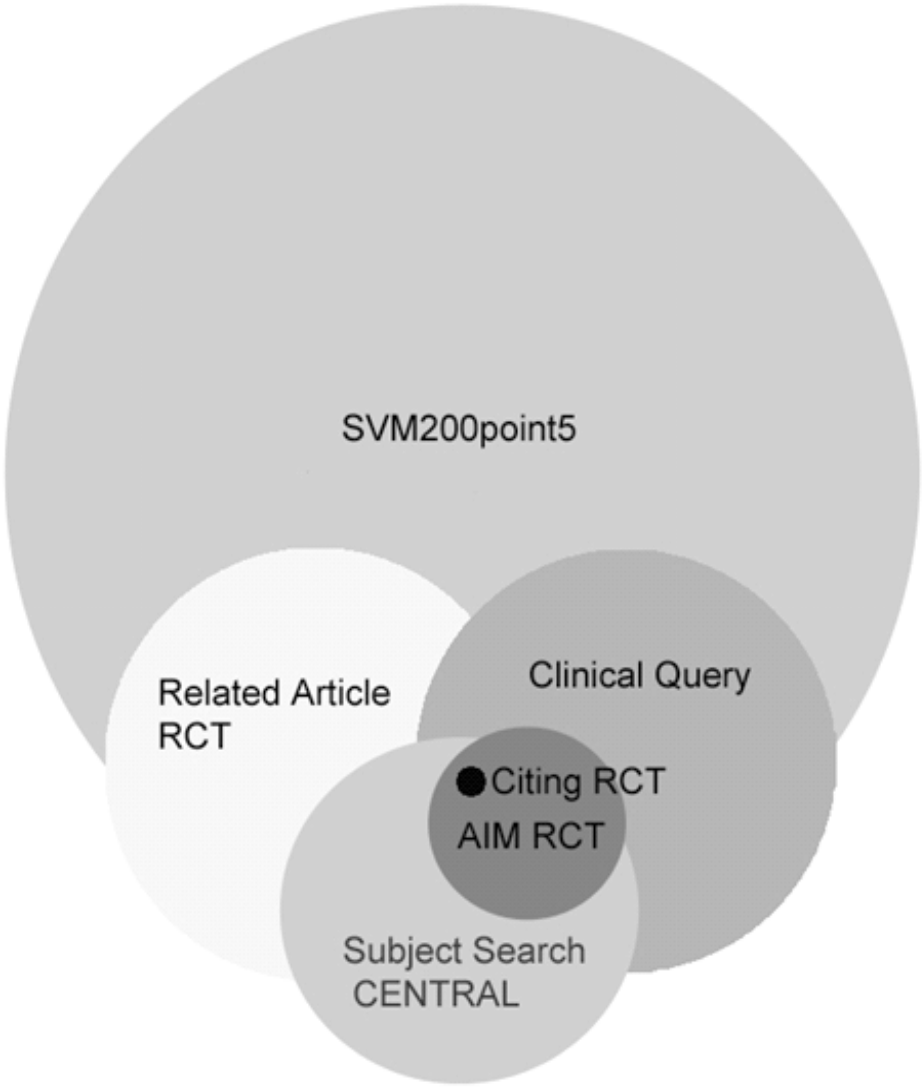


Figure 26. Unique Contribution and Overlap in the Cochrane Cohort

6.18.2 Unique Contribution and Overlap in the AHRQ Evidence Reports Cohort

Clinical Queries, Related Articles RCT, Citing RCT and Support Vector Machine searches all identified some material not retrieved by any other search method. The AIM RCT search is narrower than the Clinical Query, thus that search had no unique component. This cohort provides the only opportunity, in these experiments, to assess the unique contribution of Vector Machine. In addition, unlike the Cochrane cohort, data is

available on whether the unique material was *On Topic, Eligible,* and whether it identified novel material (reports with new n attributed). Overlap of retrievals is shown in Figure 27, - this is based on the entire search retrieval, regardless of relevance status of the records retrieved. Exact figures for unique material, unique *On Topic* records, *Eligible* records and records with new N attributed are given in Table 50.

Table 51. Unique Component of Each Search for AHRQ Evidence Reports

	Unique retrievals n=1945	Unique on topic retrievals n=452	Unique eligible retrievals n=101	Unique eligible studies with N attributed n=55	Number needed to screen (for new n)
Clinical Query	1060	272	40	26	40.77
Abridged Index Medicus RCT	0	0	0	0	-
Citing RCT	31	14	0	0	-
Related Article RCT	466	103	35	12	38.83
SVM200point5	388	63	26	17	22.82

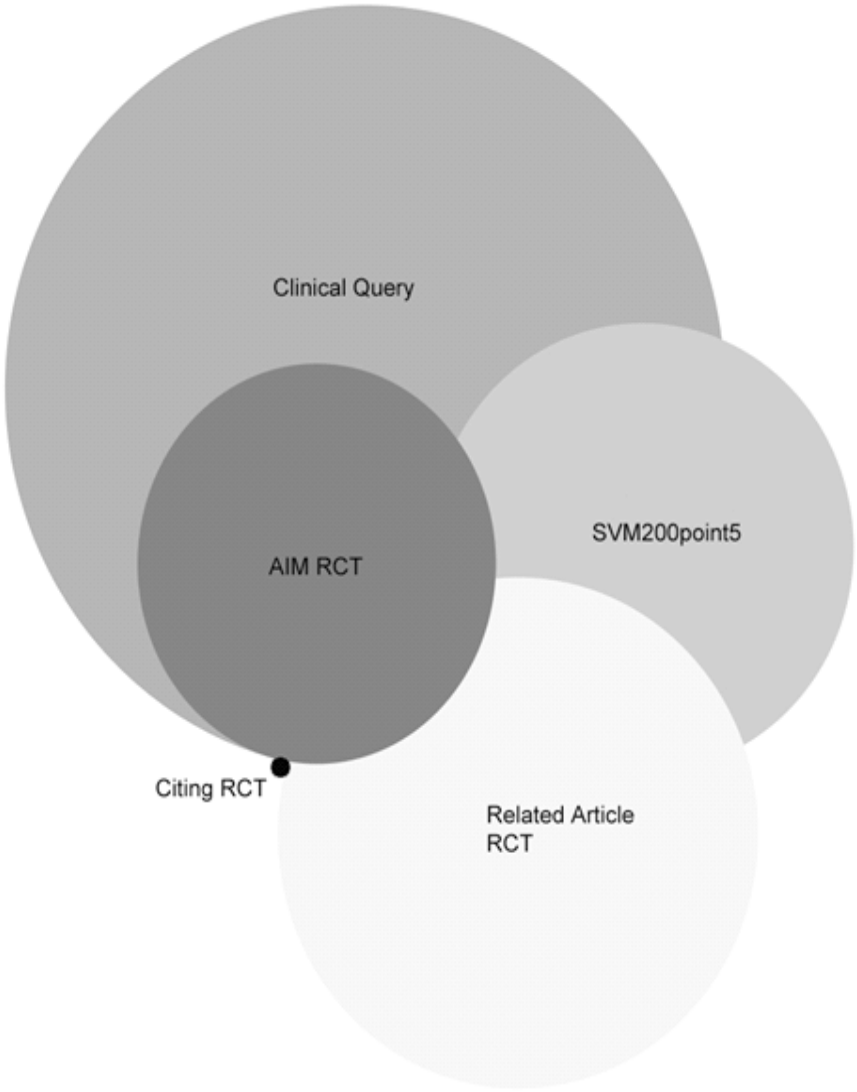


Figure 27. Overlap and Unique Component of Test Searches in the AHRQ Cohort

In total, there were 928 *On Topic* new records and 283 *Eligible* new records, and 35.69% of these *Eligible* new records (n=101) were detected by only one of the search methods. The proportion of unique material that was eligible ranged from a high of 0.0751 for the Related Article RCT to 0.0377 for Clinical Queries.

In terms of impact on updating, the identification of previously unidentified studies is arguably more important - of the *Eligible* new retrievals, 45.54% had no new N attributed as the study had previously been identified through another report.

It is notable that while SVM methods provided the largest unique retrievals for the Cochrane cohort, Clinical Query is the more significant contributor in the AHRQ cohort.

6.19 CONVERGENCE OF MULTIPLE RETRIEVAL METHODS

In the main cohort of 77 the mean number of methods retrieving any one relevant record was 1.88 (standard deviation 1.23), the mean number of methods retrieving any one irrelevant record was 1.10 (standard deviation 0.40). Association between number of retrieving searches and relevance was examined (Methods 4.2.2). Convergence measured either by the number of search methods retrieving a candidate or the number of search types (where two methods based on subject searching, Clinical Query and Abridged Index Medicus RCTs, are considered a single type of search) retrieving a candidate was associated with eligibility (For searches χ_{26} = 1283.5 $P_{2\ sided}$ <.001, for search types χ_{24}=754.8, $P_{2\ sided}$ <.001). Thus the more times an item was retrieved, the more likely it was that it was relevant. Table **52** through Table 54 show the distributions of off topic, *On Topic* and *Eligible* retrievals by number of times the record was retrieved. The distributions are also presented graphically in Figure 28, which shows complete details, and Figure 29 which focuses on *On Topic* (labeled as Not Relevant) and *Eligible* records (labeled as Relevant) using a larger scale.

When the four searches are considered (Table **52**), we see the very high proportion of relevant records among those nominated by the review team and not found by any of the test searches, but otherwise there is a steady and large gain in proportion of retrieved records that are eligible as the number of searches finding the item increases. Although overall, only 0.05 of the records considered in this analysis were *Eligible,* over a third of those retrieved by all search methods were relevant (precision=0.37). The pattern is similar for the cases in which the number of types of searches, where types are similarity search, Boolean search and citing reference search (Table 53).

Recall declines as precision increases. Because so few relevant records are Citing RCTs, this limits the recall that can be achieved by this convergence approach. When three of the four searches find the record, recall is 0.19 and precision is 0.149. When two

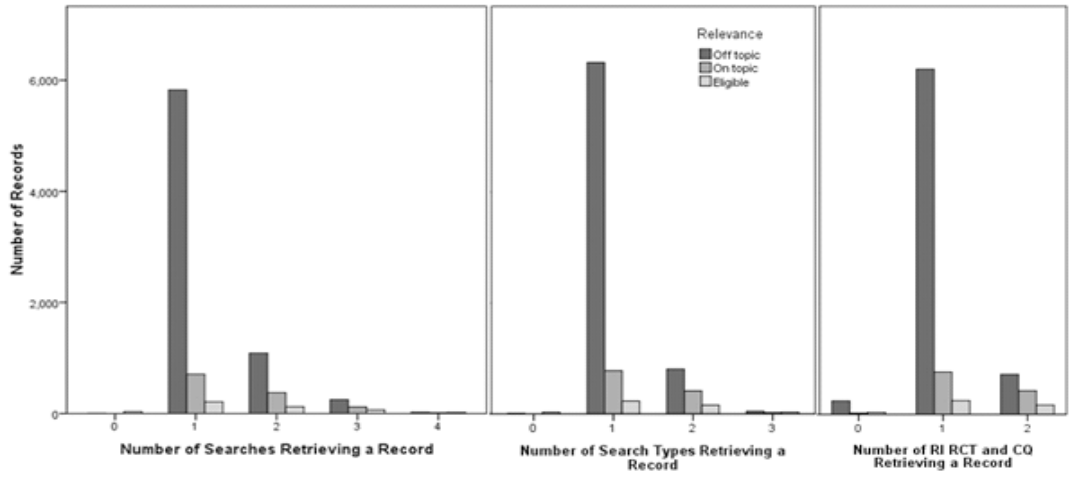


Figure 28. All Relevance Categories by Number of Searches and by Number of Methods Retrieving the Record

of the three search approaches find the record, recall is 0.42 and precision is 0.115. When both Related article RCT and Clinical Query find the record, recall is 0.37 and precision is 0.128.

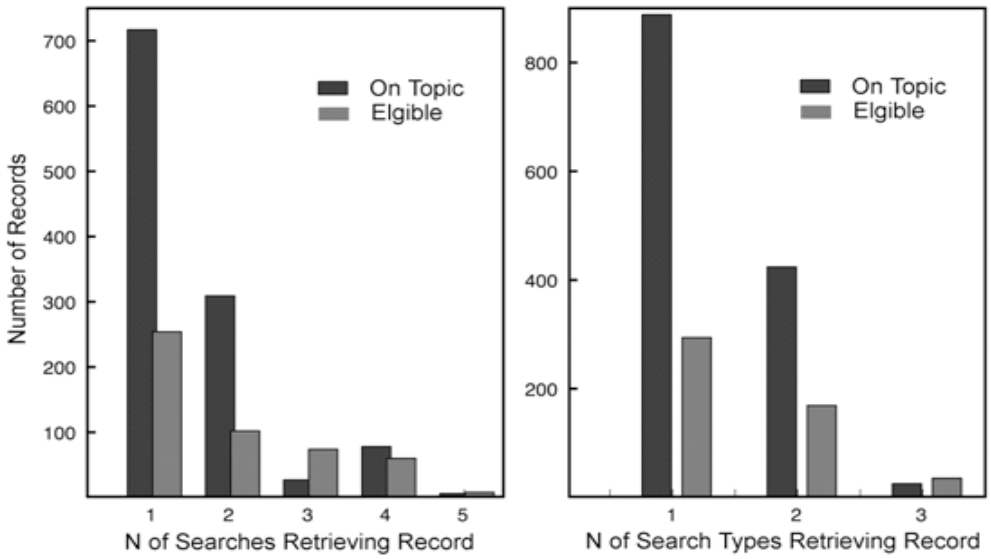


Figure 29. On Topic and Eligible Status by Number of Searches and by Number of Methods Retrieving the Record

Table 52. Relevance by Number of Searches Retrieving a Candidate in the Main Cohort

Found by x of 4 searches*	Off topic	On topic	Eligible	Total	Recall†	Precision
4	25	13	22	60	0.05	0.367
3	251	120	65	436	0.19	0.149
2	1090	376	124	1590	0.46	0.078
1	5827	708	212	6747	0.93	0.031
0	5	0	33	38	1.00	0.868
Total	7198	1217	456	8871		0.051

*Searches are Related Article RCT, Clinical Query, Abridged Index Medicus RCT, and Citing RCT

†Recall is cumulative – that is, while only 5% of records were retrieved by 4 searches, 19% were retrieved by at least 3 of the searches, 46% were retrieved by at least 2 of the searches, and so on.

Table 53. Relevance by Number of Types of Searches Retrieving a Candidate in the Main Cohort

Found by x of 3 search types*	Off Topic	On Topic	Eligible	Total	Recall†	Precision
0	5	0	33	38	1.00	0.868
1	6337	777	231	7345	0.93	0.031
2	809	415	159	1383	0.42	0.115
3	47	25	33	105	0.07	0.314
Total	7198	1217	456	8871		0.051

*Search types are similarity search, Boolean search and citing reference search.

†Recall is cumulative – that is, while only 7% of records were retrieved by all 3 search types, 42% were retrieved by at least 2 of the types, and so on.

Table 54. Relevance by Number of Related Article RCT and Clinical Query Searches Retrieving a Candidate in the Main Cohort

Found by x of Related Article RCT and Clinical Query	Off Topic	On Topic	Eligible	Total	Recall*	Precision
2	721	427	169	1317	0.37	0.128
1	6233	767	249	7249	0.92	0.034
0	244	23	38	305	1.00	0.125
Total	7198	1217	456	8871		0.051

*Recall is cumulative – that is, while 37% of records were retrieved by both searches, 92% were retrieved by one or the other or both.

The figure shows that, across the five search types considered here (Clinical Query, Abridged Index Medicus RCT, Citing Reference, Related Article RCT and CENTRAL)

6.20 POPULATION ESTIMATION THROUGH CAPTURE – RECAPTURE

Capture-Recapture analysis was undertaken to estimate whether the search methods were likely to have missed relevant studies and to explore the dependencies between searches (See Methods 4.2.5).

Applying this method to the current data, the best estimate should be obtained by the two methods with highest recall and most likely to be independent. These are Related Article RCT and Clinical Query. *Eligible* records retrieved by either Related Article RCT or Clinical Query were selected from all sets, and a cross tabulation was performed (Table 55).

Table 55. Cross-tabulation of Retrieval Status for Clinical Query and Related Article RCT.

		Clinical Query		
		Retrieved	Not Retrieved	Total
Related Article RCT	Retrieved	282	246	528
	Not Retrieved	148	0	148
	Total	430	246	676

The number of *Eligible* studies missed by both is nil in this table, as only records found by one or the other are considered, but the expected number missed is calculated using Bailey's estimator and Jensen's confidence intervals:

M = 430

n = 528

m = 282

$$N = \frac{M(n+1)}{m+1}$$

$$= ((430 * 529)/283)$$

The total population size is estimated to be 804 (95% confidence interval: 768, 843). The number of relevant records that appear to be missing after these two searches have been conducted is 128. Recall after two methods, Related Article RCT and Clinical Query, would therefore be 0.84. The predicted total population size can be compared to the number of *Eligible* records identified by any method, including reviewer nomination. Across the three cohorts, 752 *Eligible* records were found, below the lower level of the confidence interval of the population estimate, so approximately 52 *Eligible* records were still unidentified (confidence interval; 16, 90). Thus true recall could range from 0.89 – 0.98.

In the sample, the true number of relevant articles is likely most closely approximated in the AHRQ Evidence Report. All records retrieved by the searches were assessed directly, with no reliance on identification through author searches in updated systematic reviews (this reliance occurs in the large set as the first approach was to look at newer systematic reviews on the same or similar topics). In the other samples, it is possible or even likely that there were additional records in the retrievals that were relevant.

6.20.1 Positive Dependence

Capture-recapture should underestimate the true population size (here, the actual number of relevant records) when the samples are not independent. This allows us to

explore the dependence between searches in these cohorts. Positive dependence occurs when the probability of overlap is equal to the average probabilities of appearing in A and of appearing in B, and positive dependence of sources will tend to produce an underestimate of the true population size.[218] If the probability of ascertainment (of being retrieved by a search method) is taken to equal recall, than if the product of recall of method 1 and recall of method 2 is equal to the overlap between the two methods, there is a potential for underestimation. When one search result is a subset of another, dependence is complete. This can be demonstrated by examining estimates derived from Abridged Index Medicus RCT and Clinical Query, as the Abridged Index Medicus RCT retrieval is a subset of Clinical Query retrieval.

Relative recall of Abridged Index Medicus RCT is estimated at 0.22 across the three cohorts, and the recall of Clinical Query is estimated at 0.56 (Table 31). The product of these is 0.22 x 0.56 or 0.12. The overlap is =2936/13539 records or 0.22, much greater than the product. Calculating the capture-recapture statistics, the estimated number of targets is 430 (95% confidence interval; 430, 430) while the actual number of *Eligible* records identified by all searches was 752. Therefore, capture-recapture in this case where one search is a subset of the other, underestimates the true population size considerably.

6.20.2 Negative Dependence

Hook continues by demonstrating that if, the probability of being in both samples is less than the product of the probability of being in any one of them, as is the case when there is no overlap between methods, than there is negative dependence and capture-recapture methods will tend to overestimate the true population size. When the upper boundary of the confidence interval for a population estimate is less than the actual number of eligible records, we can attribute the result to underestimation, and infer positive dependence between methods. When the actual eligible records found was outside the lower limit of the estimate, this could be due to missed records, or due to negative dependence between sets. While the existence of missed records was the more plausible explanation, negative dependence cannot be ruled out.

6.20.3 Capture – Recapture Population Estimates when Recall of the Ascertainment Methods is Low

What happens when searches can be expected to be independent, but recall is low? This would be the case when Abridged Index Medicus RCT and Citing RCTs searches are used to estimate population size. Based on the capture-recapture calculations for all three cohorts combined, the estimate of total *Eligible* records is only 412 (confidence interval; 325, 562), while there are 752 known *Eligible* records across these three cohorts. Therefore, capture-recapture underestimates the true population size in this case. Substituting Clinical Query for Abridged Index Medicus RCT, the estimated population becomes 713 (confidence interval; 599, 880), and the number of known *Eligible* records falls within the estimate. Although it would seem that low recall of both searches impairs estimation, one is still faced with a competing hypothesis that there is a positive dependence between the Abridged Index Medicus RCTs and Citing RCTs. This could occur if the journals in the Abridged Index Medicus journal set are more likely to insist on a complete literature review, with relevant prior evidence cited, than other MEDLINE-indexed journals. This is easily explored. Across all three cohorts there were 838 citing RCTs, of which 835 were indexed in MEDLINE. Of these 835 records, 321 were from Core Clinical Journals (38.4%). Compare this to the 13364 records retrieved by the Clinical Query searches, of which only 2936 (22%) were in Core Clinical Journals.

Comparing the estimate based on Abridged Index Medicus and a similarity search with the estimate of Clinical Query and a similarity search, we see that using the lower recall Boolean search method does lower the overall estimate, although the number is plausible – the estimate of 770 records (confidence interval; 716, 834) is very similar to the 752 *Eligible* records actually found whereas the estimate with Clinical Query was 826 (confidence interval; 792, 863).

In a less extreme example, a capture-recapture population estimate based on Abridged Index Medicus RCT and Related Article RCT can be compared with estimates obtained using Clinical Query and Related Article RCT as the ascertainment methods.

The contrast is between lower and higher recall is less marked when only one of the searches is low recall; the lower recall pairing results in a population estimate of 719 (confidence interval; 667, 781) while substituting the higher recall Boolean search yields an estimate of 804 (confidence interval; 768, 843). The known population falls within the range of the first estimate, but is lower than the second.

Thus, it seems that the capture – recapture method is somewhat sensitive to the recall of the searches used to make the estimates. Use of capture-recapture method to estimate completeness of retrieval seems not to be warranted unless dependencies between methods are understood and recall of the methods used is expected to be high.

6.20.4 Capture-Recapture Estimates from Various Ascertainment Methods

The population estimates derived from other pairs of ascertainment methods are shown in Table 56 and Figure 30.

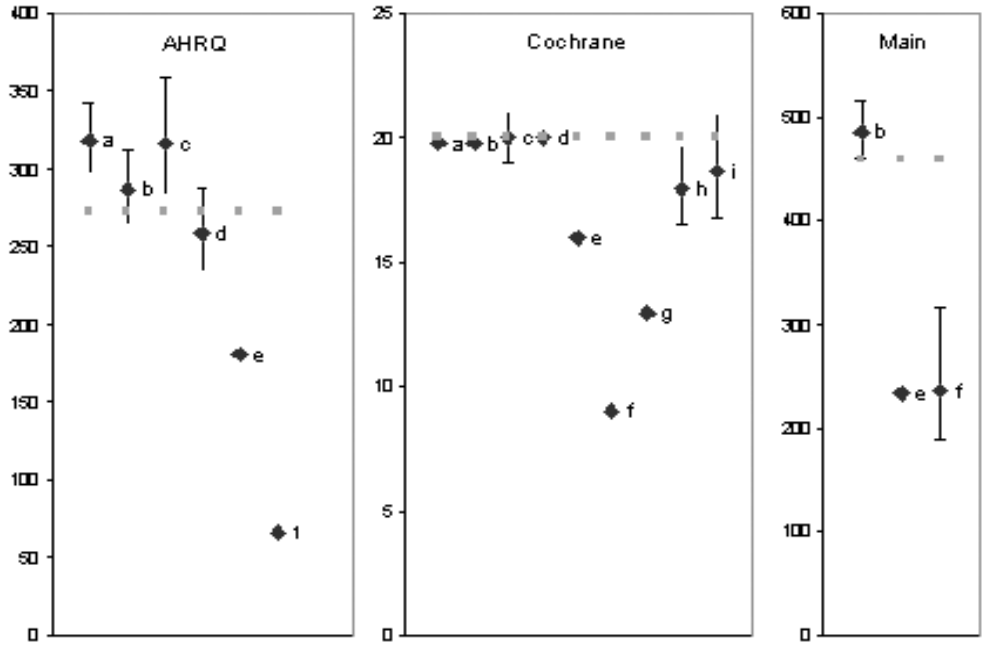


Figure 30. Capture-Recapture Population Estimates Based on Pairs of Search Methods

Diamonds represent the estimated number of eligible records, whiskers show the confidence intervals. Squares represent the eligible studies found by all methods including searches, studies included in updates, and nominations.

Search pairings are; a - Clinical Query and RI_RCTorSVM200point5, b - Clinical Query and Related Article RCT, c - Clinical Query and SVM200point5, d - Related Article RCT and SVM200Point5, e - Clinical Query and AIM RCT, f - AIM RCT and Citing RCT, g -

Author's actual search for the original review and authors' update search in CENTRAL, h - Author's actual search for the original review and our subject search in CENTRAL, i - Clinical Query and update search in CENTRAL.

Table 56. Comparison of Eligible Records and Capture-Recapture Population Estimates in the Three Cohorts

	Ascertainment Methods for Capture-Recapture Estimates	Eligible Identified from all Methods	Capture-Recapture Population Estimate from these Searches (Confidence Interval)	Estimated Records Missed by Searches Used*	Estimated Missed Eligible Records by all Methods
			AHRQ Evidence Reports		
a	CQ, RI RCTorSVM200point5	273	319 (298, 343)	49	46
b	CQ, RI RCT	273	286 (265, 312)	16	13
c	CQ, SVM200point5	273	317 (284, 359)	82	44
d	RI RCT, SVM200point5	273	258 (234, 288)	53	-
e	CQ, AIM RCT	273	181 (181, 181)	†	-
f	AIM RCT, CR	273	66 ‡	†	-
			Updated Cochrane		
a	CQ, RI RCT&SVM200point5	20	20 (20, 20)	0	0
b	CQ, RI RCT	20	20 (20, 20)	0	0
c	CQ, SVM200point5	20	20 (19, 21)	0	0
d	RI RCT, SVM200point5	20	20 (20, 20)	0	0
e	CQ, AIM RCT	20	16 (16, 16)	†	-
f	AIM RCT, CR	20	9 (9, 9)	†	-
g	ACTUAL, UCENTRAL	20	13 (13, 13)	†	-
h	ACTUAL, SSCENTRAL	20	18 (17, 20)	†	-
i	CQ, UCENTRAL	20	19 (17, 21)	1	1
			Main Cohort		
b	CQ, RI RCT **	459	487 (461, 516)	69	28
e	CQ, AIM RCT	459	233 (233, 233)	†	-
f	AIM RCT, CR	459	236 (188, 317)	†	-

Abbreviations: CQ is Clinical Query, AIM RCT is Abridged Index Medicus RCT, RI RCT is Related Article RCTs, SVM200point5 is the support vector machine retrieval limited to the first 200 records or those with relevance of .5 or greater, CR is Citing RCT, RI RCT&SVM200point5 is the joint retrieval of the RI RCT and SVM200point5 searches, ACTUAL is the actual subject search of the original review, UCENTRAL is the update search used in CENTRAL by the original reviewers, SSCENTRAL is the test subject search run in CENTRAL

*This may indicate that there are missed relevant records, or may be an overestimate due to source independence.
† Underestimates
‡ The confidence interval can't be calculated as the second source (citing RCTs) did not identify any eligible records.
**CQ & RI RCT and CQ & RI RCTorSVM200point5 would yield identical estimates in this cohort as SVM200pint5 was not run.

The ascertainment pairings *a, b, c*, and *d*, involving combinations of Clinical Query, Related Article RCT, and SVM200point5 all yielded population estimates relatively consistent with observed values. There is indication of some missed records in the AHRQ Evidence Reports and in the main cohort. Only in the two combinations involving SVM200point5 did the actual number of *Eligible* records fall below the confidence interval of the population estimate from capture-recapture. The first case is the pairing of Clinical Query with "RIRCTorSVM200point5", where "RI RCTorSVM200point5" represents records retreived by either similarity search method, and Clinical Query with SVM200point5. This could indicate negative dependence between Clinical Query and SVM200point5 or it could be that there is positive dependence between Clinical Query and Related Article RCT, which is lower the other estimates.

Searches pairings *e* (Clinical Query and AIM RCT) and *f* (AIM RCT and Citing RCT) underestimate the actual number of *Eligible* records in all cohorts, indicating positive dependence between those pairs, possibly with the relatively low recall of Abridged Index Medicus and Citing RCTs contributing.

Looking at the Updated Cochrane reviews, the three ascertainment pairings involving CENTRAL - *g* (Author's actual MEDLINE search for the original review and authors' update search in CENTRAL), *h* (Author's actual search for the original review and our subject search in CENTRAL) and *i* (Clinical Query and update search in CENTRAL) - all underestimate the actual number of *Eligible* records to some degree. In the case of pairing *g*, the number of *Eligible* studies is well above the confidence interval for the capture- recapture estimate.

6.21 MULTIDIMENSIONAL SCALING

Modeling the similarities between search results (see Methods 4.3.1), the derived stimulus configurations for the combined AHRQ and Cochrane Cohorts are shown in Figure 31 and for the main cohort in Figure 32. These are two dimensional multidimensional scaling analyses, with the number of dimensions was decided *a priori* due to the limited number of input variables (called stimuli in these analyses). Stress (a measure of goodness of fit) with this number of dimensions was very low at .00092 for the AHRQ and Cochrane cohort and 0.00381 for the analysis of the main cohort, considered excellent. Examination of distance in the model compared to the original similarity scores did not identify outliers for either analysis, again suggesting that the data are well described by the two dimensional model.

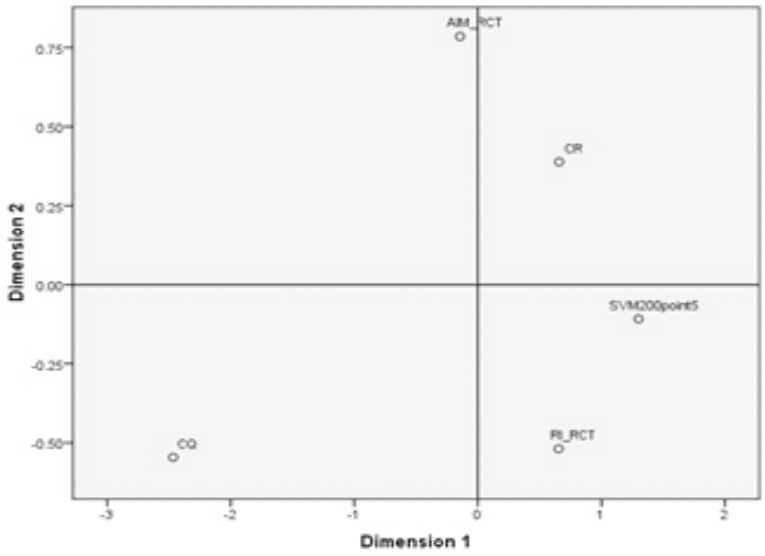


Figure 31. Derived Stimulus Configuration from Multidimensional Scaling for the Updated Cochrane Reviews and AHRQ Evidence Reports

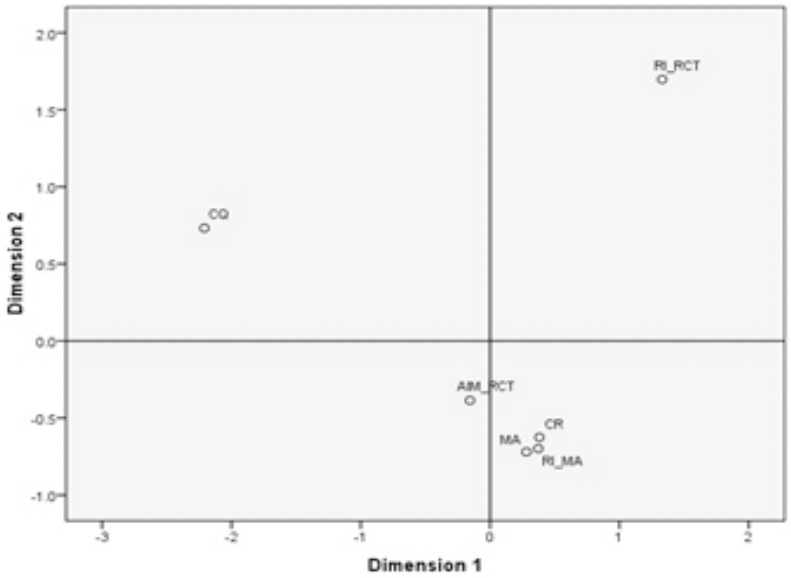


Figure 32. Derived Stimulus Configuration from Multidimensional Scaling for the Main Cohort

Examining the plots, the two subject searches are alone on the negative scale of the first dimension. Related articles RCT is alone in a quadrant for the main cohort, and shares a quadrant only with the Support Vector Machine in the cohorts where SVM was used.

Citing RCT is in its own quadrant except in the main cohort when it is joined by the searches for newer meta-analyses. Although one meta-analysis search approach (labeled MA) is the subject search with the MEDLINE meta-analysis publication type limit applied, and the other (labeled RI_MA) is the Related Article search with the meta-analysis publication type instead of the randomized controlled publication type limit, these to not group with their search types. Of 818 distinct citing RCTs, only 12 have the publication type Meta-analysis, and only four were retrieved by the subject search with the MA publication type limit, and four by the related article search with the publication type limit, so overlap between the sets is not an explanatory factor in the close grouping of these searches. One aspect that could explain the proximity is recency (we enforced a one year lag from the search date of the original review for the MA searches, and citing RCTs would be offset from the publication of the original reviews by a publication lag). Another aspect is that both the MAs and the citing RCTs would both be expected to cite prior RCTs. Finally, both the searches for newer meta-analyses and the searches for citing RCTs had smaller retrievals than the other search methods.

The capture-recapture analysis suggested a positive dependency between Citing RCT and Abridged Index Medicus RCT search methods. In both these multidimensional scaling analyses, the test search closest to Citing RCTs is Abridged Index Medicus RCT (although the distance from Citing RCTs to SVM is only slightly greater than the distance from Citing RCTs is Abridged Index Medicus in the AHRQ-Cochrane analysis).

The most parsimonious interpretation of Dimension 1 in both analyses is type of search, with subject searches at one end of the scale, and similarity methods on the other. Searches for publications that cite RCTs are in the middle. The second dimension is less clear-cut, but is likely influence heavily by retrieval size and precision (which are correlated) but not recall. These results are congruent with the dependencies implied by the capture-recapture analysis.

6.22 CORRELATIONS OF ACTUAL AND TEST SEARCHES

The poor performance of the MEDLINE searches used by the authors in updating these systematic reviews makes it difficult to assess the hypothesis arising from Cohen's work[15] that searches where the librarian is able to achieve precision are also where the classifiers are able to obtain good precision, however, this was examined briefly (see Methods 4.3.2). The authors' Actual searches only identified relevant included studies in two of the six updated systematic reviews. As precision is calculated with relevant retrievals as the numerator, precision is zero when no relevant studies are found.

The precision of the various test searches is shown in Figure 33. It does seem that precision of both the Boolean searches (shown with solid lines) tends to rise and fall as the precision of the SVM searches (shown with dashed lines) and Related Article RCT (shown with a dotted line).

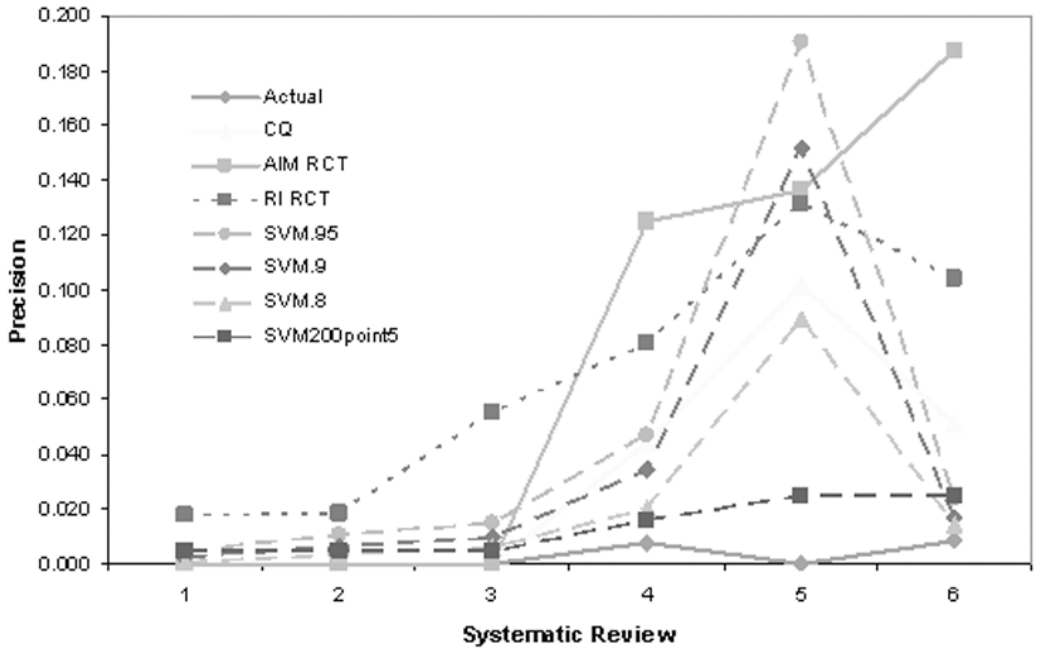


Figure 33. Correlations of Actual and Test Searches in the Updated Cochrane Review

6.23 RELATED ARTICLE SEED REFINEMENT

6.23.1 Optimizing the Selection Criteria for Seeds

Characteristics of related article seed articles were studied in the Cochrane cohort for indications of how the search algorithm could be refined (see Methods 4.2.3). Seeds were classified based on qualifications as seeds (independent variable) and performance (dependent variable). Seeds could qualify by being one of the three largest or one of the three newest studies included in the original review, and some met both criteria. Their performance could be productive, when one or more new included studies were related to that seed, or unproductive, when none of the related articles were new included studies. Super seeds were productive seeds that were related to all new included studies in the updated reviews. In some cases, an included study was eligible as a seed, but was not indexed in MEDLINE and so could not be searched for related articles.

Each systematic review could have a maximum of six related article seed articles. In this cohort, there were only 26 distinct MEDLINE-indexed seeds (Table 57). Seven seeds qualified by being both among both the newest and largest, two were super seeds, two were otherwise productive seeds and thee were unproductive seeds. Three articles that qualified as seeds were not indexed in MEDLINE and so could not be used in the Related Article RCT search.

Table 57. Productivity of the Related Article Seeds in the Cochrane Cohort

	Frequency	Percent
Super seeds	5	17.2
Other productive seeds	9	31.0
Unproductive productive seed	12	41.4
Non-MEDLINE	3	10.3
Total	29	100.0

Mean number of related RCTs for each seed was 262.2 (Standard deviation 163.1), median number of related RCTS for each seed was 222.5 (1st and 3rd quartile

130.3, 307.5). It was previously seen that none of the related article RCT searches for this cohort resulted in null retrievals (Table 22). In fact, all MEDLINE-indexed seeds yielded related RCTs newer than the search date of the original review, with a minimum retrieval of 103 related RCTs and a maximum of 666 related RCTs. There was overlap between the retrievals for each systematic review, the mean retrieval size was 952.5 (standard deviation; 220.7). There was no difference in retrieval size between seeds that included one or more new studies added in the update and those that did not retrieve any of the relevant new studies. (Table 58**,** $t_{24} = -0.378$, $P_{2\ sided} = 0.709$).

Table 58. Number of Newer Related Randomized Controlled Trials for Seeds that Did or Did not Retrieve New Studies Added in the Cochrane Updates

	Productive	N	Mean	Std. Deviation	Std. Error Mean
N of Related	Unproductive	12	248.92	146.030	42.155
	Productive	14	273.57	181.212	48.431

Turning to the qualifications of the seeds as determinants of productivity (Table 59), it is notable that all five super seeds were new studies (two were also large studies) but qualifications were not significantly associated with productivity when the MEDLINE-indexed seeds were considered ($\chi^2_2 = 1.5$, $P_{2\ sided} = 0.462$).

Table 59. Productivity of Largest and Newest Related Article Seeds for the Cochrane Cohort

	Qualification		
	Large	New	Total*
Super Seed	2	5	7
Other Productive Seed	6	6	12
Unproductive Seed	8	6	14
Non-MEDLINE	2	1	3
Total	18	18	36

*Seeds qualifying by being among both the newest and biggest are counted in each category.

Position within newest or largest was assessed to see if the number of seeds of each type could be trimmed (Table 60). Non-MEDLINE seeds where excluded from the analysis. MEDLINE-indexed records that were qualified both as one of the newest and as one of the largest studies were included in the analysis one for each qualification. Again, there was no association between position and productivity (χ^2_4=4.1, $P_{2\ sided}$ =0.397). Overall, analysis of variance with qualification and position as independent variables and productivity as the dependent variable was not significant.

Seven seeds qualified by being both amongst the newest and amongst the largest studies included in the original Cochrane reviews. Five of the six Cochrane reviews in this cohort have at least one doubly-qualified seed. To examine whether these seeds were more productive than others that met only one criterion, seeds were classified dichotomously as Productive (with Super and Other productive seeds combined) vs. Unproductive (unproductive or non-MEDLINE seeds combined) and also as Super vs. Other. The association between these variables and whether seeds were doubly qualified was assessed through χ^2, which did not approach significance in either case (Productive/Unproductive χ^2_1=2.0, $P_{2\ sided}$ =0.159, Super/Other χ^2_1=0.8, $P_{2\ sided}$ =0.362).

Table 60. Productivity of Related Article Seeds by Position for the Cochrane Cohort

Position	Super	Other Productive	Unproductive	Total
1	3	2	7	12
2	2	5	5	12
3	2	5	2	9
Total	7	12	14	33

6.23.2 MEDLINE Misses as Seed Articles

Twelve of the articles that qualified as seeds for the related article search were indexed in MEDLINE but missed by the MEDLINE search used by the authors in the original systematic review (Table 61). These MEDLINE misses were productive in almost equal proportion to those seeds found by the authors' search in the original review $\chi^2_1=0.1$, $P_{2\ sided}=0.716$.

Table 61. Productivity of Seeds by Retrieval Status in the Original Cochrane Review

Productivity	Found	Missed	Total
Unproductive	6	6	12
Productive	8	6	14
Total	14	12	26

6.23.3 Performance of Related Article Searching when Retrieval Size is Limited

Retrieval sizes for the three cohorts are shown in Table 62 and these numbers have been drawn from Table 18, Table 22 and Table 25 where it can be seen that the related article search method yield larger retrievals, similar overall to Clinical Queries, but larger than the other search methods.

Table 62. Retrieval Sizes for Related Article RCT Searches in the Three Cohorts

	Median	1st Quartile	3rd Quartile	Maximum
Main Cohort	100	62	142	376
Updated Cochrane Reviews	67	56	88	174
AHRQ Evidence Reports	79	45	106	172

The feasibility of improving precision of the Related Article searches by limiting the retrieval size was explored in the updated Cochrane reviews and the AHRQ Evidence Reports. Position in the ranking of related article retrieval was recorded for each *Eligible* study, and the number of *Eligible* new studies that would have been missed had the retrieval been truncated at various sizes was determined. Recall and precision of the truncated searches was computed.

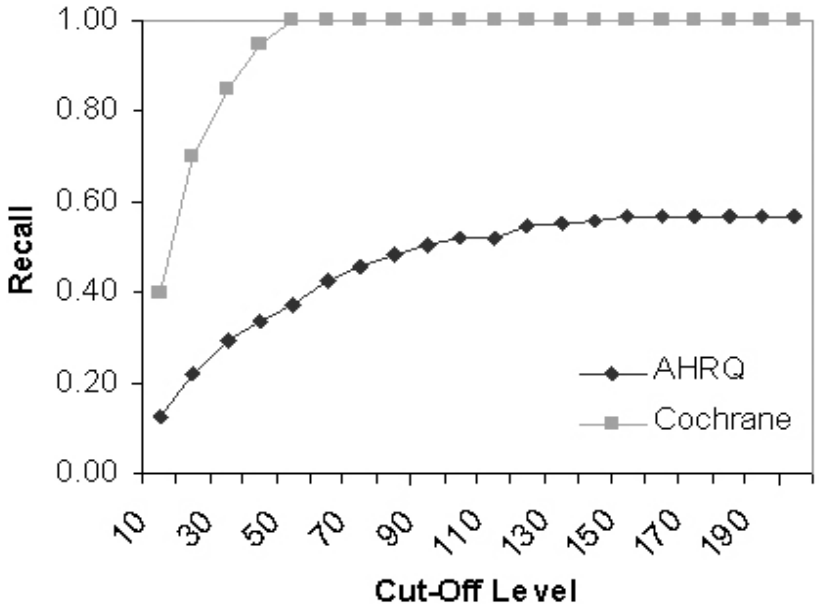


Figure 34. Recall at various Related Article Rankings for the AHRQ and Cochrane Cohorts

Complete recall of new studies added in the updates was achieved in the updated Cochrane reviews by the time the Related Article RCT set size reached 50, as the lowest ranking new study in the updates was in position 43 (Figure 34). In the AHRQ Cohort, where the Related Article RCT search has relative recall of 0.57, that maximum would

have been obtained had set size been limited to 150 records, as the lowest ranking new eligible study was in position 146 (Figure 34).

Precision increased considerably when the retrieval size was restricted (Figure 35). In the Cochrane cohort, full recall was obtained within the top 50 records, and precision would have been above 0.200 at that point. At a retrieval size of 100, precision would have been 0.033, which is a typical value for systematic reviews (see Section 6.25). At an arbitrary retrieval size of 200, precision would have dropped to 0.017, but this is artificial, as the Related Article RCT search returned a maximum of 174 records. In the AHRQ cohort, gains in precision are even greater, with precision exceeding 0.100 at rank 180, by which point all relevant records had be identified.

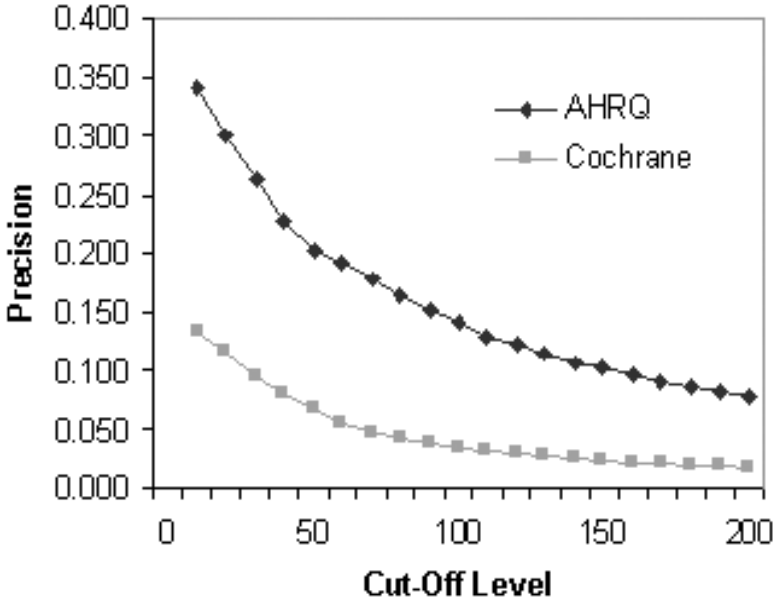


Figure 35. Precision at Various Related Article Rankings for the AHRQ and Cochrane Cohorts

Looking at the position of the lowest ranked relevant record in the Cochrane and AHRQ sets, it does not seem feasible to set a cut point as a proportion of the retrieval size – for instance, assessing only those with top half ranking (Table 63).

Table 63. Lowest Related Article RCT Rank of Eligible and Ineligible Records

Systematic Review	Lowest Ranked Eligible Record	Lowest Ranked Ineligible Record
Cochrane		
Alejandria, 2001[274]	9	56
Arroll,1999[275]	43	60
Cody, 2001[276]	2	20
Fisher 1994[277]	36	50
Demicheli, 2001[278]	28	53
Lengeler, 1999[279]	26	38
AHRQ		
Santaguida,2004[280]	116	117
Shekelle, 2003[201]	155	154
Berkman, 2000[281]	37	75
Duscemi, 2004[282]	96	116
Grady, 2003[283]	145	172
Guise, 2003[284]	--	15
McNamara, 2000[285]	50	58
Shekelle, 2003[254]	69	88
Shekelle, 2004[286]	24	41
Velmahos, 2000[202]	12	15

While there is the problem with larger set sizes in the Main cohort (the largest retrieval was 376 records), the set size exceeded 200 records in only 12 systematic reviews (15.6%) and exceeded 150 records in 21 systematic reviews (27.2%). The precision and recall seen at various limits of retrieval size are shown in Table 64.

Table 64. Search Performance of Related Article RCT at Various Limits to Set Size in the AHRQ Evidence Reports

Retrieval Size Cut Off	Retrieval size	Eligible Retrieved Total= 270	Recall	Precision
Unlimited	853	154	0.57	0.181
200	853	154	0.56	0.183
150	826	151	0.56	0.183
100	692	143	0.53	0.207
50	421	102	0.38	0.242

6.24 SUFFICIENT STRATEGIES

Seventy-two systematic reviews in the main cohort (those for which new *Eligible* studies were identified) were examined to determine which combinations of searches would have been sufficient to identify all of the *Eligible* studies found through any of the searches (see Methods 4.2.4).

Table 65. Search Strategies Sufficient to Retrieve All Eligible New Studies

	N*	%
One Strategy (N=59, 81.9%)		
Related Article RCT	46	63.9%
Clinical Query	34	47.2%
CENTRAL	14	19.4%
Abridged Index Medicus RCT	8	11.1%
Citing RCT	3	4.2%
Two Strategies Needed (N=12, 16.6%)		
Related Article RCT with Clinical Query	10	13.9%
Related Article RCT with CENTRAL	4	5.6%
Related Article RCT with Abridged Index Medicus RCT	1	1.4%
Related Article RCT with Citing RCT	1	1.4%
Three Strategies needed (N=1, 1.4%)		
Related Article RCT with Citing RCT with (Clinical Query or CENTRAL)	1	1.4%

* Some systematic reviews had more than one sufficient search, therefore figures sum to more than the number of cases (n=72).

In 59 of the systematic reviews, a single search strategy would have been sufficient to retrieve all *Eligible* new report found by any method. Related Article RCT was sufficient in 46 cases, Clinical Query in 34, Abridged Index Medicus RCT in 8, CENTRAL in 14 and Citing RCT in 3. Systematic reviews with few studies seem to be more likely to have multiple sufficient strategies.

In 12 cases, a combination of two search strategies was needed to find all *Eligible* new reports found by any method. Related Article RCT with Clinical Query was sufficient in ten cases, Related Article RCT with CENTRAL in four, Related Article RCT with Abridged Index Medicus RCT in one and Related Article with Citing RCT in one.

In only one case was it necessary to use three methods to find all found by any method. In that case, Related Article RCT with Citing RCT and either Clinical Query or CENTRAL would have been sufficient.

In 69 of the 72 systematic reviews examined (96%), searching Related Article RCT and Clinical Query would have retrieved all studies either because one strategy or the other was sufficient or because the two together was sufficient. For the three reviews that needed additional searches to retrieve all relevant new studies identified through any strategy, the combination needed was different in each case.

6.25 PRECISION OF SEARCHES IN A CROSS-SECTIONAL SAMPLE

Ninety-four of 300 (31%) systematic reviews in the cross-section sample of December 2006 reported their search results in enough detail that precision could be calculated[3] (see Methods 4.4). As a quality measure, the search was replicated for ten systematic reviews, mostly in cases where precision seemed very high.

In the cohort of 300, 43% of the systematic reviews studies were Cochrane reviews, while in the 94 cases where precision could be calculated, 45 (47%) were Cochrane reviews (χ^2_1 = 0.984 $P_{2\ sided}$ =.321). Although reporting is generally more complete in Cochrane reviews relative to journal published reviews, the greater use of QUOROM flow diagrams in journal-published reviews results in more complete reporting of the number of retrieved, screened and included records and studies.[3] QUOROM is "Quality of Reporting of Meta-analyses", a reporting standard published in 1999.[42] The flow diagram documents the number of possibly relevant reports identified, the number included, and the number excluded and the reasons for exclusion. Figures 17-19 at the beginning of chapter are examples of QUOROM flow diagrams.

Overall, there were 5,734 relevant included studies across the 94 systematic reviews and a total of 189,334 retrieved records were screened. The overall precision is therefore 0.030. The median precision, calculated by taking the median of the precision scores for the individual systematic reviews, is 0.029 (1st and 3rd quartiles; 0.013, 0.081). The maximum precision achieved by any search was 0.368 while the minimum was 0.0007. Median number of records screened was 634 (1st and 3rd quartiles; 169, 1612). Median number of included studies was 15.5 (1st and 3rd quartiles; 5, 30).

Six of the 94 for which precision could be calculated were described as updated systematic reviews. Here, the number of studies added in the update was used as the numerator and the number of records retrieved by the update search was used as the denominator. Median search precision for these was 0.023 (1st and 3rd quartiles; 0.016, 0.045).

The number of records screened is compared to the number of included studies in Figure 37 and the distribution of precision by retrieval size is shown in Figure 36.

In these figures, two outliers are excluded, one case with 10,578 records screened and 254 included studies and precision of 0.038,[287] and another with 64,586 records retrieved and 2443 included studies for precision of 0.024.[288] The Gulmezoglu review had the largest retrieval of any review of which I am personally aware. Not surprisingly, they published an article on the experience.[289]

All systematic reviews included in our cohorts and searched with the test searches were systematic reviews of interventions (See Methods 3.3). Intervention studies may benefit from the existence of search aids for identifying clinical trials, such as the Cochrane Central Register of Controlled Trials (CENTRAL), the Cochrane Highly Sensitive Search Strategies for MEDLINE, and the MEDLINE Randomized Controlled Trial publication type tag, and so may achieve higher precision than is possible in searches for other types of systematic reviews where such tools still need to be developed. Precision was calculated for each type of systematic review represented in the Epidemiology of Systematic Reviews cohort (Table 66, **Error! Reference source not found.**). Searches for systematic reviews of interventions show no better precision than other types of reviews, although numbers are very small in the other categories.

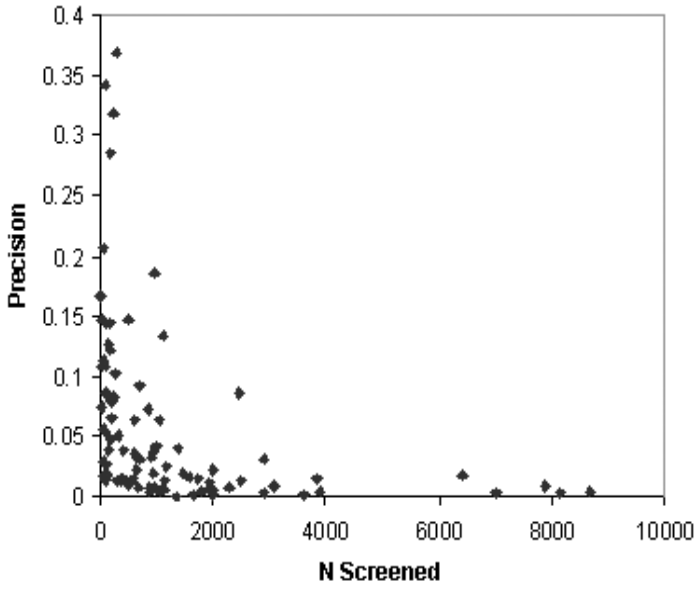


Figure 36. Relationship Between Screening Volume and Precision of the Search

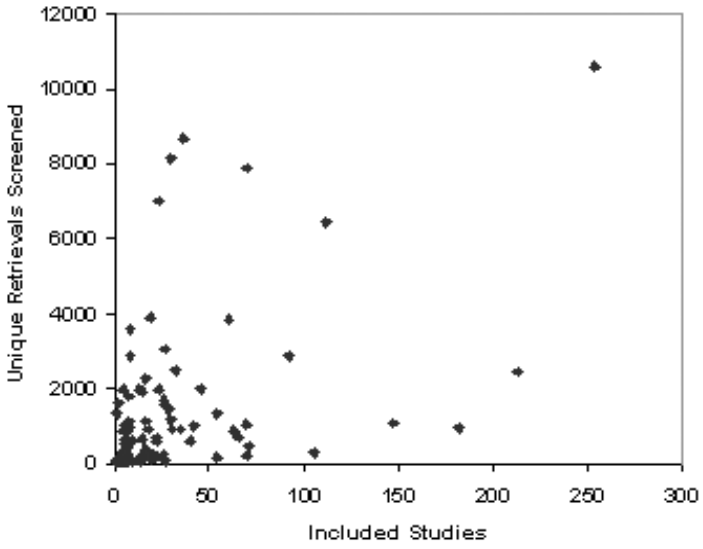


Figure 37. Relationship Between Screening Volume and Number of Included Studies

Table 66. Precision for Different Types of Systematic Reviews

Focus of the Review	N	Median	1st Quartile	3rd Quartile
Intervention (Treatment or Prevention)	71	0.027	0.011	0.076
Diagnosis	10	0.028	0.016	0.251
Epidemiology	4	0.040	0.029	0.103
Other*	9	0.065	0.014	0.103

* Other includes harms, education, associational studies, instrumentation and research methods.

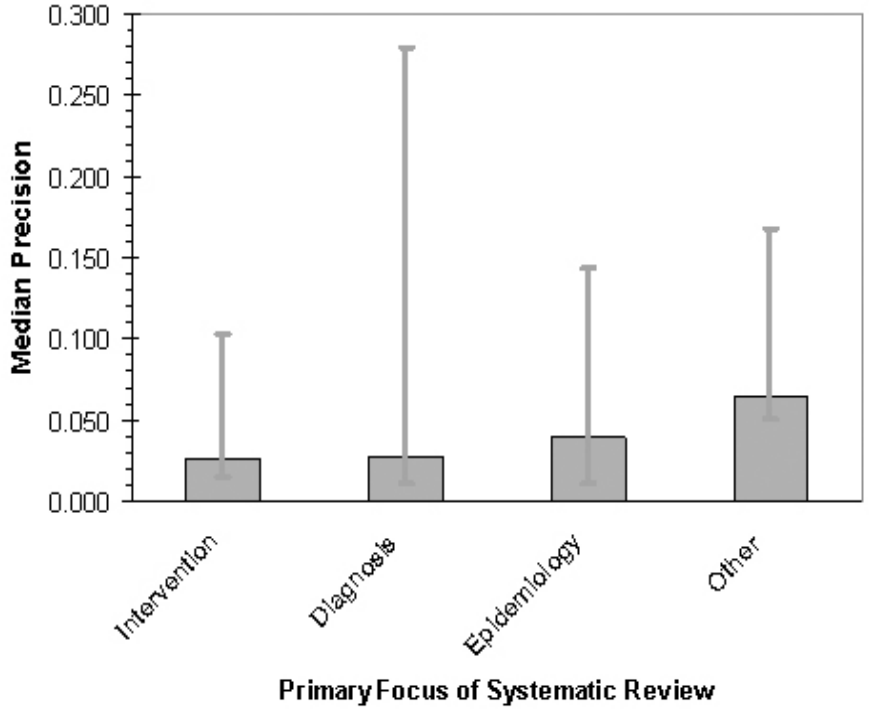


Figure 38. Median Precision by Review Focus

Sixty-eight of the 94 (72.3%) were described as restricted to randomized controlled trials. Median precision for these searches was 0.024 (1st and 3rd quartile; 0.013, 0.068) while median precision for the 25 systematic reviews not restricted to randomized controlled trials was 0.034 (1st and 3rd quartile; 0.015, 0.107).

6.26 RECALL OF THE AUTHORS' SEARCHES FROM THE ORIGINAL REVIEWS FOR UPDATING

In the Cochrane updates, all studies added in the updates were indexed in MEDLINE. Performance of the author search was assessed and found to be variable,

achieving perfect recall of new relevant studies in two Cochrane updates while failing to recall any of the new relevant studies in the other four. Overall recall of new evidence was 0.60.

AHRQ author searches showed a similar pattern of results compared to the new evidence found by the test searches. Four achieved perfect recall but the remaining six searches had recall below 0.30, and in two cases failed to recall any of the new evidence. Overall recall of new evidence was 0.52 for the AHRQ Evidence Report author searches.

The author search strategy used for the original review was modified for update in two Cochrane reviews (33.3%), and unchanged from the original in the other four reviews (66.6%). One modification improved recall of the MEDLINE-indexed studies included in the original review from 0.50 to 0.67, the other reduced it from 0.42 to nil.

Table 67. Recall of New Evidence by the Actual MEDLINE Search Used in the Original Review

	Recall of New Records	Recall of New N
Cochrane Updated Systematic Reviews (n=6)	0.60	0.15
AHRQ Evidence Reports (n=10)	0.52	0.77

Our practice of allocating all N to a single report, in those cases where multiple reports of the same study had been identified, could have lowered recall of new N. Therefore, the relevant new evidence was re-examined and, if any report of a relevant study was found, the N was credited. All studies added in the Cochrane updates had N attributed, so the allocation practice had no impact on those results. For the ten AHRQ reports, three had to be reexamined. In the other seven reports either all of the new studies were found by the authors' ACTUAL search (n=4), none of the new studies were found (n=2) or all relevant new studies had new N attributed (n=1). For the three systematic reviews that were re-examined, there no cases where re-attribution of N would change the result. In all cases, either the main report of the study had been found, or none of the reports of the same study had been found.

6.27 Recall of Original Evidence by Searches Used in the Original Reviews

The performance of the authors' MEDLINE searches in the original reviews was examined when the true positive and true negative training sets were being developed for SVM testing. First, known-item searches of included studies for each systematic review showed that the number of MEDLINE-indexed records was 97 for the Cochrane cohort and 972 for the AHRQ cohort. Of these, 396 (37%) were not found by MEDLINE searches used in the review – these are false negatives. Conversely, the 673 that were found represent recall of 0.63. AHRQ Evidence Reports had many more included studies than did Cochrane reviews (Table 68). Although the median number of included studies not retrieved by the MEDLINE search was higher for the AHRQ Evidence Reports, this was offset by the larger size of the reviews. Median recall of MEDLINE-indexed studies was 0.60 for AHRQ Evidence Reports and 0.49 for Cochrane reviews. Overall recall of the authors' search in AHRQ Evidence Reports was 0.64 and for Cochrane Reviews, 0.50.

Table 68. Number of Included Reports from the Original Reviews Indexed in Medline and not Retrieved by the Authors' Search for the Cochrane and AHRQ Cohorts

	MEDLINE indexed		Not retrieved by the MEDLINE search	
	Median	1st, 3rd Quartile	Median	1st, 3rd Quartile
Cochrane	17	13.75, 19.5	6.5	3.5 , 11
AHRQ	96	32.0, 121.7	23.5	15.25, 28.5
Overall	30	18.75, 111	15.5	7.25, 24.75

6.28 Related Article Searching as an Adjunct Search in the Original Review

In the cohort of updated Cochrane reviews, the reviewers' MEDLINE searches missed included studies that were indexed in MEDLINE in all six of the original systematic reviews. The search used by the authors in the update did not improve upon this performance (Table 67). The related article search was run to see how many of the

MEDLINE-indexed included studies could be detected, and how many of the studies missed by the authors' search would be detected (see Methods 4.6). Performance of the Related Article RCT search was good in five of the six Cochrane reviews (Table 69). Retrieval of MEDLINE misses by the Related Article RCT searches was 0.81 overall (43 of 53 retrieved), and recall was 0.95-1.00 in five of the six systematic reviews. Precision, estimated based on trimming the Related Article RCT set to those records entering MEDLINE prior to 2005, was adequate, ranging from 0.028 to 0.282 in the individual reviews, and was 0.068 overall.

Thus, related article searching appears to be a useful adjunct to the MEDLINE search undertaken by these review authors.

Table 69. Related Article RCT Performance in Retrieving Studies in the Original Reviews of the Updated Cochrane Cohort

	Cohort Review						
	Alejandria[274]	Arroll[275]	Cody[276]	Del Mar[177]	Demicheli[278]	Lengeler[279]	Overall
N of MEDLINE indexed studies included in the original review	23	6	12	15	20	21	97
Retrieved by Related Article RCT search	19	6	10	8	14	20	77
Recall of Related Article RCT search	0.83	1.00	0.83	0.53	0.70	0.95	0.79
Missed by Original Search	12	3	8	5-10†	1	19	53
Miss by Original Search but Retrieved by Related Article RCT	12	3	8	1	1	18	43
Recall of MEDLINE misses by Related Article RCT	1.00	1.00	1.00	0.10	1.00	0.95	0.81
Size of Related Article RCT retrieval*	362	213	162	152	169	71	1121
Precision of RI search	0.052	0.028	0.062	0.053	0.083	0.282	0.068

*All were date limited to 2005 or earlier
†High and low estimates under various interpretations of the authors' search

6.29 PERFORMANCE OF THE HSSS REVISED

The performance of the revised Highly Sensitive Search Strategy ($HSSS_{2006}$) was tested in the updating dataset (see Methods 4.7). Three sources of "relevant evidence" were used; the signaling evidence, targets from the six updated Cochrane reviews used to test SVM, and all *Eligible* evidence from the screened material, including candidate studies from the test searches and reviewer nominations. Removing overlap, there are 695

pieces of relevant evidence with PMIDs. The $HSSS_{2006}$ retrieved 689 of these (recall = 0.991).

Of six PMIDs not retrieved, five were from the candidate and reviewer nomination lists (n=687), none were from the new included studies in the updated Cochrane reviews (n=27), and two were from the final signaling evidence set (n=58).

In comparison, the original HSSS missed 23 records from the set of 687 *On Topic, Eligible* candidates and reviewer nominations (3.3%), two of the 27 added studies from the updated Cochrane reviews (7.4%) and one record from the final evidence set of 58 new studies (1.7%). Overall recall was 0.965.

The total retrieval size of the $HSSS_{2006}$ was 2,060,797 records and of the original HSSS was 784,592 records when tested in Ovid MEDLINE as of January 26, 2009.

6.30 CHARACTERISTICS OF THE EVIDENCE IN UPDATED SYSTEMATIC REVIEWS

6.30.1 Where Does The New Evidence Come From?

There were 421 unique new MEDLINE-indexed records found eligible for the systematic reviews in main cohort. These were examined to determine some bibliographic characteristics (see Method 4.8.1). There were 270 different *Eligible* PubMed records for the AHRQ Evidence reports, and 20 new *Eligible* records for the cohort of updated Cochrane reviews. There was overlap between sets, and when duplicate records were removed, the number of records dropped from 711 to 690. The broad bibliographic characteristics of the new evidence are shown in Table 70.

Table 70. Bibliographic Characteristics of New Evidence

Characteristic	N
Number of references	690
Authors	4045
MeSH terms	1528
Periodicals	245
Publication years	16

6.30.2 Journals Contributing New Evidence

The Abridged Index Medicus list includes 119 journal titles. Forty-five of those journals (37.8%) yielded new *Eligible* evidence, and Abridged Index Medicus Journals account for 43.3% of the new eligible records but make up only 18.3% of the journals represented.

Examining the distribution of journals by productivity, 11 journals each with 10 or more *Eligible* articles account for a third of the new MEDLINE-indexed evidence. This is the first zone of a three zone Bradford distribution, and in this case includes 241 articles, 34.9% of the total. The next rough third encompasses 51 journals each with between 3 and 9 articles. This second zone accounts for 226 articles, 32.7% of the total. The final third encompasses 183 journals each with 1 or 2 articles. This final zone has 223 articles, 32.2% of the total.

The distribution, shown in Figure 39, is convex through all three zones. The classic distribution describing a field, at least part of zone one (shown at the bottom of the graph) rises above the diagonal, and this represents the core or nucleus. The middle zone would be linear, while the third zone or part of it, (here at the top of the graph) droop below the line plotted on the linear part of the semi log distribution.[290] In this distribution, we see a large core rising above the diagonal. All of the first zone is in the core, and even part of the second zone, when a diagonal is plotted along the most linear part of the semi-log curve (Figure 39). The third zone rises above the diagonal, due to the large number of journals contributing the same number of articles. All journals in the third zone contributed either 1 or 2 pieces of new evidence.

Taking the nucleus as those journals in zone 1 and the part of zone 2 rising above the diagonal, the first 13 data points are included. These represent 13 journals all contributing 9 or more pieces of new evidence. These journals, along with the number of *Eligible* studies from them, are shown in Table 71.

Table 71. Journals Publishing the Most Eligible New Evidence

Zone	Title	AIM*	Impact Factor	Articles	Cumulative % of New Evidence
1	JAMA	Yes	23.175	38	5.5
	New England Journal of Medicine	Yes	51.296	37	10.9
	Circulation	Yes	10.940	34	15.8
	Lancet	Yes	25.800	33	20.6
	American Heart Journal	Yes	3.514	23	23.9
	American Journal of Cardiology	Yes	3.015	20	26.8
	Journal of the American College of Cardiology		9.701	14	28.8
	American Journal of Geriatric Psychiatry		2.894	12	30.6
	Alimentary Pharmacology & Therapeutics		3.287	10	32.0
	Diabetes Care		7.912	10	33.5
	European Heart Journal		7.286	10	34.9
2	American Journal of Gastroenterology		5.608	9	36.2
	International Journal of Geriatric Psychiatry		2.197	9	37.5

*AIM is Abridged Index Medicus

Bradford's explanation of the concentration in zone 1 and scattering in zone 2 related to the organization of the literature – the core are journals devoted to a subject.[291] In this case, we probably have many topics, defined by population and intervention, which we have classified into 12 medical specialties (see Methods 3.35), several of which are reflected by specialty journal in Zone 1. i.e. cardiology and psychiatry. As well as subject concentration, we are seeing a concentration of certain types of evidence, randomized controlled trials in important topics, concentrating in a few journals. The correspondence between Abridged Index Medicus membership and Bradford zones is quite good, 54% of titles in zone 1 are Abridged Index Medicus journals, as are 35% of zone 2 and 11% of zone 3.

Of the 245 journals represented, 204 had journal impact factors in the 2007 Journal Citation Reports.[241] Journal impact factors ranged from 51.296 for the New England Journal of Medicine to 0.329 for the Saudi Medical Journal, with a median of 2.864 (first and third quartiles; 1.769, 4.644). Looking at the journal impact factors across the zones (Table 72 and Figure 39) it is apparent that the higher impact journals concentrate in zones 1 and 2, but that lower impact journals are well represented in all zones. Journal Citation Reports does not assign an impact factor to all journals. Most journals without impact factors were found in the third zone, contributing either one or two pieces of new evidence. Fifty-five studies (8% of new evidence) came from journals with no impact factor.

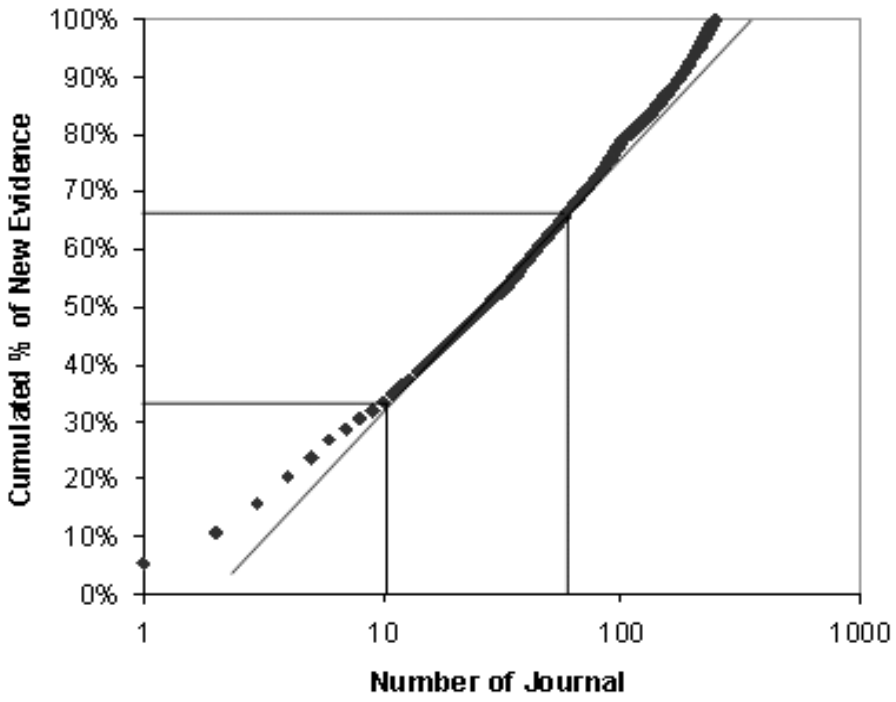


Figure 39. Journal Contribution of New Evidence

Drop lines indicate the limits of the zones. The diagonal represents the segment with linear growth on a semi-log scale.

Table 72. Journal Impact Factors of New Evidence by Zone

	Median	1st Quartile	3rd Quartile	N of Journals without Impact Factors
Zone 1	7.912	3.401	17.058	0
Zone 2	4.021	2.529	5.981	4
Zone 3	2.523	1.487	3.801	37
Overall	2.864	1.769	4.664	41

In summary, the journals publishing new evidence for this cohort followed a Bradford distribution, with a few journals contributing a disproportionate amount of evidence and a long tail of journals contributing one or two new studies. Journals included in Abridged Index Medicus and journals with high journal impact factors were strong contributors, although evidence came from journals not in Abridged Index Medicus and with no journal impact factor. It would be interesting to examine the scattering of evidence from systematic reviews such as those undertaken for management and policy making and described by Greenhalgh as being based on "complex and heterogeneous evidence"[292] rather than studies of intervention effectiveness examined here.

6.31 DOES OLD EVIDENCE PERSIST?

Eighteen systematic reviews,[177,177,274-276,278,279,279,293-304] all of which were Cochrane reviews, had explicit updates and these were examined to determine the persistence of evidence (see Methods 4.8.2). Six of the reviews (33.3%) had references to included studies that were present in the original but not in the update.[275,296,297,299-301] Such references ranged from one to eight in number and made up between 4.5% and 87.5% of the total trials in the original review. Overall, in this subset of 18 updated reviews, 25 of the 329 (7.6%) references in the original reviews were excluded from the updates.

Looking at the distribution of these "Original Only" references across the quintiles, the proportion of the updated systematic reviews in the age quintile where the

update excluded references that were in the original review increased as the age of the oldest trials increased (Figure 40).

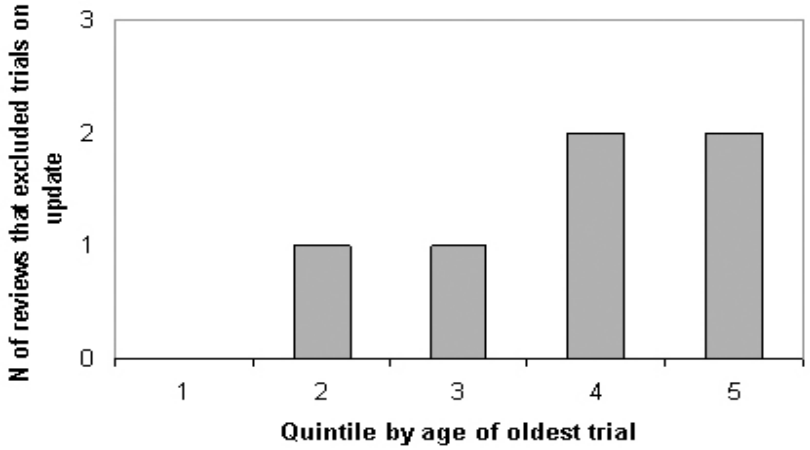


Figure 40. Updated Systematic Reviews that, on Update, Excluded References to Trials Originally Included – Distribution by Age of Oldest Evidence

6.31.1 Characteristics of References to Studies Excluded on Update

The excluded references were neither the oldest nor newest of the studies included in the original. They tended to have somewhat smaller N than studies retained in update. One exception was a single study excluded from CohortID28[275] – the excluded study was the second newest and the largest in the trial. It was excluded from the update as it was not placebo controlled. Overall, the most notable characteristic of these excluded trials is that only 11 of 25 (44%) of them were indexed in MEDLINE.

6.31.2 Reasons for Exclusion

All exclusions were acknowledged and explained by the investigators, with the exception of one systematic review that excluded a single study on update.[296] This accounting occurred either in the "Reasons for Exclusions" list or the "What's New" section of these Cochrane reports.

The most common reason for exclusion was that the original study had been reported as an abstract, but was replaced by a report in a full publication – that is, only the report was excluded, not the study (n=4 reports representing three studies). Cochrane reviews use study-based reporting, where references to studies are grouped by study, rather than being presented in a standard reference list in order of appearance in the text or by date. An asterisk is used to designate the major publication of the study. Thus, these excluded references could still have been reported although the "major" report would presumably become the full publication. An additional two reports of a pilot study, one of which was an abstract, the other a journal publication, were replaced by another publication released the following year by the same author. It is not clear if this was the same or a different study. Treating these six cases as partial results being subsequently published more fully, the mean lag in publication between versions was 2.3 years. The median difference was 2.5 years with a minimum of one and maximum of four years. Scherer *et al.*[247] show that approximately 45% of reports of randomized controlled trials initially published as abstracts are published as peer-reviewed journal articles within this time frame, with just over 60% ever being fully published (Figure 2). Positive results favour publication.

Four (16%) studies in two systematic reviews were excluded from the updates for design reasons. One was not placebo controlled and three were open label. One (4%) study was excluded based on intervention. It used a higher dose of the intervention than was eventually licensed. Six studies (24%) were excluded based on population. These came from one systematic review of TENS in the treatment of chronic low back pain.[301] The definition of "chronic" had been changed from pain persisting eight or more weeks to pain persisting 12 or more weeks. These six studies became ineligible as they included a mix of patients with chronic pain and some with duration of symptoms less than 12 weeks. Only one trial from the original review met the revised criteria and was retained. One additional trial was added in the update. The update found the intervention less

effective than had the original, and one commentator (in the feedback section of the review) argued that the redefinition and exclusion biased the review against the treatment.

In one case, the systematic review was split into two separate reviews on update.[305] The topic was pressure sores, and studies of treatment and prevention were treated separately. The eight studies not appearing in the update (which was the prevention study) appeared to deal with treatment. As of The Cochrane Library Issue 4, 2007, the treatment review had not been published so inclusion of those eight reports in that systematic review could not be verified.

In summary, few papers included in original reviews are excluded from explicit updates. Such exclusions occur for various reasons including replacement by a more complete report of the study and changes in the eligibility criteria for the systematic review. Study-based reporting of references is helpful, and it seems preferable to retain references to abstracts even when a more complete version of the study has available at the time of update, unless those preliminary results are discordant with the final results. Similarly, a section detailing changes in inclusion criteria for the review is helpful when such changes occur between versions of reviews. This group of Cochrane reports demonstrates that such exclusions can be accounted for. Full accounting is helpful to the reader who can assess the influence of these changes on the results of the review. Although reviews with older evidence were more likely to have reports included in the original review but excluded on update, this does not seem informative.

6.32 Is Maturity of the Literature a Predictor of Survival?

The final avenue of exploration is the age of the literature reviewed, using the oldest included study to represent the maturity of the evidence base (see Methods 4.9). The cohort of 100 reviews considered in the survival analysis were divided into five groups (quintiles) based on the age of the oldest included study in the original review (Table 73).

Table 73. Distribution of Systematic Reviews by Age of First Evidence

		Maximum Age of First Evidence at Search		Reviews with Major or Potentially Invalidating New Evidence (n=59)	
Quintile		Years	N in Quintile	N	% of Quintile
Youngest	1	6	25	20	80.0
	2	10	17	13	76.5
	3	16.4	18	9	50.0
	4	22	23	10	43.5
Oldest	5	47	17	7	41.2

In keeping with the more limited time since the first trial, systematic reviews in the lower quintiles tended to have fewer included studies, and fewer included study participants – with the exception of the average included N in the systematic reviews of the fourth quintile (Table 74).

Table 74. Size of Systematic Reviews by Age of First Evidence

Quintile		Mean Number of Included Trials in Original Review	Mean Total N of All Trials Included in Original Review
Youngest	1	12.6	9247
	2	13.5	10068
	3	17.7	12212
	4	16.7	3780
Oldest	5	39.6	24757

One indicator of the stability of reviews relative to the age of the included evidence is the number of systematic reviews in each quartile with potentially invalidating or major new evidence. There are 100 reviews in the cohort, 58 of these had potentially invalidating or major new evidence that would have altered the clinical application of the evidence. The array of these reviews, by quintile, is shown in Figure 41 and Table 73.

The criteria of potentially invalidating new evidence could be met through one of three qualitative signal; opposing findings; substantial harm; or a superior new treatment.

Four qualitative signals for major changes in evidence were; important changes in effectiveness short of 'opposing findings'; clinically important expansion of treatment; clinically important caveat and finally; opposing findings from discordant meta-analysis or non-pivotal trial. These are fully defined in the on-line Appendix A of the AHRQ technical report.[16] Fifty-six of the 59 reviews with signals for updating had qualitative signals. The final three reviews[306-308] met statistical criteria only.

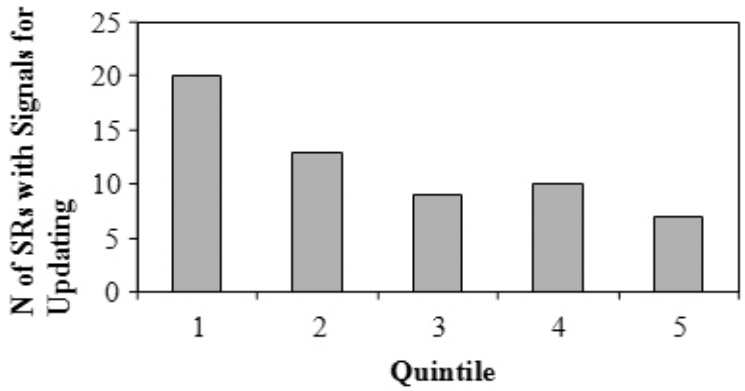


Figure 41. Distribution of Reviews with Signals for Updating by Quintile of Age of First Evidence

Only three of the qualitative criteria occurred with any frequency; potentially invalidating opposing findings (n=8, 14.3% of qualitative signals) major changes in effectiveness (n=27, 48.2% of qualitative signals) and major caveats (n=16, 28.6% of qualitative signals). The other four qualitative signals occurred only five times in this cohort, and accounted for only 8.9% of qualitative signals. They are not considered further. Distribution of the three common qualitative signals are shown in Figure 42.

Caveats and changes in effectiveness, both considered major not necessarily invalidating new evidence, occurred more frequently in reviews with a more recent evidence base. Opposing findings, considered by us as potentially invalidating evidence, occurred sporadically – that is, infrequently, and with no discernable pattern. This can be characterized as a trend in the evidence (Table 75).

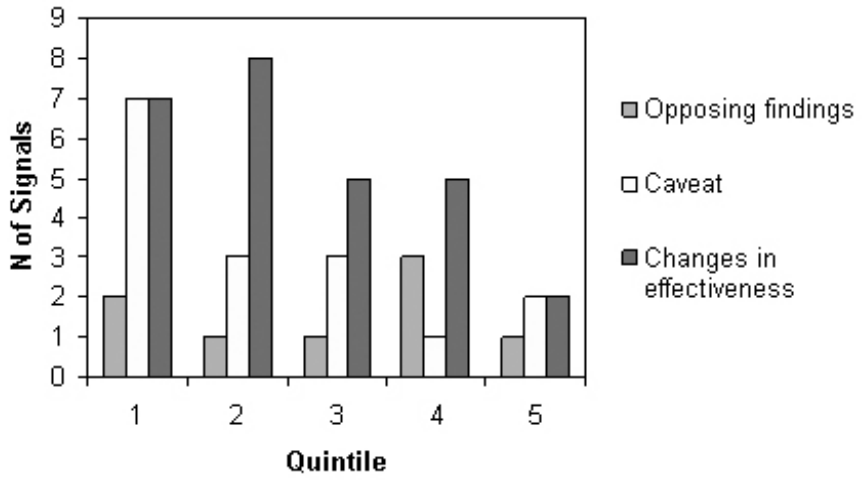


Figure 42. Distribution of Qualitative Signals by Quintile Based on Age of Evidence where the First Quintile Represents Reviews Including the Newer Trials

In 12 of the 54 (22.2%) systematic reviews with signals for updating, the signal involved harm. These signals were distributed evenly through the quintiles of age of oldest evidence (χ^2_8= 5.71, $P_{2\ sided}$ =.680).

Table 75. Quintile of Age of First Evidence by Specific Qualitative Criteria

		Specific Qualitative Criteria			
Quintile		Opposite Characterization of Effect	Important Caveat	Movement of One Point Toward More or Less Effective than Thought	Total
Youngest	1	2	7	7	16
	2	1	8	3	12
	3	1	5	3	9
	4	3	5	1	9
Oldest	5	1	2	2	5
	Total	8	27	16	51

Eta not significant (p with quintile as dependent measure = 0.180). When only caveats and small movements in effectiveness are examined, eta remains non-significant at 0.099

Heterogeneity, publication bias and ongoing trials refer to issues of the evidence arising in the original review. These may be influenced by the maturity of the oldest evidence.

Table 76. Attribute of the Evidence by Quintile of Age of First Evidence

		Proportion of Quintile Showing Attribute		
Quintile		Heterogeneity (Known or Suspected) in Original Review	Publication Bias (Known or Suspected) in Original Review	Ongoing Trials Identified by Original Reviewers
Youngest	1	0.76	0.24	0.28
	2	0.47	0.06	0.34
	3	0.39	0.17	0.22
	4	0.65	0.26	0.17
Oldest	5	0.71	0.06	0.29
	Overall	0.61	0.17	0.26

For these indicators, the total number of systematic reviews in the quintile forms the denominator. Observed publication bias is low and fluctuates. This may be more a factor of incomplete assessment for the presence of publication bias by the original reviews. It was assessed in only 40% of these systematic reviews.[17] The percent of systematic reviews reporting awareness of ongoing trials is likewise low, but quite stable, as would be expected under linear growth of trials. Still, if later trials involved longer-term follow-up of participants, one might expect to see an increase in this measure.

The presence or suspicion of heterogeneity was a significant predictor of survival in the multivariate analysis.[17] Heterogeneity was common in all quintiles, but less so in the second and third quintile. This could be a chance observation, but could also be based on changes in the application of the intervention in the context of trials – possibly early work lacks refinement in terms of selection of the population most apt to benefit from the intervention, or variation in dosing strategy until optimal doses are established. Later trials, in part as a function of their larger size, could explicitly study subgroups or particular populations as the application of the intervention becomes more refined. Such

trends should be reflected in two of the categories of qualitative evidence, caveats and expansions of treatment.

> Caveat (A4): Pivotal trial, new meta-analysis, "discordant" meta-analysis, trial indexed in ACP Journal Club, more recent practice guideline, or recent textbook does not contradict the previous review, but adds an important caveat about the patient populations who benefit.

Such caveats emerged for 27 of the 100 systematic reviews in the cohort. They were common in the reviews with the newest evidence and declined by cohort.

> Expansion of treatment (A5): Instead of a caveat, there has been expansion of the role of the treatment (e.g., the treatment has now been shown to be of benefit in primary prevention, not just secondary; or now shown to be of benefit in children or aged population etc).

In fact, the expansion of the role of treatment was observed only three times in the cohort of 100, once in each of the first, second and fifth quintile. Thus, these two signals potentially arising from heterogeneity in the evidence do not show the same bimodal pattern as the heterogeneity variable.

6.32.1 Survival by Age of Oldest Included Evidence

When the 100 systematic reviews included in the original survival analysis were divided into quintiles based on the age of the oldest included evidence, signals were more common in the systematic reviews with younger initial trials (χ^2_4= 9.51, $P_{2\ sided}$ =.050, Table 77).

Table 77. Quintile for Age of Oldest Evidence by Signal

Quintile for age of oldest evidence		Signal				
		No		Yes		Total
		N	%	N	%	
Youngest	1	8	32.0	17	68.0	25
	2	4	23.5	13	76.5	17
	3	10	55.6	8	44.4	18
	4	13	56.5	10	43.5	23
Oldest	5	11	64.7	6	35.3	17
Total		46	46.0	54	54.0	100

Figure 43 shows the survival curve of the complete sample. Figure 44 shows the survival curve for each quintile based on age of oldest included evidence. In these figures,

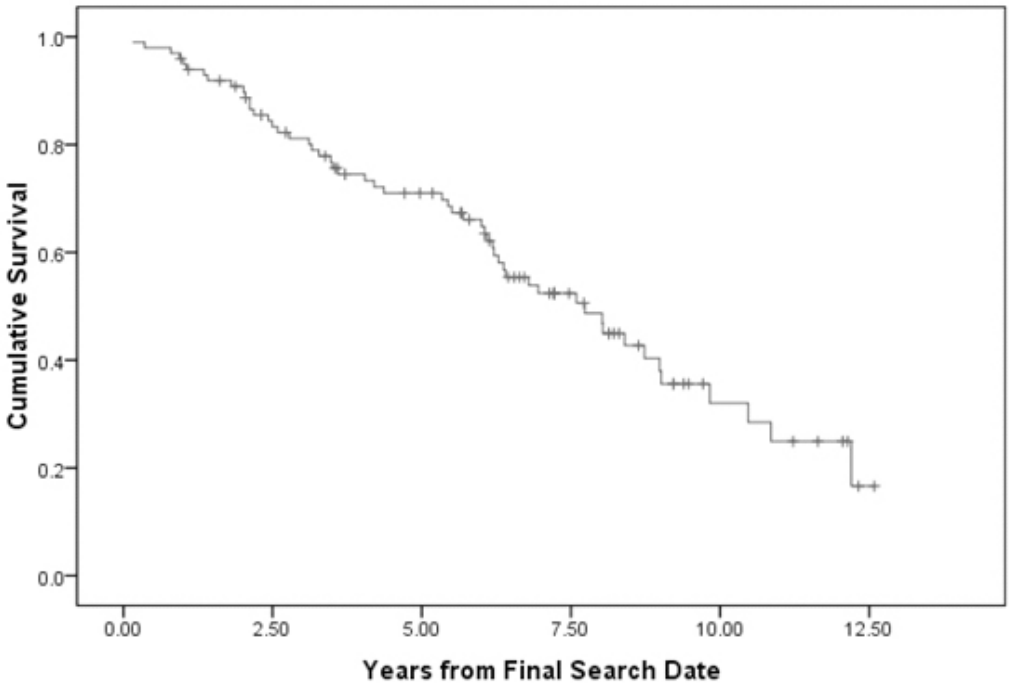


Figure 43. Survival of 100 Systematic Review

each step in the line indicated the point in time where one or more system reviews went out of date based on a signal from new evidence. Each tick on the horizontal line

represents the point where one systematic review was censored, that is, survived to the end of the follow period before going out of date.

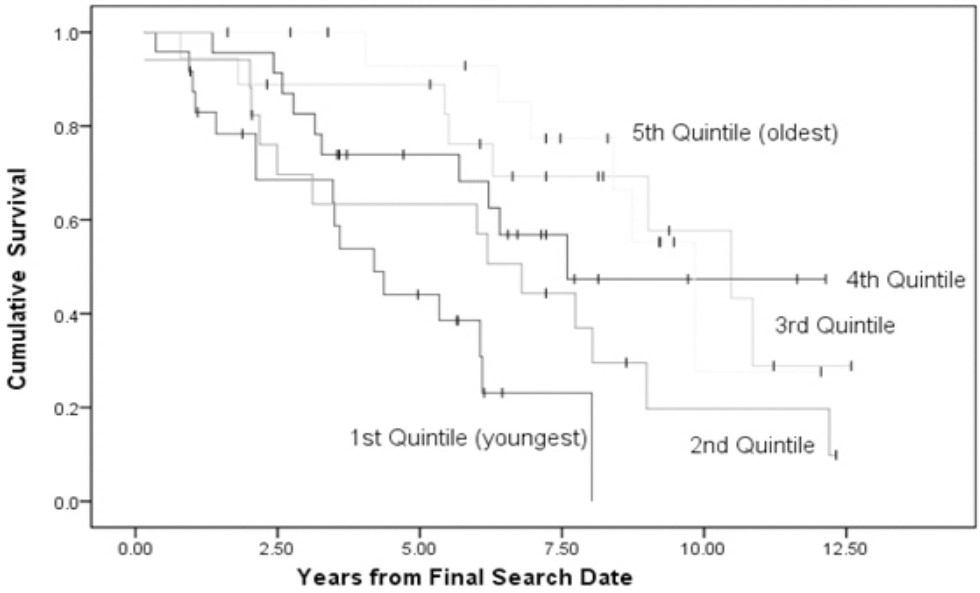


Figure 44. Survival of 100 Systematic Reviews by Age of Oldest Evidence

The first quintile, those systematic reviews where the age of the oldest included evidence was six years or less at the time of the search, showed the shortest survival. All were out of date or censored by eight years. While the second quintile (maximum age of included studies of ten years) generally showed the next shortest survival, some reviews survived until 12.5 years. The fourth quintile actually tends towards shorter survival - any systematic reviews in the fourth quintile that survived eight years survived until censored at the end of our follow-up period. The oldest evidence in the systematic reviews in the fourth quintile ranged between 16.5 and 22 years. The third and fifth quintiles both show a pattern with most reviews surviving to the five year mark and a few surviving for the duration of follow-up, to September 2006.

The median search date of all cohorts was 1998, except for the youngest cohort, with a median search date of 2000, so this cohort was at risk for slightly less time than the others. The protocol stipulated that no systematic review with a search date later than

2004 was eligible (see Section 4.1.2.1), and all quintiles but the oldest has one systematic review with a 2004 search date – the latest search date in the oldest quintile was 2003.

6.32.2 Nature of the original and subsequent findings: youngest and most mature cohorts

Titles and major conclusions of reviews signals for updating are presented in Table 79. This is followed by the classification of the qualitative signal, and the major supporting evidence for that classification.

Table 78. Reviews Including the Oldest Evidence with Major or Invalidating New Evidence

Review	Title
Antithrombotic Trialists' Collaboration, 2002[309]	Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients "In patients at high risk for occlusive vascular events because of preexisting disease, antiplatelet therapy reduces the risk for nonfatal myocardial infarction, nonfatal stroke, or death from vascular or unknown causes."[309] Movement of one point toward more or less effective than thought: Promising results from the original systematic review are now confirmed. For a select high risk population with previous ischaemic cereberovascular event, the combination regimen of aspirin plus dipyridamole over aspirin alone as antithrombotic therapy reduces incidence of vascular death, stroke, or myocardial infarction.[310]
Bucher, 1998[311]	Effect of HMGcoA reductase inhibitors on stroke: A meta-analysis of randomized, controlled trials "HMGcoA reductase inhibitors reduce risk for fatal and nonfatal stroke, coronary heart disease mortality, and all-cause mortality. Resins decrease risk for fatal coronary heart disease, but this effect has not been shown for fibrates or diet."[311] Expansion of the role of treatment: "In patients with recent stroke or TIA and without known coronary heart disease, 80 mg of atorvastatin per day reduced the overall incidence of strokes and of cardiovascular events, despite a small increase in the incidence of

	hemorrhagic stroke".[312] Thus a population that could not be examined in the original cohort is now examined in this major new trial and new information of treatment efficacy in this population is available.
Cullum, 2000[297]	Beds, mattresses and cushions for pressure sore prevention and treatment "In patients at high risk for pressure ulcers, foam-based, constant low-pressure (CLP) mattresses are better than standard hospital mattresses for reducing pressure ulcers. High-technology, CLP, alternating pressure, and other surfaces have varied success in reducing pressure ulcers."[297] Movement of one point toward more or less effective than thought: The alternating pressure devices appeared less effective than originally thought in the 2004 update to this review[313] and a subsequent large trial.[314]
Freemantle, 1999[315]	Beta Blockade after myocardial infarction: systematic review and meta regression analysis The original review concluded that beta blockers were underused following myocardial infarction, leading to avoidable mortality and morbidity.[315] Important Caveat: The new evidence, from a large trial published in the Lancet, found that while beta-blockers increase the risk of cardiogenic shock, at least initially. "Consequently, it might generally be prudent to consider starting beta-blocker therapy in hospital only when the haemodynamic condition after MI has stabilised."[316]
Jefferson, 2001[302]	Amantadine and rimantadine for preventing and treating influenza A in adults "Amantadine and rimantadine are similarly effective in the prevention and treatment of influenza in healthy adults, but rimantadine is associated with fewer adverse effects."[302] Important Caveat: The author of the original review subsequently recommended that use of these drugs for prophylaxis and treatment should be limited to pandemics due to marked increase in drug resistance.[317]

Lefering, 1995[303]	Steroid controversy in sepsis and septic shock: a meta-analysis "Corticosteroids do not reduce mortality in patients with sepsis, septic shock, or severe infections."[303] Opposite characterization of effect: a subsequent definitive trial demonstrated important gains in patients with inadequate adrenal reserves.[318]

These systematic reviews involve serious conditions, high risk populations, and often mortality as an outcome. While several of the changes involve effectiveness, the signals also involve harms or caveats regarding the application of the intervention.

Two of the six had potentially invalidating new evidence. In one,[315] the emergence of a harm added an important caveat to the original treatment recommendations. The original review concluded that beta blockers were underused following myocardial infarction, leading to avoidable mortality and morbidity. The new evidence, concluded "it might generally be prudent to consider starting beta-blocker therapy in hospital only when the haemodynamic condition after MI has stabilised."[316] In the other review with new evidence potentially invalidating the original findings,[303] the new evidence resulted in an opposite characterization of the effect. Rather than being ineffective in treating septic shock, a subsequent definitive trial of corticosteroids demonstrated important gains in some patients.[318]

The other four had major new evidence, although new findings did not appear to invalidate the original review. In two cases, antiplatelet therapy in seriously ill patients,[309] and beds, mattresses and cushions for the prevention at treatment of bed sores[297] the intervention appeared less effective than originally thought. In one, HMGcoA reductase inhibitors on stroke,[311] the new evidence supported an expansion of treatment, to include those with prior stroke - a group excluded from the original review due to the amount of heterogeneity in the evidence. In the final case,[302], the author of the original review subsequently recommended that use of these drugs for prophylaxis and treatment should

be limited to pandemics due to marked increase in drug resistance.[317] This recommendation is not without controversy.[319-321]

Four of the systematic reviews demonstrated a change in effect size with the addition of new evidence. One showed a change from non-significant to significant (but it became clear the intervention was less effective than originally thought), while another showed a gain with a decrease in the confidence interval of at least 50%. Four met one or more criteria involving the amount of new evidence. The four all had an increase in the number of patients at least 50% while one has a new study emerge that was at least three times larger than the previous largest trial. Three also had an increase in the number of trials by at least 50%.

Titles and major conclusions of reviews without signals for updating are presented in Table 79. Where there was new evidence not fully consistent with the findings of the original review it is noted in the table.

Table 79. Titles and Conclusions of Reviews Including the Oldest Evidence but Without Major or Invalidating New Evidence

Review	Title
Adams, 2001[293]	Inhaled beclomethasone versus placebo for chronic asthma "In adults and children with chronic asthma, inhaled beclomethasone diproprionate reduces airflow limitation, need for rescue bronchodilators, and symptoms. In patients dependent on oral corticosteroids, beclomethasone reduces use of oral prednisolone."[293]
Anand, 1999[306]	Oral anticoagulant therapy in patients with coronary artery disease: a meta-analysis "In the presence of aspirin, low intensity OA does not appear to be superior to aspirin alone, while moderate to high intensity OA and aspirin vs. aspirin alone appears promising and the bleeding risk is modest, but this requires confirmation from ongoing trials".[306] This systematic review is guarded about the benefit of this combination of intervention as it acknowledges that the analysis is based on smaller number of patients and the confidence interval is wide. Our review team concluded that, after adding newer eligible trials, the estimate of benefit on composite outcome is now more precise, albeit a bit less strong, and the risk of major

	bleed is confirmed as significant. This was not considered a major change as it confirmed the conclusions of the original review.
Arroll, 1999[275]	Antibiotics versus placebo in the common cold "In patients with acute upper respiratory tract infection, antibiotics are no more beneficial in terms of general improvement than placebo and are associated with a nonsignificant increase in adverse effects."[275] With the addition of new evidence, a previously non-significant harm – gastrointestinal side effects, became significant, but this was not considered to have major implications as the original reviews had not recommended the treatment.
Chronicle,2004[322]	Anticonvulsant drugs for migraine prophylaxis "In patients with migraine, anticonvulsants as a class are more effective than placebo for reducing the frequency of migraine attacks. The most common adverse events were nausea, asthenia or fatigue, tremor, weight gain, and dizziness or vertigo."[322]
Colman, 2004[323]	Parenteral metoclopramide for acute migraine: meta-analysis of randomised controlled trials "In patients with acute migraine, metoclopramide reduces headache pain more than placebo. Compared with other single agents, metoclopramide shows variable effectiveness for other migraine symptoms."[323]
de Feranti, 1998[324]	Are amoxycillin and folate inhibitors as effective as other antibiotics for acute sinusitis? A meta-analysis "In patients with acute sinusitis, any antibiotic reduces clinical failure. Newer, more expensive antibiotics are not superior to amoxycillin or folate inhibitors."[324]
del Mar, 1997a[177]	Antibiotics for the symptoms and complications of sore throat "Antibiotics (oral or intramuscular) modestly reduce the complications and duration of sore throat, but the absolute risk reduction is small."[177]
Del Mar, 1997b[325]	Are antibiotics indicated as initial treatment for children with acute otitis media? A meta-analysis "In children with acute otitis media, the use of antibiotics decreases pain at 2 to 7 days after presentation and reduces contralateral acute otitis media and deafness at 3 months. Antibiotics do not reduce pain within 24 hours or prevent recurrent otitis media."[325]

Demicheli, 2001[278]	Vaccines for preventing influenza in healthy adults "In healthy persons 14 to 60 years of age, vaccines reduce the incidence of serologically confirmed influenza."[278]
Furkawa, 2002[326]	Meta-analysis of effects and side effects of low dosage tricyclic antidepressants in depression: systematic review "In adults, low-dose (75 to 100 mg/d) tricyclic antidepressants are more effective than placebo and as effective as standard-dose (>= 100 mg/d) tricyclic antidepressants and are associated with fewer dropouts from side effects than standard-dose regimens."[326]
Heidenreich, 1999[327]	Meta-analysis of trials comparing β-blockers, calcium antagonists, and nitrates for stable angina "In patients with stable angina, β-blockers and calcium antagonists have similar clinical outcomes. β-blockers result in fewer withdrawals because of adverse events."[327]

Conditions being studied here are less serious, the interventions are mostly well established and in some cases the review is considering whether a new drug in a class is more effective than existing therapies. Although quite a bit of new evidence emerged, it was in accordance with the finding of the original review and mainly functioned to narrow the confidence interval around the point estimate.

In eight of 11 (73%) the number of new trials increased by at least 50% and in seven of 11 systematic reviews (64%), the number of new patients increased by 50% or more, but in only one case was there a new trial at least three times larger than the largest trial in the original review. The accumulated new evidence reduced the width of 95% confidence interval by at least 50% in six cases (55%). In all but one of these six cases, there had been significant heterogeneity suspected or identified by the original reviews, while heterogeneity was a factor in one other review that did not meet any size criteria.

Looking at reviews where the age of the oldest included trials were in the newest quintile, we see that more had signals for updating due to major or potentially invalidating new evidence. Of the 25 reviews in the quintile, 20 (80%) had major or

potentially invalidating new evidence signaling the need for updating. That percent declines with each quintile, falling to 41.2% in the quintile with the oldest included evidence (Table 73). Survival was also shortest in this quintile.

Table 80. Reviews Including the Youngest Evidence with Major or Invalidating New Evidence

Review	Title
Birck, 2003[294]	Acetylcysteine for prevention of contrast nephropathy: meta-analysis "In patients with chronic renal insufficiency, adding prophylactic acetylcysteine to hydration reduces the incidence of contrast nephropathy more than hydration alone."[294] Movement of one point toward more or less effective than thought: Several subsequent meta-analyses found this effect non-significant after the addition of new trials.
Blood Pressure Lowering Treatment Trialists' Collaboration, 2000[328]	Effects of ACE inhibitors, calcium antagonists, and other blood-pressure-lowering drugs: results of prospectively designed overviews of randomised trials "In patients with hypertension, diabetes mellitus, coronary artery disease, or renal disease, angiotensin-converting enzyme inhibitors and calcium antagonists are more beneficial than placebo. Calcium antagonists lead to a lower risk for stroke but to a slightly higher risk for coronary artery disease than do diuretics or β-blockers."[328] Important Caveat: A subsequent meta-analysis overturned this finding, concluding that "the relative risk of stroke was 16% higher for BB than for other drugs ".[329]
Blumenauer, 2003[330]	Etanercept for the treatment of rheumatoid arthritis "Etanercept (25 mg subcutaneously twice weekly) reduces symptoms and disease activity in patients with rheumatoid arthritis."[330] Important Caveat: The intervention appears to be more effective if delivered as combination therapy.[331]
Boucher, 2002[332]	Efficacy of rosiglitazone and pioglitazone compared to other anti-diabetic agents: systematic review and budget impact analysis "In patients with type 2 diabetes, little evidence exists to support

	rosiglitazone or pioglitazone being more effective monotherapy than existing antidiabetic agents. When added to a nonthiazolidinedione agent, both drugs reduce glycosylated hemoglobin and fasting plasma glucose levels more than monotherapy with either agent."[332] Important Caveat: Significant harms were identified in a new trial.[333]
Brown, 2001[334]	Meta-analysis of effectiveness and safety of abciximab versus eptifibatide or tirofiban in percutaneous coronary intervention "The use of glycoprotein IIb/IIIa inhibitors in patients who are having percutaneous coronary interventions does not reduce mortality but does decrease the need for urgent revascularization. Abciximab, but not eptifibatide or tirofiban, is associated with a reduced rate of recurrent myocardial infarction."[334] Movement of one point toward more or less effective than thought: A subsequent trial provided enough new evidence to make a trend toward efficacy significant. [335]
Bucher, 2000[336]	Percutaneous transluminal coronary angioplasty versus medical treatment for non-acute coronary heart disease: meta-analysis of randomised controlled trials "Percutaneous coronary angioplasty was associated with a lower rate of angina and a higher rate of coronary artery bypass grafting in patients with nonacute coronary artery disease." [336] Movement of one point toward more or less effective than thought: Where the original review concluded that the intervention was probably not effective, additional trials led a subsequent meta-analysis to conclude "In patients with chronic stable CAD, in the absence of a recent myocardial infarction, PCI does not offer any benefit in terms of death, myocardial infarction, or the need for subsequent revascularization compared with conservative medical treatment." [337]
Crouse, 1997[338]	Reductase inhibitor monotherapy and stroke prevention "Lowering LDL cholesterol levels by using statin monotherapy reduces stroke in patients with coronary heart disease."[338] Movement of one point toward more or less effective than thought: Subsequent research extended the role of this intervention to primary prevention, even in patients without diagnosed coronary disease.[339,340]

Ducharme, 2001[299]	Anti-leukotriene agents compared to inhaled corticosteroids in the management of recurrent and/or chronic asthma in adults and children "In patients with chronic asthma, daily antileukotrienes are not as effective as inhaled corticosteroids and increase asthma exacerbations requiring systemic corticosteroids."[299] Opposite characterization of effect: The review team felt that the author of the original review had understated his findings to some degree, stating "In the end the review could not clearly rule in or out anti-leukotrienes as first line agents: "Reliable conclusions cannot yet be drawn regarding the efficacy of this treatment due to the paucity of trials published in full text". A subsequent update, with new evidence, is clearer about the superiority of steroids.[298]
Etminan, 2002[341]	Efficacy of angiotensin II receptor antagonists in preventing headache: a systematic overview and meta-analysis "Angiotensin II receptor antagonists prevent headache in patients with mild-to-moderate hypertension."[341] Important Caveat: Reservations expressed in the original review due to lack of a clear mechanism of action are set aside by subsequent reviewers, although they were working with essentially the same data. The change is that the later review found this effect occurred with other hypertensive drugs.[342]
Evans, 2004[343]	Tegaserod for the treatment of irritable bowel syndrome "Tegaserod is more effective than placebo for reducing some symptoms in women with the constipation-predominant irritable bowel syndrome."[343] Important Caveat: Cases of severe diarrhea associated with dehydration and syncope have been described in 0.04 percent of patients, prompting the FDA to issue an advisory in April 2004.
Ezekowitz, 2003[344]	Implantable cardioverter defibrillators in primary and secondary prevention: a systematic review of randomized, controlled trials "In patients at risk for sudden cardiac death, implantable cardioverter defibrillators reduce sudden cardiac death, all-cause mortality, and cardiac mortality."[344] Important Caveat: The ICD malfunction replacement rate is significantly

	higher than that for pacemakers, which must be considered when selecting between the new treatment (ICD) and the standard of care (pacemaker).[345]
Gotzsche, 1995[346]	Somatostatin v placebo in bleeding oesophageal varices: randomised trial and meta-analysis "Compared with placebo, somatostatin did not increase survival or decrease the number of blood transfusions in patients with bleeding esophageal varices."[346] Movement of one point toward more or less effective than thought: While the original review was unable to show benefit, a later systematic review found this to be a safe and effective adjuvant therapy.[347]
Jong, 2002[200]	Angiotensin receptor blockers in heart failure: meta-analysis of randomized controlled trials "In patients with heart failure, angiotensin-receptor blockers do not reduce mortality or hospitalization rates more than do angiotensin-converting enzyme inhibitors or placebo."[200,200] Movement of one point toward more or less effective than thought: The change did not involve the main conclusion, but rather the role of Angiotensin receptor blockers as monotherapy, which the original reviewers found probably beneficial. Additional evidence makes them clearly beneficial.[348]
Keenan, 1997[349]	Effect of noninvasive positive pressure ventilation on mortality in patients admitted with acute respiratory failure: a meta-analysis "Standard therapy plus NIPPV reduces mortality and the need for endotracheal intubation in adults in the intensive care unit with acute respiratory failure. The benefits are greater when only patients with COPD are analyzed; patients without COPD may need further study."[349] Expansion of the role of treatment: Additional evidence supported the expansion of treatment to patients with acute hypoxemic respiratory failure not due to cardiogenic pulmonary edema and chronic obstructive pulmonary disease.[350]
Kjaergard, 2001[351]	Interferon alfa with or without ribavirin for chronic hepatitis C: systematic review of randomised trials "In patients with chronic hepatitis C, interferon-α plus ribavirin is more

	effective than interferon-α alone for improving the hepatitis C virologic response, but not for reducing liver-related morbidity and mortality."[351] Another treatment is superior: A significantly more effective combination was subsequently discovered.[352]
Laine, 2001[353]	Therapy for Helicobacter pylori in patients with nonulcer dyspepsia: A meta-analysis of randomized, controlled trials "Helicobacter pylori therapy in patients with nonulcer dyspepsia and Helicobacter pylori infection does not reduce symptoms."[353] Movement of one point toward more or less effective than thought: A subsequent systematic review found a small, but statistically significant, effect.[354]
Lord, 2003[355]	Metformin in polycystic ovary syndrome: systematic review and meta-analysis "Metformin used alone or combined with clomifene is effective for improving ovulation rates in women with the polycystic ovary syndrome."[355] Opposite characterization of effect: A subsequent much larger trial overturned this finding.[356]

Many have conclusions that cast doubt on the efficacy of the treatments.[200,200,332,334,336,346,351,353]

Table 81. Titles and Conclusions of Reviews Including the Youngest Evidence but Without Major or Invalidating New Evidence

Review	Title
Barr, 2005[357]	Inhaled tiotropium for stable chronic obstructive pulmonary disease "In patients with stable chronic obstructive pulmonary disease, tiotropium reduces exacerbations and hospitalizations, and improves health-related quality of life."[357]
Dalby, 2003[307]	Transfer for primary angioplasty versus immediate thrombolysis in acute myocardial infarction: a meta-analysis "In patients with acute ST-segment elevation myocardial infarction, transfer to a percutaneous coronary intervention (PCI) center for primary PCI is more effective than immediate thrombolysis for reducing all-cause mortality, reinfarction, or stroke."[307]
Ducharme, 2003[298]	Addition of anti-leukotriene agents to inhaled corticosteroids for chronic asthma "In patients with chronic asthma who are symptomatic while receiving moderate-to-high doses of inhaled beclomethasone, the addition of 2 to 4 times the licensed dose of antileukotriene (AL) agents reduces the rate of exacerbations that require systemic corticosteroids. Insufficient evidence exists that AL confers benefit over doubling the dose of corticosteroids or that it has an inhaled corticosteroid-sparing effect."[298]
Edmonds, 2001[300]	Inhaled steroids in acute asthma following emergency department discharge "Inhaled corticosteroids alone appear to be as effective as oral corticosteroids after discharge from the emergency department in patients with mild asthma exacerbations. Evidence is insufficient on the benefit of addition of inhaled corticosteroids to oral corticosteroids in this setting."[300]
Eikelboom, 2001[358]	Extended-duration prophylaxis against venous thromboembolism after total hip or knee replacement: a meta-analysis of the randomised trials "In patients who have received total hip or knee replacement, extended-duration prophylaxis with heparin is more effective than placebo or no treatment for preventing deep venous thrombosis. The

	preventive effect is associated with increased minor bleeding."[358]
Fink, 2002[359]	Sildenafil for male erectile dysfunction: a systematic review and meta-analysis "Sildenafil improves erectile dysfunction and is well tolerated."[359]
Kong, 1998[308]	Clinical outcomes of therapeutic agents that block the platelet glycoprotein IIb/IIIa integrin in ischemic heart disease "Platelet glycoprotein IIb/IIIa-receptor antagonists reduce the combined end points of death or MI and death, MI, or revascularization in patients with ischemic heart disease."[308]
Lane, 1995[360]	Endoscopic ligation compared with sclerotherapy for treatment of esophageal variceal bleeding: A meta-analysis "Compared with sclerotherapy, endoscopic ligation has lower rates of rebleeding, mortality, and complications and requires fewer treatments for obliteration of esophageal varices."[360]

These all had clear statements of efficacy, although two [298,300] claim insufficient evidence to draw conclusions for some comparisons.

6.32.3 Summary

Systematic reviews done closer to the date of the first included trial tend to have fewer trials, fewer participants were enrolled in those trials, and major and potentially invalidating changes in evidence were seen more often than those reviews with a longer record of trials. Changes such as caveats involving certain study populations and shifts in strength of the evidence were seen in the same proportion as in the oldest cohort, as did signals involving harms.

Finding of reviews done closer to the date of the first included trial seemed more robust when they supported the use of the intervention under consideration than when findings were negative or inconclusive. Reviews that were overturned sometimes went from non-significant findings to significant findings, but rarely did the direction of effect

reverse. Thus new evidence could be largely overcoming the heterogeneity commonly found in these reviews and causing the meta-analysis to gain power through the addition of studies and participants. Still, one intervention was superseded by another more effective intervention, and another intervention thought to be effective was shown ineffective by a large trial.

Findings of reviews done in topics with a more established evidence base tended to be overturned when the treatment had been considered effective in the initial review. These reviews had much larger evidence bases than the first quintile reviews, both in terms of patients and studies. These results most commonly overturned because more was learned about the use of the intervention in complex conditions involving seriously ill patients.

Chapter 7: Discussion and Conclusion

7.0 INTRODUCTION

In the discussion, I will review the major findings of this research. First, I will review the performance of the test searches, and the patterns of the dependence between them. I will discuss two search methods that performed below expectations. Later I will discuss how the complementary nature of some of the search methods can be used in update searching to replace some of the more time consuming aspects of standard systematic review searches. Before that discussion, however, I will discuss the implications of the historical context in which systematic review searching developed, and the opportunities and challenges that history creates for update searching. Next, I will review the performance of the searches used in the original reviews, and in the Cochrane sample, their adaptation and performance in the updates.

From there, I argue that the performance of the test searches supports replacing some of the multi-database, multi-modal aspects of searching used (and apparently needed) in original reviews with a more streamlined approach in the update. Namely, I will argue that Boolean searches and similarity searches represent two independent approaches to the literature, and that they are sufficient to achieve optimal recall from a single database in most cases. By examining the bibliometric characteristics of the original evidence base, and leveraging the original evidence base through similarity search methods, we will often find that we can simplify the update search, reducing the number of databases and eliminating the systematic non-database portion of the search.

Finally, I will discuss barriers to implementation, including limitations of the generalizability of these results, and suggest areas for further research and development.

7.1 SUMMARY OF MAIN FINDINGS

The main experiment examined the performance of searches in three cohorts. Two Boolean searches, two similarity searches and one non-database search approach were tested. The two Boolean searches were based on a simple subject search that used a mean of 3.6 terms and 1.5 search features (Section 6.3.1). The subject search was paired with a filter selecting only RCTs from Abridged Index Medicus journals and with the balanced Clinical Query.

Clinical Query provided good recall but with large retrievals. Abridged Index Medicus RCT had smaller retrieval sizes and identified fewer new eligible studies, but did detect many large studies, so performed well when recall of new participants was considered. The two similarity searches were Support Vector Machine (SVM), where various cut points were tested in the Cochrane and AHRQ cohorts, and an algorithm based on related article searching in PubMed. The Related Article search showed the highest recall of new eligible studies overall. Recall with SVM was lower, and it trailed the Clinical Query. SVM's advantage was smaller retrievals. The non-database approach was citing reference searches, where the citing references were randomized controlled trials that cited the systematic review being updated. A high proportion of Citing RCTs were eligible for inclusion in updated reviews, but this method identified only a small proportion of all relevant new studies (Section 6.8.1).

Relative performance of the test searches was stable across the three cohorts, although there was a ceiling effect in the updated Cochrane reviews, where most searches performed strongly. Relative performance was also stable regardless of whether the intervention was a drug, device or procedure (Section 6.9), and whether the evidence appeared before or after the signal that an update was necessary (Section 6.8). All search approaches showed variable performance across clinical areas, but Related Articles RCT showed the most consistency (Section 6.8.1).

Precision, tested in the AHRQ cohort, ranged from 0.111 to 0.192 for the more productive test searches but reached 0.515 for the Citing RCT method. (Section 6.10)

Similarity methods showed higher precision than did the two Boolean approaches. Precision and recall showed a strong negative correlation whether recall of new studies or recall of new participants was considered (Section 6.11)

Relevance ranking was examined for the similarity searches. Placement of eligible studies near the top exceeded chance for both the SVM and Related Article RCT searches. (Section 6.12.1 and 6.12.2) When receiver operating characteristic curves were examined for these searches, SVM outperformed Related Article RCT in the Cochrane set, although the area under the curve was significantly greater than chance for both. Area under the curve was less impressive in the AHRQ Evidence Report searches, with only the Related Article RCT search improving over chance (Section 6.13).

The structural relationship between search methods was explored using capture-recapture and multidimensional scaling. The Boolean and similarity searches appeared to be independent, based on capture-recapture population estimates (Section 6.20.4) and appeared to operate on different dimensions (Section 6.21). The approaches complemented each other and the pairing of Related Article RCT and Clinical Query gave excellent recall of new relevant material (Section 6.24).

7.1.1 Performance of SVM

Support Vector Machine has an excellent record of accomplishment[174,361,362] and was expected to perform well in this task. It was surprising that the Related Article search outperformed it. It could be that the PubMed Related Articles feature is more effective than has been recognized simply because it has been overlooked for study. Still, Cohen's work with SVM, which appeared while this research was underway, found the MeSH terms figured prominently in successful ranking.[111] We limited emphasis on MeSH due to technical limitations with refreshing the database to ensure the MeSH indexing was completely up to date. It may be that SVM performance would have been improved had we given more weight to MeSH terms. On the other hand, because the role of MeSH was restricted, these results give some indication of the sort of performance that could be achieved by using Support Vector Machine with databases other than MEDLINE.

7.1.2 Performance of Citing Reference RCT

The Citing RCT approach was not useful in these tests. In part, this is due to the short update interval over which we were working. Pao and Lee noted that the most recent articles would not have had time to be cited.[162] I tested a very simple approach to using citing references. I only examined citations to the systematic review itself. Larson[363] suggests using network analysis of overlapping citations. Such a network could be constructed from the subsequent citations of all included studies. A more complex approach could broaden the scope of retrieval, and could possibly overcome the citation lag problem inherent in looking only at citations to the systematic review itself.

7.2 EVOLUTIONARY INFLUENCES

The original vision of The Cochrane Library was for it to be a library of living documents. These documents would be systematic reviews with meta-analysis, and as each relevant new study was added, the statistical computation and graphical presentation of the results would be updated in real time. The importance of regular updating was recognized, and the early Cochrane Collaboration Reviews were to be updated "at least annually". In that vision, new material would be found through centralized and ongoing efforts based on hand searching of journals and electronic searches. Those studies using a controlled clinical trial design would be added to the Cochrane CENTRAL register of controlled trials, and to systematic reviews for which they were relevant (see Section 2.2).

Two main influences prevented the realization of that vision. First, awareness of the complexity of managing bias in systematic reviews appears to have increased rapidly. It would soon have become clear that simply adding one or more new studies, updating the results in real time, and applying that new result to clinical practice was naïve. The original approach was likely based on the idea that setting the bar for eligible study designs at randomized controlled trials, as Archie Cochrane proposed, would provide adequate protection against methodological bias.[22] In fact, not all randomized controlled trials are created equal and quality assessment of included studies is now an important

part of conducting a systematic review. For example, The Cochrane Collaboration has introduced a "Risk of Bias" table to their reviews in which each included study is assessed on several dimensions that could potentially introduce bias.[364] This assessment must be completed for each existing Cochrane Collaboration Review when it is next updated.

Second, although CENTRAL is still considered the single best source of new studies,[41,79] there is no evidence that reviewers anywhere have come to consider it a sufficient source.[365] Thus, broader searches, involving multiple databases, are used. The large retrievals from more comprehensive updating searches paired with the continual advances in methodological and reporting complexity make updating a much more complex task than just adding several new studies. Large volumes of new material may be identified and require screening and assessment. If studies with stronger research designs, better diagnostic criteria, less subjective outcomes or other improvements appear, the inclusion criteria may need to be refined and all included studies may need to be re-assessed and some may be excluded (see Section 3.2.5). The report structures become more complex and detailed over time, increasing the complexity of updating as all evidence may be re-evaluated and additional data may need to be extracted from studies included in the original review.[96] These two influences, the greater attention to bias and the complex searches, make periodic updating more feasible than continual updating.

7.3 MULTIPLE DATABASE SEARCHING AS THE NORM

7.3.1 Actual *versus* Potential Contribution of Databases

In the absence of any other proven approach, current guidance from the Cochrane Updating Group has the reviewers repeat the search in each database used in the original review.[118] Any decision to search multiple databases has important implications for updating a systematic review, so alternatives should be considered.

The notion that only about half of the relevant material for reviews will be found from a search of MEDLINE stems from the article that introduced the original highly sensitive search strategy, the Randomized Controlled Trial publication type, and the Cochrane/National Library of Medicine re-tagging effort.[64] This article is still heavily cited today.[***] In a commentary on that paper, close to the time of its publication, Hersh pointed out that the introduction of the Randomized Controlled Trial tag "also invalidates the results of the studies reported in this paper, since the sensitivity of finding trials is no doubt improving with the new indexing tags."[66] Was this anticipated improvement in recall realized?

Let us consider the evidence base for the multiple database imperative. The contribution of databases depends first on the proportion of relevant studies indexed in the database, and second, on the proportion retrieved by the searches used. Some studies examining database contributions also consider whether excluding those studies found only in the additional sources would bias the estimate of intervention effectiveness. Examples of this approach include my own work in the contribution of Embase to meta-analyses of randomized controlled trials[366] and Egger's study of MEDLINE *versus* non-MEDLINE indexed trials in meta-analyses of randomized controlled trials.[75] Bias occurs if the studies found in the additional sources differ consistently from those found in the main source. If no bias is present, the loss of the additional studies will not change the estimate of how well the treatment works. However, adding more information that is consistent with previous information increases confidence in the finding. In meta-analysis, this is reflected in a narrower confidence interval.

MEDLINE indexing of eligible RCTs is 90% or greater in studies other than case reports of single systematic reviews.[203,367,368] Figures of 86% indexing of diagnostic studies[92] and 62% coverage of prevalence and incidence studies[289] have been reported.

[***] According to Web of Science, this paper had been cited 780 times of April 10, 2009, and 101 of these citations occurred in articles published in 2008 or 2009. Of course, the paper is multifaceted and not all of these references would have been to the retrieval rate from MEDLINE. However, by 2008 most systematic reviews citing the highly sensitive search strategy in support of their search methods should have been citing the revised HSSS that was published in 2006, not the 1994 study.

The discrepancy between potential and perceived database contributions indicates that research into how to optimize retrieval from the most complete sources is needed.

7.3.2 Precision Is Reduced In Multiple Database Searches

While recall may be improved by searching multiple databases, precision will be reduced. We have seen that the precision in systematic reviews is quite low – slightly below 3% of retrieved records are eligible (see Section 6.25). Material identified from multiple sources is more likely to be relevant than material found from only one source (see Sections 5.1.5 and 6.19). Therefore, we would expect that the incremental yield from each additional source would come at the expense of precision, as few of the uniquely identified articles would be expected to be eligible.

The most precise searches will most likely be those done in CENTRAL, as its included studies have been pre-selected by study design. However, as it is a secondary database, indexing lag is inevitable and limits recall in the short run. Further, its lack of a searchable entry date field limits precision for updating searches at present. The next most precise searches will be those that come from MEDLINE. MEDLINE has numerous features that enhance precision, including the highly-developed MeSH thesaurus, the validated Clinical Queries, the Randomized Controlled Trial publication type, the Core Clinical Journals subset, accurate indexing of age and gender, and for the case of updating, entry dates indicating when a record was added to MEDLINE.

Given the strengths and precision enhancing tools available in MEDLINE, particularly for identifying controlled clinical trials, the low precision of RCT-base systematic reviews is most easily explained by hypothesizing that these reviews use searches that are more exhaustive than the searches used for other designs. Precision is undermined by backfilling from other databases that cannot be searched with as much precision as MEDLINE. As there is no evidence that I am aware of that indicates that these non-RCT reviews are incomplete, despite their higher precision, it may be that RCT-focused systematic reviews are searching more extensively than necessary, at the expense of precision. Still, the retrieval sizes for the cohort described in Section 6.25 are

larger for non-RCT systematic reviews than for systematic reviews of RCTs (data not shown), so screening burden is higher although precision is similar.

The searches tested in the updates of the cohort of AHRQ reviews have much higher precision than was seen in full systematic reviews (see Section 6.10). While whole-review precision was below 0.030 for most types of reviews, the test searches showed precision above 0.100. Some of this increase is because the test searches were designed to be of higher precision (and this would account for variability between the test searches, which ranged from 0.111 to 0.192 for the approaches involving MEDLINE). However, much of the increase in precision is achieved by restricting the search to MEDLINE.

7.4 SUPPLEMENTAL SEARCHES AS THE NORM

Another issue to be considered is the role of non-database searches in both the original review and in the update. Specifically, do these efforts need to be repeated in the search for the update? These non-database efforts include contacting investigators to identify unpublished studies or published studies not otherwise found, checking reference lists to find additional relevant studies that may have been cited, and hand searching journals and conference abstracts. I refer to searches using these techniques to supplement database searches as multi-modal searches. Most published guidance on searches for systematic reviews advocates such measures.[40,242,369,370] However, a systematic review of evaluations of the contribution of checking reference lists identified that the efficacy of such measures has not been either proved or disproved. In fact, the issue has not been studied systematically using prospective research designs. Only case reports, cross sectional and retrospective observational studies were identified.[371] Case reports are susceptible to publication bias, since a paper about how much time was spent in unproductive efforts to find relevant articles is not likely to hold much appeal for journal editors. None of the studies identified for that review controlled for the quality of the

electronic searches. As recall of the electronic searches improves, the apparent contribution of alternate approaches will appear to decline.

Unpublished trials are considered grey literature.[372] Van Driel *et al.* examined the contribution of unpublished trials to Cochrane Collaboration Reviews. They called into question the role of unpublished trials in original reviews, which they found to be limited at best and possibly introduced bias.[373] Unpublished trials are typically found through methods other than searching bibliographic databases. Van Driel *et al.* found that few Cochrane Collaboration Reviews published since 2000 included any unpublished trials, and when they were included, they were few in number. Those trials also tended to included fewer participants and be of lower methodological quality than published trials. Their analysis indicated that trials that were eventually published were of better methodological quality, and therefore less susceptible to bias, than those trials that were truly unpublished. Further, other researchers have found discrepancies between preliminary results, such as conference abstracts, and results presented in full publication.[374] Van Driel *et al.* suggest that a more efficient and scientifically sound approach is to forego extensive searching for unpublished studies, and instead include those studies in updates after they have been vetted by the publication process.[373] In my examination of studies included in original reviews but excluded from the update, the most common reason for exclusion was that the original study had been reported in an abstract, which was replaced by a full publication in the update (Section 6.31.2).

My results from testing the performance of authors' searches to find MEDLINE-indexed studies included in the original review may provide the best evidence for the role of non-database supplemental search methods in the original reviews, although it is indirect evidence. In the six updated Cochrane Collaboration Reviews, recall of MEDLINE-indexed included studies was only 0.50 and in the ten AHRQ Evidence Reports it was only 0.64 (Section 6.27). This begs the question of how the remaining 36-50% of studies was found. There are several possibilities. These studies may have already been known to the investigators or their advisors, they may have been found through

checking reference lists, hand searching or contacting experts, or they may have been found through the other databases searched. It is apparent that these MEDLINE misses were found somehow, but we do not know how.

Importantly, these systematic reviews were selected, in part, because they used comprehensive searches, including supplemental search methods. This improved the chances that a high proportion of all relevant studies would be found and included, even if the MEDLINE searches were not particularly effective (Section 4.1.2.2). We do not know which aspects of the multi-database, multi-modal search was productive, and each review may have had a different combination of productive approaches. What seems clear is that when searches are done using methods that depend, in part, on operator skill, it is necessary to build in multiple opportunities to find relevant studies. Non-database methods, such as checking reference lists or contacting authors, provide approaches to the literature that do not require skill in Boolean searching.

7.5 ROLE OF REDUNDANCY

In summary, it would seem that in the original reviews, multi-database, multi-modal searches serve two functions. First, they provide opportunities to identify material uniquely available through only one of the sources. In addition, they provide second, third and fourth chances to retrieve material available through several sources. We have seen that the number of sources identifying a study was correlated with its probability of relevance (Section 5.1.5 and 6.19). Therefore, it is important to find these studies. Lemeshow *et al.* give an account of multiple search modes as second chances in their analysis of searches for observational studies; "The publications found during the cross-check of the reference lists of the reviews and meta-analyses were not accessible solely through this method of searching. The titles either originally appeared in one or more of the databases but were prematurely discarded because of seemingly irrelevant titles, or the databases had access to the publications but did not yield them because of our choice of search terms or the miscoding of keywords."[375] Savoie also acknowledged that many studies identified through supplemental methods were indexed in major databases, but

were missed by the searches.[376] That is, redundancy appears necessary in the searches for the original reviews in order to identify as many relevant studies as possible.

7.6 INDEPENDENT APPROACHES TO THE LITERATURE ARE NECESSARY

I argue that the overriding conclusion that can be drawn from the research evidence and actual practice is that multiple independent approaches to the literature are needed. By searching different databases, for instance, MEDLINE and Embase, a reviewer may be able to identify a higher proportion of relevant trials than by searching MEDLINE alone. This need not be because there were additional trials indexed in Embase but not in MEDLINE – my previous work has demonstrated that the majority of systematic reviews that searched both did not include any Embase-unique trials.[366] Instead, it seems likely that the real contribution was that trials were found through Embase that were missed from MEDLINE. The gain would come because those trials were indexed differently (MEDLINE and Embase use different subject headings), and the search may have used a subject heading other than that assigned by the MEDLINE indexers but may have achieved a match in Embase. Title, abstract, author and journal information would typically be the same between databases – the gain would due to differences in indexing. In this scenario, it is the independent approach to the literature rather than the additional coverage of the second database that results in more complete recall.

On the other hand, searches of MEDLINE and CENTRAL cannot really be considered independent approaches to the literature. Most CENTRAL records come from either MEDLINE or Embase. If searchers use their MEDLINE subject search in CENTRAL, they are unlikely to identify additional records; in particular, they will not have the opportunity to identify overlapping Embase records. The MEDLINE and CENTRAL searches would be independent is a) CENTRAL retained overlapping records from Embase and MEDLINE, and b) the searcher consulted the Embase thesaurus and added EMTREE terms to the search. In fact, Embase records that represent articles

already identified from MEDLINE are not added to CENTRAL.[79†††] My impression, after reviewing many CENTRAL searches for research purposes, is that it is not a common practice for EMTREE terms to be included in the CENTRAL search.

The Cochrane Reviewer Handbook states that CENTRAL includes "310,000 trial reports are from MEDLINE, 50,000 additional trial reports are from EMBASE and the remaining 170,000 are from other sources such as other databases and handsearching."[79] Thus, CENTRAL will be useful for identifying the material unique to sources not searched directly, but will not help reviewers compensate for a sub-optimal selection of terms in MEDLINE. In fact, good retrieval from these additional sources will depend on a well-developed free text component to the search.

Related Article RCT and Clinical Queries do constitute two independent approaches to the literature. The capture-recapture results (see Section 6.20.4), and multidimensional scaling results (Section 6.21) both support this. In addition, the two searches together resulted in nearly complete retrieval of relevant new studies (Section 6.8.1). Taken together, these findings provide strong evidence that these two searches represent independent approaches to the literature. These two searches of MEDLINE represent the most practical and efficient approach to update searches.

7.7 STREAMLINING SOURCE SELECTION AND SEARCH LIMITS FOR THE UPDATE

When a new review is started, the bibliometric characteristics of the relevant literature will not be known. This includes the language distribution, where or how the published literature will be indexed and the contribution grey literature will make to the evidence base. Since one of the functions of a comprehensive search is to guard against potential bias, it may be inevitable that reviewers will wish to make the initial search as inclusive as possible, regardless of the evidence for or against the global prospect of bias. A recent systematic review of the impact of language restrictions stated, "We could not find evidence of a systematic bias from the use of language restrictions in systematic

††† Handbook section 6.3.2.2

reviews/meta-analyses of conventional medicine." Yet they went on to conclude, "it seems that systematic reviewers of conventional medicine who hope to minimize the risk of producing a biased summary effect estimate should search for foreign language studies when resources and time are available."[377] By the time a systematic review with comprehensive search methods is updated, quite a bit will be known about where and how the evidence can be found. Rather, quite a bit could be known, if the productivity of sources and the performance of the MEDLINE search for the original review were evaluated *post hoc*.

These data provide numerous examples where multi-database searching could be replaced by a search of MEDLINE, if recall from MEDLINE could be maximized. The original review provides a real world test of the performance of the subject search and the potential contribution of MEDLINE. If retrieval from MEDLINE was sub-optimal, the search can be revised for the update. The related article protocol complements the subject search - it is independent from the subject search, and does not rely on operator skill. The similarity search methods, such as Support Vector Machine or Related Article RCT, have a clear role in the update, but they may not be so practical in the original review. This is because these similarity searches are informed by the true positive examples from the original review in the case of Related Article RCT, and in the case of Support Vector Machine, by the true positives and true negatives from the original review. These complementary approaches (Boolean subject search plus similarity search) provided quite complete identification of evidence for the update.

There are numerous examples of methodological research into search factors or bias factors in which the authors note that the results may not hold for some specialties. Egger, who studied several aspects of comprehensive literature searches and trial quality concluded "our results indicate that unpublished trials are also important in oncology, whereas non-English language trials are particularly prevalent in psychiatry, rheumatology and orthopaedics."[75] One very important result of these findings is the relative stability of recall of the Related Article RCT search across medical specialties –

greater stability than was seen with the Clinical Query, or Abridged Index Medicus RCT search, both of which more closely resemble standard search methods (see Section 6.8.1). Thus, the related articles approach tested here may help protect against variable retrievability of evidence between specialties.

Bibliometric characteristics of the evidence base of the original review can inform other aspects of the search plan for the update. If all eligible evidence in the original review was published in English, even though records in any language were sought and were eligible, one could restrict the update search to English language studies. Similarly, if grey literature was sought through a comprehensive approach such the CADTH protocol,[378] but no relevant grey literature was found, one could restrict the update search to the published literature. It could be that a small number of unpublished studies were included in the original review, but all appeared early in the life cycle of the intervention being studied, and a substantial body of published evidence appeared subsequent to grey studies. In this case, reviewers might also streamline the update search to focus on the published evidence, which is much less time consuming to find.

The advantage of using the bibliometric characteristics of the original included studies to inform the update search is that this evidence is derived from the exact context under study. This may assuage reviewer concerns that their topic may be an exception to the global findings that language restrictions or exclusion of non-MEDLINE articles, for example, *tends* not to bias results. Returning to the evidence hierarchies (Section 2.11), Guyatt placed N-of-1 trials at the pinnacle of evidence for guiding treatment decisions. This is because the results of most randomized controlled trials represent average treatment effects, and any given patient may or may not experience that degree of effect. The N-of-1 trial provides the best and most direct evidence of how that patient individual will respond.[379]

The execution of a full N-of-1 design in the search context might include blinding of searcher and reviewer, random selection of the search methods to be used, and a standard-search control condition (adapted from [36]). This is more than the average

systematic reviewer will wish to undertake in their review, but observational evidence from their initial systematic review can help reviewers decide if research findings will be applicable in their case.

Of course, if analysis of the included studies and their indexing showed that the evidence base was multilingual, that grey literature was important or that two or more databases were necessary for the complete identification of material, the update search can be broadened. However, results in these cohorts indicate that a more streamlined approach is usually possible.

Given the evidence gaps pointed out by McAllister where "the therapeutic options for many diseases which cause substantial disability adjusted life years lost worldwide have not been systematically examined at all,"[142] it does not seem like a good use of research resources either to do more updates than needed, or to use an exhaustive approach to the search for new material when analysis of that topic indicates that a more parsimonious approach is possible.

7.8 WAYS FORWARD

The information retrieval problem of updating is quite distinct from the problem of creating the original evidence base. Decisions around the update search can be informed by the results of that original search and the included studies from the original review can be leveraged through similarity searches.

In the update, I recommend the following practice:

- Validate the MEDLINE search against MEDLINE indexed articles, assuming the original search was comprehensive;
- If recall is low, amend that search, structuring it after the approach used here to develop the clinical query searches. That is, describe the population and intervention under study using the appropriate MeSH subject headings, adding free text terms only when suitable MeSH headings are unavailable (see Section 4.2.4.1);
- Have the amended search peer reviewed;[83]

- Complement this with a related article search based on the three newest and three largest studies included in the original review, limited with the Randomized Controlled Trial publication type tag;
- Date limit both searches to the material added to MEDLINE since the original search date;
- Forego non-database search methods and grey literature searching unless grey literature has been present and influential beyond the initial stages. However, if unpublished trials were excluded from the original review, following van Driel's recommentations[373] then full published versions of the excluded trials could be sought for the update; and
- Forego other databases unless they have been demonstrated to contribute material not also indexed in MEDLINE.

The advantages of such an approach are simplicity, lack of variability across clinical area, high levels of identification of new material and that the search is tailored to the particular review.

One of the most important aspects of this recommendation may be the validation of MEDLINE search by testing the proportion of MEDLINE-indexed material that it retrieves. It should be stressed that validation of the original MEDLINE search is only possible when multiple databases and supplemental methods not dependent of searching skill were use to form the original evidence base.

If such testing were done routinely,‡‡‡ MEDLINE searching might improve. First, feedback, either knowledge of results or knowledge of performance, improves performance of skilled activity.[380] Examples of this phenomenon range from the use of audit with feedback to influence physician behaviour[381] to machine learning like Support Vector Machine.[361] Techniques like peer review[83] can provide knowledge of performance (or how well the skill was executed). In the case of feedback on the recall of the MEDLINE search, knowledge of result is the salient aspect, and this type of feedback may allow searchers to improve their searching skills. Second, upon finding that the

‡‡‡ I have seen only one published systematic review in which *post hoc* testing of MEDLINE retrieval was reported, but I have not been able to locate it again to cite it.

results of their MEDLINE searches were poor, searchers may be inclined to have their future searches peer reviewed prior to use, or may seek out someone with more extensive training and expertise to create the searches.

Curiously, there has been very little attention paid to validating subject searches other than the methodological filters. A recent review of reporting standards for the searches of systematic reviews[382] found only one of 11 standards called for evidence of the effectiveness of the search strategy used.[20] Patrick states, "any reporting of a retrieval strategy that does not also report evidence of the effectiveness of that strategy is similarly at odds with the basic tenets of evidence-based medicine."[39] I have demonstrated a simple validation procedure that should be widely used. I have proposed it as a test of the search before it is re-used in updating, but it could be done as soon as final inclusion decisions are made for an original systematic review.

7.9 SURVEILLANCE: SETTING TARGETS

In Chapter 3, we have seen a movement from updating at fixed intervals to updating based on need. Our own and other studies fail to find many predictive factors of survival time, and where found, they are quite general.[17,124] Our major predictor, for example, was that the topic involved cardiovascular disease.[17] However, the concept of optimum information size does provide some guidance on when to update (see Section 3.6.2), although none of those working with the concept of updating based on the accumulation of new evidence have suggested how this accumulation can best be tracked.

For efficient surveillance, some automation would be needed. My results shed some light on how effectively various searches detect new N. Automated extraction of N from abstracts of randomized controlled trials has been demonstrated with 97% accuracy.[383]

Studies retrieved from ongoing searches could be sorted from largest to smallest N, or by some weighting of N and relevance score, and the most promising candidates flagged for immediate relevance assessment. Such screening would continue until a predetermined threshold of new N is reached. Not all candidates will be assessed under

such an approach; therefore, the threshold should be set below the optimum information size. For example, if the optimum information size for a review was 1,000 new participants, formal screening of all remaining candidate studies could begin when 800 new participants had been detected through surveillance.

The best surveillance approach for use with such a scheme may be Related Article RCT, which provides useful relevance scores (see Section 6.11) paired with Abridged Index Medicus RCT, which has better precision than Clinical Queries and preferentially finds larger trials. Once the threshold for updating is reached, Clinical Queries could be run, and material not already screened from the Abridged Index Medicus results could be added.

7.10 IMPLEMENTATION CHALLENGES

Systematic reviewers tend to be very conservative in adopting research findings into their own practice, as has been discussed. However, two factors may aid adoption of these findings. First, it is clear that updating goals have been consistently unmet. Reviewers must choose between leaving potentially outdated evidence in the public arena, finding significantly more resources to support updating, adopting efficiencies or implementing some combination of these strategies. The efficiencies suggested here are evidence-based and so worthy of consideration. Second, these results are derived from the same cohort as the Annals of Internal Medicine survival paper,[17] which has seen good uptake. It has been cited 24 times in a relatively brief period, and some of these articles,[147,384-386] may be influential as they, in turn, have already been cited two or more times. This encouraging initial response suggests that the systematic review community accepts its findings as valid.

Still, widespread adoption may depend on groups such as the Cochrane Collaboration endorsing such an approach. Current recommendations from the updating working groups provide an exhaustive protocol for monitoring for new studies (see Section 3.2.1). As well, the Cochrane Handbook sets limited expectations for identification of studies through MEDLINE based on the 1994 article that introduced the original highly sensitive

search strategy, the Cochrane handsearching initiative and the National Library of Medicine re-tagging of randomized controlled trials.[64] The handbook cites, in section 6.1.1.2, that "only 30% - 80% of all known published randomized trials were identifiable using MEDLINE."[79] Changes to those expectations may be needed to pave the way for acceptance of a search protocol that would involve, in many cases, searching only MEDLINE.

7.11 LIMITATIONS OF THIS RESEARCH

7.11.1 Screening Method

We screened new records up to the point where we could confidently say the systematic review needed to be updated. It is possible that search performance results would have been different had all records available during the observation period been reviewed. However, for the eight AHRQ Evidence Reports where all retrieved records were assessed, recall of eligible studies was very similar for the periods before and after the signal (sees Section 6.8).

7.11.2 Precision

One very real limitation imposed by our screening method is that precision of the searches could only be established in the cohort of ten AHRQ Evidence Reports (see Section 4.5.4). The estimates may not generalize to Cochrane Collaboration Reviews or journal published systematic reviews. As one objective of this research is to find more efficient ways to conduct update searches, more reliable precision figures would be helpful. Retrieval size for the searches was examined in each cohort (see Sections 6.5.1, 6.6.1 and 6.7.1) and these give some indication of what levels of precision might be expected. Median retrieval size was substantially lower for Related Articles RCT than for the other searches in the AHRQ cohort where the precision scores were derived. This did not hold for the largest cohort - median retrieval size was larger there for Related Article RCTs than for Clinical Queries, the search approach closest to it in terms of recall. This suggests that the precision for Related Articles seen in the AHRQ cohort may be

optimistic. However, in all three samples, the third quartile was lower for Related Article RCT than for Clinical Queries, suggesting that un-manageably large retrievals would be relatively uncommon. The other search approaches with data available in all three cohorts were Abridged Index Medicus RCT and Citing RCTs. These had relatively higher precision in the AHRQ sample (see Section 6.10), and small retrievals in all three samples, but that is at the expense of recall (see Section 6.11).

7.11.3 Generalizability

The selection of the sample may have implications for generalizability. Strengths of the selection criteria were that they ensured the topics were clinically important and the quality of the reviews met certain minimum criteria. However, we excluded systematic reviews of complementary and alternative medicines. These findings may still be applicable there as the most recent evidence suggests that most randomized controlled trials of complementary and alternative medicines are indexed in MEDLINE.[368]

I considered systematic reviews of drugs, devices or procedures, and excluded reviews of diagnostic measures, epidemiology, educational interventions, associational studies, instrumentation or research methods, for example. The searches were tested for their ability to retrieve randomized controlled trials, and therefore will be most applicable to searches of randomized controlled trials. A reasonable approach in other areas would be to use the appropriate clinical query – they have been developed and validated to find clinically important studies of causation (including harm), diagnosis or prognosis, for example.[14] This would substitute directly for the Clinical Query tested here, which was balanced query for studies of treatment and prevention. The Related Article search, based on a similar seed selection protocol, could also be limited by the appropriate clinical query. Pairing it with a validated query may still provide good performance. A more conservative approach would be to run the Related Article protocol using the three newest and three largest included studies as seeds, but without any methodological or study design limits. This unrestricted approach may be quite feasible for topics where there are few studies or where precision is not a concern.

The finding that the major search innovation presented here, Related Article RCT, shows consistent recall across clinical area is a very positive sign that these results may be generally applicable. However, exceptions cannot be ruled out.

7.11.4 Goals Not Achieved

An early goal of this research was to find a search strategy that would have precision and recall around 0.80 to enable efficient surveillance. I have developed a deeper understanding of the performance measures of a search during the course of this research and this has helped my understanding of why this goal remains elusive.

Using the analogy of the search as a diagnostic test for document relevance, the standard two-by-two epidemiology table where the marginals are the prevalence of the disease and prevalence of the attribute.[387] This corresponds to the prevalence of relevant articles and prevalence of retrieved articles. Sensitivity and specificity are unaffected by the prevalence of the disease, or the frequency of relevant documents within the collection. Sensitivity is calculated in the same way as recall. Sensitivity, recall and specificity "condition on the test." Precision is the positive predictive value of a test, or in our case, how likely it is that a record is relevant give that it is retrieved. Predictive values, whether positive or negative, are influenced by underlying prevalence of the disease. They "condition on the disease". The disease we wish to diagnose is not a disease of course, but rather relevance of an article for a particular systematic review. Precision will be influenced by the prevalence of relevant documents in the collection. This means the precision of a search will be higher when the search is run in a database with a high concentration of relevant documents than it will be if the same search is run in a database with a lower concentration of relevant documents.[388] This is predicted by the functional relationship between recall, precision and fall-out that van Rijsberen calls generality, or G (see Section 4.4.4.3)[195] In any search of MEDLINE, it is highly probable that most of the records in the database will be on some topic other than the one you are searching, this limits precision, except in the simplest searches, like the search for a named entity.

I now understand that the most effective way to increase precision is not to refine the search strategy, but rather to restrict the search to sources with a higher concentration of relevant documents. Thus, better precision will be obtained in CENTRAL, which contains only controlled trials, than in MEDLINE. Precision will be better if the search is restricted to MEDLINE Core Clinical Journals, as the concentration of trials is greater here than in all of MEDLINE.

7.12 FUTURE WORK

7.12.1 Related Articles

The most exciting finding of this research is the complementary nature of the Related Article and Clinical Query searches. Further exploration and refinement of similarity searching as an adjunct to subject searching would be very useful.

PubMed's Related Article feature has been widely adopted by those searching PubMed. Based on query logs gathered during a one-week period in June 2007 roughly a fifth of all non-trivial PubMed user sessions contain at least one related article search.[168] However, I found few evaluations of its performance.[169,389] It has not been widely adopted for systematic review searches. Using an automated MEDLINE query run throughout this research period, I identified only five systematic reviews that reported a related article search in the methods section of their abstract.[389-393] The PubMed Related Articles feature warrants further research and greater use by systematic reviewers.

The Related Articles RCT out-performed Support Vector Machine in this sample. However, Related Articles is available only for PubMed. Further work to optimize SVM and deploy it to search other databases would be very useful for updating those systematic reviews where databases other than MEDLINE have a demonstrable contribution. If only a few included studies where indexed in those databases it would be impossible to establish an adequate training set of true positives and true negatives. Ideally, those would be studies not also indexed in MEDLINE, so that overlap might be minimized when SVM was used to find new relevant material. Realistically, the MEDLINE component of almost

all systematic reviews is too large to make such an approach practical. Other challenges would be publisher agreement to access the database, and costs associated with such access.

Computing nearest neighbour scores for other databases using the computational strategy employed by National Library of Medicine for PubMed Related Articles searches would be another approach. Other search interfaces provide related article-like features. Ovid has a “find similar” feature that seems to be based solely on words in the title of an article, and this impression is confirmed by the Ovid Online Help.[394] EBSCO*host* provides a feature labeled "Find Similar Results using SmartText Searching”. In CINAHL, the system appears to use the title and subject headings as the basis for the search, although the EBSCO*host* Help indicates that the abstract will be the basis for the search in some databases, and subject headings will be used in others.[395] The EBSCO*host* feature returns a prohibitive number of records – over 700,000 similar articles to a randomized controlled trial of antioxidants, although the results were relevance ranked and top-ranked articles appeared useful.

7.12.2 Maturity of the Literature

The search for factors that predict the survival of systematic reviews has had limited success to date, although it is still a new area of study. My finding that reviews of evidence of new drugs, devices or procedures went out-of-date more quickly than reviews of more established therapies may be useful (see Section 6.32.1). This makes some sense when one considers the “regression toward the truth”[142] whereby early promise of some therapies is not maintained (see Section 3.2.7). Altman considers this a natural phenomenon as “only those treatments that appear promising in early studies are likely to be the subject of extensive further study. We might expect, therefore, that treatments will perform somewhat less well in further studies simply as a consequence of regression to the mean.”[34] Altman used this as an argument for rigorously controlled trials. We have seen that early results from abstracts may not reflect final findings.[374] Publication bias and selective reporting of the most favourable results are likely contribute to distortion in the

early literature.[396] Finally, the number of new participants needed to overturn a result will be smaller in these early systematic reviews, due to the smaller number of included participants.

It may be that reviews can be characterized as early exploratory reviews that attempt to arrive at a definitive answer as soon as possible, as Lau *et al.* argue should have happened with streptokinase as thrombolytic therapy for acute infarction.[135] These exploratory reviews may require frequent updating. Middle stage reviews may attempt to address treatment complexities and resolve differences between individual studies, and explore sub-populations. AHRQ Evidence Reports, with their multifaceted questions and large numbers of included studies are examples of these. Finally, there may be confirmatory systematic reviews, such as those on the efficacy of bed nets for prevention of malaria.[279] Few surprises and few new primary studies are expected in these areas, but the systematic review may still play an important role by providing a single source of synthesized information for reference purposes. Such reviews may need little updating.

If the updating thresholds are set using optimum information size, then the distinction between these types of reviews becomes less important. The exploratory reviews will need fewer new participants to trigger updating; the middle stage reviews will vary in the amount of new information needed depending how conclusive their results are. Additional studies will be unlikely to overturn the findings of confirmatory reviews – the cumulative slope described by Muellerleile and Mullen will approach zero,[146] suggesting that they will need little or no ongoing attention.

7.12.3 Interagency Collaboration to Automate Support for Updating

Given the current interest of major national and international groups in updating (see Section 3.2.9), this may be an opportune time for agencies to collaborate to develop infrastructure to support automating some aspects of surveillance and signaling of the need to update. Sutton suggests that instead of updating all systematic reviews in a collection annually, thresholds should be established for each review based on the amount of new evidence that is likely to be needed to influence the result in ways that have

implications for practice. Calculation of these thresholds could be paired with automated systems to search for new studies, rank them by priority for review, and extract new N. Components of such a system have been developed.[102,112] Collaboration in developing such systems could support the goal of harmonization of updating efforts between agencies (see Section 3.2.9).

Even in the absence of such automation and integration, an automated procedure for capturing related article would be extremely helpful. At present, if the Related Articles search is done through the PubMed interface, a second step is needed to apply the limits – including date limits – and relevance ranking is lost in that second step. Searching through the eLink utility allows date limiting and the addition of a single limit, but converting the results from the XML output to a form more amenable to systematic reviews is cumbersome. The programming of such a query interface is likely less complex than the programming involved in the SYRIAC system.[112] Although the relevance ranking seems to successfully sort relevant material toward the beginning of the result list, it is not yet clear how best to include relevance ranking into the systematic review itself. However, for surveillance purpose, assessing the most relevant material first will allow efficient signal detection.

7.13 CONCLUSION

This research conceptualizes systematic review searches as being made up of multiple approaches to the literature, which need to be independent in order to achieve good recall. An overlooked aspect of the traditional approach of using both database and non-database methods is that such multimodal searches combine methods that require Boolean search skills with those that are independent of such skill.

I introduce the need to test the performance of the MEDLINE subject search, both to provide feedback to the searcher and to determine the potential contribution of MEDLINE to the update of the evidence base. The biblometrics of the initial evidence base can guide the update search by informing the selection of databases and limits, and by providing seed articles for similarity searching.

This is the first demonstration of the structural relationship between different modes of information retrieval – traditional Boolean searches and similarity-based methods – in the context of the very high levels of recall needed for systematic reviews. A similarity search and a Boolean subject search, used together, were shown to provide comprehensive identification of relevant new evidence in several cohorts. A similarity search using the PubMed Related Articles feature is effective and is accessible to all systematic reviewers since it requires no special resources and no proficiency in Boolean searching.

This thesis also introduces the idea of considering the information density of documents, here measured as the number of patients enrolled in studies, rather then number of studies retrieved, as the unit of measure in assessing the performance of searches in the systematic review context. For surveillance purposes, the identification of new study participants permits tracking of progress toward optimum information size.

REFERENCE LIST

(1) Levin A. The Cochrane Collaboration. Ann Intern Med 2001; 135(4):309-312.

(2) National Library of Medicine. MeSH database. Available at: http://www.ncbi.nlm.nih.gov/sites/entrez. 2009. Accessed: 3-12-2009.

(3) Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007; 4(3):e78.

(4) Garritty C, Tsertsvadze A, Tricco AC, Sampson M, Moher D. Updating systematic reviews: An international survey. PLoS ONE (Submitted) 2009.

(5) Hopewell S, Clarke M, Stewart L, Tierney J. Time to publication for results of clinical trials (Cochrane Methodology Review). The Cochrane Library 2004;(3).

(6) Moher D, Tsertsvadze A, Tricco A, Eccles M, Grimshaw JM, Sampson M et al. A systematic review identified few methods and strategies describing when and how to update systematic reviews. J Clin Epidemiol 2007; 60:1095-1104.

(7) Tricco AC, Tetzlaff J, Sampson M, Fergusson D, Cogo E, Horsley T et al. Few systematic reviews exist documenting the extent of bias: a systematic review. J Clin Epidemiol 2008; 61(5):422-434.

(8) Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials for systematic reviews. Online Journal of Current Clinical Trials 1993; Doc No 33:3973.

(9) Barrowman NJ, Fang M, Sampson M, Moher D. Identifying null meta-analyses that are ripe for updating. BMC Med Res Methodol 2003; 3(1):13.

(10) Topfer LA. Summary of responses re Updating of HTA literature searches. owner-spig-ir@stonebow.otago.ac.nz, (ed.). SPIG-IR listserv . 3-18-2004. Internet Communication.

(11) Garritty C, Tricco AC, Sampson M, Tsertsvadze A, Shojania KG, Eccles MP et al. A framework for updating systematic reviews. in preparation 2008.

(12) Moher D, Eccles M, Grimshaw JM, Garritty C, Sampson M, Tsertsvadze A et al. A provisional model for when to update systematic reviews. XIV Cochane Colloquium, Dublin, Ireland 2006.

(13) National Library of Medicine. Abridged Index Medicus (AIM) journal titles. Available at: http://www.nlm.nih.gov/bsd/aim.html. 2-19-2005. National Library of Medicine, Bibliographic Services Division. Accessed: 9-3-2006.

(14) Zaroukian MHM. PubMed clinical queries: a Web tool for filtered retrieval of citations relevant to evidence-based practice. ACP Journal Club 2001; 134(2):A15.

(15) Cohen A, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc 2006; 13(2):206-219.

(16) Shojania KG, Sampson M, Ji J, Ansari MT, Garritty C, Rader T et al. Updating systematic reviews. Evid Rep Technol Assess (Summ) 2007.

(17) Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date: a survival analysis. Ann Intern Med 2007; 147(4):224-233.

(18) Sampson M, Shojania KG, McGowan J, Daniel R, Rader T, Iansavichene AE et al. Surveillance search techniques identified the need to update systematic reviews. J Clin Epidemiol 2008; 61(8):755-762.

(19) Sampson M, Shojania KG, Garritty C, Horsley T, Ocampo M, Moher D. Systematic reviews can be produced and published faster. J Clin Epidemiol 2008; 61(6):531-536.

(20) Patrick TB, Demiris G, Folk LC, Moxley DE, Mitchell JA, Tao D. Evidence-based retrieval in evidence-based medicine. J Med Libr Assoc 2004; 92(2):196-199.

(21) Glanville JM, Lefebvre C, Miles JN, Camosso-Stefinovic J. How to identify randomized controlled trials in MEDLINE: ten years on [Erratum in: J Med Libr Assoc. 2006 Jul;94(3):354]. J Med Libr Assoc 2006; 94(2):130-136.

(22) Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. Eval Health Prof 2002; 25(1):12-37.

(23) Starr M, Chalmers I. The evolution of The Cochrane Library, 1988-2003. Available at: www.update-software.com/history/clibhist.htm. 2003. Oxford, Update Software. Accessed: 10-24-2006.

(24) Systematic reviews of all the relevant evidence [Editorial commentary]. Available at: The James Lind Library (www.jameslindlibrary.org). 2007. Accessed: 1-31-2009.

(25) Bosch FX, Molas R Archie Cochrane: Back to the front. Barcelona: Ajuntament de Barcelona : Cochrane Collaboration : Institut Catala` d'Oncologia, 2003.

(26) Cochrane AL Effectiveness and Efficiency. Random Reflections on Health Services. London: Nuffield Provincial Hospitals Trust, 1972.

(27) Chalmers I, Enkin M, Keirse MJ. Preparing and updating systematic reviews of randomized controlled trials of health care. Milbank Quarterly 1993; 71(3):411-437.

(28) The Cochrane Collaboration. Archie Cochrane: the name behind the Cochrane Collaboration. Available at: http://www.cochrane.org/docs/archieco.htm. 2009. Accessed: 3-7-2009.

(29) Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996; 312(7023):71-72.

(30) Last JM, International Epidemiological Association. A dictionary of epidemiology. New York: Oxford University Press, 2001.

(31) Mulrow CD. The medical review article: state of the science. Ann Intern Med 1987; 106(3):485-488.

(32) Huth EJ. The move toward setting scientific standards for the content of medical review articles. Available at: The James Lind Library (www.jameslindlibrary.org). 2008. Accessed: 3-7-2009.

(33) Lind J A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the subject. Edinburgh: Printed by Sands, Murray and Cochran for A Kincaid and A Donaldson, 1753.

(34) Altman DG. What randomized trials and systematic reviews can offer decision makers. Hormone Research 1999; 51:36-43.

(35) School of Health and Related Research (ScHARR). Systematic reviews: what are they and why are they useful? Hierarchy of evidence. Available at: http://www.shef.ac.uk/scharr/ir/units/systrev/hierarchy.htm. 2009. University of Sheffield. Accessed: 3-12-2009.

(36) Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD et al. Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. Evidence-Based Medicine Working Group. JAMA 2000; 284(10):1290-1296.

(37) McKinlay RJ, Cotoi C, Wilczynski NL, Haynes RB. Systematic reviews and original articles differ in relevance, novelty, and use in an evidence-based service for physicians: PLUS project. J Clin Epidemiol 2008; 61(5):449-454.

(38) Haynes RB. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. Evid Based Med 2006; 11(6):162-164.

(39) Altman DG, Chalmers I Systematic reviews. London: BMJ, 1996.

(40) Egger M, Davey-Smith G, Altman DG Egger M, Davey-Smith G, Altman DG (eds). Systematic reviews in health care: Meta-analysis in context. London: BMJ, 2001.

(41) Lefebvre C, Clarke MJ. Identifying randomised trials. In: Egger M, Davey-Smith G, Altman DG, editors. Systematic reviews in health care: Meta-analysis in context. London: BMJ Books, 2001: 69-89.

(42) Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF et al. Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Lancet 1999; 354:1896-1900.

(43) Wen J, Ren Y, Wang L, Li Y, Liu Y, Zhou M et al. The reporting quality of meta-analyses improves: a random sampling study. J Clin Epidemiol 2008; 61(8):770-775.

(44) Moher D, Tetzlaff J, Liberati A, Altman DG, for the PRISMA group. Recommendations for reporting systematic reviews of randomized trials: the PRISMA Statement. PLoS Medicine (Submitted) 2009.

(45) Kagolovsky Y, Moehr JR. Current status of the evaluation of information retrieval. J Med Syst 2003; 27(5):409-424.

(46) Neufeld ML, Conog M. Database history: From dinosaurs to compact discs. J Am Soc Inf Sci 1986; 37(4):183-190.

(47) Dee CR. The development of the Medical Literature Analysis and Retrieval System (MEDLARS). J Med Libr Assoc 2007; 95(4):416-425.

(48) Adams S, McCarn DB. On-line terminal: one hundred years of medical indexing. Available at: National Library of Medicine Annual Report (http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1975.pdf). 1976. National Library of Medicine. Accessed: 3-9-2009.

(49) Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature [erratum appears in J Am Med Inform Assoc 2001 Nov-Dec;8(6):597]. J Am Med Inform Assoc 2001; 8(4):317-323.

(50) Pizer IH. A regional medical library network. Bull Med Libr Assoc 1996; 57(2):101-115.

(51) National Library of Medicine. Programs and Services: Financial Year 1974. Available at: http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1974.pdf. 1975. Accessed: 3-9-2009.

(52) National Library of Medicine. Communication in the Service of American Health: A Bicentennial Report from the National Library of Medicine. Available at: http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1975.pdf. 1976. Accessed: 3-9-2009.

(53) National Library of Medicine. Programs and Services: Financial Year 1977. Available at: http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1977.pdf. 1978. Accessed: 3-9-2009.

(54) National Library of Medicine. Programs and Services: Financial Year 1979. Available at: http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1979.pdf. 1980. Accessed: 3-9-2009.

(55) National Library of Medicine. Programs and Services: Financial Year 1981. Available at: http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1981.pdf. 1982. Accessed: 3-9-2009.

(56) National Library of Medicine. Programs and Services: Financial Year 1984. Available at: http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1984.pdf. 1985. Accessed: 3-9-2009.

(57) National Library of Medicine. Programs and Services: Financial Year 1987. Available at: http://www.nlm.nih.gov/hmd/manuscripts/nlmarchives/annualreport/1987.pdf. 1988. Accessed: 3-9-2009.

(58) Shojania KG, Bero LA. Taking advantage of the explosion of systematic reviews: an efficient MEDLINE search strategy. Eff Clin Pract 2001; 4(4):157-162.

(59) United States Congress. Medlars. MEDLARS and Health Information Policy. Washington, D.C.: Congress of the U.S., Office of Technology Assessment, 1982.

(60) Kagolovsky Y, Moehr JR. Terminological problems in information retrieval. J Med Syst 2003; 27(5):399-408.

(61) Cummings MM. History of the National Library of Medicine. Presented at Central/Northern Florida HIMSS Chapter Meeting - Sarasota, Florida. Available at: http://www.smh.com/sections/services-procedures/medlib/cmeonline/cummingsHIMSS/cummingsHIMSS.html. 2009. Accessed: 1-30-2009.

(62) Lindberg DA. Testimony on the Fiscal Year 2001 President's Budget Request for the National Library of Medicine. Available at: http://www.nlm.nih.gov/od/fy2001testimony.html. 2-29-2000. National Library of Medicine. Accessed: 3-7-2009.

(63) Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, McDonald S. Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials. Eval Health Prof 2002; 25(1):38-64.

(64) Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994; 309(6964):1286-1291.

(65) The Cochrane Collaboration Methods Working Groups Newsletter. Olsen KL (ed). Vol. 1. 1997.

(66) The Cochrane Collaboration Methods Working Groups Newsletter. Olsen KL (ed). Vol. 2. 1998.

(67) The Cochrane Collaboration Methods Groups Newsletter. Clarke M., Olsen KL (eds). Vol. 3. 1999.

(68) The Cochrane Collaboration Methods Group Newsletter. Clarke M. (ed). 2000.

(69) The Cochrane Collaboration Methods Groups Newsletter. Hopewell S, Clarke M. (eds). Vol. 5. 2001.

(70) Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials. Stat Med 2002; 21(11):1625-1634.

(71) The Cochrane Collaboration Methods Groups Newsletter. Hopewell S, Clarke M. (eds). Vol. 6. 2002.

(72) McDonald S, Taylor L, Adams C. Searching the right database. A comparison of four databases for psychiatry journals. Health Libr Rev 1999; 16(3):151-156.

(73) Hopewell S, Clarke M, Lefebvre C, Scherer R. Handsearching versus electronic searching to identify reports of randomized trials. Cochrane Methodology Review, The Cochrane Library 2003; Issue 4.

(74) White VJ, Glanville JM, Lefebvre C, Sheldon TA. A statistical approach to designing search filters to find systematic reviews: objectivity enhances accuracy. J Inf Sci 2001; 27(6):357-370.

(75) Egger M, Juni P, Bartlett C, Holenstein F, Sterne JA. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003; 7(1):1-76.

(76) Royle P. Obtaining published errata to randomised controlled trials: Is it worth the effort? 10th Cochrane Colloquium, Stavanger 2002;20-21.

(77) Royle P, Waugh N. Should systematic reviews include searches for published errata? Health Info Libr J 2004; 21(1):14-20.

(78) The Cochrane Collaboration Methods Groups Newsletter. Hopewell S, Clarke M. (eds). 1-44. 2003.

(79) Lefebvre C, Manheimer E, Glanville J. Searching for studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. www.cochrane-handbook.org: Wiley, 2008.

(80) The Cochrane Collaboration Methods Groups Newsletter. Hopewell S, Clarke M. (eds). Vol. 8. 2004.

(81) The Cochrane Collaboration Methods Groups Newsletter. Hopewell S, Clarke M. (eds). Vol. 5, 1-39. 2005.

(82) Sampson M, McGowan J, Lefebvre C, Moher D, Grimshaw JM. PRESS: Peer Review of Electronic Search Strategies. CADTH Technical Report Available at: http://cadth ca/index php/en/publication/781 Appendices available at http://www cadth ca/media/pdf/477_PRESS-Peer-Review-Electronic-Search-Strategies_tr_Appendices pdf 2008.

(83) Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol (epub ahead of print) 2009.

(84) Golder S, McIntosh HM, Duffy S, Glanville J. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Info Libr J 2006; 23(1):3-12.

(85) The Cochrane Collaboration Methods Groups Newsletter. Hopewell S, Clarke M. (eds). Vol. 10, 1-35. 2006.

(86) Doust J, Sanders S, Glasziou P, Pietrzak E. Identifying studies for systematic reviews of diagnostic tests. XI Cochrane Colloquium, Barcelona, Spain, 26th - 31st October 2003 2003;63.

(87) Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. J Clin Epidemiol 2005; 58(5):444-449.

(88) Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound treatment studies in EMBASE. J Med Libr Assoc 2006; 94(1):41-47.

(89) Schmidt LM, Gotzsche PC. Of mites and men: reference bias in narrative review articles: a systematic review. J Fam Pract 2005; 54(4):334-338.

(90) French SD, McDonald S, McKenzie JE, Green SE. Investing in updating: how do conclusions change when Cochrane systematic reviews are updated? BMC Med Res Methodol 2005; 5:33.

(91) The Cochrane Collaboration Methods Groups Newsletter. Hopewell S, Clarke M. (eds). Vol. 11, 1-34. 2007.

(92) Whiting P, Westwood M, Burke M, Sterne J, Glanville J. Systematic reviews of test accuracy should search a range of databases to identify primary studies. J Clin Epidemiol 2008; 61(4):357-364.

(93) The Cochrane Collaboration Methods Groups Newsletter. Hopewell S, Clarke M. (eds). Vol. 12, 1-39. 2008.

(94) Moher D, Tsertsvadze A, Tricco A, Eccles M, Grimshaw JM, Sampson M et al. When and how to update systematic reviews. The Cochrane Library 2008;(1).

(95) Ervin AM. Motivating authors to update systematic reviews: practical strategies from a behavioural science perspective. Paediatr Perinat Epidemiol 2008; 22 Suppl 1:33-37.

(96) Linde K. Updating systematic reviews. Explore (NY) 2006; 2(4):363-364.

(97) Moher D, Tsertsvadze A. Systematic reviews: when is an update an update? Lancet 2006; 367(9514):881-883.

(98) Schmidt FL, Raju NS. Updating meta-analytic research findings: Bayesian approaches versus the medical model. Journal of Applied Psychology 2007; 92(2):297-308.

(99) Shea B, Boers M, Grimshaw JM, Hamel C, Bouter LM. Does updating improve the methodological and reporting quality of systematic reviews? BMC Med Res Methodol 2006; 6:27.

(100) Soll RF. Updating reviews: the experience of the Cochrane Neonatal Review Group. Paediatric & Perinatal Epidemiology 2008; 22:29-32.

(101) Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. Stat Med 2006.

(102) Sutton AJ, Donegan S, Takwoingi Y, Garner P, Gamble C, Donald A. An encouraging assessment of methods to inform priorities for updating systematic reviews. J Clin Epidemiol 2009; 62(3):241-251.

(103) Lan KKG, Hu M, Cappelleri JC. Applying the law of iterated logarithm to cumulative meta-analysis of a continuous endpoint. Statistica Sinica 2003; 13:1135-1145.

(104) Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Control Clin Trials 1997; 18(6):580-593.

(105) Ioannidis J, Lau J. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. Proc Natl Acad Sci U S A 2001; 98(3):831-836.

(106) Mullen B, Muerllereile P, Bryant B. Cumulative meta-analysis: a consideration of indicators of sufficiency and stability. Pers Soc Psychol Bull 2001; 27:1450-1462.

(107) Shekelle PG, Ortiz E, Rhodes S, Morton SC, Eccles MP, Grimshaw JM et al. Validity of the Agency for Healthcare Research and Quality clinical practice guidelines: how quickly do guidelines become outdated? JAMA 2001; 286(12):1461-1467.

(108) Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 [updated May 2005]. Chichester, UK: John Wiley & Sons, Ltd 2005.

(109) The Cochrane Collaboration. Maintaining your review. The Cochrane Collaboration open learning material Module 19 2002; available at: http://www.cochrane-net.org/openlearning/HTML/mod19.htm.

(110) Bergerhoff K, Ebrahim S, Paletta G. Do we need to consider 'in process citations' for search strategies? [abstract]. 12th Cochrane Colloquium: Bridging the Gaps; 2004 Oct 2-6; Ottawa, Ontario, Canada 2004;126.

(111) Cohen A. Optimizing feature representation for automated systematic review work prioritization. AMIA Annu Symp Proc 2008;121-125.

(112) Yang JJ, Cohen A, McDonagh MS. SYRIAC: The Systematic Review Information Automated Collection System: A data warehouse for facilitating automated biomedical text classification. AMIA Annu Symp Proc 2008;825-829.

(113) Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005; 12(2):207-216.

(114) Aphinyanaphongs Y, Statnikov A, Aliferis CF. A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. J Am Med Inform Assoc 2006; 13(4):446-455.

(115) Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc 2009; 16(1):25-31.

(116) National Library of Medicine. Unified Medical Language System. Available at: http://www.nlm.nih.gov/research/umls/. 2009. Accessed: 3-19-2009.

(117) Cohen AM, Bhupatiraju RT, Hersh WR. Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. The Thirteenth Text Retrieval Conference (TREC13). The Thirteenth Text Retrieval Conference (TREC13) 2004.

(118) Loudon K, Hopewell S, Clarke M., Moher D, Scholten R, Eisinga A et al. Checklist for Updating Cochrane Reviews. 3-20-2008. Unpublished Work.

(119) Cochrane Pain PaSCG. Information to help review authors update their Cochrane review. Available at: http://www.liv.ac.uk/evidence/CIDG/checklist-review-update.doc. 2008. Accessed: 2-16-2009.

(120) Clinical Evidence. Tovey D, (ed.). Available at: http://www.clinicalevidence.com. 2007. London, BMJ Publishing Group.

(121) Roach J, Thomas A, Tovey D. Exploring the need to update systematic reviews. XIV Cochane Colloquium, Dublin, Ireland 2006.

(122) Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull 1979; 86(3):638-641.

(123) Koch GG. No improvement -still less than half of the Cochrane reviews are up to date. XIV Cochane Colloquium, Dublin, Ireland 2006.

(124) Peterson K, McDonagh MS, Chan B, Fu R, Thakurta SG. Development of a practical and efficient methodology for updating [abstract]. XV Cochrane Colloquium; 2007 Oct 23 27; Sao Paulo, Brazil 2007;113.

(125) Peterson K, Chan B, McDonagh MS. Empirical evaluation for a methodology for determing when a comparative drug effectiveness review has become out of date. XVI Cochrane Colloquium, Freiberg Germany 2008.

(126) Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M et al. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. JAMA 1998; 280(3):278-280.

(127) Henderson S, Hampson L, Atherton D, Neilson J. Ten years of updating Cochrane reviews: the experiences of the Pregnancy and Childbirth Group. XIV Cochane Colloquium, Dublin, Ireland 2006.

(128) Jacquerioz FA, Belizan JM, Buekens P. Recommendations to increase the impact of maternal and childbirth health systematic reviews in the Americas. Paediatr Perinat Epidemiol 2008; 22 Suppl 1:61-66.

(129) Brok J, Thorlund K, Gluud C, Wetterslev Jr. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. J Clin Epidemiol 2008; 61(8):763-769.

(130) Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive--Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. Int J Epidemiol 2009; 38(1):287-298.

(131) Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2009; 38(1):276-286.

(132) Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008; 61(1):64-75.

(133) Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol (epub ahead of print) 2009.

(134) Hu M, Cappelleri JC, Lan KKG. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. Clinical Trials 2007; 4(4):329-340.

(135) Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med 1992; 327(4):248-254.

(136) Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL et al. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol 2008; 61(9):857-865.

(137) Cecil WT, Kasteridis P, Barnes JW, Jr., Mathis RS, Patric K, Martin S. A meta-analysis update: percutaneous coronary interventions. Am J Manag Care 2008; 14(8):521-528.

(138) Dahabreh IJ, Economopoulos K. Meta-analysis of rare events: an update and sensitivity analysis of cardiovascular events in randomized trials of rosiglitazone. Clin Trials 2008; 5(2):116-120.

(139) Peter JV, Moran JL, Phillips-Hughes J, Warn D. Noninvasive ventilation in acute respiratory failure--a meta-analysis update. Crit Care Med 2002; 30(3):555-562.

(140) Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005; 294(2):218-228.

(141) Gehr BT, Weiss C, Porzsolt F. The fading of reported effectiveness. A meta-analysis of randomised controlled trials. BMC Med Res Methodol 2006; 6:25.

(142) McAlister FA, Mohamed R. The evolution of evidence: Cautionary notes for the clinician and the meta-analyst. Am Heart J 2007; 153(2):156-158.

(143) Vaitkus PT, Brar C. N-acetylcysteine in the prevention of contrast-induced nephropathy: publication bias perpetuated by meta-analyses. Am Heart J 2007; 153(2):275-280.

(144) Thomas A, McNeil A. Decreasing effectiveness of treatments with increased evidence: results from updating BMJ Clinical Evidence systematic reviews. XV Cochrane Colloquium; 2007 Oct 23 27; Sao Paulo, Brazil 2007;148.

(145) Thomas A, McNeil A. Decreasing effectiveness of treatments with increased evidence: results from updating BMJ Clinical Evidence systematic reviews [abstract]. HTAi 2008, 6-9 July, Montréal, Canada 2008.

(146) Muellerleile P, Mullen B. Sufficiency and stability of evidence for public health interventions using cumulative meta-analysis. Am J Public Health 2006; 96(3):515-522.

(147) Whitlock EP, Lin JS, Chou R, Shekelle P, Robinson KA. Using existing systematic reviews in complex systematic reviews. Ann Intern Med 2008; 148(10):776-782.

(148) Jones L. Recommendations Report for the Strategic Review of The Cochrane Collaboration: Prioritising the Review's Recommendations. Available at: http://ccreview.wikispaces.com/Prioritising+the+Review%27s+recommendations. 2009. The Cochrane Collaboration. Accessed: 4-30-2009.

(149) Purpose and Procedure. ACP Journal Club. Available at: http://www.acpjc.org/shared/purpose_and_procedure.htm. 2007. Accessed: 2-7-2007.

(150) Shea B, Dube C, Moher D. Assessing the quality of reports of systematic review: the QUOROM statement compared to other tools. In: Egger M, Smith GD, Altman DG, editors. Systematic reviews in health care: meta-analysis in context. London: BMJ Publishing Group, 2001: 122-139.

(151) Whitener BL, Sampson M, Crumley E, Bonnell CJ, Dorgan M, McGowan J et al. Identifying studies for AHRQ Evidence Reports: Is searching MEDLINE enough? (research proposal). 2003. Unpublished Work.

(152) Ovid MEDLINE [Electronic database]. Ovid Technologies, 2005.

(153) National Library of Medicine. Search strategy used to create the systematic reviews subset on PubMed. Available at: http://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html. 2003. National Library of Medicine. Accessed: 12-3-2004.

(154) Reference Manager [Computer program]. Carlsbad, California: Thomson Researchsoft Inc., 2005.

(155) SRS [Computer program]. Ottawa, Ontario: Trialstat Corp., 2007.

(156) Armour T, Dingwall O, Sampson M. Contribution of checking reference lists to systematic reviews. Poster presentation at: XIII Cochrane Colloquium 2005.

(157) Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994; 1(6):447-458.

(158) Rada G, Corbalan J, Barrios G, Candia R, Jaime F, Larrondo F et al. Pragmatic comparison of the Cochrane Highly Sensitive Search Strategy and the HEDGES Team strategies in four electronic databases. XIV Cochrane Colloquium, Dublin, Ireland 2006.

(159) Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ 2004; 328(7447):1040.

(160) Wilczynski NL, Haynes RB, Lavis JN, Ramkissoonsingh R, Arnold-Oatley AE. Optimal search strategies for detecting health services research studies in MEDLINE. CMAJ 2004; 171(10):1179-1185.

(161) Tsay MY, Yang YH. Bibliometric analysis of the literature of randomized controlled trials. J Med Libr Assoc 2005; 93(4):450-458.

(162) Pao ML, Worthen DB. Retrieval effectiveness by semantic and citation searching. J Am Soc Inf Sci 1989; 40(4):226-235.

(163) Patsopoulos NA, Analatos AA, Ioannidis JP. Relative citation impact of various study designs in the health sciences. JAMA 2005; 293(19):2362-2366.

(164) CIHR Randomized Controlled Trials Program - Guidelines for Completion. Available at: http://www.irsc.ca/e/3448.html. 7-14-2006. Canadian Institutes for Health Research.

(165) Bernstam E. MedlineQBE (Query-by-Example). Proc AMIA Symp 2001;47-51.

(166) Saracevic T, Kantor P. A study of information seeking and retrieving. 3. Searchers, searches, and overlap. J Am Soc Inf Sci 1988; 39(3):197-216.

(167) National Library of Medicine. PubMed Help. Appendix: Computation of Related Articles. Available at: Bethesda (MD): National Library of Medicine (US). 2005. Accessed: 11-1-2008.

(168) Lin J, DiCuccio M, Grigoryan V, Wilbur WJ. Navigating information spaces: A case study of related article search in PubMed. Information Processing & Management 2008; 44(5):1771-1783.

(169) Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics 2007; 8:423.

(170) Liu X, Altman RB. Updating a bibliography using the related articles function within PubMed. Proc AMIA Symp 1998;750-754.

(171) White HD. Toward automated search strategies. Proceedings, 13th International Online Information Meeting 1989; Oxford, Learned Information:33-47.

(172) Schlosser RW, Wendt O, Bhavnani S, Nail-Chiwetalu B. Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive Pearl Growing. A review. International Journal of Language & Communication Disorders 2006; 41(5):567-582.

(173) de Bruijn B, Martin J. Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inf 2002; 67(1-3):7-18.

(174) Pavlidis P, Wapinski I, Noble WS. Support Vector Machine classification on the web. Bioinformatics 2004; 20(4):586-587.

(175) O'Blenis P, Matwin A, Armour Q, Yimin A. Using machine learning to automate the broad screening process -a research update. XIV Cochane Colloquium, Dublin, Ireland 2006.

(176) Herbinson P, Hay-Smith J. Response to systematic reviews. Cochrane Colloquium 2005.

(177) Del Mar CB, Glasziou PP. Antibiotics for the symptoms and complications of sore throat. Cochrane Database Syst Rev 1997; (3).

(178) Sampson M, Barrowman NJ, Moher D, Clifford TJ, Platt RW, Morrison A et al. Can electronic search engines optimize screening of search results in systematic reviews: an empirical study. BMC Med Res Methodol 2006; 6:7.

(179) Tague-Sutcliffe J. The pragmatics of information retrieval experimentation, revisited. Information Processing & Management 1992; 28(4):467-490.

(180) Saracevic T. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. Journal of the American Society for Information Science & Technology 2007; 58(13):1915-1933.

(181) Hersh WR. System evaluation. In: Springer-Verlag, editor. Information retrieval: A health and biomedical perspective. New York: 2003: 83-113.

(182) Borlund P. The concept of relevance in IR. Journal of the American Society for Information Science and Technology 2003; 54(10):913-925.

(183) Blair DC. STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years after. J Am Soc Inf Sci 1996; 47(1):4-22.

(184) van der Weide T, van Bommel P. Measuring the incremental information value of documents. Information Sciences 2006; 176(2):91-119.

(185) Mizzaro S. Relevance: The whole history. J Am Soc Inf Sci 1997; 48(9):810-832.

(186) White HD. Combining bibliometrics, information retrieval, and relevance theory, Part 1: First examples of a synthesis. Journal of the American Society for Information Science & Technology 2007; 58(4):536-559.

(187) Kagolovsky Y, Moehr JR. A new look at information retrieval evaluation: P roposal for solutions. J Med Syst 2004; 28(1):103-116.

(188) Tramer MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. BMJ 1997; 315(7109):635-640.

(189) Hersh W. Relevance and retrieval evaluation: Perspectives from medicine. J Am Soc Inf Sci 1994; 45(3):201-206.

(190) Derogatis LR, Lynn LL. Screening and monitoring psychiatric disorders in pediatric primary care setting. In: Maruish ME, editor. Handbook of psychological assessment in primary care settings. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 2000: 115-152.

(191) Benn P, Wright D, Cuckle H. Practical strategies in contingent sequential screening for Down syndrome. Prenat Diagn 2005; 25(8):645-652.

(192) Sampson M, Zhang L, Morrison A, Barrowman NJ, Clifford TJ, Platt R et al. An alternative to the hand searching gold standard: Validating methodological search filters using relative recall. BMC Med Res Methodol 2006; 6:33.

(193) Jenkins M. Evaluation of methodological search filters--a review. Health Info Libr J 2004; 21(3):148-163.

(194) Lendgrebe TCW, Paclik P, Duin, .P.W. Precision-recall operating characteristic (P-ROC) curves in imprecise environments. IEEE Computer Society. Proceedings of the 18th International Conference on Pattern Recognition 2006; 4:123-127.

(195) Van Rijsbergen CJ Information Retrieval. Newton, MA: Butterworth-Heinemann, 1979.

(196) Weiss GM. The effect of small disjuncts and class distribution on decision tree learning. Rutgers The State University of New Jersey - New Brunswick, 2003.

(197) SPSS [Computer program]. Chicago, Illinois: SPSS Inc., 2007.

(198) Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001; 323(7305):157-162.

(199) Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club 1995; 123(3):A12-A13.

(200) Jong P, Demers C, McKelvie RS, Liu PP. Angiotensin receptor blockers in heart failure: meta-analysis of randomized controlled trials. J Am Coll Cardiol 2002; 39:463-470.

(201) Shekelle P, Morton SC, Hardy M. Effect of supplemental antioxidants vitamin C, vitamin E, and coenzyme Q10 for the prevention and treatment of cardiovascular disease. Evid Rep Technol Assess 2003; (83):1-3.

(202) Velmahos GC, Kern J, Chan L, Oder D, Murray JA, Shekelle P. Prevention of venous thromboembolism after injury. Evid Rep Technol Assess (Summ) 2000;(22):1-3.

(203) Day D, Furlan A, Irvin E, Bombardier C. Simplified search strategies were effective in identifying clinical trials of pharmaceuticals and physical modalities. J Clin Epidemiol 2005; 58(9):874-881.

(204) Plot [Computer program]. Berlin, Germany: 2007.

(205) CurveExpert [Computer program]. Tenessee: 2001.

(206) Adobe Photoshop Elements [Computer program]. San Jose, California: Adobe Systems Inc., 2002.

(207) Neway JM, Lancaster FW. The correlation between pertinence and rate of citation duplication in multidatabase searches. J Am Soc Inf Sci 1983; 34(4):292-293.

(208) Pao ML. Perusing the literature via citation links. Comput Biomed Res 1993; 26(2):143-156.

(209) Wu S, McClean S. Performance prediction of data fusion for information retrieval. Information Processing & Management 2005; 42(4):899-915.

(210) Beitzel SM, Frieder O, Jensen EC, Grossman D, Chowdhury A, Goharian N. Disproving the fusion hypothesis: an analysis of data fusion via effective information retrieval strategies. Proceedings of the 2003 ACM Symposium on Applied Computing 2003;823-827.

(211) Lee J. Analyses of multiple evidence combination. Proceedings of the 20th annual international ACM SIGIR conference on Research and Development in Information Retrieval 1997;267-276.

(212) Tutorials in biostatistics. Volume 1, Statistical methods in clinical studies D'Agostino RB (ed). Tutorials in biostatistics. Volume 1, Statistical methods in clinical studies. John Wiley & Sons, 2004.

(213) Spoor P, Airey M, Bennett C, Greensill J, Williams R. Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. BMJ 1996; 313(7053):342-343.

(214) Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. Am J Epidemiol 1992; 135(9):1060-1067.

(215) Bennett DA, Latham NK, Stretton C, Anderson CS. Capture-recapture is a potentially useful method for assessing publication bias. J Clin Epidemiol 2004; 57(4):349-357.

(216) Kastner M, Straus SE, McKibbon KA, Goldsmith CH. The capture-mark-recapture technique can be used as a stopping rule when searching in systematic reviews. J Clin Epidemiol 2009; In Press, Corrected Proof.

(217) Booth A. Cochrane or cock-eyed? how should we conduct systematic reviews of qualitative research? Qualitative Evidence-based Practice Conference, Taking a Critical Stance 2001.

(218) Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995; 17(2):243-264.

(219) Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996; 276(8):637-639.

(220) Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technol Assess 2000; 4(10):1-115.

(221) Jensen AL. Confidence intervals for nearly unbiased estimators in single-mark and single-recapture experiments. Biometrics 1989; 45(4):1233-1237.

(222) Tabachnick BG, Fidell LS Using multivariate statistics. New York: Harper & Row, 1989.

(223) Multidimensional Scaling. Electronic Statistics Textbook. Lewicki P, Hill T, (eds.). Available at: Statsoft Inc. 2008. Statsoft Inc. Accessed: 10-19-2008.

(224) Leydesdorff L. On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. Journal of the American Society for Information Science & Technology 2008; 59(1):77-85.

(225) Parker J. Evaluating Bibliographic Database Overlap for Marine Science Literature Using an Ecological Concept. Issues in Science & Technology Librarianship 2005; 42(Spring):e1.

(226) Egghe L, Rousseau R. Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve. Information Processing & Management 2006; 42(1):106-120.

(227) Wilson FL, McGrath WE. Cluster analysis of title overlap in twenty-one library collections in Western New York. In: Egghe L, Rousseau R, editors. Proceedings 1st International Conference on Bibliometrics and Theorectical Aspects of Information Retrieval. Diepenbeek: Belgium, 1990: 335-354.

(228) Hamers L, Hemeryck Y, Herweyers G, Janssen M, Keters H, Rousseau R et al. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Information Processing & Management 1989; 25(3):315-318.

(229) Allen RE, Fowler HW, Fowler FG The Concise Oxford Dictionary of Current English. Oxford; New York: Clarendon Press; Oxford University Press, 1990.

(230) Borgatti S. Social Network Analysis: Multidimensional Scaling. Available at: http://www.analytictech.com/borgatti/mds.htm. 1997. Accessed: 10-24-2008.

(231) Beckstead JW, Beckstead LG. A multidimensional analysis of the epistemic origins of nursing theories, models, and frameworks. International Journal of Nursing Studies 2006; 43(1):113-122.

(232) Jadad AR, Moher D, Klassen TP. Arch Pediatr Adolesc Med 1998; 152(8):812-817.

(233) McCord JF, Michelinakis G. Systematic review of the evidence supporting intra-oral maxillofacial prosthodontic care. Eur J Prosthodont Restor Dent 2004; 12(3):129-135.

(234) Fokkinga WA, Kreulen CM, Vallittu PK, Creugers NH. A structured analysis of in vitro failure loads and failure modes of fiber, metal, and ceramic post-and-core systems. Int J Prosthodont 2004; 17(4):476-482.

(235) Evers JL, Collins JA. Surgery or embolisation for varicocele in subfertile men. Cochrane Database Syst Rev 2004;(3):CD000479.

(236) Henry C, Ghaemi SN. Insight in psychosis: a systematic review of treatment interventions. Psychopathology 2004; 37(4):194-199.

(237) Bruscoli M, Lovestone S. Is MCI really just early dementia? A systematic review of conversion studies. Int Psychogeriatr 2004; 16(2):129-140.

(238) Fricke M. Measuring recall. J Inf Sci 1998; 24(6):409-417.

(239) Kent A Encyclopedia of library and information science. New York: Dekker, 1987.

(240) Potter WG. "Of Making Many Books There Is No End": bibliometrics and libraries. Journal of Academic Librarianship 1988; 14(4):238a-238c.

(241) Anonymous. ISI Journal Citation Reports. Available at: http://isiknowledge.com/wos. 2007. Thompson ISI. Philadelphia, Pa.

(242) Cochrane Handbook for Systematic Reviews of Interventions Higgins JPT, Green S (eds). Cochrane Handbook for Systematic Reviews of Interventions. 5.0.0 ed. www.cochrane-handbook.org: The Cochrane Collaboration, 2008.

(243) DerSimonian R, Levine RJ. Resolving discrepancies between a meta-analysis and a subsequent large controlled trial. JAMA 1999; 282(7):664-670.

(244) Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ 1994; 309:1351-1355.

(245) Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990; 263(10):1385-1389.

(246) Egger M, Smith GD. Bias in location and selection of studies. BMJ 1998; 316(7124):61-66.

(247) Scherer RW, Langenberg P, von Elm E. Full publication of results initially presented in abstracts. Cochrane Database of Methodology Reviews. Cochrane Database Syst Rev 2007;(2).

(248) Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ 1997; 315(7109):640-645.

(249) Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol 2001; 54(10):1046-1055.

(250) Copas J, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. Biostatistics 2000; 1(3):247-262.

(251) Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001; 135(11):982-989.

(252) Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. Stat Methods Med Res 2001; 10(4):251-265.

(253) U.S.Department of Health and Human Services. Evidence-based Practice. Available at: Agency for Healthcare Research and Quality. 2008. Accessed: 11-11-2008.

(254) Shekelle P, Hardy ML, Coulter I, Udani J, Spar M, Oda K et al. Effect of the supplemental use of antioxidants vitamin C, vitamin E, and coenzyme Q10 for the prevention and treatment of cancer. Evid Rep Technol Assess (Summ) 2003;(75):1-3.

(255) Ridker PM, Cook NR, Lee IM, Gordon D, Gaziano JM, Manson JE et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. N Engl J Med 2005; 352(13):1293-1304.

(256) Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D et al. The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. Arch Intern Med 2004; 164(21):2335-2342.

(257) Lee IM, Cook NR, Gaziano JM, Gordon D, Ridker PM, Manson JE et al. Vitamin E in the primary prevention of cardiovascular disease and cancer: the Women's Health Study: a randomized controlled trial.[see comment]. JAMA 2005; 294(1):56-65.

(258) Buring JE. Aspirin prevents stroke but not MI in women; vitamin E has no effect on CV disease or cancer. Cleve Clin J Med 2006; 73(9):863-870.

(259) Liu S, Lee IM, Song Y, Van DM, Cook NR, Manson JE et al. Vitamin E and risk of type 2 diabetes in the women's health study randomized controlled trial. Diabetes 2006; 55(10):2856-2862.

(260) Song Y, Manson JE, Buring JE, Liu S. A prospective study of red meat consumption and type 2 diabetes in middle-aged and elderly women: the women's health study. Diabetes Care 2004; 27(9):2108-2115.

(261) Schaumberg DA, Sullivan DA, Buring JE, Dana MR. Prevalence of dry eye syndrome among US women. Am J Ophthalmol 2003; 136(2):318-326.

(262) Christen W, Glynn R, Sperduto R, Chew E, Buring J. Age-related cataract in a randomized trial of beta-carotene in women. Ophthalmic Epidemiol 2004; 11(5):401-412.

(263) Cook NR, Lee IM, Gaziano JM, Gordon D, Ridker PM, Manson JE et al. Low-dose aspirin in the primary prevention of cancer: the Women's Health Study: a randomized controlled trial. JAMA 2005; 294(1):47-55.

(264) Song Y, Manson JE, Buring JE, Liu S. Dietary magnesium intake in relation to plasma insulin levels and risk of type 2 diabetes in women. Diabetes Care 2004; 27(1):59-65.

(265) Galan P, Briancon S, Favier A, Bertrais S, Preziosi P, Faure H et al. Antioxidant status and risk of cancer in the SU.VI.MAX study: is the effect of supplementation dependent on baseline levels? Br J Nutr 2005; 94(1):125-132.

(266) Hercberg S, Galan P, Preziosi P, Malvy M, Briancon S, Ait HM et al. The SU.VI.MAX trial on antioxidants. IARC Sci Publ 2002; 156:451-455.

(267) Galan P, Favier A, Preziosi P, Bertrais S, Arnault N, Hercberg S. [The bank of biological material in the SU.VI.MAX study]. [French]. Rev Epidemiol Sante Publique 2003; 51(1 Pt 2):147-150.

(268) Czernichow S, Bertrais S, Blacher J, Galan P, Briancon S, Favier A et al. Effect of supplementation with antioxidants upon long-term risk of hypertension in the SU.VI.MAX study: association with plasma antioxidant levels. J Hypertens 2005; 23(11):2013-2018.

(269) Zureik M, Galan P, Bertrais S, Mennen L, Czernichow S, Blacher J et al. Effects of long-term daily low-dose supplementation with antioxidant vitamins and minerals on structure and function of large arteries. Arterioscler Thromb Vasc Biol 2004; 24(8):1485-1491.

(270) Meyer F, Galan P, Douville P, Bairati I, Kegle P, Bertrais S et al. Antioxidant vitamin and mineral supplementation and prostate cancer prevention in the SU.VI.MAX trial. Int J Cancer 2005; 116(2):182-186.

(271) Malvy DJ, Favier A, Faure H, Preziosi P, Galan P, Arnaud J et al. Effect of two years' supplementation with natural antioxidants on vitamin and trace element status biomarkers: preliminary data of the SU.VI.MAX study. Cancer Detect Prev 2001; 25(5):479-485.

(272) Hercberg S, Estaquio C, Czernichow S, Mennen L, Noisette N, Bertrais S et al. Iron status and risk of cancers in the SU.VI.MAX cohort. J Nutr 2005; 135(11):2664-2668.

(273) Biondi-Zoccai GG, Lotrionte M, Abbate A, Testa L, Remigi E, Burzotta F et al. Compliance with QUOROM and quality of reporting of overlapping meta-analyses on the role of acetylcysteine in the prevention of contrast associated nephropathy: case study. BMJ 2006; 332(7535):202-209.

(274) Alejandria MM, Lansang MA, Dans LF, Mantaring JB. Intravenous immunoglobulin for treating sepsis and septic shock. Cochrane Database Syst Rev 2001; (1)(CD001090).

(275) Arroll B, Kenealy T. Antibiotics versus placebo in the common cold. Cochrane Database Syst Rev 1999; (1).

(276) Cody J, Daly C, Campbell M. Recombinant human erythropoietin for chronic renal failure anaemia in pre-dialysis patients. Cochrane Database Syst Rev 2001;((4)):CD003266.

(277) Fisher M, Friedman SB, Strauss B. The effects of blinding on acceptance of research papers by peer review [published erratum appears in JAMA 1994 Oct 19;272(15):1170]. JAMA 1994; 272(2):143-146.

(278) Demicheli V, Rivetti D, Deeks JJ, Jefferson TO. Vaccines for preventing influenza in healthy adults. Cochrane Database Syst Rev 2001; (1):CD001269.

(279) Lengeler C. Insecticide treated bednets and curtains for malaria control. Cochrane Database Syst Rev 1999; 1.

(280) Santaguida P, Raina P, Booker L. Pharmacological treatment of dementia. Evid Rep Technol Assess 2004; Apr:1-16.

(281) Berkman ND, Thorp JM, Jr., Hartmann KE, Lohr KN, Idicula AE, McPheeters M et al. Management of preterm labor. Evid Rep Technol Assess (Summ) 2000;(18):1-6.

(282) Buscemi N, Vandermeer B, Pandya R, Hooton N, Tjosvold L, Hartling L et al. Melatonin for treatment of sleep disorders. Evid Rep Technol Assess (Summ) 2004;(108):1-7.

(283) Grady D, Chaput L, Kristof M. Diagnosis and treatment of coronary heart disease in women: systematic reviews of evidence on selected topics. Evid Rep Technol Assess (Summ) 2003;(81):1-4.

(284) Guise JM, McDonagh MS, Hashima J, Kraemer DF, Eden KB, Berlin M et al. Vaginal birth after cesarean (VBAC). Evid Rep Technol Assess (Summ) 2003;(71):1-8.

(285) McNamara RL, Bass EB, Miller MR, Segal JB, Goodman SN, Kim NL et al. Management of new onset atrial fibrillation. Evid Rep Technol Assess (Summ) 2000;(12):1-7.

(286) Shekelle PG, Morton SC, Maglione M, Suttorp M, Tu W, Li Z et al. Pharmacological and surgical treatment of obesity. Evid Rep Technol Assess (Summ) 2004;(103):1-6.

(287) Nichol G, Huszti E, Rokosh J, Dumbrell A, McGowan J, Becker L. Impact of informed consent requirements on cardiac arrest research in the United States: exception from consent or from research? Resuscitation 2004; 62(1):3-23.

(288) Gulmezoglu AM, Say L, Betran AP, Villar J, Piaggio G. WHO systematic review of maternal mortality and morbidity: methodological issues and challenges. BMC Med Res Methodol 2004; 4:16.

(289) Betran AP, Say L, Gulmezoglu AM, Allen T, Hampson L. Effectiveness of different databases in identifying studies for systematic reviews: experience from the WHO systematic review of maternal morbidity and mortality. BMC Med Res Methodol 2005; 5(1):6.

(290) Eto H. Rising tail in Bradford distribution: its interpretation and application. Scientometrics 1988; 13(6):271-288.

(291) Bradford SC. "Sources of information on specific subjects" - reprinted and introduced by B.C. Brookes. J Inf Sci 1985; 10(4):173-180.

(292) Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ 2005; 331(7524):1064-1065.

(293) Adams NP, Bestall JB, Jones PW. Inhaled beclomethasone versus placebo for chronic asthma. Cochrane Database Syst Rev 2001; (2):CD002738.

(294) Birck R, Krzossok S, Markowetz F. Acetylcysteine for prevention of contrast nephropathy: meta-analysis. Lancet 2003; 362:598-603.

(295) Bischoff-Ferrari HA, wson-Hughes B, Willett WC. Effect of vitamin D on falls: a meta-analysis. JAMA 2004; 291:1999-2006.

(296) Cullum N, Fletcher AW, Nelson EA, Sheldon TA. Compression bandages and stockings in the treatment of venous leg ulcers. Cochrane Database Syst Rev 1998; (2).

(297) Cullum N, Deeks J, Sheldon TA, Song F, Fletcher AW. Beds, mattresses and cushions for pressure sore prevention and treatment. Cochrane Database Syst Rev 2000; (4):CD001735.

(298) Ducharme F, Hicks G, Kakuma R. Addition of anti-leukotriene agents to inhaled corticosteroids for chronic asthma. Cochrane Database Syst Rev 2003; (3):CD003133.

(299) Ducharme FM, Hicks GC. Anti-leukotriene agents compared to inhaled corticosteroids in the management of recurrent and/or chronic asthma in adults and children. Cochrane Database Syst Rev 2001; (1):CD002314.

(300) Edmonds ML, Camargo CA, Jr., Saunders LD, Brenner BE, Rowe BH. Inhaled steroids in acute asthma following emergency department discharge. Cochrane Database Syst Rev 2001; (1):CD002316.

(301) Gadsby JG, Flowerdew MW. The effectiveness of transcutaneous electrical nerve stimulation (TENS) and acupuncture-like transcutaneous electrical nerve stimulation (ALTENS) in the treatment of patients with chronic low-back pain. Cochrane Database Syst Rev 1997; 3.

(302) Jefferson TO, Demicheli V, Deeks JJ, Rivetti D. Amantadine and rimantadine for preventing and treating influenza A in adults. Cochrane Database Syst Rev 2001; (1).

(303) Lefering R, Neugebauer EA. Steroid controversy in sepsis and septic shock: a meta-analysis. Crit Care Med 1995; 23:1294-1303.

(304) McIntyre PB, Berkey CS, King SM. Dexamethasone as adjunctive therapy in bacterial meningitis: a meta-analysis of randomized clinical trials since 1988. JAMA 1997; 278:925-931.

(305) Langhorne P, Dennis M Stroke units: an evidence-based approach. London: BMJ Books, 1998.

(306) Anand SS, Yusuf S. Oral anticoagulant therapy in patients with coronary artery disease: a meta-analysis. JAMA 1999; 282:2058-2067.

(307) Dalby M, Bouzamondo A, Lechat P, Montalescot G. Transfer for primary angioplasty versus immediate thrombolysis in acute myocardial infarction: a meta-analysis. Circulation 2003; 108:1809-1814.

(308) Kong DF, Califf RM, Miller DP. Clinical outcomes of therapeutic agents that block the platelet glycoprotein IIb/IIIa integrin in ischemic heart disease. Circulation 1998; 98:2829-2835.

(309) Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. BMJ 2002; 324:71-86.

(310) Halkes PH, van GJ, Kappelle LJ, Koudstaal PJ, Algra A. Aspirin plus dipyridamole versus aspirin alone after cerebral ischaemia of arterial origin (ESPRIT): randomised controlled trial. Lancet 2006; 367(9523):1665-1673.

(311) Bucher HC, Griffith LE, Guyatt GH. Effect of HMGcoA reductase inhibitors on stroke: A meta-analysis of randomized, controlled trials. Ann Intern Med 1998; 128:89-95.

(312) Amarenco P, Bogousslavsky J, Callahan A, III, Goldstein LB, Hennerici M, Rudolph AE et al. High-dose atorvastatin after stroke or transient ischemic attack. N Engl J Med 2006; 355(6):549-559.

(313) Cullum N, Nelson EA, Nixon J. Pressure sores. Clin Evid 2004;(11):2565-2575.

(314) Nixon J, Cranny G, Iglesias C, Nelson EA, Hawkins K, Phillips A et al. Randomised, controlled trial of alternating pressure mattresses compared with alternating pressure overlays for the prevention of pressure ulcers: PRESSURE (pressure relieving support surfaces) trial. BMJ 2006; 332(7555):1413.

(315) Freemantle N, Cleland J, Young P, Mason J, Harrison J. [beta] Blockade after myocardial infarction: systematic review and meta regression analysis. BMJ 1999; 318:1730-1737.

(316) Chen ZM, Pan HC, Chen YP, Peto R, Collins R, Jiang LX et al. Early intravenous then oral metoprolol in 45,852 patients with acute myocardial infarction: randomised placebo-controlled trial. Lancet 2005; 366(9497):1622-1632.

(317) Jefferson T, Demicheli V, Rivetti D, Jones M, Di PC, Rivetti A. Antivirals for influenza in healthy adults: systematic review. Lancet 2006; 367(9507):303-313.

(318) Annane D, Sebille V, Charpentier C, Bollaert PE, Francois B, Korach JM et al. Effect of treatment with low doses of hydrocortisone and fludrocortisone on mortality in patients with septic shock. JAMA 2002; 288(7):862-871.

(319) Brasseur D. Antivirals for influenza in healthy adults. Lancet 2006; 367(9522):1572-1573.

(320) Smith J, Dutkowski R, Ward P. Antivirals for influenza in healthy adults. Lancet 2006; 367(9522):1571.

(321) Monto AS. Antivirals for influenza in healthy adults. Lancet 2006; 367(9522):1571-1572.

(322) Chronicle E, Mulleners W. Anticonvulsant drugs for migraine prophylaxis. Cochrane Database Syst Rev 2004; (3):CD003226.

(323) Colman I, Brown MD, Innes GD. Parenteral metoclopramide for acute migraine: meta-analysis of randomised controlled trials. BMJ 2004; 329:1369-1373.

(324) de Ferranti SD, Ioannidis JP, Lau J, Anninger WV, Barza M. Are amoxycillin and folate inhibitors as effective as other antibiotics for acute sinusitis? A meta-analysis. BMJ 1998; 317:632-637.

(325) Del Mar CB, Glasziou P, Hayem M. Are antibiotics indicated as initial treatment for children with acute otitis media? A meta-analysis. BMJ 1997; 314:1526-1529.

(326) Furukawa TA, McGuire H, Barbui C. Meta-analysis of effects and side effects of low dosage tricyclic antidepressants in depression: systematic review. BMJ 2002; 325:991-995.

(327) Heidenreich PA, McDonald KM, Hastie T. Meta-analysis of trials comparing [beta]-blockers, calcium antagonists, and nitrates for stable angina. JAMA 1999; 281:1927-1936.

(328) Blood Pressure Lowering Treatment Trialists' Collaboration. Effects of ACE inhibitors, calcium antagonists, and other blood-pressure-lowering drugs: results of prospectively designed overviews of randomised trials. Lancet 2000; 356:1955-1964.

(329) Lindholm LH, Carlberg B, Samuelsson O. Should beta blockers remain first choice in the treatment of primary hypertension? A meta-analysis. Lancet 2005; 366(9496):1545-1553.

(330) Blumenauer B, Judd M, Cranney A. Etanercept for the treatment of rheumatoid arthritis. Cochrane Database Syst Rev 2003; (4):CD004525.

(331) Klareskog L, van der HD, de Jager JP, Gough A, Kalden J, Malaise M et al. Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial. Lancet 2004; 363(9410):675-681.

(332) Boucher M, McAuley L, Brown A, Keely E, Skidmore B. Efficacy of rosiglitazone and pioglitazone compared to other anti-diabetic agents: systematic review and budget impact analysis. CCOHTA Technology Report 2002; 29.

(333) Dormandy JA, Charbonnel B, Eckland DJ, Erdmann E, Massi-Benedetti M, Moules IK et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial. Lancet 2005; 366(9493):1279-1289.

(334) Brown DL, Fann CS, Chang CJ. Meta-analysis of effectiveness and safety of abciximab versus eptifibatide or tirofiban in percutaneous coronary intervention. Am J Cardiol 2001; 87:537-541.

(335) de Queiroz Fernandes Araujo JO, Veloso HH, Braga De Paiva JM, Filho MW, Vincenzo De Paola AA. Efficacy and safety of abciximab on acute myocardial infarction treated with percutaneous coronary interventions: a meta-analysis of randomized, controlled trials. Am Heart J 2004; 148(6):937-943.

(336) Bucher HC, Hengstler P, Schindler C, Guyatt GH. Percutaneous transluminal coronary angioplasty versus medical treatment for non-acute coronary heart disease: meta-analysis of randomised controlled trials. BMJ 2000; 321:73-77.

(337) Katritsis DG, Ioannidis JP. Percutaneous coronary intervention versus conservative therapy in nonacute coronary artery disease: a meta-analysis. Circulation 2005; 111(22):2906-2912.

(338) Crouse JR, III, Byington RP, Hoen HM, Furberg CD. Reductase inhibitor monotherapy and stroke prevention. Arch Intern Med 1997; 157:1305-1310.

(339) Amarenco P, Labreuche J, Lavallee P, Touboul PJ. Statins in stroke prevention and carotid atherosclerosis: systematic review and up-to-date meta-analysis. Stroke 2004; 35(12):2902-2909.

(340) Baigent C, Keech A, Kearney PM, Blackwell L, Buck G, Pollicino C et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet 2005; 366(9493):1267-1278.

(341) Etminan M, Levine MA, Tomlinson G, Rochon PA. Efficacy of angiotensin II receptor antagonists in preventing headache: a systematic overview and meta-analysis. Am J Med 2002; 112:642-646.

(342) Law M, Morris JK, Jordan R, Wald N. Headaches and the treatment of blood pressure: results from a meta-analysis of 94 randomized placebo-controlled trials with 24,000 participants. Circulation 2005; 112(15):2301-2306.

(343) Evans BW, Clark WK, Moore DJ, Whorwell PJ. Tegaserod for the treatment of irritable bowel syndrome. Cochrane Database Syst Rev 2004; (3):CD003960.

(344) Ezekowitz JA, Armstrong PW, McAlister FA. Implantable cardioverter defibrillators in primary and secondary prevention: a systematic review of randomized, controlled trials. Ann Intern Med 2003; 138:445-452.

(345) Maisel WH, Moynahan M, Zuckerman BD, Gross TP, Tovar OH, Tillman DB et al. Pacemaker and ICD generator malfunctions: analysis of Food and Drug Administration annual reports. JAMA 2006; 295(16):1901-1906.

(346) Gotzsche PC, Gjorup I, Bonnen H. Somatostatin v placebo in bleeding oesophageal varices: randomised trial and meta-analysis. BMJ 1995; 310:1495-1498.

(347) Corley DA, Cello JP, Adkisson W, Ko WF, Kerlikowske K. Octreotide for acute esophageal variceal bleeding: a meta-analysis. Gastroenterology 2001; 120(4):946-954.

(348) Lee VC, Rhew DC, Dylan M, Badamgarav E, Braunstein GD, Weingarten SR. Meta-analysis: angiotensin-receptor blockers in chronic heart failure and high-risk acute myocardial infarction. Ann Intern Med 2004; 141(9):693-704.

(349) Keenan SP, Kernerman PD, Cook DJ. Effect of noninvasive positive pressure ventilation on mortality in patients admitted with acute respiratory failure: a meta-analysis. Crit Care Med 1997; 25:1685-1692.

(350) Keenan SP, Sinuff T, Cook DJ, Hill NS. Does noninvasive positive pressure ventilation improve outcome in acute hypoxemic respiratory failure? A systematic review. Crit Care Med 2004; 32(12):2516-2523.

(351) Kjaergard LL, Krogsgaard K, Gluud C. Interferon alfa with or without ribavirin for chronic hepatitis C: systematic review of randomised trials. BMJ 2001; 323:1151-1155.

(352) Fried MW, Shiffman ML, Reddy KR, Smith C, Marinos G, Goncales FL, Jr. et al. Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. N Engl J Med 2002; 347(13):975-982.

(353) Laine L, Schoenfeld P, Fennerty MB. Therapy for Helicobacter pylori in patients with nonulcer dyspepsia: A meta-analysis of randomized, controlled trials. Ann Intern Med 2001; 134:361-369.

(354) Moayyedi P, Soo S, Deeks J, Delaney B, Harris A, Innes M et al. Eradication of Helicobacter pylori for non-ulcer dyspepsia. Cochrane Database Syst Rev 2006;(2):CD002096.

(355) Lord JM, Flight IH, Norman RJ. Metformin in polycystic ovary syndrome: systematic review and meta-analysis. BMJ 2003; 327:951-956.

(356) Moll E, Bossuyt PM, Korevaar JC, Lambalk CB, van d, V. Effect of clomifene citrate plus metformin and clomifene citrate plus placebo on induction of ovulation in women with newly diagnosed polycystic ovary syndrome: randomised double blind clinical trial. BMJ 2006; 332(7556):1485.

(357) Barr RG, Bourbeau J, Camargo CA, Ram FS. Inhaled tiotropium for stable chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2005; (2):CD002876.

(358) Eikelboom JW, Quinlan DJ, Douketis JD. Extended-duration prophylaxis against venous thromboembolism after total hip or knee replacement: a meta-analysis of the randomised trials. Lancet 2001; 358:9-15.

(359) Fink HA, MacDonald R, Rutks IR, Nelson DB, Wilt TJ. Sildenafil for male erectile dysfunction: a systematic review and meta-analysis. Arch Intern Med 2002; 162:1349-1360.

(360) Laine L, Cook D. Endoscopic ligation compared with sclerotherapy for treatment of esophageal variceal bleeding: A meta-analysis. Ann Intern Med 1995; 123:280-287.

(361) Joachims T. Text categorization with Support Vector Machines: Learning with many relevant features. Machine Learning: ECML-98. Berlin: Springer, 1998: 137-142.

(362) Cristianini N, Shawe-Taylor J. Applications of Support Vector Machines. An introduction to Support Vector Machines : and other kernel-based learning methods. Cambridge; New York: Cambridge University Press, 2000: 149-161.

(363) Larsen B. Exploiting citation overlaps for information retrieval: generating a boomerang effect from the network of scientific papers. Scientometrics 2002; 54(2):155-178.

(364) Higgins JPT, Altman DG. Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. www.cochrane-handbook.org: The Cochrane Collaboration, 2008.

(365) Crumley ET, Wiebe N, Cramer K, Klassen TP, Hartling L. Which resources should be used to identify RCT/CCTs for systematic reviews: a systematic review. BMC Med Res Methodol 2005; 5:24.

(366) Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R et al. Should meta-analysts search Embase in addition to Medline? J Clin Epidemiol 2003; 56(10):943-955.

(367) Suarez-Almazor ME, Belseck E, Homik J, Dorgan M, Ramos-Remus C. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. Control Clin Trials 2000; 21(5):476-487.

(368) Cogo E, Sampson M, Ajiferuke I, Manheimer E, Campbell K, Daniel R et al. Searching for controlled trials of complementary and alternative medicine: a comparison of 15 databases. Evidence Based Complementary and Alternative Medicine (in press) 2007.

(369) NHS Centre for Reviews & Dissemination. Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or

commissioning reviews. 4 (2nd Edition)1-91. 2001. York, York Publishing Services Ltd. Report.

(370) Khan KS, Kunz R, Kleijnen J, Antes G Systematic reviews to support evidence-based medicine: How to review and apply findings of healthcare research. London: Royal Society of Medicine Press Ltd., 2003.

(371) Horsley T, Dingwall O, Tetzlaff J, Sampson M. Checking reference lists to find additional studies for systematic reviews. Cochrane Database of Systematic Reviews (Protocol) 2009;(1).

(372) McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet 2000; 356(9237):1228-1231.

(373) van Driel ML, De Sutter A, De Maeseneer J, Christiaens T. Searching for unpublished trials in Cochrane reviews may not be worth the effort. J Clin Epidemiol 2009; In Press, Corrected Proof.

(374) Toma M, McAlister FA, Bialy L, Adams D, Vandermeer B, Armstrong PW. Transition from meeting abstract to full-length journal article for randomized controlled trials. JAMA 2006; 295(11):1281-1287.

(375) Lemeshow AR, Blum RE, Berlin JA, Stoto MA, Colditz GA. Searching one or two databases was insufficient for meta-analysis of observational studies. J Clin Epidemiol 2005; 58(9):867-873.

(376) Savoie I, Helmer D, Green CJ, Kazanjian A. Beyond Medline: reducing bias through extended systematic review search. Int J Technol Assess Health Care 2003; 19(1):168-178.

(377) Morrison A, Moulten K, Clark M, Polisena J, Fiander M, Mierzwinski-Urban M et al. English-Language Restriction When Conducting Systematic Review-based Meta-analyses: Systematic Review of Published Studies. CADTH Technical Report Available at http://www cadth ca/media/pdf/H0478_Language_Restriction_Systematic_Review_Pub_Studies_e pdf 2009.

(378) Information Services - Canadian Agency for Drugs and Technology in Health (CADTH). Grey Matters: a practical tool for evidence-based searching. Available at http://www.cadth.ca/media/pdf/Grey-Matters_A-Practical-Search-Tool-for-Evidence-Based-Medicine.doc. Available at: http://www.cadth.ca/media/pdf/Grey-Matters_A-Practical-Search-Tool-for-Evidence-Based-Medicine.doc. 2009. Accessed: 4-11-2009.

(379) Guyatt G, Sackett D, Taylor DW, Chong J, Roberts R, Pugsley S. Determining optimal therapy--randomized trials in individual patients. N Engl J Med 1986; 314(14):889-892.

(380) Chiviacowsky S, Wulf G. Feedback after good trials enhances learning. Research Quarterly for Exercise & Sport 2007; 78(2):40-47.

(381) Davis DA, Thomson MA, Oxman AD, Haynes RB. Changing physician performance. A systematic review of the effect of continuing medical education strategies. JAMA 1995; 274(9):700-705.

(382) Sampson M, McGowan J, Tetzlaff J, Cogo E, Moher D. No consensus exists on search reporting methods for systematic reviews. J Clin Epidemiol 2008; 61(8):748-754.

(383) Hansen MJ, Rasmussen NO, Chung G. A method of extracting the number of trial participants from abstracts describing randomized controlled trials. J Telemed Telecare 2008; 14(7):354-358.

(384) Laupacis A, Straus S. Systematic reviews: Time to address clinical and policy relevance as well as methodological rigor. Ann Intern Med 2007; 147(4):273-274.

(385) Hay MC, Weisner TS, Subramanian S, Duan N, Niedzinski EJ, Kravitz RL. Harnessing experience: Exploring the gap between evidence-based medicine and clinical practice. J Eval Clin Pract 2008; 14(5):707-713.

(386) Chou R. Using evidence in pain practice: Part I: Assessing quality of systematic reviews and clinical practice guidelines. Pain Med (USA) 2008; 9(5):518-530.

(387) Murphy JR. The relationship of relative risk and positive predictive value in 2 x 2 tables. Am J Epidemiol 1983; 117(1):86-89.

(388) Wilczynski NL, McKibbon KA, Haynes RB. Response to Glanville et al.: How to identify randomized controlled trials in MEDLINE: ten years on. J Med Libr Assoc 2007; 95(2):117-118.

(389) O'Leary N, Tiernan E, Walsh D, Lucey N, Kirkova J, Davis MP. The pitfalls of a systematic MEDLINE review in palliative medicine: symptom assessment instruments. Am J Hosp Palliat Care 2007; 24(3):181-184.

(390) Blikman MJ, Le TM, Bruinse HW, van der Heijden GJ. Ultrasound-predicated versus history-predicated cerclage in women at risk of cervical insufficiency: a systematic review. Obstet Gynecol Surv 2008; 63(12):803-812.

(391) Theoharidou A, Petridis HP, Tzannas K, Garefis P. Abutment screw loosening in single-implant restorations: a systematic review. Int J Oral Maxillofac Implants 2008; 23(4):681-690.

(392) Sergeant G, Penninckx F, Topal B. Quantitative RT-PCR detection of colorectal tumor cells in peripheral blood--a systematic review. J Surg Res 2008; 150(1):144-152.

(393) Alcaraz A, Hammerer P, Tubaro A, Schroder FH, Castro R. Is there evidence of a relationship between benign prostatic hyperplasia and prostate cancer? Findings of a literature review. Eur Urol [Epub ahead of print] 2008.

(394) Ovid SP Online Help - OvidSP Find Similar Articles. http://ovidsp.tx.ovid.com/spa/ovidweb.cgi: Ovid Technologies, 2009.

(395) EBSCO*host* Online Help - Search Modes. http://support.ebsco.com: EBSCO Support Site, 2009.

(396) Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ 2005; 330(7494):753.

(397) SUNY Downstate Medical Center.Medical Research Library of Brooklyn. Guide to Research Methods: The Evidence Pyramid. Available at: Evidence-based Medicine Tutorial (http://library.downstate.edu/EBM2/2100.htm). 2006. Accessed: 3-7-2009.

APPENDIX 1 – ACKNOWLEDGEMENTS

I received outstanding guidance and support in this project. My co-advisors were Christine Urquhart and David Moher – both gave focused and insightful guidance throughout. Kaveh Shojania led the updating project, it was a privilege to work with him and I learned an enormous amount from the experience. I thank Berry de Bruijn, National Research Council, Canada, for his efforts running the Support Vector Machine experiments. Special thanks go to Jessie McGowan for reading and commenting on the entire thesis and Carol Lefebvre and Elise Cogo for chapters. I would like thank my family and friends for their great patience, and I extend a special note of appreciation to Dr. Jan Bormanis, my physician, for keeping me healthy throughout.

I wish to acknowledge and thank the following researchers for sharing their work prior to publication:

- Alan Thomas, Senior Scientific Editor, BMJ Evidence Centre, for providing copies of conference slides and posters presenting results from the Clinical Evidence updates.[121,144,145]
- Aaron Cohen for sharing copies of manuscripts reporting their work on automatic collection and classification of documents.[111,112]
- Kim Peterson for sharing conference posters and presentations for the Drug Effectiveness Review Project (DERP) team's work on updating.[124,125]

I wish to acknowledge and thank the agencies who provided funding in support of the research presented here:

- Agency for Healthcare Research and Quality who provided financial support through Contract No. 290-02-0021 for the University of Ottawa Evidence-based Practice Center (EPC) to perform the updating project[16] from which these data are derived.
- Canadian Agency for Drugs and Technologies in Health who provided financial support for the original systematic review on updating[6] through their Health Technology Assessment Capacity Building Grants Program.

I would like to acknowledge the guidance and expertise contributed to this project by the Technical Expert Panel members for the AHRQ Updating Project:

- David Atkins – Chief Medical Officer, Agency for Healthcare Research and Quality
- Paul Shekelle – RAND (USA)
- Evelyn P. Whitlock – Kaiser Permanente Center for Health Research (USA)
- Cynthia Mulrow – University of Texas, San Antonio and Annals of Internal Medicine (USA)
- Doug Altman – Center for Statistics in Medicine (UK)
- Martin Eccles – Center for Health Services Research, Newcastle University (UK)
- P.J. Devereaux – McMaster University Health Sciences Centre (CAN)

I would like to acknowledge the contributions of members of the University of Ottawa Evidence-based Practice Center and others. Many of these researchers made significant contributions to portions of the updating project:

- Keith O'Rourke and Nicholas J Barrowman, for statistical advice
- Andrea Tricco for guidance with survey methods
- Alexander Tsertsvadze, for early methodological discussions and assistance with screening
- Tanya Horsley, for mentorship of team members
- Jessie McGowan, for guidance with search method development
- Raymond Daniel, for search assistance and document acquisition
- Alla E. Iansavichene, for search assistance and data quality management
- Mary Ocampo, for technical and administrative assistance
- Alison Jenkins, for development of the meta-analytic Excel worksheet

I would like to acknowledge the following for permission to reprint figures:

- Annals of Internal Medicine granted permission to reproduce the flow diagram of our process for the review of new evidence, Figure 12.[17] Request number WAA0916220.
- BMJ granted permission to reproduce Figure 11, a receiver operator curve, from Deeks' paper on systematic reviews of diagnostic and screening tests.[198] License Number 2155020810533.
- SUNY Downstate Medical Centre granted permission to reproduce Figure 1, The Evidence Pyramid, from their website.[397]