



# Systematic review automation methods

Christopher Norman

## ► To cite this version:

Christopher Norman. Systematic review automation methods. Information Retrieval [cs.IR]. Université Paris-Saclay; Universiteit van Amsterdam, 2020. English. NNT : 2020UPASS028 . tel-03060620

**HAL Id: tel-03060620**

**<https://theses.hal.science/tel-03060620>**

Submitted on 14 Dec 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Systematic Review Automation Methods

## Thèse de doctorat de l'université Paris-Saclay

n°580, sciences et technologies de l'information et de la communication  
(STIC)

Spécialité de doctorat: Informatique

Unité de recherche : Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France

Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Amsterdam, le 11 février  
2020, par

**Christopher NORMAN**

### Composition du Jury

**Alexandre ALLAUZEN**

Professeur, Université Paris-Saclay

President

**Patrice BELLOT**

Professeur, Aix-Marseille Université

Rapporteur & Examineur

**Maroeska M. ROVERS**

Professeure, Radboud Universiteit  
Nijmegen

Rapporteur & Examinatrice

**Nicolette F. DE KEIZER**

Professeure, AMC, Universiteit van  
Amsterdam

Examinatrice

**Sandra BRINGAY**

Professeure, Université Paul-Valéry,  
Montpellier 3

Examinatrice

**Aurélie NEVEOL**

Docteure, CNRS, LIMSI, Université Paris-  
Saclay

Directrice de thèse

**Patrick M.M. BOSSUYT**

Professeur, AMC, Universiteit van  
Amsterdam

Co-Directeur de thèse

**Mariska M.G. LEEFLANG**

Docteure, AMC, Universiteit van  
Amsterdam

Co-encadrante de thèse



Recent advances in artificial intelligence have seen limited adoption in systematic reviews, and much of the systematic review process remains manual, time-consuming, and expensive. Authors conducting systematic reviews face issues throughout the systematic review process. It is difficult and time-consuming to search and retrieve, collect data, write manuscripts, and perform statistical analyses. Screening automation has been suggested as a way to reduce the workload, but uptake has been limited due to a number of issues, including licensing, steep learning curves, lack of support, and mismatches to workflow. There is a need to better align current methods to the need of the systematic review community.

Diagnostic test accuracy studies are seldom indexed in an easily retrievable way, and suffer from variable terminology and missing or inconsistently applied database labels. Methodological search queries to identify diagnostic studies therefore tend to have low accuracy, and are discouraged for use in systematic reviews. Consequently, there is a particular need for alternative methods to reduce the workload in systematic reviews of diagnostic test accuracy.

In this thesis we have explored the hypothesis that automation methods can offer an efficient way to make the systematic review process quicker and less expensive, provided we can identify and overcome barriers to their adoption. Automated methods have the opportunity to make the process cheaper as well as more transparent, accountable, and reproducible.

CHRISTOPHER NORMAN      Systematic Review Automation Methods

# Systematic Review Automation Methods

Christopher R. Norman

# SYSTEMATIC REVIEW AUTOMATION METHODS

## ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor  
aan de Universiteit van Amsterdam  
op gezag van de Rector Magnificus  
prof. dr. ir. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,  
in het openbaar te verdedigen in de Agnietenkapel  
op dinsdag 11 februari 2020, te 12.00 uur

door CHRISTOPHER ROBIN NORMAN  
geboren te Norberg

**Promotiecommissie:**

Promotor	prof. dr. P.M.M. Bossuyt dr. A. Névél	Amc-UvA Université Paris-Saclay
Copromotores	dr. M.M.G. Leeftang	Amc-UvA
Overige leden	prof. dr. N.F. de Keizer prof. dr. M. de Rijke prof. dr. M.M. Rovers prof. dr. P. Bellot prof. dr. S. Bringay prof. dr. A. Allauzen	Amc-UvA Universiteit van Amsterdam Radboud Universiteit Nijmegen Aix-Marseille Université Université Paul-Valéry, Montpellier 3 Université Paris-Sud
Faculteit der Geneeskunde		

Dit proefschrift is tot stand gekomen in het kader van het European project 'MiRoR' GA N° 676207, met als doel het behalen van een gezamenlijk doctoraat. Het proefschrift is voorbereid in de Faculteit der Geneeskunde van de Universiteit van Amsterdam en in het Centre National de la Recherche Scientifique (CNRS) van de Université Paris-Saclay.

This thesis has been written within the framework of the European project 'MiRoR' GA N° 676207, with the purpose of obtaining a joint doctorate degree. The thesis was prepared in the Faculty of Medicine at the University of Amsterdam and in the Centre National de la Recherche Scientifique (CNRS) at the Université Paris-Saclay.

*To Ingemar Norman*

*in memoriam*



---

## ACKNOWLEDGMENTS

When I arrived in Paris to start my PhD in 2016, France was in a state of emergency which was to last until November 2017. Civil unrest have since been a constant backdrop to this PhD project, occasionally without, but primarily within France, giving the macabre feel of an Orwellian joke to the travel advisories given by the CNRS when the author went to the ‘at risk country’ Japan for the LREC conference. Those two weeks in Japan may have been the only period of true peace encountered during these three years. The last pages of this thesis were written under military curfew barricaded in a hotel room in Santiago, during the early days of the 2019-20 Chilean protests. The author hopes that the thesis may be read in less interesting times.

This PhD would not have been possible without the help of a large number of people.

First, I want to thank my eminent supervisors, Aurélie Névéol and Mariska Leeftang. Thank you for your infinite patience, your encouragement, your unwavering support, for all the advice I have received, and for all the discussions we have had. I have grown – both academically and as a person – over the last three years, all thanks to you.

I would like to thank Patrick Bossuyt for being my promotor, and for welcoming me into the BIRE group at AMC.

I am very honored to have Nicolette de Keizer, Maarten de Rijke, Maroeska Rovers, Patrice Bellot, Sandra Bringay, and Alexandre Allauzen serving on my committee. Thank you very much for taking the time to read and approve this lengthy thesis.

I would like to thank René Spijker, who hosted me during my secondment at Cochrane Netherlands. Your knowledge and keen interest in data science methods have been an incommensurate help to me.

I would like to thank Evangelos Kanoulas, who welcomed me to the ILPS group at the University of Amsterdam. Your advice during our many discussions have been invaluable.

I would like to thank all the MiRoR PhD students, who have been the true rays of light in the MiRoR project: Thank you Linda Nyanchoka, Maria Olsen, Alice Biggane, Efstathia Gkioni, Van Thu Nguyen, Camila Olarte Parra, Lorenzo Bertizzolo, Thang Vo Tat, Melissa Sharp, Mona Ghannad, Anna Koroleva, Kete-van Glonti, David Blanco, and Cecilia Superchi. Thank you for being amazing colleagues, and for being awesome in general. It is beyond my ability to express how much I have enjoyed working with you these past three years.

I would like to thank all past and current members of the ILES group at LIMS1 in Paris. Thank you Julien Tourille, Sanjay Kamath, Swen Ribeiro, Arnaud Ferre, Zheng Zhang, Hicham el-Boukkouri, Arthur Boyer, Timothy Miller, Pierre Zweigenbaum, Rashedur Rahman, Pierre Magistry, Leonardo Campillos, Eva D’hondt, Charlotte Rudnik, Catherine Thomas, Yuming Zhai, and Patrick

---

Paroubek. I would like to extend a particular thanks to Arthur, who helped me when I had difficulty finding accommodation in Orsay after returning from Amsterdam, and putting up with me as a flatmate for almost a year. Also, thank you for introducing many kinds of productivity-boosting activities to LIMSI.

Likewise, I would like thank all past and current members of the BITE group at KEBB, as well as other colleagues in the department, who made my stays at the AMC a wonderful experience. Thank you Attila Csala, Bada Yang, Jenny Lee, Allard van Altena, Amber Boots, Marileen Wiegersma, Sandeep Singh, Hugo van Mens, and again Mona Ghannad and Maria Olsen.

I would like to thank Paula Williamson and Elizabeth Gargon at the University of Liverpool for giving me the chance to work on applying screening automation in real-life systematic reviews, and for the chance to collaborate. Writing manuscripts with you has been a pleasure.

I would also like to thank Darko Hren at the University of Split for very enjoyable collaborations, and for being a great guy in general. Thank you for the chance to let me see how data science and NLP methods can be of help in other domains.

One of the included papers would not have been possible without the help of Raphaël Porcher. Thank you for all the help with the statistics, and for improving the statistical rigour of the paper, and for making it easier to read. I would also like to thank the administrative staff at LIMSI, in particular Laurence Rostaing and Blanche Gonzalez, for your never-ending patience with my travel plans and expense claims, as well as for always putting up with my still stunted French. Thank you Isabelle Boutron, Francois Yvon, Anne Vilnat, and Laura de Nale for your prompt responses to the never-ending issues encountered through the years. Thank you AMIC, for procuring and setting up computers for me and for your prompt administrative support.

Finally, I would like to acknowledge my family. Thank you, mom for your unconditional love and support. I thank my sister, Magdalena, and my brother-in-law, Dan, my brother, Mikael, and my sister-in-law, Maria for always being there when I need it.

Christopher R. Norman  
Bjurfors Manor, Epiphany AD 2020

# CONTENTS

---

1	Introduction	1
1.1	Objectives	2
1.2	Research Questions	3
1.3	Contributions	4
1.4	Outline	5
1.4.1	Background & Context	5
1.4.2	Screening Automation Systems	6
1.4.3	The Impact of Screening Automation	6
1.4.4	Data Extraction & Synthesis	7
1.4.5	Quick Guide to the Thesis Contents	7
I	BACKGROUND & CONTEXT	9
2	What is a Systematic Review?	11
2.1	Evidence Based Medicine	12
2.1.1	The Issues with Opinion	13
2.1.2	The Need for Evidence in Medicine	14
2.1.3	Hierarchies of Evidence	15
2.1.4	Systematic Reviews & Evidence Based Medicine	16
2.1.5	Systematic Reviews Answer Research Questions	17
2.2	Systematic Reviews of Diagnostic Test Accuracy	18
2.3	Issues in Systematic Review Production	20
2.3.1	The Publication of Systematic Reviews is Growing	20
2.3.2	The Workload in Systematic Reviews is Growing	21
2.3.3	Systematic Reviews Take a Long Time to Complete	22
3	The Systematic Review Process	25
3.1	Writing the Protocol	28
3.1.1	Formulate Review Question	29
3.1.2	Determine Eligibility Criteria	29
3.1.3	Determine Target Databases	30
3.1.4	Construct Search Queries	31
3.2	Searching and Screening	32
3.2.1	Run Search Queries	32

---

3.2.2	Deduplicate	33
3.2.3	Screen Abstracts	33
3.2.4	Retrieve Full Text	34
3.2.5	Screen Full Text	35
3.2.6	Search Reference Lists	36
3.3	Writing the Systematic Review	37
3.3.1	Extract Data	37
3.3.2	Homogenize Data	38
3.3.3	Synthesize Data	40
3.3.4	Re-check Literature	41
3.3.5	Meta-Analyze	41
3.3.6	Write Review	42
4	Systematic Review Automation	43
4.1	Text Mining Methods	44
4.1.1	Feature Representations	46
4.2	An overview of Screening Automation Methods	47
4.2.1	Database Searches	47
4.2.2	Inclusion Criteria	48
4.2.3	Results	49
4.3	Summary of Screening Automation Approaches	49
4.4	Metrics Used in Previous Literature	52
4.5	Available datasets	54
4.6	List of Individual Studies	56
II	SCREENING AUTOMATION SYSTEMS	79
5	Introduction	80
5.1	Screening Automation Methods	80
5.2	Designing a Screening Automation Model	81
5.3	Using training data from different stages of screening	84
5.4	Evaluation of Performance	85
5.5	Objective	86

## 6 How to Use Training Data Effectively 89

Published as: **(Norman et al., 2018c)**: Norman, C., Leeflang, M., Zweigenbaum, P., and Névéol, A. (2018c). Automating document discovery in the systematic review process: How to use chaff to extract wheat. In Calzolari, N. et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA)

6.1	Introduction	89
6.1.1	Related Work	91
6.1.2	Objective	91
6.2	Datasets	92
6.2.1	The Yearbook Dataset	94
6.2.2	The Cohen Dataset	94
6.3	Document Ranking Method	94
6.3.1	Experimental Setup	95
6.4	Results	97
6.5	Discussion	97
6.5.1	Performance of Our System	97
6.5.2	Can We Separate the Screening into Two Stages?	97
6.5.3	Do We Need Examples from All Stages of Screening ( $\gamma$ , $m$ , $n$ )?	98
6.5.4	Can We Use $m$ as Positive Examples for Training?	98
6.5.5	Strategies for Ranking Articles	99
6.5.6	Limitations	100
6.5.7	Future work	100
6.6	Conclusion	100

## 7 Ranking Performance for DTA Reviews (2017) 103

Published as: **(Norman et al., 2017b)**: Norman, C., Leeflang, M., and Névéol, A. (2017b). LIMS@CLEF eHealth 2017 task 2: Logistic regression for automatic article ranking. *Working Notes of CLEF*

7.1	Introduction	103
7.2	Datasets	104
7.3	Method	106
7.3.1	Overview	106
7.3.2	Classification Approach	107
7.3.3	Features	107
7.3.4	Class Imbalance	107
7.3.5	Relevance Feedback	109
7.3.6	Use of the CLEF Development Dataset	113

---

7.4	Results	113
7.5	Discussion	114
7.5.1	Datasets	114
7.5.2	Performance	114
7.5.3	Metrics	117
7.5.4	Reliability of the Experiments	117
7.5.5	General Remarks on the Shared Task Model	118
7.6	Conclusions	119
8	Ranking Performance for DTA Reviews (2018)	121
	Published as: <b>(Norman et al., 2018b)</b> : Norman, C., Leeflang, M., and Névél, A. (2018b). LIMS@CLEF eHealth 2018 task 2: Technology assisted reviews by stacking active and static learning. <i>Working Notes of CLEF</i> , pages 10–14	
8.1	Introduction	121
8.2	Material	122
8.3	Methods	122
8.3.1	Overview	124
8.3.2	Static Ranking Model	124
8.3.3	Active Learning	125
8.3.4	Stacked Model	126
8.4	Results	129
8.5	Discussion	129
8.5.1	Datasets	129
8.5.2	Performance	131
8.6	Conclusions	132
9	Discussion & Conclusions	133
9.1	Summary	133
9.1.1	Differences Between Studies Included by Abstract and Full-Text	133
9.1.2	Using Training Data from Both Screening Stages	134
9.1.3	Screening Approaches	134
9.2	Screening Performance	136
9.2.1	Static Model (Intratopic)	136
9.2.2	Static Model (Intertopic)	137
9.2.3	Active Learning	138
9.2.4	Stacked Model	139
9.3	Conclusions	139

III THE IMPACT OF SCREENING AUTOMATION	141
10 Introduction	142
10.1 How Can We Measure Review Integrity?	142
10.2 Measuring Integrity Prospectively	144
10.3 Reducing Bias	144
10.4 Determining Safe Screening Thresholds	146
10.5 Objectives	146
11 Prospective Use of Screening Automation	149
<p>Published as: (<b>Norman et al., 2019f</b>): Norman, C. R., Gargon, E., Leeftang, M. M. G., Névél, A., and Williamson, P. R. (2019f). Evaluation of an automatic article selection method for timelier updates of the COMET core outcome set database. <i>Database</i>. Oxford University Press</p>	
11.1 Introduction	149
11.1.1 Screening Automation in Systematic Reviews	150
11.2 Material and Methods	151
11.2.1 Data Preprocessing	153
11.2.2 Document Ranking Method	153
11.3 Implementation	157
11.3.1 Evaluation	158
11.4 Results	158
11.5 Discussion	159
11.6 Conclusions	160

---

12 When is it Safe to Stop Screening? 163

Published as: **(Norman et al., 2019c)**: Norman, C., Leeflang, M., Porcher, R., and Névelo, A. (2019c). Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. *BMC Systematic Reviews*. Springer Nature

12.1	Background	164
12.1.1	Meta-analyses of Diagnostic Test Accuracy	165
12.1.2	Systematic Reviews Require Perfect Recall	166
12.1.3	The Impact of Rapid Reviews on Meta-Analysis Accuracy	167
12.1.4	Related Methods for Screening Prioritization	168
12.1.5	Objectives	169
12.2	Methods	170
12.2.1	Data Used in the Study	170
12.2.2	Automated Screening Method	171
12.2.3	Evolution of a Summary Estimate	173
12.2.4	Finding a Balance Between Loss and Effort	174
12.2.5	Calculation of Summary Statistics	177
12.3	Results	177
12.3.1	Characteristics of the Systematic Reviews	177
12.3.2	How Many Studies Does it Take to Make a Meta-analysis?	179
12.3.3	Contribution of Screening Prioritization	179
12.4	Discussion	182
12.4.1	Estimates Converge Faster Using Screening Prioritization	182
12.4.2	Sufficiently Large Meta-analyses Can be Stopped Prematurely	183
12.4.3	Data Saturation is Seldom Reached in DTA Systematic Reviews	184
12.4.4	External validity	186
12.4.5	Recommendations	186
12.4.6	Limitations of this Study	188
12.4.7	Future Work	188
12.5	Conclusions	189
13	Discussion & Conclusions	191
13.1	Screening Reduction Compatible with Current Practice	191
13.2	Better Metrics for Screening Automation	192
13.3	Conclusions	194

IV DATA EXTRACTION & SYNTHESIS	195
14 Introduction	196
14.1 Automated Data Extraction	197
14.2 Automated Data Homogenization	200
14.3 Automated Data Synthesis	201
14.4 Automated Meta-Analysis	201
14.5 Objectives	202
15 Data Extraction and Synthesis in DTA Systematic Reviews	205
Published as: <b>(Norman et al., 2018a)</b> : Norman, C., Leeflang, M., and Névéol, A. (2018a). Data extraction and synthesis in systematic reviews of diagnostic test accuracy: A corpus for automating and evaluating the process. In <i>AMIA Annual Symposium Proceedings</i> , volume 2018, page 817. American Medical Informatics Association	
15.1 Introduction	205
15.1.1 Automated Data Extraction and Synthesis	206
15.1.2 Systematic Review Reproducibility	207
15.1.3 Related Work	207
15.2 Objectives	208
15.3 Material and Methods	209
15.3.1 The Cochrane DTA Data Form Corpus	209
15.3.2 Summary Score Replication	212
15.4 Results	212
15.4.1 Dataset Construction	212
15.4.2 Summary Score Replication	213
15.5 Discussion	216
15.5.1 Dataset Construction	216
15.5.2 Summary Score Replication	216
15.5.3 Dataset Applications and Future Work	217
15.6 Conclusion	218

---

16 Automated data extraction for systematic reviews of diagnostic accuracy 221

Published as: **(Norman et al., 2019e)**: Norman, C., Spijker, R., Kanoulas, E., Leeflang, M., and Névéol, A. (2019e). A distantly supervised dataset for automated data extraction from diagnostic studies. *ACL BioNLP*)

16.1	Background	221
16.1.1	BERT	222
16.1.2	Objectives	223
16.1.3	Related Work	224
16.2	Material	226
16.3	Method	228
16.3.1	Evaluation	232
16.4	Results	232
16.5	Discussion	233
16.5.1	Limitations	233
16.6	Conclusions	234
16.6.1	Future Work	234
17	Discussion & Conclusions	235
17.1	Automated Data Extraction from DTA Studies	235
17.2	Automated Meta-Analyses for DTA systematic reviews	238
17.3	Conclusions	241

V SUMMARY	243
18 English Summary	244
18.1 Screening Automation	244
18.1.1 Differences Between Studies Included by Abstract and Full-Text	245
18.1.2 Using Training Data from Both Screening Stages	245
18.1.3 Screening Approaches	245
18.2 Screening Performance	246
18.2.1 Static Model (Intratopic)	246
18.2.2 Static Model (Intertopic)	246
18.2.3 Active Learning	247
18.2.4 Stacked Model	247
18.3 Screening Reduction Compatible with Current Practice	247
18.3.1 Better Metrics for Screening Automation	248
18.4 Data Extraction & Synthesis	249
18.4.1 Automated Data Extraction from diagnostic test accuracy Studies	249
18.4.2 Automated Meta-Analyses for diagnostic test accuracy systematic reviews	250
18.5 Conclusions	250

19	Nederlandse Samenvatting	252
19.1	Automatisering van screening	253
19.1.1	Verschillen tussen referenties geselecteerd op basis van de samenvatting en op basis van de volledige tekst	253
19.1.2	Gebruik van trainingsgegevens uit beide screeningsfasen	253
19.1.3	Screening benaderingen	254
19.2	Resultaten	255
19.2.1	Statisch model (intrathematisch)	255
19.2.2	Statisch model (interthematisch)	255
19.2.3	Actief leren	255
19.2.4	Stackingmodel	255
19.3	Screeningreductie compatibel met de huidige praktijk	256
19.3.1	Betere metrieken voor screeningautomatisering	257
19.4	Gegevensextractie & synthese	258
19.4.1	Geautomatiseerde gegevensextractie van diagnostische onderzoeken	258
19.4.2	Geautomatiseerde meta-analyses voor systematische reviews van diagnostische accuratesse	259
19.5	Conclusies	260
20	Résumé Français	262
20.1	Automatisation de la sélection d'articles	263
20.1.1	Différences entre les articles inclus d'après le résumé et le texte intégral	263
20.1.2	Utilisation des données d'entraînement des deux étapes de sélection	263
20.1.3	Approches de sélection d'articles	264
20.2	Performance de la sélection automatique	265
20.2.1	Modèle statique (intra-thématique)	265
20.2.2	Modèle statique (inter-thématique)	265
20.2.3	Apprentissage actif	265
20.2.4	Modèle de Stacking	265
20.3	Réduction de la charge de travail associée à la sélection d'articles compatible avec la pratique	266
20.3.1	Meilleures mesures pour l'automatisation de la sélection	267
20.4	Extraction & synthèse de données	268
20.4.1	Extraction de données des études portant sur l'exactitude des tests diagnostiques	268
20.4.2	Méta-analyses automatisées pour les revues systématiques portant sur l'exactitude des tests diagnostiques.	269

20·5	Conclusions	270
VI	APPENDICES	273
A	Publication List	275
A·1	Peer Reviewed Publications	275
A·1·1	Letters to Editor	275
A·1·2	Journal Publications	275
A·1·3	Shared Task Papers	275
A·1·4	Conference Papers	276
A·2	Other Publications	276
A·2·1	Peer Reviewed Conference Abstracts	276
A·2·2	Other Conference Abstracts	276
A·2·3	System Demonstrations	277
	Bibliography	279



## 1

## INTRODUCTION



AT THE TIME OF WRITING, the Epistemonikos database lists 290,604 systematic reviews, and PubMed alone is indexing over 17,000 new systematic reviews every year (section 2.3.1). While the number of annually produced systematic reviews is staggering, the cost of producing these is also growing, and the cost to produce a single systematic review today may reach as high as \$300,000 USD (Lau, 2019). Furthermore, methodologically rigorous systematic reviews typically take years to complete. Systematic reviews are thus difficult to complete in response to urgent policy needs – such as the recent Ebola outbreak (Schünemann and Moja, 2015) – and may no longer be up-to-date by the time they are completed (Shojania et al., 2007). Consequently, there is a need to evaluate alternative methods to cope with the cost, workload, and delay between inception and completion of the review.

Artificial intelligence has seen huge advances over the last few decades, and many tasks which once required human intelligence can today be automated. However, these advances have so far seen little to no adoption in systematic reviews (Thomas, 2013), and much of the systematic review process remains manual, time-consuming, and expensive. Authors conducting systematic reviews face issues throughout the systematic review process. It is difficult and time-consuming to search and retrieve, collect data, write manuscripts, and perform statistical analyses (Allen and Olkin, 1999; Pham et al., 2018). Dozens of studies have been published since 2006 arguing that screening automation can reduce the workload (O'Mara-Eves et al., 2015), but due to issues including licensing, steep learning curves, lack of support, and mismatches to workflow (Van Altena et al., 2019), uptake by the systematic review community has been limited (Thomas, 2013). There is a need to better align current methods to the need of the systematic review community.

Diagnostic tests are any kind of procedures performed to assist clinicians with the diagnosis of specific health conditions. Diagnostic tests can be invasive (e.g. amniocentesis), minimally invasive (e.g. blood test) or non-invasive (e.g. urine analysis). It is crucial to weigh the benefits of more accurate tests against the financial and psychological burden associated with specific tests and their resulting follow-up. However, accurate information on the utility and accuracy of diagnostic tests are commonly buried in free text articles. Single diagnostic test accuracy studies seldom definitely resolve their utility (Davidoff et al., 1995a, quoted by Cook et al., 1997), and systematic reviews are typically necessary to combine the results from multiple studies.

One of the main challenges for identifying diagnostic test accuracy studies is that such studies are often not indexed in any easily retrievable way, with variable terminology and missing or inconsistently applied database labels. Methodological search queries to identify diagnostic studies therefore tend to have low accuracy (Beynon et al., 2013; Leeflang et al., 2008, 2006), and are discouraged for use in systematic reviews (De Vet et al., 2008, in Deeks et al., 2013a). Consequently, a typical search strategy for diagnostic test accuracy will retrieve around 5,000 initial hits, of which a couple of hundred will have to be read as full-text and only around 10 to 20 will be included in the review.

In this thesis we have explored the hypothesis that automation methods can offer an efficient way to make the systematic review process quicker and less expensive, provided we can identify and overcome barriers to their adoption. At the same time, there is also a need to document and monitor the information trail of diagnostic tests through the evidence synthesis workflow, on the one hand to inform clinicians and patients, but also to gather data to train automated methods. Automated methods have the opportunity to make the process cheaper as well as more transparent, accountable, and reproducible.

## 1.1 OBJECTIVES

As outlined in the project description,<sup>1</sup> written before the start of this project, this project aimed to:

1. Develop natural language processing (NLP) techniques to:
  - a) Automatically identify diagnostic test accuracy publications among candidate references; and
  - b) Automatically determine study characteristics necessary to perform diagnostic test accuracy systematic reviews from article text
2. Implement recommendation systems that will retrieve diagnostic test studies for inclusion in systematic reviews, as well as identify specific study characteristics
3. Populate a knowledge base with comprehensive information about diagnostic tests which will inform researchers and clinicians.
4. Update automatic predictions over time using the supervised data obtained through interactive annotation

---

<sup>1</sup> <http://miror-ejd.eu/individual-research-projects/>

## 1.2 RESEARCH QUESTIONS

When this project was started, there were no datasets available to train screening automation for DTA systematic reviews. Furthermore, there were unresolved questions regarding how such datasets should be constructed and used. Should gold standard decisions be based on inclusion decisions based on abstract and title, or only on inclusion decisions based on full-text? Do we need training data from the same topic, or can we use similar topics for training? How much data do we need? How do screening automation methods cope in a review where there are very small numbers of included studies?

Thus, before implementing machine learning methods we must first address the following question:

RQ 1 *What kind of data should we use to train screening automation methods?*

A number of machine learning algorithms, approaches, and parameter settings exist for screening automation, and it is not clear which would work best for systematic reviews of diagnostic test accuracy reviews. Thus we also address:

RQ 2 *How do different screening automation approaches compare with each other for DTA screening?*

One of the main aims of this project is to develop functional tools that can be used in live systematic reviews. These should be as performant as possible. Thus:

RQ 3 *Are the screening automation methods we develop competitive with the current state-of-the-art?*

At the same time, it is also important that the way we use these methods do not invalidate the findings of the systematic review. An automated systematic review should be just as rigorous as a review using conventional methods. The straightforward way to accomplish this is to adhere as closely as possible to the accepted practice:

RQ 4 *Can we use screening automation in a live systematic review while keeping the same rigorous methodology?*

However, it has previously simply been assumed that any divergence from accepted practice will result in an inherently flawed process. It has never been investigated whether, to what extent, and under what conditions this is true:

- 
- RQ 5 *What are the minimum conditions for a systematic review to guarantee the same results and conclusions using screening prioritization as with the conventional process?*

In particular, we will seek to determine if there are any elements of current practice that can be safely discarded.

There is also a need to more fully understand the current, manual process, including its strengths and shortcomings. Previous work to this end has primarily focused on screening methodology, and less on the later stages in the process:

- 4 RQ 6 *How are the current data extraction, data synthesis, and meta-analysis stages of DTA systematic reviews performed by human authors?*

And lastly, it would also aid if we could automate not just the screening process, but also other resource-intensive stages of the process. In particular:

- RQ 7 *Can we extract important study characteristics automatically from primary DTA studies?*

### 1.3 CONTRIBUTIONS

In this project, we have introduced

- ❖ *A screening prioritization and screening reduction system.* The system is empirically validated to perform well across a range of systematic review approaches and topics, including systematic reviews of diagnostic test accuracy.
- ❖ *An empirical prospective validation of screening reduction,* where we evaluate the use of the screening reduction to facilitate the COMET Core Outcome Set systematic review update in 2019.
- ❖ *A retrospective validation of the impact of screening reduction in diagnostic test accuracy systematic reviews.* To our knowledge, this is the first attempt to investigate whether screening reduction methods leads to a loss of accuracy of the results and conclusions of the review. We also evaluate the impact of the commonly used 95% recall requirement in systematic reviews. This has to our knowledge never been done.
- ❖ *A dataset describing the data flow from individual studies to individual meta-analyses.*
- ❖ *An empirical study of the current meta-analysis process,* where we investigate the current process as performed by human reviewers to investigate potential shortcomings in the process.

- ❖ *A pilot study to extract study characteristics from DTA primary studies.* We here attempt to extract the *target condition*, *reference standard*, and *index test*, which have not been attempted by previous literature.

## 1.4 OUTLINE

This thesis is multi-disciplinary, and concerns a very specific cross-section of natural language processing, machine learning, information retrieval, meta-research, life science, and evidence based medicine. Reading the entire thesis from cover to cover therefore only makes sense for readers whose interests align precisely with the scope of this thesis. Most readers are better served by choosing a selection – and order – of chapters to suit their interests. This section will attempt to guide readers in this endeavor.

The thesis is divided into five parts, starting with the background and context, and ending with a summary of the entire thesis work in English, Dutch, and French. Parts II–IV lists the seven studies included in this thesis. Each part is preceded by an introduction that sets the context of the studies in the section, and is succeeded by a summary of the main results of the studies. Both the introductions and summaries have been written to be standalone, as well as easier to read by non-experts than the individual studies. Readers who want to have an overview of the studies may therefore want to read these before – or instead of – reading the actual material.

### 1.4.1 *Background & Context*

In part I we will briefly go through the background necessary to understand the remainder of the thesis, as well as set the context in which the thesis work was performed. This thesis is multidisciplinary, and is intended for audiences ranging from computer science to the life sciences. The background is therefore intended to be exhaustive, and cover any background material that may not be known to some part of the audience. Conversely, it will necessarily cover some amount of material familiar to most readers.

Part I is in turn broken up into three chapters.

In chapter 2 we will introduce systematic reviews. In section 2.1 will introduce the background of evidence based medicine and the role systematic reviews play in current evidence based policy. We will then in section 2.3 introduce the issues faced in systematic review production that have prompted the work in this thesis. In chapter 3 we will introduce the systematic review process. Readers unfamiliar with systematic reviews are invited to read through this chapter, and make note of the relevant background material. The intricacies of the review process will shape

the later work of this thesis, and particular the design concerns necessary to make screening automation viable in practice. In particular, readers who are uncertain about the differences between literature reviews and systematic reviews are invited to read this chapter in detail.

In chapter 4 we will give an overview of previous research on screening automation. We will first (section 4.1) give a brief overview of relevant text mining methods. In section 4.3 we will summarize the relevant literature and the types of approaches and methods that have been addressed by previous researchers. In section 4.4 we will summarize the metrics and evaluation methods that have been used by previous literature. In section 4.5 we will summarize the publicly available datasets that have previously been used by multiple authors. In the final section (4.6), we will list the identified studies, with a brief summary of each study.

#### 1.4.2 *Screening Automation Systems*

In part II, we will examine how screening automation methods can be used to reduce the workload in systematic reviews. We will look at how these methods can be made to work, and how the performance of these methods compare with each other. All performance comparisons will concern intrinsic performance. In other words, we here seek to evaluate the performance of the component models in reproducible laboratory settings. We will examine the extrinsic performance of the methods – i.e. how the methods influence the systematic review process – in part III. We will however start thinking about how different approaches fit into different systematic review contexts and settings.

These inquiries and the evaluations will be predominantly technical. Readers not interested in the technical details of the screening methods can safely skip to the summary.

#### 1.4.3 *The Impact of Screening Automation*

In part III we seek to put the performance of the screening automation into perspective. Instead of measuring numbers, we will attempt to ask how screening automation impacts the systematic review. The main purpose of this part of the thesis is to establish criteria to automatically exclude records with screening reduction methods, while still resulting in the ‘same’ systematic review. In order to do this however, we also need to formalize what it means for two reviews to be the ‘same.’ A systematic review that uses screening automation should not be methodologically inferior to one conducted according to the established systematic review process. The systematic review should remain reproducible, transparent, and free of bias.

The first study in this part (chapter 11) documents the screening automation used in the 2019 update of the COMET Core Outcome Set database (Gargon et al., 2019). We here attempted to adhere as closely as possible to the conventional process, down to screening in randomized order in EndNote. Since the process is fundamentally unaltered we argue that is as unbiased as the conventional process.

In the second paper in this part (chapter 12) we replaced the conventional process with one using screening prioritization, and demonstrated that this results in negligible changes to the results and conclusions of the meta-analyses.

#### 1.4.4 *Data Extraction & Synthesis*

7

In part IV we will look at how software can be used in the later stages of the systematic review process.

Previous work on systematic review automation in DTA systematic reviews have focused exclusively on screening automation (Kanoulas et al., 2017b, 2018). Datasets are thus available for training screening automation methods, but no such datasets are available describing any other review stage. One of the purposes of this part of the thesis is to partially fill these gaps.

In chapter 15 we therefore present a dataset documenting the data extraction, data synthesis, and meta-analysis stages of systematic reviews of diagnostic test accuracy. This dataset is to our knowledge the first of its kind. We hope it will be of aid for better understanding how the process is undertaken by human reviewers, as well as for modelling the process with automated methods.

We will use this in order to perform cumulative meta-analyses to measure the impact screening automation methods have on the results of the systematic review (see chapter 12).

This dataset was also used for the data extraction work performed in chapter 16, where we explored methods to extract data automatically from DTA reports. Almost all relevant extracted data elements in diagnostic studies have been overlooked by previous work on automated data extraction. We identified the target condition, index test and reference standard as the primary data elements to extract, since these were felt to be those that would lead to the largest work savings.

#### 1.4.5 *Quick Guide to the Thesis Contents*

- ☞ Readers interested in an overview of why systematic reviews are important are invited to read chapter 2
- ☞ Readers interested in the current practice for performing systematic reviews are invited to read chapter 3

- 
- ❖ Readers interested in an overview of text mining methods are invited to read section 4.1
  - ❖ Readers interested in an overview of previous screening automation methods are invited to read sections 4.3–4.5
  - ❖ Researchers interested in the technical aspects of screening automation are invited to read chapter 4 and part II. Chapter 4 furthermore serves as reference material for the concepts encountered in part II. A brief familiarity with chapter 3 will be expected.
  - ❖ Systematic reviewers interested in adopting screening automation methods are invited to read the introduction and summary of part II, and the entire part III. The reader will be expected to be familiar with the contents of chapter 2 and 3
  - ❖ Researchers with an interest in how screening automation impacts systematic reviews are invited to read part III and particularly chapter 12. A brief familiarity with the subject of chapter 2 and 3 will be expected
  - ❖ Researchers with an interest in how the data extraction, data synthesis, and meta-analyses are performed in current Cochrane DTA systematic reviews are invited to read chapter 15 and the summary of part IV. Background material is provided in section 3.3.1–3.3.5.
  - ❖ Researchers with an interest in data extraction methods for systematic reviews are invited to read the introduction and summary of part IV, as well as chapter 16. Additional background material is also available in section 3.3.1.

# PART I

## BACKGROUND & CONTEXT



## WHAT IS A SYSTEMATIC REVIEW?



ONE EARLY SYSTEMATIC REVIEW has been made iconic by being enshrined in the Cochrane logotype. The graphics in the logo depict the combined results of eight<sup>1</sup> randomized control trials (RCTs) of the use of corticosteroids, a simple, inexpensive, and very effective treatment improving lung development in fetuses. Administered to women about to give preterm birth, the treatment substantially improves the chances of survival for the newborn, and reduces complications. The first such study was published in 1972, but it would take until 1990 – and the publication of a literature review combining the evidence – for the effectiveness of the treatment to become widely known to obstetricians. Before then, corticosteroids were not widely administered for preterm birth. By 1996, six years after the publication of the review, the use of antenatal corticosteroids in relevant cases had risen from 20% to 65%, with a substantial decrease in infant mortality.

Why was the evidence overlooked? Not due to lack of dissemination – the studies were in fact known to the obstetrics community. However, as can be seen in the logo itself (Figure 2) the confidence intervals of all but two of the studies cross the line of no effect. In other words, the results were – considered in isolation – not significant. The results are significant when combined over all eight studies, denoted by the diamond below the bars, which sits well apart from the vertical line (Figure 2).

This literature review has subsequently been expanded and improved in methodological quality. The most recent version, published in 2017, includes 22 randomized control trials measuring the relative risk of neonatal mortality. These subsequent results are confirming the effectiveness of the treatment, but fail to reach significance individually. Out of 22 randomized control trials of neonatal mortality, only 4 are statistically significant by themselves (figure 2). Furthermore, the trials report quite heterogeneous results, with a single study reporting an almost five-fold average increase in mortality after corticosteroid treatment.

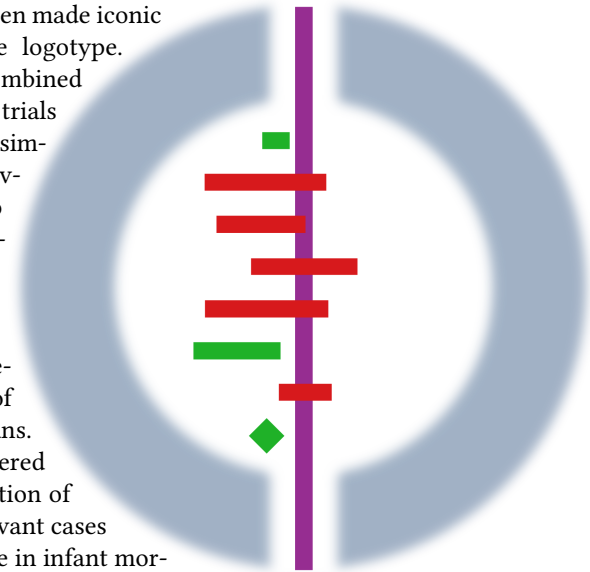
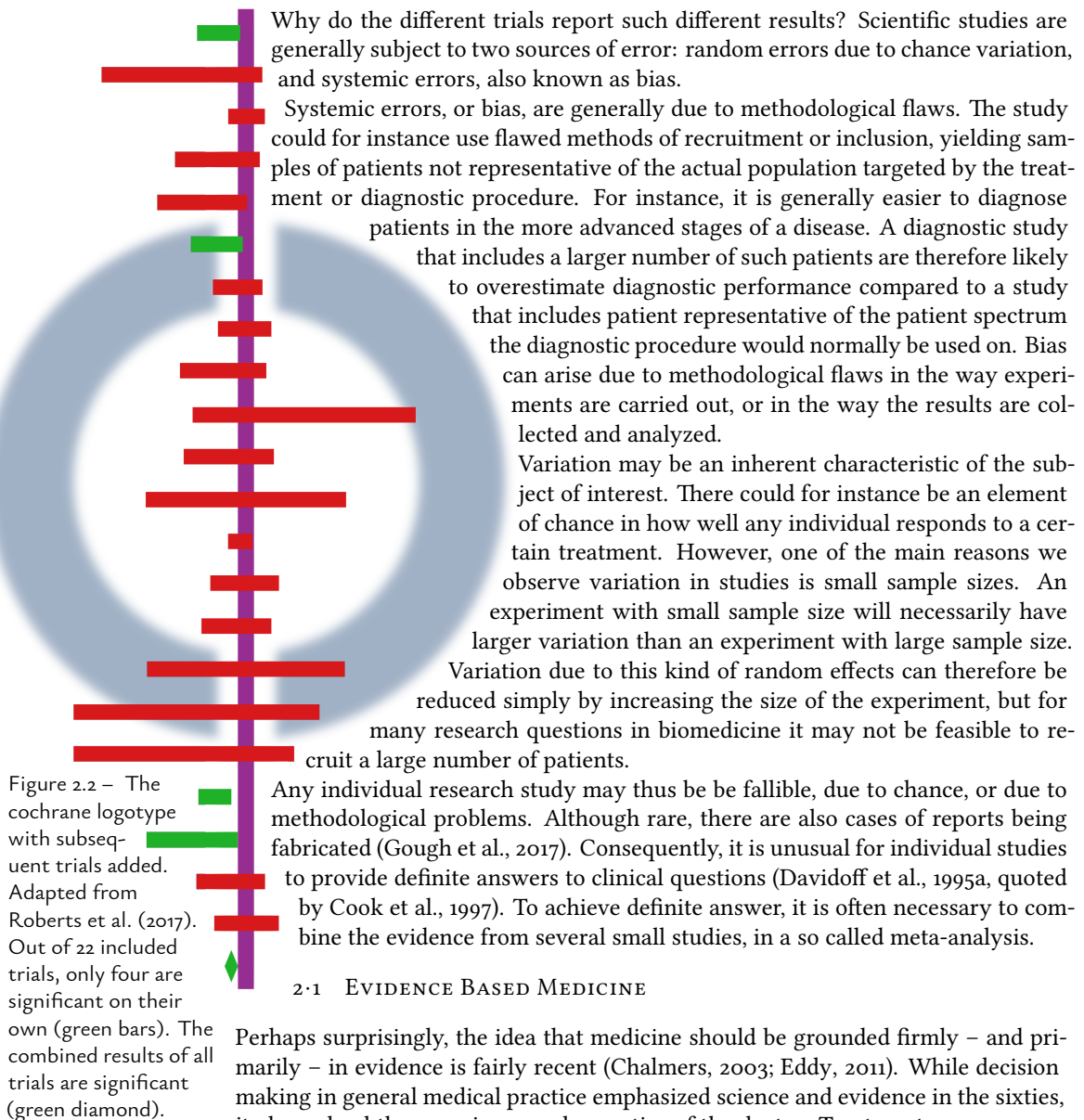


Figure 2.1 – The Cochrane logotype. The logo depicts seven studies examining corticosteroids administered before preterm birth. Five of the studies (red bars) are non-significant on their own, two (green bars) are significant. The combined results of all studies are significant (green diamond).



### 2.1 EVIDENCE BASED MEDICINE

Perhaps surprisingly, the idea that medicine should be grounded firmly – and primarily – in evidence is fairly recent (Chalmers, 2003; Eddy, 2011). While decision making in general medical practice emphasized science and evidence in the sixties, it also valued the experience and expertise of the doctor. Treatments were recommended if physicians believed them to be effective, rather than based on evidence, and healthcare decisions were frequently made based on simple rules-of-thumb – under the guise of accepted practice – without scientific backing (Eddy, 2011). These empirical observations may be explained by the complexity of medical decision making – medical decision making is too complex and involves too many variables for anyone to accurately process all the information in their heads for a complex medical decision (Eddy, 2005).

2.1.1 *The Issues with Opinion*

By the late sixties, the effect of biases on experts' decision making process became more apparent with the publication of Alvan Feinstein's *Clinical Judgment* in 1967. In 1972, Archie Cochrane published *Effectiveness and Efficiency*, where he described several common practices not supported by clinical trials, and called for the British National Health Services (NHS) to base its practices on randomized control trials (RCTs) rather than expert opinion (Cochrane et al., 1972). In 1973, Wennberg et al. documented wide variations in practice patterns among clinics (Eddy, 2011; Wennberg and Gittelsohn, 1973), and by the 1980s it became apparent that a large proportion of the procedures being performed by physicians were considered inappropriate even by the standards of their own experts (Chassin et al., 1987; Eddy, 2011).

However, inconsistencies between experts do not mean that expert opinion should be discarded outright. Expert opinion can cover issues that may be difficult to address by trials, and may be more up-to-date than the best synthesized evidence (Booth, 2016a; Sackett et al., 1996). Experts often have knowledge, practical experience, insight, and implicit knowledge difficult to formalize in research. Unfortunately, such expertise is often subject to biases, and the lack of transparency means that it is difficult or impossible to determine what sources an expert opinion is based on (Gough et al., 2017):

- ❖ The opinion of the experts may be affected by their ideological and theoretical perspectives, which may not be explicitly stated. The perspectives of the experts may be influenced by personal interests.
- ❖ The boundaries of the experts' knowledge may not be transparent
- ❖ The experts may know some studies better than others, so not all research has equal representation in the conclusions they draw
- ❖ It may not be clear to what extent experts' conclusions are based on practice wisdom rather than evidence
- ❖ It may be difficult to assess expertise in a field, and to what extent the credibility of the expert is based on research. The credibility of an expert may for instance stem from their esteem as a practitioner, rather than as a researcher

To some extent, the dangers of expert opinion or expert panels are the same as those of an unsystematic review. The conclusions may be based on great insight, but due to the lack of transparency, it is difficult or impossible to determine to what extent the conclusions are drawn from evidence, preconceptions, personal beliefs, or accepted practices in the field (Gough et al., 2017).

2.1.2 *The Need for Evidence in Medicine*

In 1978, the Office of Technology Assessment of the US Congress estimated that only 10–20% of all procedures then used in medical practice were based on controlled trials (Banta et al., 1978). The Office's followup report in 1983 repeated the same estimate (Gelband, 1983). These estimates were independently confirmed in 1979, with 10% of common medical practices for three subspecialties of internal medicine lacking any foundation in published research (Williamson et al., 1979, cited by Sackett et al., 1995). In 1990, 21% of treatments and diagnostic procedures were firmly based on scientific evidence (Dubinsky and Ferguson, 1990, cited by Sackett et al., 1995). The view that less than 20% of general practice was not based on RCTs was common through the 1970s and 1980s, and quoted repeatedly by leading physicians and laymen alike (Eddy, 2005; Sackett et al., 1995). Simultaneously, the US Food and Drug Administration (FDA) had been requiring proof of efficacy of new drugs since 1962, and many other countries soon after (Bastian et al., 2010).

	Treatment Effect	Diagnosis	Prognosis
I	Systematic review of randomized trials or <i>n</i> -of-1 trials	Systematic review of cross sectional studies with consistently applied reference standard and blinding	Systematic review of inception cohort studies
II	Randomized trial or observational study with dramatic effect	Individual cross sectional studies with consistently applied reference standard and blinding	Inception cohort studies
III	Non-randomized controlled cohort/follow-up study	Non-consecutive studies, or studies without consistently applied reference standards	Cohort study or control arm of randomized trial
IV	Case-series, case-control studies, or historically controlled studies	Case-control studies, or poor or non-independent reference standard	Case-series or case-control studies, or poor quality prognostic cohort study
V	Mechanism-based reasoning	Mechanism-based reasoning	

Table 2.1 – Levels of evidence according to the Oxford Centre for Evidence-Based Medicine (The Oxford Centre for Evidence-Based Medicine, 2016). The boundaries between the levels are undulated to indicate that the order of the ranking is not absolute.

Even if good scientific evidence for treatment was hard to come by, healthcare policy in the 1980s had started to call for solid evidence, rather than informed opinion (Eddy, 2011). For instance, in its new recommendations published in 1980, the American Cancer Society insisted that ‘there must be good evidence that each test or procedure recommended is medically effective in reducing morbidity or mortality’ (Eddy, 1980). Over the next two decades, a number of organizations adopted guidelines based on evidence, including the American College of Physicians (ACP) in 1985, the Council of Medical Specialty Societies (CMSS) in 1987, the American Medical Association (AMA) in 1987, the US Preventive Services Task Force (USPSTF) (The US Preventive Services Task Force, 1989) in 1989, the Agency for Healthcare Research and Quality (AHRQ, then known as the Agency for Health Care Policy and Research, AHCPR) in 1993, the BMJ Publishing Group in 1995, and the American Association of Health Plans (now America’s Health Insurance Plans) in 1997 (Eddy, 2005).

The philosophical roots of evidence based medicine go back to mid 19th century Paris (Sackett et al., 1996), but the term was first used in 1990, in the context of guidelines (Eddy, 1990). It calls for a ‘conscientious, explicit, and judicious use of current best evidence’ and ‘integrating individual clinical expertise with the best available external clinical evidence from systematic research.’ (Sackett et al., 1996). In the mid-1990s evidence based medicine had received widespread endorsement, as well as the publicational outlet of its own journal (Davidoff et al., 1995a,b; Feinstein and Horwitz, 1997). By 1995, 82% of treatment given in general practice was evidence based, if not supported by RCTs (53%), at least by convincing non-experimental evidence (27%) (Sackett et al., 1995).

### 2.1.3 Hierarchies of Evidence

Evidence based medicine calls for guidelines to preferentially use evidence of ideal methodological quality, and if such evidence is not available, to use the best evidence at hand (Sackett et al., 1996). The gold standard evidence for interventions is the randomized controlled trials (Cochrane et al., 1972; Hariton and Locascio, 2018), which has become a cornerstone of judging the effectiveness of treatments (Sackett et al., 1996). Nonrandomized trials, cohort studies, or case studies can provide evidence, but are more susceptible to bias and random variation.

The idea of a hierarchy of evidence of this kind was first made explicit in the guidelines published by the Canadian Task Force on the Periodic Health Examination in 1979 (Canadian Medical Association, 1979). Its purpose was to develop recommendations on the periodic health exam based on evidence from the medical literature (Burns et al., 2011; Canadian Medical Association, 1979). Such hierarchies are intended as rules-of-thumb for quickly assessing the available evidence.

Hierarchies of evidence have been inflexibly used, and criticized for decades (The Oxford Centre for Evidence-Based Medicine, 2016). The relative strength of different levels of evidence is not cut in stone. In practice, the methodological quality of the study and the consistency of the results should be critically assessed whether the evidence should be graded up or down. For instance, an observational study with very dramatic effects – cf. the original discovery of penicillin – may be convincing even in the absence of higher level evidence, such as RCTs or systematic reviews (Cochrane et al., 1972). Conversely, a systematic review may provide weak evidence, for instance if it has identified too few studies (The Oxford Centre for Evidence-Based Medicine, 2016), or if the identified studies report conflicting results (Deeks et al., 2019, in Higgins et al., 2019). Similarly, while RCTs are commonly considered among the gold standard for interventions, in practice many RCTs suffer from methodological problems, and their evidence may need to be graded down due to poor randomization or blinding, large numbers of withdrawals, or wide confidence intervals (Burns et al., 2011).

However, all else being equal, systematic reviews are consistently placed at the top of evidence hierarchies (e.g. see table 2.1). Systematic reviews are better at assessing strength of evidence than single studies and are recommended over single studies if available (Chalmers, 2007; The Oxford Centre for Evidence-Based Medicine, 2016). Systematic reviews allow for results with narrower confidence intervals than most single studies (Leeftang et al., 2008), are more likely to give conclusive answers to research questions (Chalmers, 2007), and stronger generalizability than individual studies (Leeftang et al., 2008). Furthermore, systematic reviews can address research questions that are difficult to address by single studies, and can better assess to what extent the findings of the included studies have been affected by methodological issues and bias.

#### 2.1.4 *Systematic Reviews & Evidence Based Medicine*

A systematic review is a meta-research study type that attempts to answer research questions by identifying and analyzing all published empirical evidence relevant to the question. Unlike e.g. narrative reviews, the systematic review uses controlled, systematic methods in order to minimize bias. This results in generally very reliable evidence, and is typically placed at the top of evidence hierarchies.<sup>1</sup> For instance, systematic reviews of randomized controlled trials are often considered the gold standard for intervention research.

<sup>1</sup> It is sometimes held that systematic reviews produce *the* strongest evidence, but this is a simplification, and will depend on the evidence included in the review. For instance, a large, well-conducted RCT is likely more reliable than a systematic review of case studies.

This systematic and controlled procedure is what distinguishes a systematic review from the traditional literature review. A traditional literature review summarizes what is known on a topic, and the studies that have been published addressing it, but do not explicitly specify the criteria used to identify and include studies. Relevant studies may not have been included because the review authors were unaware of them, or have been excluded for reasons known only to the author. Unless the identification process is explicit, it is not possible to judge whether the inclusion of studies is appropriate, or whether the inclusion has been consistent (Gough et al., 2017). The review is typically performed by domain experts, whose expertise guides the selection of studies and analysis of them, and their expertise may in turn have been shaped by the studies they are aware of.

17

### 2.1.5 Systematic Reviews Answer Research Questions

A systematic review is a scientific project in its own right (Chalmers, 2003), and can be used to answer any question that can be answered by primary research (Thomas et al., 2019, in Higgins et al., 2019). Like other types of research, it requires a systematic and methodological approach to adequately address its research question, without being misled by systematic biases or random chance (Chalmers, 2003; Gough et al., 2017).

Roughly 50 years ago, Sir Austin Bradford Hill summarized the structure of a scientific study as four questions: Why did you start? What did you do? What answer did you get? And what does it mean anyway? (Hill, 1965) These questions are reflected in the ubiquitous structure of scientific reports: the *introduction*, *method*, *results*, and *discussion* (sometimes called IMRaD) (Clarke et al., 2002).

Scientific methods must be systematic and rigorous, and their reporting explicit and transparent so that the results can be interpreted and assessed in the light of how the results were produced. We should rightly be suspicious of results produced by flawed or poorly reported methods (Altman, 1994; Chalmers, 2003). The same applies to reviews of research (Gough et al., 2017). The systematic review consequently involves the same basic steps as any other scientific inquiry (Chalmers, 2003):

- ❖ Defining the objectives of the research
- ❖ Defining the methodology to gather data and perform analyses
- ❖ Analyzing the data, using prespecified statistical methods if appropriate
- ❖ Interpreting the findings and preparing a structured report of the research

As outlined in the Cochrane handbook for interventions, the key characteristics of systematic reviews are:

- ✦ A clearly stated set of objectives with pre-defined eligibility criteria for studies
- ✦ An explicit, reproducible methodology
- ✦ A systematic search that attempts to identify all studies that meet the eligibility criteria
- ✦ An assessment of the validity of the findings of the included studies, for example through the assessment of risk of bias
- ✦ A systematic presentation, and synthesis, of the characteristics and findings of the included studies

18

The crucial difference between a systematic review and primary research is the unit of interest in the study – in e.g. an intervention or diagnostic study the unit is a patient, in a systematic review the unit is a primary study (Schünemann and Moja, 2015).

Literature searches are performed for a number of reasons, including traditional literature reviews, but can also be performed to populate literature databases, or identify relevant work or prior art. However, the systematic review as a study design shares all methodological qualities of an observational study, while having few methodological similarities with traditional literature reviews, and the comparison to literature reviews may therefore ultimately be unhelpful. The systematic review may be better understood as an observational study, where the data collection takes the form of a (systematic) literature search (Cook et al., 1997; Schünemann and Moja, 2015). Consequently, the same methodological requirements of planning, rigour, and transparency apply in a systematic review, as in any research endeavor (Chalmers et al., 2013).

## 2.2 SYSTEMATIC REVIEWS OF DIAGNOSTIC TEST ACCURACY

Diagnostic test accuracy studies measure how accurate tests are in detecting the presence or absence of a medical conditions. Such medical conditions may commonly be a disease, but can also be non-malign conditions, such as twin pregnancy (Monni et al., 2014). Tests may be conventional laboratory procedures, such as biochemical, immunological, omic technologies, or imaging tests such as ultrasound or MRI scans. The tests may also comprise other measurements that may help in distinguishing the healthy from the diseased, such as signs and symptoms from

patient history and examination, questionnaires, scores and decision rules, or physiological measurements (Chandler and Hopewell, 2013).

The accuracy of a diagnostic test, measurement, or procedure is one of the key criteria for recommending the test for use in clinical practice, but not the only one. The accuracy of a procedure must also be balanced against for instance its intrusiveness, its safety, its cost, and whether it is practical to adopt in clinical practice. For instance, biopsy is a highly accurate procedure to diagnose many diseases, but is highly invasive, and its detrimental effects on the well-being of patients is limiting its clinical use.

Diagnostic studies are often reported in studies with small sample sizes, and the accuracy measurements are therefore often imprecise, with wide confidence intervals (Bachmann et al., 2006). Systematic reviews are typically necessary to achieve precise accuracy measurements (Leeflang et al., 2008). Systematic reviews are also useful for analyzing the variability of the results across subgroup, identify risks of bias in individual studies, and address questions not covered by the original studies, such as the differences between different tests (Bachmann et al., 2006; Leeflang et al., 2008).

Unlike randomized control trials, which typically report results as a single measure of effect (e.g. as a relative risk ratio), diagnostic test accuracy necessarily involves a trade-off between sensitivity and specificity depending on the threshold for positivity for the test (Leeflang et al., 2008; Macaskill et al., 2010, in Deeks et al., 2013a). Diagnostic test accuracy studies therefore usually report results as two or more statistics: e.g. sensitivity and specificity, negative and positive predictive value, or the Receiver Operating Characteristic (ROC) curve.

Meta-analyses of diagnostic test accuracy pool the  $2 \times 2$  tables reported in multiple DTA studies together to form a summary estimate of the diagnostic test performance. The results of DTA studies are expected to be heterogeneous, and the meta-analysis thus needs to account for both inter- and intra-study variance (Macaskill et al., 2010, in Deeks et al., 2013a). This is commonly accomplished using hierarchical random effects models, such as the bivariate method, or the hierarchical summary ROC model (Reitsma et al., 2005; Rutter and Gatsonis, 2001). Pooling sensitivity and specificity separately to calculate separate summary values is discouraged, as it may give an erroneous estimate, e.g. a sensitivity/specificity pair not lying on the ROC curve (Leeflang et al., 2008).

## 2.3 ISSUES IN SYSTEMATIC REVIEW PRODUCTION

2.3.1 *The Publication of Systematic Reviews is Growing*

A large number of systematic reviews is being published every year, and the number is rising steadily.

An examination of records indexed in PubMed between Jan 1, 2000–Dec 31, 2018 shows 122,979 records tagged as systematic reviews (figure 2.3).<sup>1</sup> PubMed indexed 17,254 new systematic reviews in 2018 alone, compared to 3,336 in 2008 – more than a five-fold increase over 10 years (figure 2.3).

The publications of systematic review may be increasing rapidly, but it is being out-paced by the growing number of annual registrations in PROSPERO. There were 15,667 new registrations in PROSPERO in 2018 – almost as many as the number of new systematic reviews in PubMed. The rapidly increasing number of registration is likely partly due to the growing rate of publication of systematic reviews. At the same time, there is also a growing awareness of the need for prospective registrations of systematic reviews, and funding organizations increasingly require new systematic reviews to be registered as a precondition for funding.

Why is the number of systematic reviews increasing? On the one hand the production of more and more systematic reviews is a response to their increased reliance by guidelines and policy-makers. Furthermore, not only systematic reviews are being produced in ever larger quantities. The same trends are true for most of science, including the primary studies systematic reviews are based on. Coping with the increasing number of primary studies is an often cited reason to perform systematic reviews in the first place.

However, there may be less benign reasons behind the increase. Like elsewhere in science, researchers performing systematic reviews are subject to the same incentives of publish-or-perish. Systematic reviews may increasingly be pursued in order to advance the careers of the researchers, rather than the state of the science (Ioannidis, 2016). Meanwhile, publishers and journals are incentivized to publish systematic reviews, since these are often more highly cited than other types of literature (Bastian et al., 2010). Multiple independent systematic reviews have been

1 PubMed publications were collected using the query 'systematic [sb]' in accordance with the US National Library of Medicine recommendations:

[https://www.nlm.nih.gov/bsd/pubmed\\_subsets/sysreviews\\_strategy.html](https://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html)

These numbers are less than half of what has previously been reported by Ioannidis (2016), who reported 28,959 PubMed systematic reviews in 2014 alone. The likely explanation for the discrepancy is that the US National Library of Medicine has since updated its criteria for tagging systematic reviews to be more consistent, including introducing a dedicated publication type tag in February 2019 (Collins, 2019). PROSPERO registrations were collected using the PROSPERO search function with a date limit set to Jan 1–Dec 31 for each year.

identified for a number of topics, including gastric ulcer prophylaxis, dosing of aminoglycosides, selective decontamination of the digestive tract, orthopedic procedures, and wound healing (Ioannidis, 2016). Independent replication is useful in any field of research (Siontis et al., 2013), but is also a potential for research waste and usually discouraged (Ioannidis, 2016; Lasserson et al., 2019, in Higgins et al., 2019).

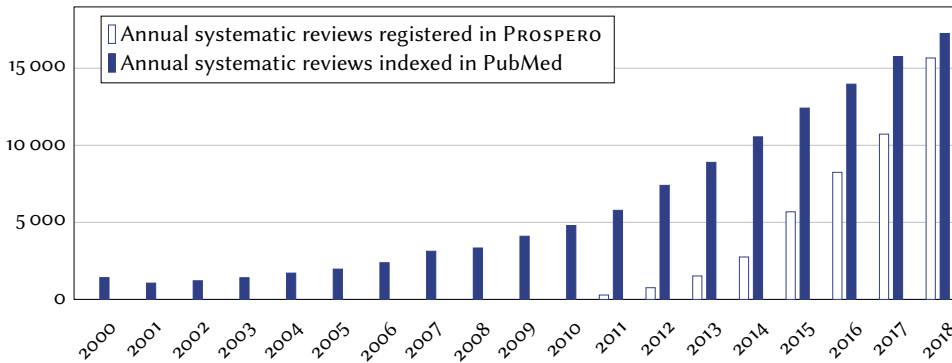


Figure 2.3 – The number of annual systematic reviews indexed in PubMed, and the number of annual registrations of started systematic reviews in the PROSPERO database. The searches were performed in August 2019.

### 2.3.2 The Workload in Systematic Reviews is Growing

While the demand for systematic reviews is rapidly increasing, the workload involved in their production is also growing. Part of the reason is simply that an increasing number of studies is being published each year. In 2010, 75 trials were being published per day, compared to 14 trials per day in 1979 (Bastian et al., 2010). Consequently, systematic reviews now need to consider a much larger body of published literature.

At the same time, the scope and methodological rigour of systematic reviews have increased since the 1970s, in response to increasing awareness of risks of bias. Systematic reviews are more complex than previously, and have stricter expectations of screening methodology and explicit quality assessment of included studies. Early systematic reviews were typically 10–20 pages long, even when they included several studies. Today it is not unusual for a review by a health technol-

ogy agency to be several hundred pages long. Systematic reviews are now often longer than the combined length of their included reports (Bastian et al., 2010). There is also an increasing expectation that systematic reviews include study types other than RCTs (Bastian et al., 2010), and other sources than published literature. For instance, case studies are often necessary to detect adverse effects (Bastian et al., 2010). In addition, systematic reviews are increasingly expected to consider gray literature and unpublished trials to mitigate publication bias, further increasing the workload.

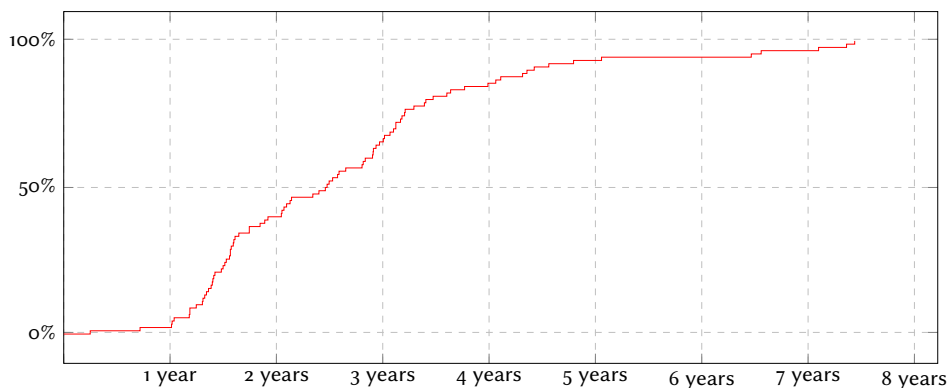


Figure 2.4 – Cumulative plot of the delays between the publication of the protocol and the first publication of the review for the systematic reviews in the ‘Diagnosis’ section of the Cochrane Library. Out of the 120 publications in this section (August 2019), this data includes the 90 for which the publication dates were recorded in the published article. One article was excluded due to being a systematic review of DTA systematic review methods.

### 2.3.3 Systematic Reviews Take a Long Time to Complete

Most systematic reviews take years to complete, and require considerable expense. Systematic reviews from the Agency for Healthcare Research and Quality (AHRQ) addressing comparative effectiveness with five key questions and the need to review about 10,000 citations are reported to cost upward of \$300,000 (Lau, 2019). There are however large variations between reviews. Extreme examples can take as little as 16 weeks, and as much as 12 years (Schünemann and Moja, 2015). The systematic review process is reported to take an average of 67 weeks to complete for systematic reviews of interventions (Borah et al., 2017). Typical timeframes

for systematic reviews in general have been reported to fall within 12–24 month (Tsafnat et al., 2018), within 6–24 months (Beller et al., 2018; Khangura et al., 2012, quoting Tsertsvadze et al., 2015), or within 6 months to several years (Tsertsvadze et al., 2015). Systematic reviews in different fields and topics may vary in time, but there can also be large variations within the same area depending e.g. on how many references need to be considered in each stage through the review process (Allen and Olkin, 1999; Pham et al., 2018).

Cochrane systematic reviews account for an estimated 20% of the annual output of systematic reviews (Moher et al., 2007), but are known for uniformly high methodological quality (Jadad et al., 1998). Due to the stricter methodology, commonly cited estimates of the time taken to perform a review have been based on general systematic review, and may therefore not be representative for Cochrane systematic reviews. The median delay between publication of protocol and publication of review for DTA systematic reviews in the Cochrane Library is 904 days (2.47 years) with the longest taking 2,715 days (7.44 years) (figure 2.4). The same delay for Cochrane intervention reviews has previously been reported as 2.24 years (range: 0.25–7.75 years) (Tricco et al., 2008). Since writing and publishing the protocol often takes months (Lasserson et al., 2019, in Higgins et al., 2019), the total timeframe of the review, from start to finish, will be even longer.

The long production times of systematic reviews has several implications. First, much of the work is performed by human reviewers, and the process is therefore costly. Second, the delay means that many reviews will not be up to date at the time of publication, with a median delay between the time of the last database search and publication of 5.1 months (Beller et al., 2013). Any more recent publications will not have been considered in the systematic review. A somewhat older study by Shojania et al. (2007) found that due to this delay, 7% of systematic reviews were outdated at the time of publication, and would have reached different conclusions had they included all studies available at the time of publication.

The conventional countermeasure to the combination of short review half-life and slow review turnover is to perform an abridged database search just before publication to identify studies published after the initial search. For instance, Cochrane systematic reviews are required to rerun searches if the initial search was performed more than 12 months (preferably 6 months) before the intended publication date of the review (Lefebvre et al., 2019, in Higgins et al., 2019). Prevention is better than cure however, and rerunning searches would be unnecessary if the screening could be completed quicker.

#### SUMMARY

SYSTEMATIC REVIEWS ARE A META-RESEARCH STUDY TYPE that address research questions by analyzing all published relevant studies. Systematic reviews often produce higher quality evidence than individual studies

SYSTEMATIC REVIEWS ARE IMPORTANT for evidence based medicine, and are today the main sources of evidence for clinical guidelines

SYSTEMATIC REVIEWS ARE EXPENSIVE to produce, and the cost is rising. The long delays involved in their production mean that they are difficult to produce in response to urgent policy needs



**G**IVEN THEIR CENTER-STAGE ROLE in evidence based medicine, systematic reviews are today important to a range of stake-holders. They are important to investigators to summarize existing data, refine hypotheses, estimate sample sizes, and define future research agendas (Cook et al., 1997). Without systematic reviews researchers may chase the wrong leads, study questions that have already been adequately addressed, or fail to interpret findings in the light of available evidence (Clarke et al., 2002; Ioannidis, 2016; Lasserson et al., 2019, in Higgins et al., 2019). They are important to policy-makers to optimize outcomes with available resources (Cook et al., 1997). They are important to law-makers and regulators to guide public policy (Lasserson et al., 2019, in Higgins et al., 2019). Not least, they are important for patients – healthcare decisions affecting individual patients are increasingly taken based on conclusions drawn from systematic reviews, and may directly influence what treatment and healthcare patients receive.

For these reasons it is imperative that systematic reviews are as high-quality, relevant, unbiased, and up-to-date as possible. To ensure this, reviews follow a systematic and highly controlled procedure to minimize sources of bias (Cook et al., 1997), resulting in a robust but cumbersome and time-consuming process (figure 3-1). This process takes on average 67 weeks to complete for systematic reviews of interventions, but there are large variations (Borah et al., 2017). It is not uncommon for a review to take years to complete, with extreme examples needing 12 years (Schünemann and Moja, 2015).

Most steps in the process require manual processing, but the majority of the time and work is required in a few of the stages. According to Allen and Olkin (1999), on average 52% of the time is spent on search and extraction, 18% is spent on statistical analysis, 18% is spent on administrative tasks, and 13% is spent on manuscript writing. A newer study by Pham et al. (2018) found similar numbers by studying event logs from systematic reviews: 26% was spent on search and retrieval, 24% on data collection, 23% on manuscript writing, and 17% on statistical analysis. In both accounts, over half the workload was spent searching, screening, and extracting. Unsurprisingly, the steps associated with the largest workload may be those that involve the most mechanical and repetitive components of the work, those most amenable to automation.

In this work we will focus on the process of a Cochrane systematic review. Cochrane systematic reviews typically follow a more rigorous procedure, and is often

seen as a gold standard for systematic reviews (Chandler and Hopewell, 2013; Jadad et al., 1998). However, Cochrane systematic reviews only make up approximately 20% of all published systematic reviews (Moher et al., 2007). Other systematic reviews may skip some of the steps, or relax part of the procedure. For instance, while all Cochrane systematic reviews are required to report the assessed quality of all included studies, this is omitted in about a third of published systematic reviews (Moher et al., 2007).

Conceptually, the individual steps are performed independently and sequentially. This rigid compartmentalization is considered a feature intended to minimize bias (Kellermeyer et al., 2018). In practice the distinction between some of the steps may be blurred, and the process may be backtracked to correct for issues encountered in later stages (Page et al., 2019a, in Higgins et al., 2019). Changes may be appropriate, for instance if the search filters are determined to be inadequate after known relevant references are not returned by the database search (Higgins and Deeks, 2011). Such post-hoc changes should be avoided and – when unavoidable – described and justified in the final report (Lasserson et al., 2019; Page et al., 2019a, in Higgins et al., 2019).

To the greatest extent possible, all judgements should be made during the planning stage, and before the available studies are known. Consequently, the review questions, the inclusion and exclusion criteria as well as the statistical methods used for analysis are specified before starting the review. Post-hoc decisions – made after seeing the evidence – to include or exclude studies, or to change statistical methods, are highly susceptible to bias, and therefore to be avoided (Lasserson et al., 2019, in Higgins et al., 2019).

Due to a lack of alternatives meeting the high recall requirements, as well the high stakes associated with errors, virtually all of the work is performed manually. Specialized software is used to support much of the process, as well as to assist in documenting the decisions and conduct and to distribute the tasks among the authors, who may be in different countries.

Dedicated systematic review managers are software that aims to streamline a broad range of tasks within the systematic review process. This may include preparation of the protocol, assistance in writing the protocol and review manuscripts, keeping track of study characteristics and study data, performing meta-analyses, and presenting graphical results.

The use of dedicated software for preparing systematic reviews is not just to facilitate the work, but also to document the decisions to ensure transparency and reproducibility (Lasserson et al., 2019, in Higgins et al., 2019). Examples of review managers include RevMan, Rayyan, Covidence, and DistillerSR (Cochrane, 2014; Evidence Partners, 2019; Ouzzani et al., 2016; Veritas Health Innovation, 2019). RevMan is required for preparing, writing, and maintaining Cochrane systematic reviews.

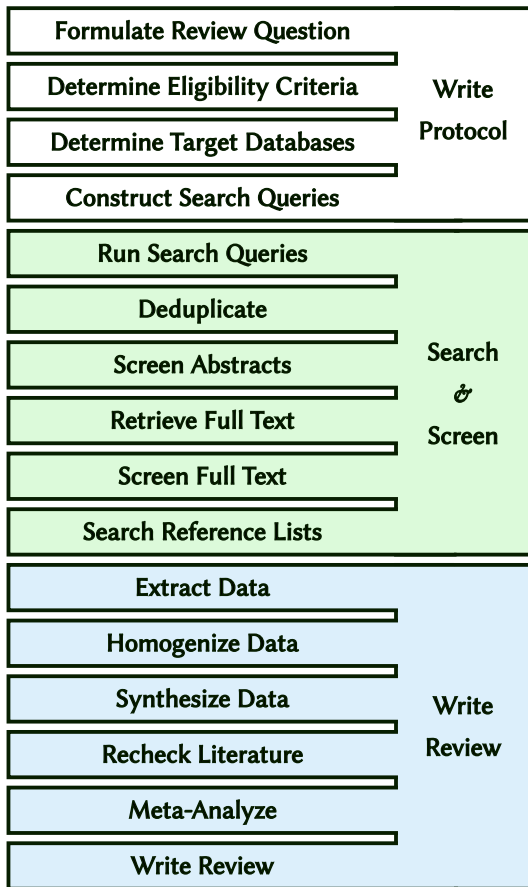


Figure 3.1 – The systematic review process, adapted from (Deeks et al., 2019; Lasserson et al., 2019; Lefebvre et al., 2019; Li et al., 2019; McKenzie et al., 2019a,b; Thomas et al., 2019, in Higgins et al., 2019)).

lack of specialized features – typically only in a limited capacity, for instance to keep track of screening decisions (Lorenzetti and Ghali, 2013; Roth et al., 2018). Meta-analyses in systematic reviews typically rely on general statistical software, such as Stata or SAS.

However, while these software tools are helpful in streamlining and documenting the process, they are better described as tools rather than automation. By automation we mean that the software is performing tasks normally performed by review

Reference managers are software used to store and manage bibliographic records. Dedicated reference managers for use in systematic reviews include EndNote, RefWorks, F1000 Workspace, and Zotero, as well as legacy reference managers such as ProCite and Reference Manager (Center for History and New Media, George Mason University., 2019; Clarivate Analytics, 2019; F1000, 2019; ProQuest, 2019; Thomson Reuters, 1999, 2008). EndNote appears to have been the most dominant reference manager in 2013 (Lorenzetti and Ghali, 2013), and is frequently used in systematic reviews. It is difficult to get an accurate and up-to-date view of the current relative market share however, since current reporting guidelines such as PRISMA do not require explicitly reporting software usage. In practice, only 4.8% of reference manager usage in systematic reviews is reported (Lorenzetti and Ghali, 2013). Spreadsheet software, such as Microsoft Excel, is occasionally used to handle references, but – due to its

authors, and which involves some kind of non-trivial decision-making (Van Altena et al., 2019). The main exception where software is ‘making decisions’ is the deduplication process, which is often performed using heuristic or machine learning methods, for example the *find duplicates* feature in EndNote. Even this process is typically only semi-automated, and the review authors still have to arbitrate unclear cases.

Why is the systematic review process not automated to a greater degree? Not because of a lack of methods. Automation tools have in fact been developed since at least 2005, targeting several aspects of the review process. According to a 2018 survey, licensing, steep learning curves, lack of support, and mismatch to workflow were cited as the main reasons review authors have not used automated tools (Van Altena et al., 2019)

### 3.1 WRITING THE PROTOCOL

Both registration of systematic reviews in databases like PROSPERO and the publication of protocols in journals avoid duplication, by allowing researching to search for ongoing systematic reviews (Rombey et al., 2019; Stewart et al., 2012). Publication also allows the review methodology to be peer reviewed prior to starting the review, and increases the chance that methodological problems are corrected (Rombey et al., 2019).

The protocol writing process in a systematic review is complex and may undergo several rounds of peer review and revisions before publication. Peer reviewed publication of the protocol is mandatory in systematic reviews by Cochrane (Deeks et al., 2013b, in Deeks et al., 2013a), the Campbell Collaboration (Campbell Collaboration, 2019), and the Joanna Briggs Institute (Aromataris and Munn, 2017). For other systematic reviews, peer review and publication of the protocol is typically optional (Rombey et al., 2019).

Publishing the protocol involves substantial effort, but is important to avoid making judgements based on the findings of the review. Publishing the protocol of the review mitigates review authors’ bias, promotes transparency of methods and processes, and avoid duplicate reviews (Rombey et al., 2019). Systematic reviews with published protocols are associated with higher standards of reporting and methodological quality than systematic reviews without published protocols (Allers et al., 2018). Even for systematic reviews where the protocol is not published, it is strongly recommended that a protocol is prepared before the systematic review is started (Lasserson et al., 2019, in Higgins et al., 2019).

3.1.1 *Formulate Review Question*

A systematic reviews can address any research question that can be addressed by primary studies (Thomas et al., 2019, in Higgins et al., 2019). Just like for primary studies, a well-formulated research question is key to a properly conducted review, and is integral to ensure the relevance and novelty of the results, as well for reducing and mitigating biases.

High quality systematic reviews have a purpose statement detailing the question of the systematic review (Jackson and Kuriyama, 2018). To avoid ad-hoc decisions made after seeing the evidence, and associated bias, the question should be specified a priori (Lasserson et al., 2019, in Higgins et al., 2019). Review authors' prior knowledge of the evidence may influence the definition of the research question, eligibility criteria, or the analysis (Lasserson et al., 2019, in Higgins et al., 2019).

Novel systematic reviews address research gaps. To avoid duplicate reviews, authors should search for previous systematic reviews in published literature, and check the PROSPERO register of systematic reviews before starting the review (Thomas et al., 2019, in Higgins et al., 2019).

3.1.2 *Determine Eligibility Criteria*

The review question in a systematic review must clearly delineate its scope and what studies are eligible for inclusion (Higgins et al., 2019; Lasserson et al., 2019, in Higgins et al., 2019). This is one feature that distinguishes a systematic review from narrative reviews and scoping reviews, and serves to mitigate bias (McKenzie et al., 2019b). The review question needs to specify what the authors are willing to accept as evidence, i.e. what populations, what clinical settings, which tests or treatments, and which study designs are eligible for the review.

For systematic reviews of interventions, eligibility criteria are often expressed in terms of PICO elements. The participants, intervention, and comparison often translate directly into eligibility criteria for a review (McKenzie et al., 2019b, in Higgins et al., 2019). Eligible studies should match the target population, intervention, and comparison specified by the review question. Some reviews of interventions may restrict eligibility to specific outcomes, but determining the range of potential outcomes is typically part of the aim of an intervention review (McKenzie et al., 2019b, in Higgins et al., 2019).

Systematic reviews of diagnostic test accuracy do not have an equivalent formalization, but the index test, target condition (population), and reference standard are often used as eligibility criteria, similarly to PICO for interventions (Norman et al., 2019e).

3.1.3 *Determine Target Databases*

Systematic reviews require a thorough, objective, and reproducible search (Lefebvre et al., 2019, in Higgins et al., 2019).

Searching a single database is generally inadequate to identify all relevant studies, and may miss up to 40% of relevant studies (Bahaadinbeigy et al., 2010; Betrán et al., 2005; Egger et al., 2003; Halladay et al., 2015; Lemeshow et al., 2005; Lorenzetti et al., 2014; Marshall et al., 2019; Nussbaumer-Streit et al., 2018; Parkhill et al., 2011; Royle and Milne, 2003; Royle and Waugh, 2005; Royle et al., 2005; Sampson et al., 2003; Slobogean et al., 2009; Stevinson and Lawlor, 2004; Subirana et al., 2005). In practice, most or all studies relevant to a given systematic review may be indexed in a single database<sup>1</sup> (Booth, 2016b; Halladay et al., 2015), but there is no way to tell whether this is true without searching all appropriate databases.

A number of literature databases are available, including PubMed (MEDLINE), EMBASE, CINAHL, IEEE Xplore, the ACM Digital Library, ISI Web of Knowledge, Scopus, CiteSeer, arXiv, DBLP, and Google Scholar. The archival of literature is balkanized, with little overlap between databases (figure 3-2) (Hull et al., 2008). Different databases cover different subject areas, and database selection is therefore often guided by the review topic (Lorenzetti and Ghali, 2013).

Failure to search multiple databases may influence the quality of the review, and is a known source of potential bias (Marshall et al., 2019; Nussbaumer-Streit et al., 2018; Royle et al., 2005). Different databases often index studies with different characteristics, and missing studies showing e.g. negative findings may skew the results of the systematic review (Sterne et al., 2001).

For instance, studies with small sample sizes are more likely to be published in lower impact journals. Negative results are more likely to take longer to publish, be published as gray literature, be published in other languages than English, or not be published at all (Chalmers, 2003; Dickersin et al., 1992; Easterbrook et al., 1991; Egger et al., 1997; Stern and Simes, 1997; Sterne et al., 2001). A systematic reviews that only include studies from high impact journals is therefore at risk of overestimating treatment effects or diagnostic performance. It is important to search multiple databases so that the set of references that are identified are not systematically different from those that would have been identified by a more comprehensive search (Sterne et al., 2001).

Google Scholar provides almost complete coverage of the published literature (Gehanno et al., 2013; Hull et al., 2008). Unfortunately, Google Scholar does not provide a search interface allowing boolean queries, nor bulk download of its database

<sup>1</sup> Including the iconic systematic review of corticosteroid treatment for pre-term birth, where all studies were indexed in Medline (Booth, 2016b).

(Martín-Martín et al., 2018). Consequently, while all relevant studies are usually indexed in Google Scholar, there are no methods to systematically and comprehensively identify these through Google Scholar in a systematic review.

Removing language restrictions from searches in English language databases is not a good substitute for searching non-English language journals and databases (Lefebvre et al., 2019, in Higgins et al., 2019).

### 3.1.4 Construct Search Queries

Search strategies should be planned in advance, and should be motivated by the eligibility criteria of the review, such that all studies meeting the eligibility criteria are identified. This includes tailoring the search to relevant PICO criteria, as well as to publication status and language of publication (Lefebvre et al., 2019, in Higgins et al., 2019).

In a Cochrane systematic review, search query development is a complicated process that may involve several rounds of revisions. Cochrane guidelines strongly recommend that search strategies are ‘peer reviewed by a suitably qualified and experienced medical / healthcare librarian or information specialist’ (Lefebvre et al., 2019, in Higgins et al., 2019). The final search strategies including search queries are published as part of the protocol to promote transparency. While peer review and publication of search strategies or search queries is required by Cochrane reviews (Lefebvre et al., 2019, in Higgins et al., 2019), this is uncommon for non-Cochrane systematic reviews (Rombey et al., 2019).

Where applicable, systematic reviews are recommended to use published, validated search queries (Cochrane Community, 2019), which have been validated empirically and demonstrated to yield acceptable sensitivity guarantees (Lefebvre et al., 2019, in Higgins et al., 2019). Such filters are available to identify RCTs, quasi-RCTs, and CCTs in MEDLINE, EMBASE, and CINAHL. No similar filters are known for identifying DTA studies with acceptable sensitivity (Beynon et al., 2013).

Database search queries used in systematic reviews are typically designed for high

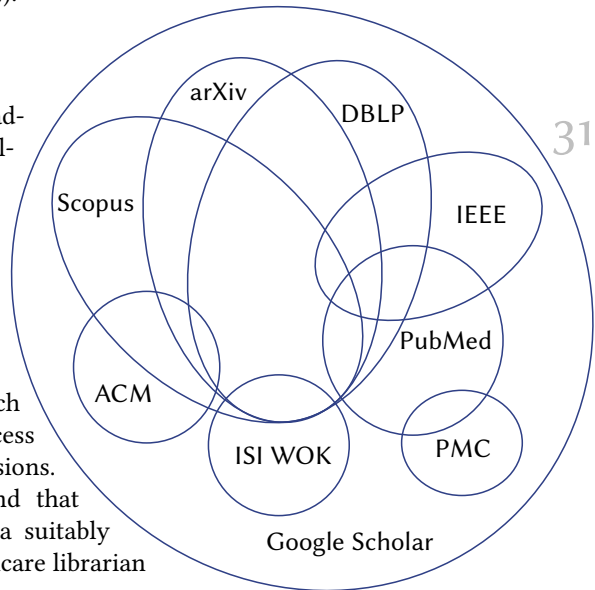


Figure 3.2 – The overlap in several literature databases, adapted from Hull et al. (2008). The size of the regions are not proportional to the number of references indexed by each database or their relative overlap.

sensitivity rather than specificity (Lefebvre et al., 2019, in Higgins et al., 2019; Lorenzetti et al., 2014). For instance, published validated search filters to identify RCTs and CCTs may yield 100% recall (sensitivity) for 4.3% precision (Eisinga et al., 2007). For some topics – notably diagnostics – it may not be possible to construct study type search filters that are both sensitive and specific at the same time (Beynon et al., 2013; Leeftang et al., 2008, 2006). Consequently, DTA systematic reviews omit study type filters entirely, searching only by target condition (De Vet et al., 2008, in Deeks et al., 2013a). While this keeps sensitivity high, precision may be as low as 0.42% (Kanoulas et al., 2018; Norman et al., 2018b).

### 3.2 SEARCHING AND SCREENING

#### 3.2.1 *Run Search Queries*

To ensure a comprehensive and unbiased search, systematic reviews must systematically search several sources (Greenhalgh and Peacock, 2005; Lefebvre et al., 2019, in Higgins et al., 2019), including dedicated literature databases (e.g. MEDLINE, EMBASE, and CINAHL), databases of gray literature (e.g. databases of dissertations and conference abstracts), trial registries (e.g. ClinicalTrials.gov (The US National Library of Medicine, 2019), the WHO's International Clinical Trials Registration Platform (ICTRP) (The World Health Organization, 2019), and the EU Clinical Trials Register (The European Medicines Agency, 2019)), as well as hand searching relevant journals not covered by the literature databases.

Less than half of all trials are published (Chalmers et al., 2013), and negative results are more likely to remain unpublished, or published as gray literature (Sterne et al., 2001). Failure to search sources beyond standard databases may therefore be at the mercy of publication bias, and may cause a systematic review to overestimate treatment effects (Lefebvre et al., 2019, in Higgins et al., 2019) or diagnostic performance (De Vet et al., 2008, in Deeks et al., 2013a).

For these reasons, a systematic review may need to consider dozens of databases, each with its own search engine, metadata, vocabulary, and query syntax (Tsafnat et al., 2014). Different databases generally use different syntax to specify logical operators such as OR, AND, and NOT, different syntax to specify fields such as *authors*, *year*, *title*, *abstract*, et c., and different controlled vocabularies, such as the Medical Subject Headings (MeSH) for MEDLINE (PubMed), or Emtree for EMBASE (Tsafnat et al., 2014).

Interoperability between databases is rare (Hull et al., 2008; Tsafnat et al., 2014). Running search queries for a systematic review therefore require specialized expertise in a number of database systems, and is recommended to be performed by – or with the assistance of – dedicated information specialists (Lefebvre et al., 2019, in Higgins et al., 2019).

### 3.2.2 *Deduplicate*

When retrieving records from multiple databases, the same records may be retrieved from more than one database. To avoid wasting time and resources considering the same reports multiple times, duplicates are conventionally removed, a process known as deduplication. However, unless the databases have substantial overlap, the number of duplicate reference may be small, and the gains from deduplication may vary. If the screening stages were substantially automated the benefit would largely disappear and deduplication could be performed later in the process, or omitted entirely.

Deduplication is one of the few steps of the systematic review process where the use of automation methods is widespread.

Obstacles for deduplication include variant spellings and formatting of titles or author names, as well as inconsistently indexed meta-data across databases. Deduplication methods typically use fuzzy matching with heuristic methods, or machine learning, and are available in common reference managers such as EndNote (Clarivate Analytics, 2019).

Current systems – based on heuristics or machine learning – perform sufficiently well that deduplication is ubiquitous in Cochrane systematic reviews (De Vet et al., 2008, in Deeks et al., 2013a). The cost associated with false negatives (i.e. screening duplicate records) is typically small. While deduplication has room for future improvements, benefits of better deduplication methods may therefore be relatively minor for systematic reviews.

### 3.2.3 *Screen Abstracts*

Deciding which of the candidate references ought to be included in the review is done manually, according to a pre-specified set of inclusion criteria. To ensure that the results are reproducible and consistent, and that relevant studies are not overlooked, it is recommended that the screening is performed in parallel by at least two screeners (Edwards et al., 2002; Higgins and Deeks, 2011). In the case that screening in parallel is not feasible (or not useful due to very high agreement), it is most important that the final decision as to whether to include studies is undertaken by at least two authors (Higgins and Deeks, 2011).

Screening is performed in two steps, first preliminarily based on title and abstract, then based on full-text. The reason for dividing screening into two stages is due to the often considerable difficulty inherent in locating full-texts for articles. It is typically not feasible to retrieve full-text articles for thousands of candidate references, but often the vast majority of references can be discarded simply by reading the titles and abstracts.

Screening a single study at the title and abstract level takes on average 30 seconds to one minute for an experienced screener, but may take substantially longer for inexperienced screeners (Wallace et al., 2010c). Screening a single full-text for inclusion may take around 10 minutes (Norman et al., 2019f). For a systematic review with a large number of candidate references, the screening stage may therefore take several months, or even years, to complete.

### 3.2.4 *Retrieve Full Text*

34

Retrieving the full texts of articles is generally burdensome, and difficult to automate. Obstacles include restrictions imposed by publishers, such as restrictions on access and subscription models that prohibit automated access to articles, as well as limited archival of and electronic access to articles. Many of the references included in a systematic review may lack identifiers (including DOI). In many cases, the full-texts are only possible to retrieve by contacting the authors.

Even if links to literature is available from journal webpages, journal typically do not provide standardized application programmable interfaces (API) allowing scripted access. Downloading full-texts often requires following several links, which is difficult to automate. Furthermore, scripted retrieval of literature is often against journals' terms of service, and may lead to journals blacklisting IP addresses if attempted (Elsevier, 2019).

This is also the case for Google Scholar, which includes full-text links for 54.6% of its indexed articles (Martín-Martín et al., 2018). However, Google Scholar disallows scripted access, and provides no public API as of 2019 (Martín-Martín et al., 2018). Some databases provide public APIs for scripted access, such as the cross-ref API for general articles, and the Entrez API for articles in Pubmed Central. These API provide fair coverage of full-texts for general use. Unfortunately, systematic reviews need to consider a wide range of literature, including literature from e.g. small publishers, which are less well covered by these API. The average coverage for studies in Cochrane DTA systematic reviews is less than 10% (Norman et al., 2018a, 2019e).

Some reference managers, such as EndNote have full-text retrieval functions and can retrieve articles from several databases semi-automatedly. This retrieval still requires manual input, and is rate-limited, but is able to retrieve up to 38% of references, or 58% of the references with DOI or PMIDs. While low, we are not aware of more reliable automated methods for full-text retrieval.

While journal and publisher imposed restrictions present barriers to automated retrieval of full-text articles, these generally provide alternative access mechanisms that can be used by e.g. EndNote. Articles from obscure sources may provide a more formidable obstacle, since there generally is no alternative access route not in-

volving detective work. Fully automating this process would require autonomous agents able to contact authors independently (Tsafnat et al., 2014).

### 3.2.5 *Screen Full Text*

In the conventional systematic review screening process, references are first included based on titles and abstracts only. Abstracts often omit crucial information necessary to make final decisions, and this initial screening therefore often consists of removing obviously non-relevant studies. This generally results in the initial screening being over-inclusive (O'Mara-Eves et al., 2015).

This separation of the screening into two stages is however merely a result of the human screening process, and the difficulty of performing full-text retrieval. If the full-text retrieval was adequately automated, we would be better off performing automated screening directly on the full-texts.

For the most part, screening automation for full-texts is conceptually identical to screening based on abstracts and titles, except that full-texts often includes additional information not available in abstracts, such as figures, tables, and references. There are however a few important addition differences:

First, abstracts are virtually never encumbered by copyright or other restrictions for public distribution. Full-texts, by contrast, are usually copyrighted. Consequently, there are several datasets available to train models for title/abstract screening across a range of domains (Alharbi and Stevenson, 2019; Cohen et al., 2006; Kanoulas et al., 2017a, 2018; Norman et al., 2018c). We are aware of no publicly available datasets for full-text screening. While we could envision such a dataset made publicly available if restricted to only open access articles with permissive licences, such a dataset may not represent the range of studies encountered in a real systematic review, and would therefore be biased.

Second, full-text articles include sufficient information, and sufficient details to enable data extraction. The extracted data items are typically the final arbiter for inclusion in the review. Consequently, if it were possible to extract data automatically from the articles, this could simplify, and possibly trivialize the automated screening process. Screening would simply consist of a filter selecting those articles that, e.g. evaluated the relevant index test against appropriate reference standards. Unfortunately, to make this approach realistic we would need to not only automate the data extraction to a high degree of accuracy, but also the full-text retrieval stage. This is not realistic at present.

36

Arnold 2001a		
Study characteristics		
Patient sampling	Sample size:	61
	Females:	Not stated
	Age:	Not stated
Patient characteristics and setting	Patients with potentially resectable pancreatic adenocarcinoma (after scan)	
	Setting:	Germany (setting not clear)
Index tests	Diagnostic laparoscopy	
	Criteria for positive diagnosis:	Biopsies of lesions suspicious of metastases
Target condition and reference standard(s)	Target condition:	Unresectability
	Reference standard:	Laparotomy for patients with no evidence of metastases on laparoscopy; biopsy with histopathological confirmation of spread for patients with suspected metastases
	Criteria for positive diagnosis:	Not stated
Flow and timing	Number of indeterminates for whom the results of reference standard were available:	Not stated
	Number of patients who were excluded from the analysis:	Not stated
Comparative		
Notes		

Table 3.1 – Example of the extracted data from one of the included studies in one systematic review on ‘accuracy of laparoscopy following computed tomography scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer’ (Allen et al., 2013).

3.2.6 Search Reference Lists

This step involves screening the references that have been cited by the included studies (backward search), or references that cite the included studies (forward search) (Lefebvre et al., 2019, in Higgins et al., 2019; Tsafnat et al., 2014). The forward citation search may be complemented with database alerts (also called litera-

ture surveillance services, push services or selective dissemination of information (SDI) to identify future candidate references that cite records included in the systematic review (Greenhalgh and Peacock, 2005; Lefebvre et al., 2019, in Higgins et al., 2019). This is sometimes called ‘snowballing’ (Greenhalgh and Peacock, 2005). Searching reference lists has been demonstrated to retrieve references not retrieved by database searches (Horsley et al., 2011), and have been reported to be useful for identifying high quality sources in obscure locations (Greenhalgh and Peacock, 2005). For e.g. complex interventions where relevant databases are difficult to determine at the protocol stage, following reference lists may retrieve as much as 51% of all relevant studies (Greenhalgh and Peacock, 2005). Searching reference lists is generally recommended for systematic reviews (Greenhalgh and Peacock, 2005) and required in Cochrane systematic reviews (Lefebvre et al., 2019, in Higgins et al., 2019). In practice, an estimated 92.2% of Cochrane reviews, and 64.3% of non-Cochrane reviews report following reference lists (Horsley et al., 2011).

37

### 3.3 WRITING THE SYSTEMATIC REVIEW

#### 3.3.1 *Extract Data*

Data extraction in systematic reviews refers to the identification of key characteristics of included primary studies, such as the methods used to perform the study, and the condition or population targeted (Li et al., 2019, in Higgins et al., 2019), but also involves producing assessments of the methodological quality of the included studies (Reitsma et al., 2008, in Deeks et al., 2013a). The data extraction stage is one of the more time-consuming stages of the systematic review process (Pham et al., 2018).

After a set of potentially included studies have been identified, systematic reviewers complete a so-called *data extraction form* for each study. These forms comprise a semi-structured summary of the studies, identifying and extracting a consistent, pre-specified set of data items from abstracts or full-text articles in a coherent format (tables 3.1 and 3.3). The coherent format allows the data from the studies to be synthesized qualitatively or quantitatively to address the research question of the review.

#### *Quality Assessment*

In the preface of his landmark review on scurvy, James Lind pointed out that ‘before this subject could be set in a clear and proper light, it was necessary to remove a great deal of rubbish’ (Lind, 1753). This is true more than 250 years later, and current

publishing trends in academia seem to only exacerbate the problem.

One example comes from genomics. Until the early 2000s, most studies of human genome epidemiology addressed a single or few genes at a time, and were performed by single teams, with small sample size (Chalmers et al., 2014; Ioannidis, 2016). This resulted in widespread publication bias, spurious conclusions, and lack of subsequent replication by consortia-based, genome-wide efforts (Chalmers et al., 2014). Meta-analyses collating such results almost always give statistically significant results, but this is simply due to the selective reporting of the included primary studies (Ioannidis, 2016). Such misleading results have been perpetuated in systematic reviews with insufficient quality control, many of which are still being published, and increasing in number (Chalmers et al., 2014; Ioannidis, 2016). The systematic review results in (typically) more reliable results than individual primary studies. Part of the reason is the greater amount of data available for analysis, which usually gives more reliable results with smaller confidence intervals. However, simply accumulating large amounts of data is not sufficient – if there is a consistent bias in the way results have been produced, the results of the meta-analysis may also be misleading.

This problem may be illustrated by the opposite conclusions of two systematic reviews comparing low molecular weight heparins and standard heparin in the prevention of thrombosis after surgery. One systematic review concluded that low molecular weight heparins were more effective than standard, while the other systematic review concluded no convincing evidence of a difference. The main difference between the two systematic reviews is that the former included all studies, whereas the latter only included studies of high quality (Egger and Smith, 1998; Leizorovicz et al., 1992; Nurmohamed et al., 1992).

It may be possible to mitigate the effects of low quality studies by excluding these, but even if all studies are included in the analysis and conclusions it is still recommended practice to summarize the quality appraisal of the included studies, to offer a general impression of the reliability of the available evidence (Leeftang et al., 2008). Overall scores are discouraged since different biases may generate different magnitudes, or directions of bias.

### 3.3.2 *Homogenize Data*

Studies included in a systematic review may exhibit large variations in terminology and reporting.

In some areas, language may be standardized with studies using consistent and widely understood terminology, allowing different studies to be easily compared using the labels and descriptions used in the study reports (McKenzie et al., 2019a, in Higgins et al., 2019). In many areas however, terminology may be variable, and



Figure 3.3 – Risk of bias assessments for the included studies in one systematic review on ‘accuracy of laparoscopy following computed tomography scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer’ (Allen et al., 2013).

it may be necessary to standardize the description of study characteristics across included studies to facilitate comparisons (McKenzie et al., 2019a, in Higgins et al., 2019).

Conventions in reporting of measurements may also differ in metrics, units, or formats and comparisons may therefore be facilitated by converting measures into canonical formats (Tsafnat et al., 2014).

### 3.3.3 *Synthesize Data*

40

Synthesis refers to the process of bringing together data sets of included studies with the aim of drawing conclusions about the body of evidence (Page et al., 2019b, in Higgins et al., 2019). Systematic reviews are also useful for investigating how scientific findings vary by particular subgroups (Leeflang et al., 2008), and therefore commonly perform several analysis, one for each group of interest (McKenzie et al., 2019a, in Higgins et al., 2019).

Included analyses should be reasonable in number, but should include all meaningful analyses. Systematic reviews of interventions should include both adverse and beneficial outcomes (McKenzie et al., 2019a, in Higgins et al., 2019).

Determining which studies are similar enough to be grouped into separate analyses is based on the data extraction in the previous stage. Groups in systematic reviews of interventions often consist of similar population groups, specific interventions, different control groups, or different outcomes (PICO). For instance, different groups of analysis could examine the effectiveness of an intervention at different time points or for different outcome measures. Groups in systematic reviews of diagnostic tests often consist of similar populations groups, index tests or reference standards. For instance, different groups of analysis could examine the diagnostic performance of a diagnostic at different thresholds (Leeflang et al., 2008).

How different studies will be grouped into individual analyses should to the greatest extent possible be specified in the protocol (McKenzie et al., 2019b, in Higgins et al., 2019). In practice, criteria may need to be adapted based on the evidence encountered, particularly for e.g. systematic reviews of complex interventions, where it may not be possible to define groups in the protocol (Greenhalgh and Peacock, 2005). Changes to the specifications of the groups could occur because methods for dealing with particular issues had not been identified at the time of writing the protocol, the literature search uncovered insufficient data for the analysis methods to be viable, or because preferable alternatives or more recent guidance were identified during the review (Page et al., 2019b, in Higgins et al., 2019). Planning contingencies for anticipated scenarios is however preferable to post-hoc decision making (McKenzie et al., 2019a, in Higgins et al., 2019). For transparency, any

changes between the protocol and the review must be described and justified in the review (Page et al., 2019b, in Higgins et al., 2019).

#### 3.3.4 *Re-check Literature*

A systematic review of interventions takes on average 67 months to complete, from publication of protocol to publication of the final review, and the vast majority of this time is spent on the stages prior to the meta-analysis (Allen and Olkin, 1999; Borah et al., 2017). Additional relevant studies are therefore likely to have been published after the initial database search were performed. A second search may therefore be necessary to identify these additional references.

This second search is however entirely caused by the necessarily slow manual screening process, and would be entirely unnecessary if the screening process was automated.

#### 3.3.5 *Meta-Analyze*

When the results of the individual primary studies are summarized, but not combined statistically, this is called a qualitative systematic review (Cook et al., 1997). In a quantitative systematic review the statistical data is combined in a so-called meta-analysis. The meta-analysis process is not unique to systematic reviews – some of the first known meta-analysis work was done by Karl Pearson in 1904 to combine data of the effectiveness of enteric fever inoculations from multiple military bases in South Africa and India (Simpson and Pearson, 1904). The data from the multiple studies were provided – not by the published literature – but by the British Army. Neither is the term ‘meta-analysis’ of medical origin – it was coined by American social scientist Gene Glass in 1976 (Glass, 1976).<sup>1</sup>

A systematic review seeks to answer a research question. How it will do this will depend entirely on the nature of the question. Broadly speaking, we can divide the methodology into quantitative and qualitative analyses.

A *qualitative* analysis is often undertaken if the included studies are heterogeneous. For instance, if the included studies report different measures it may not be possible to compare the studies any other way than listing their individual findings and discuss them.

A *quantitative* analysis, more commonly called a *meta-analysis* is usually undertaken if the studies report the same or similar outcomes, using the same measures, have been conducted in the same settings, et c..

<sup>1</sup> In fact, research synthesis methods were pioneered by American social scientists – Glass one of them – in the 1970s, with the medical field only starting to take significant interest in the late 1980s (Chalmers, 2003)

The meta-analysis requires systematic reviewers to make two judgements, choice of method, and choice of data. Once it has been decided what data should be synthesized, the remainder of the process is entirely deterministic. For systematic reviews of interventions, calculating analyses using the mixed random effects model can be performed entirely within RevMan. For systematic reviews of diagnostic test accuracy, the equivalent functionality is not available within RevMan, but only in external statistics software.

### 3.3.6 *Write Review*

42

A systematic review follows the same procedures as any scientific inquiry, and the publication of results largely mirror the publication of any scientific findings (Chalmers, 2003). Structured and transparent reporting is essential for any systematic review (Page et al., 2019a, in Higgins et al., 2019).

The target audience for a systematic review are a range of decision makers, including healthcare professionals, consumers, and policy and guideline developers. Thus systematic reviews are frequently read by stakeholders with a basic sense of the topic, who may not necessarily be experts in the area (Page et al., 2019a, in Higgins et al., 2019). Systematic review report usually aim for the same style and language as primary research in their topic, but Cochrane systematic reviews commonly include layman's summaries to make the findings available for non-experts.

#### SUMMARY

SYSTEMATIC REVIEWS FOLLOW A SYSTEMATIC PROCESS with rigorous checks and balanced to minimize errors and sources of bias

THE SYSTEMATIC REVIEW PROCESS IS ALMOST ENTIRELY MANUAL, with most steps involving substantial human input and oversight. Several steps are performed in parallel by multiple review authors to reduce bias

SEVERAL STAGES HAVE VERY HIGH WORKLOAD, and commonly take months or years to complete. Particularly the screening and data extraction stages involves substantial human effort

**T**HE LARGE AND GROWING number of studies published every year is making it ever harder to identify studies meeting the inclusion criteria in systematic reviews in an unbiased way. The median delay between the publication of the protocol and the final review for a Cochrane systematic review is two and a half year, with some taking over seven years. There is both a need to decrease the workload, as well as to improve the timeliness of published systematic reviews.

Screening automation has been proposed as a potential solution. By using methods from natural language processing, information retrieval, or related fields, it may be possible to decrease the amount of work that needs to be performed manually. Automated screening systems may automatically exclude some of the candidate references or decrease the amount of time that need to be spent manually reviewing each reference. Automated data extraction systems may automatically extract information from published articles or aid extractors in finding relevant information in the text.

The purpose of this section is to give an overview of the previous literature of screening automation. This overview of the literature is largely based on the literature identified in the 2015 systematic review by O'Mara-Eves et al., where we repeat the search queries to identify additional studies published between 2015–June 2019. Additional studies were also identified by searching reference lists. Some studies became known to us through the course of this project.

We will first (section 4.1) give a brief overview of relevant text mining methods. In section 4.3 we will summarize the identified literature and the types of approaches and methods that have been addressed by previous literature. In section 4.4 we will summarize the metrics and evaluation methods that have been used by previous literature. In section 4.5 we will summarize the publicly available datasets that have been used by multiple authors previously. In the final section (4.6), we will list the identified studies, with a brief summary of each study.

Several of the studies presented in this thesis (Norman et al., 2017b, 2018b,c, 2019f) are eligible for inclusion in this overview, but were omitted from this list. These will be presented in parts II and III. One of our earlier studies (Norman et al., 2017c) was omitted from this thesis, but is listed in the list of publications (appendix A). In particular, with this overview we attempt to address the following questions:

1. What approaches and methods have been used by previous literature?

2. How have different approaches and methods been evaluated by previous literature?
3. What datasets are available for performance comparisons?

#### 4.1 TEXT MINING METHODS

Broadly, machine learning attempts to make predictions about future data by identifying structure in past data. Information retrieval seek to retrieve information relevant to a query posed by a user. Collectively, these approaches are sometimes referred to as text mining.

**SUPERVISED LEARNING** is a machine learning approach that attempts to construct mathematical models from sets of data composed of the desired input and output pairs. For instance, a spam filter can be trained by showing example spam emails labeled as ‘spam’, and non-spam emails labeled as ‘non-spam’. The model should then be able to independently distinguish between spam and non-spam in future emails, even if these new emails use different language.

**CLASSIFICATION** is a form of supervised learning where the desired output are discrete classes, e.g. spam / non-spam emails, relevant / non-relevant documents, or positive / neutral / negative sentiments.

Examples of classification methods include support vector machines (svm), naive Bayes, voting perceptrons, decision trees, evolutionary svm, WAODE, Rocchia, generalized linear models, neural networks, gradient boosting machines, and random forests.

Many of these methods are available in publicly available software packages, including LibSVM,<sup>1</sup> SVM<sup>light</sup>,<sup>2</sup> scikit-learn (Pedregosa et al., 2011), or WEKA.<sup>3</sup>

**REGRESSION** is a form of supervised learning where the desired output are continuous values in a range, e.g. temperature, length, or probability. Examples include support vector regression (svr), logistic regression, and neural networks.

**UNSUPERVISED LEARNING** is a machine learning approach where no examples of desired output are provided, and the model will seek to discover structure in the input data autonomously. Clustering is one of the primary types of unsupervised learning, where machine learning model attempts to identify groups or clusters of objects that share similarities. For instance, clustering could be used in market

---

<sup>1</sup> <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

<sup>2</sup> <http://svmlight.joachims.org/>

<sup>3</sup> <https://www.cs.waikato.ac.nz/ml/weka/>

research to group similar customers into market segments. Examples of clustering methods include kNN and latent Dirichlet allocation (LDA).

SEMI-SUPERVISED LEARNING is a related approach that used both labeled and unlabeled training data. The approach is typically used when a large amount of relevant unlabeled training is available but labeling large amounts of these examples is expensive. Semi-supervised learning can then learn the target concept from the labeled training data, while simultaneously learning from the structure in the unlabeled data.

TRANSFER LEARNING is an approach that is commonly used where labeled training data is scarce, but where labeled data for a related concept is abundant. A model can then be trained to recognize the related concept in order to apply this knowledge to the target concept. The model can then be fine-tuned with training examples of the target concept. For instance, if we want to construct a model to recognize hand-written characters but lack appropriate training data, we could first try a model trained using printed characters.

RANKING is a data science problem which seek to produce and ordering of items according to some criterion. Thus the model is provided a list of items, and should produce and ordering of the items. Ranking is central to information retrieval, where this is used to obtain items (commonly documents) relevant to an information need, but it also central to machine learning, where it can be used e.g. for job scheduling to prioritize high-priority tasks over low priority ones.

Information retrieval approaches to ranking commonly use unsupervised models of document similarity to retrieve documents similar to a provided search query. Conceptually, both the candidate documents and the query are converted into vectors (often using term or word count statistics) and a similarity score is then calculated between the query and each candidate document. The candidates can then be ranked by the similarity score. Examples of document similarity measures include cosine similarity and BM25.

Ranking models can also be learned with supervised learning, i.e. by providing a machine learning model with training examples of desired orderings (Fuhr, 1992). This is called learning-to-rank.

In the most straightforward approach to learning-to-rank, a model is trained to estimate the relevance of each candidate reference (*pointwise learning*). This is thus a form of probability regression. The objects can then be sorted based on the assigned scores.

It is possible to train models to produce rankings without explicitly assigning a score to each object. The goal of the training in this case is to minimize the

number of inversions: the number of pairs that appear in the wrong order. This is known as ordinal regression, and can be done using machine learning methods by training on pairs of references (pairwise training) or on an entire list of references (listwise training) (Burges, 2010). Examples include RankNet, LambdaRank, and LambdaMART.

ACTIVE LEARNING is an iterative approach which is commonly used where unlabeled data is abundant, but manual labeling is expensive. The model then incrementally queries a human teacher for additional labels, often with a strategy to prioritize new training examples that would be the most informative. By preferentially querying for labels to data items that are the most important to learn the target concept, active learning often uses much fewer training examples than normal supervised learning.

46

#### 4.1.1 Feature Representations

Machine learning models are mathematical, and assume numerical inputs. Most of the real-world objects we want to identify structure in are not inherently numerical. Our first step is therefore to convert the objects – images, sound files, text, et c. – into mathematical representations. The components of these representations are conventionally called features in machine learning (or variables in the related field of statistics).

Text data is commonly converted into numerical representations by replacing each word with a number, with the same number for the same word. Commonly, words are first replaced with their *lemmata*,<sup>1</sup> the canonical dictionary form of the word. In this way *do*, *did*, *does* will each be turned into *do* before being turned into numbers, and will therefore have the same numerical representation. A simpler and less expensive alternative to lemmatization is stemming, where the ending of the word is truncated using simple rules. Stemming may give different results from lemmatization, in particular for morphologically rich languages. For instance the word *meeting* has the the stem *meet-*, but the lemmata *meet* or *meeting* depending on whether it is used as a verb gerund or a noun. Stemming and lemmatization also give different results for a small number of homographs with different lemmata.<sup>2</sup> Stemming is often sufficient, and typically produces similar results to lemmatization.

One of the simplest and most common feature representations for textual data is the bag-of-words (BOW) model. In natural languages word order matters: *a ma-*

---

<sup>1</sup> Singular: lemma, plural: lemmata

<sup>2</sup> I.e. words which happen to be written the same way but have different dictionary forms. The author is aware of five such homographs in English: *bustier*, *does*, *evening*, *moped*, and *number*

*chine learning model is mathematical* has a different meaning than *model is learning a mathematical machine*. The bag-of-word model takes sentences and simply notes which word occur in the sentence or document, and – optionally – how many times. The two examples sentences above would therefore have the same bag-of-word representation. The bag-of-word representation thus ignores finer nuances of meaning, but it is often sufficient for ranking or classification.

If the bag-of-word representation simply records the presence or absence of individual words, we call it a binary representation. If the bag-of-word representation records the frequency of the word in the text snippet, we call it a frequency representation. Often, frequency representations are normalized to give greater weight to words that are infrequent in the language, under the assumption that such words are more likely to be salient. One common such normalization is called *tf-idf*, which is defined as the term frequency in the sentence (*tf*) divided by the logarithm of the inverse document frequency (*idf*), the number of text snippets it does not appear in.

AN *N*-GRAM is a continuous sequence of *n* words in a sentence.<sup>1</sup> Lower order *n*-grams are commonly referred to by latin prefixes, i.e. *unigram* for 1-gram, *bigram* for 2-gram, and *trigram* for 3-gram. The bag-of-words representation may contain individual words, i.e. unigrams, but higher order *n*-grams are commonly used in order to capture more complex semantics, such as technical terms, expressions, or limited forms of word order.

## 4.2 AN OVERVIEW OF SCREENING AUTOMATION METHODS

### 4.2.1 Database Searches

To perform this overview, we attempted to repeat the systematic review performed by O'Mara-Eves et al. (2015). O'Mara-Eves et al. searched 19 databases for candidate references. Searching 19 databases is however outside the scope of this review, and we therefore focused on searching PubMed, the ACM Digital Library, and IEEE. We attempted to use the same database query as O'Mara-Eves, without modifications if possible. However, the query syntax used in ACM Digital Library has been updated since 2014, and currently does not support searching all fields. As a compromise, we search in abstracts only.

O'Mara-Eves et al. limited inclusion to studies published between 2005 and February 2014. The earlier limit was chosen because the first proposed application of natural language processing in systematic reviews was reportedly published in 2005.

<sup>1</sup> Or more generally, a continuous subsequence of items in a sequence. Character *n*-grams, i.e. sequences of *n* strings of characters in a word are also commonly used in natural language processing.

We similarly limited the search to studies published between January 2005 and June 2019, dividing the references into references published before and after December 31, 2013. We used this date as the pivot, rather than February 28, because we do not have the exact date of publication for all references, only the year. The database search was performed in June 2019. Unlike O'Mara-Eves et al., who used two reviewers to screen titles and abstracts, and one reviewer to screen full-texts, we let a single reviewer screen both stages (CN).

To limit the number of references to screen, we used a screening reduction model (Norman et al., 2019a). As training data we used all references identified through the database search that were published between January 1, 2005 and December 31, 2013, i.e. those that would have been eligible for inclusion in the Systematic review by O'Mara-Eves et al. We manually added the 43 studies included by O'Mara-Eves et al. as positive training data. Three studies included by O'Mara-Eves et al. were published in 2014 and were included in the training set. We removed duplicates, and labelled all other studies in the training set as negative.

To evaluate the performance of the method, and to establish an acceptable cut-off, we first measure the simulated performance using cross-validation on the training set. To evaluate whether abstracts are necessary for judgments, we repeated the evaluation with and without abstracts.

#### 4.2.2 Inclusion Criteria

We included studies meeting the same inclusion criteria as used by O'Mara-Eves et al., divided into a two stage screening process as follows. References were first included based on titles and abstracts using the following criteria:

1. Must be published after 2004
2. Must be relevant to natural language processing
3. Must be relevant to the screening (document selection) stage of a systematic review (or a review of the evidence that follows systematic principles, such as health technology assessment (HTA) or guidelines development)

The following criteria were used for full-text screening:

1. Must be relevant to natural language processing methods or metrics
2. Must be relevant to the screening stage of a systematic review (or similar evidence review)

3. Must not be a general discussion of the use of natural language processing in systematic reviewing screening. That is, the record must present a detailed method or evaluation of a method.

#### 4.2.3 Results

A total of 1,265 references remained after deduplication, with 770 references published between 2005–2014, which should have been included in the systematic review by O'Mara-Eves et al., and 493 references published in 2015 or later. Based on the cross-validated results, we screened the first 100 references. This overview identified 73 individual publications, of which 33 were included in the 2015 O'Mara-Eves et al. review.

#### 4.3 SUMMARY OF SCREENING AUTOMATION APPROACHES

Screening automation is an umbrella term for several disparate approaches with the common goal of reducing the workload during the screening stage in systematic reviews (O'Mara-Eves et al., 2015). These include using classification or ranking methods to automatically exclude non-relevant records (screening reduction), using natural language processing methods to aid judgements and thereby increase the rate of screening (visual text mining (VTM)), using classification or ranking methods instead of a second screener (automation as a second screener), or using screening prioritization to identify relevant records earlier in the process (screening prioritization) (O'Mara-Eves et al., 2015). Methods have been used from a range of fields, including natural language processing, machine learning, information retrieval and statistics (Beller et al., 2018; O'Mara-Eves et al., 2015; Tsafnat et al., 2014). Of these, screening prioritization is the only approach that does not automatically exclude references, and is therefore the only approach considered safe for use in systematic reviews.

SCREENING REDUCTION methods work by using automated methods to reduce the number of studies that need to be manually screened. The workload reduction in this approach comes from the reduction in number, not the order in which references are screened. Once the number is reduced, screening may proceed in any order.

The first screening reduction approach is to use a classification algorithm, where the algorithm is trained to explicitly model binary include/exclude decisions.

The second screening reduction approach is to use a ranking algorithm, where a regressor is trained to model the probability of inclusion/exclusion (Fuhr, 1992). All items falling below some threshold are then excluded from consideration (O'Mara-

Eves et al., 2015).

The main difference between a classifier and a regressor with a cut-off is how the two are trained: the classifier is typically trained to minimize the number of misclassifications, i.e. the optimal placement of the classification boundary, whereas the regressor is trained to model the scores of each item in the list. The two may therefore give slightly different results. The classifier does not need to care about the ordering of items on each side of the classification boundary. Conversely, the regressor may produce a ranking with a sub-optimal classification boundary if this gives a better overall ordering of items.

50

SCREENING PRIORITIZATION similarly uses ranking to reduce the workload, but the primary intent is to change the order of screening, so that relevant records are screened before non-relevant ones. The number of records to screen can be reduced by combining screening prioritization with a cut-off threshold. The main difference compared to screening reduction is that screening prioritization does not add an extra filtering step before the screening commences, but rather modifies the screening process to screen in descending order of likelihood of relevance.

AUTOMATION AS A SECOND SCREENER does not attempt to reduce the number of references that need to be screened, but rather to avoid having each reference screened by multiple screeners. In the conventional process, each reference is screened by at least two screeners (cf. section 3.2.3). This is to ensure the reproducibility and consistency of the results, as well as to avoid relevant studies ‘slipping through the net’. A single screener could introduce bias to the process due to their interpretation of the inclusion criteria or their understanding of the titles and abstracts (O’Mara-Eves et al., 2015). It is believed that if at least two screeners apply inclusion criteria consistently, then the process is unlikely to be biased (O’Mara-Eves et al., 2015).

Seven papers have advocated partially replacing one of the screeners with an automated system (Bekhuis and Demner-Fushman, 2010, 2012; Bekhuis et al., 2015, 2014; Frunza et al., 2010, 2011; García Adeva et al., 2014). In this approach, a single human reviewer screens all references, and an automated system works as a check that the included studies are consistent and no studies have been overlooked.

All systems evaluated in terms of recall and precision of included studies. In other words, all the system attempted to classification methods to identify relevant studies, rather than attempt to identify misclassifications by the first screener.

AUTOMATED IDENTIFICATION OF TRIAL REGISTRATIONS has been proposed by one study (Martin et al., 2019; Surian et al., 2018). The authors used matrix factorization combined with PCA and LDA to identify trial registrations in ClinicalTrials.gov.

They report a 60.9% recall at 100 screened, and propose that the method may provide an inexpensive and useful test to identify when a sufficient number of new studies have been planned to warrant updating a review.

However, the authors reported optimal  $wss@95$  scores of 99.4% which was worse than the  $tf \cdot idf$  similarity baseline (99.5%), and it is unclear what benefits the method provide over standard similarity measures.

The same group of authors (Dunn et al., 2018) similarly used machine learning to identify trial registration at ClinicalTrials.gov where such links were missing, reporting 86% recall at 50 references screened. Approximately 45% of published trials had unreported links.

VISUAL TEXT MINING methods attempt to present a graph of the connections between documents, with connections between similar documents (e.g. by document similarity or author connections). Hypothetically, this approach may allow screeners to more quickly locate similar documents, and conversely, to more quickly exclude documents dissimilar from clusters of relevant studies. All candidate references are screened – the goal of the approach is to be able to make judgements more quickly based on the visual locations of the studies in the graph in addition to abstract and title information.

Five papers have advocated this approach, all in software engineering (Felizardo et al., 2012a,b, 2011, 2013; Malheiros et al., 2007). These results suggest that reviewers can screen more quickly using this approach as with randomized order with similar accuracy (O'Mara-Eves et al., 2015).

OTHER AUTOMATED APPROACHES have also been proposed to improve the screening process. Wallace et al. (2010a) proposed *efficient citation assignment*, where an active learning model will model both relevance to the research question as well as the expected time it will take to screen individual references. By purposefully assigning easier references to junior screeners, and more difficult references to senior screeners, the authors report that more references could be screened in the same amount of time compared to conventional active learning approaches. Bannach-Brown et al. (2019) proposed to use machine learning models for *error analysis* of screened references. They applied the machine learning models on the references screened by human experts in order to identify instance where references had been inadvertently included or excluded. In this way, they identified 11 records that had been wrongly included, and 36 records that had been wrongly excluded.

#### 4.4 METRICS USED IN PREVIOUS LITERATURE

RECALL, SENSITIVITY OR TRUE POSITIVE RATE (TPR) is the percentage of relevant items identified by the model:

$$\frac{TP}{TP + FN}$$

PRECISION is the percentage of relevant items among the retrieved items:

$$\frac{TP}{TP + FP}$$

SPECIFICITY OR TRUE NEGATIVE RATE (TNR) is the percentage of non-relevant items correctly excluded:

$$\frac{TN}{TN + FP}$$

ACCURACY is the percentage of correctly classified items (both relevant and non-relevant):

$$\frac{TP + TN}{TP + TN + FP + FN}$$

F MEASURE is also known as the F score or the  $F_1$  measure, and is defined as the harmonic mean of recall and precision:

$$2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

This metric place equal importance on precision and recall, and is therefore of limited usefulness in systematic reviews, where recall is much more important than precision. Alternatives have been proposed that give more weight to recall, including the  $F_3$  score which gives recall three times the weights of precision (Bekhuis and Demner-Fushman, 2010).

ROC (AUC) is the area under the curve tracing the sensitivity against the specificity. A AUC equal to 1 is a perfect ordering, 0.5 a random ordering, and 0 is a perfect reverse ordering.

WORK SAVED OVER SAMPLING is the percentage of references not needed to screen to retrieve  $100-\alpha\%$  of the included, compared to simply sampling  $100-\alpha\%$  of the candidate references:

$$\text{wss}@\alpha = \frac{\text{TN} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} - \alpha$$

wss@95 was proposed by Cohen et al. in the earliest known example of screening automation (Cohen et al., 2006). The measure was later abandoned by the same author in favor of AUC since it fails to capture different recall requirements in different reviews.

Wss is relatively easy to interpret in terms of systematic review impact. Unfortunately, since the measure depends on the position of a single included study at the end of the ranking, the measure is strongly influenced by random effects, and therefore tend to have large variance.

A number of other performance metrics have been used in previous literature, including time, burden, yield, utility, baseline inclusion rate, performance, coverage, unit cost, classification error, error, absolute screening reduction, prioritized inclusion rate, average precision (AP), cumulative gain (CG), discounted cumulative gain (DCG), normalized discounted cumulative gain (NDCG), reciprocal rank, loss<sub>R</sub>, loss<sub>E</sub>, and reliability (Kanoulas et al., 2017a; O'Mara-Eves et al., 2015). Many of these have been used infrequently, and can therefore not be used for comparisons, or are unsuitable to measure performance of screening automation, and we will therefore not cover them in this section.

In particular, many metrics make assumptions about utility that are wrong or detrimental when used for systematic review screening. In a systematic review, false positives and false negatives are not associated with the same cost (Wallace et al., 2010b). Including an extra study typically leads to 30–60 seconds higher workload for the screeners, whereas missing a relevant study runs the risk of invalidating the results of the review and potentially recommending diagnostic tests or treatments that are poor or harmful to patients.

Several commonly used metrics, including the  $F_1$  measure and AUC, give equal weight to false positives and false negatives, and are therefore poorly suited to evaluate screening automation methods. Alternatives have been proposed to counter this, including the  $F_3$  score which gives recall three times the weights of precision (Bekhuis and Demner-Fushman, 2010), and the  $U_{19}$  score which gives recall 19 times the weight of the workload (Wallace et al., 2010b,c).

Similarly, the majority of conventional information retrieval metrics (AP, CDG with variants, reciprocal rank, loss, reliability) are predominantly 'top-heavy.' That is to say, they try to measure the ability of the retrieval methods to fill the top of the list with relevant items, rather than its ability to find all relevant items. This

typically works well for web searches, where users are unlikely to examine the retrieved results beyond the first few pages, but may be disastrous in the context of screening automation, where perfect recall is expected.

#### 4.5 AVAILABLE DATASETS

In one of the earlier papers on the subject, Cohen et al. (2006) constructed the Drug Effectiveness Review Project dataset (DERP) from 15 reviews on drug efficacy. This dataset was later extended to 18 (Cohen et al., 2010), and then to 24 reviews (Cohen et al., 2009). The smaller dataset comprising 15 reviews has been made available (Cohen et al., 2006)<sup>1</sup>. Several methods have been tested on this dataset (table 4.1), and this dataset therefore constitute among the closest things we have to a standard dataset for comparison of performance.

There are a few inconsistencies in the dataset that are not explained:

1. Over 1,000 references are included in more than one review topic. Some of these have different labels in different topics. Consequently, the dataset has a non-trivial amount of topical overlap, which might help intertopic training, since some of the references will be included in both the training and test sets.
2. In the dataset, 2,150 references have no abstracts in PubMed. It is unclear whether these should be treated the same way as the other references.
3. One reference is not indexed in PubMed (PMID 12168612).

The second broadly used dataset is the CLEF eHealth dataset of DTA systematic reviews. This data was distributed as part of CLEF eHealth Task 2: Technology Assisted Reviews in Empirical Medicine, and has therefore been used for evaluation by a number of studies. Note that the task formulation changed between 2017 and 2018. Only the studies included in the second stage were considered relevant in the first iteration of the CLEF shared task (Kanoulas et al., 2017a). In the second iteration this decision was inverted, and instead all studies included in the first stage were considered relevant (Kanoulas et al., 2018).

A shared task is a community challenge where participating systems are trained on the same training data, and evaluated blindly using pre-decided metrics (Chapman et al., 2011; Huang and Lu, 2015). The shared task model removes many of the problems inherent in performance comparisons, and normally serves to safe-guard against cheating, mistakes, the cherry-picking of metrics or data, as well as publication bias. The CLEF task diverged from standard practice for shared tasks by

<sup>1</sup> The old link has however expired. The data can now be found at <https://dmice.ohsu.edu/cohenaa/systematic-drug-class-review-data.html>

Dataset	Study
Drug Class Efficacy Project (DERP) Topic: <i>Interventions (drugs), various topics</i>	[1] Cohen et al., 2006 [6] Martinez et al., 2008 [4] Cohen, 2008 [9] Cohen et al., 2009 [17] Matwin et al., 2010 [15] Bekhuis and Demner-Fushman, 2010 [2] Cohen et al., 2010 [26] Choi et al., 2012 [33] Jonnalagadda and Petitti, 2013 [48] Khabsa et al., 2016 [47] Howard et al., 2016 [50] Ji et al., 2017 [73] Olorisade et al., 2019
TrialStat (Kouznetsov et al., 2009; Razavi et al., 2009) Topic: <i>The dissemination strategy of health care services for elderly people of age 65 and over</i>	[10] Kouznetsov et al., 2009 [8] Razavi et al., 2009 [11] Kouznetsov and Japkowicz, 2010 [13] Frunza et al., 2010 [23] Frunza et al., 2011
Chronic Obstructive Pulmonary Disease (COPD) (Castaldi et al., 2009; Wallace et al., 2011) Topic: <i>Genetic associations with COPD</i>	[25] Wallace et al., 2011 [14] Wallace et al., 2010c [25] Wallace et al., 2011 [39] Miwa et al., 2014
Proton Beam (Terasawa et al., 2009; Wallace et al., 2010c) Topic: <i>Charged particle radiation therapy for cancer</i>	[14] Wallace et al., 2010c [12] Wallace et al., 2010b [20] Wallace et al., 2010a [39] Miwa et al., 2014
Micro Nutrients (Chung et al., 2009; Wallace et al., 2010c) Topic: <i>Associations of micronutrients and disease</i>	[14] Wallace et al., 2010c [12] Wallace et al., 2010b [39] Miwa et al., 2014
CLEF eHealth (Kanoulas et al., 2017b, 2018) Topic: <i>DTA studies, various topics</i>	[37] Cormack and Grossman, 2017 [52] Van Altena and Olabarriaga, 2017 [53] Chen et al., 2017 [61] Anagnostou et al., 2017 [54] Lee, 2017 [55] Di Nunzio et al., 2017 [57] Scells et al., 2017 [58] Alharbi and Stevenson, 2017 [59] Kalphov et al., 2017 [60] Singh et al., 2017 [38] Cormack and Grossman, 2018 [63] Minas et al., 2018 [68] Wu et al., 2018 [69] Cohen and Smalheiser, 2018 [56] Di Nunzio et al., 2018 [70] Alharbi et al., 2018

Table 4.1 – List of publicly available datasets used by more than one study. Adapted from Olorisade et al. (2016), with subsequent studies added.

not blinding participants to the gold standard data used for evaluation (see section 7.5.5). This may have been a cause of biases in the reported results, and the results should be interpreted with this in mind. Even so, the CLEF dataset is still useful to compare results across different methods and implementations

In addition to the DERP and CLEF datasets, several studies have used the TriaStat, COPD, Proton Beam, and Micro Nutrient datasets, although many of these studies have been conducted by the same groups of people, using similar methods (see table 4.1).

#### 4.6 LIST OF INDIVIDUAL STUDIES

**1 (Cohen et al., 2006):** Cohen, A. M., Hersh, W. R., Peterson, K., and Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219

**2 (Cohen et al., 2010):** Cohen, A. M., Ambert, K., and McDonagh, M. (2010). A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. *AMIA Annual Symposium Proceedings*, 2010:121–125

The first known use of machine learning for screening automation was done by Cohen et al. in 2006. The authors constructed a dataset from 15 systematic reviews of drug class efficacy, later to be called the Drug Evaluation Review Project (DERP).<sup>1</sup> The same paper also introduce the wss metric.

The authors applied a standard machine learning classifier (voting perceptron) in order to retrieve references that are topically relevant and high quality. Reported wss@95 ranged from 0% to 67.95%.

This paper appears to be earliest mention of the common assumption that 95% recall is sufficient for systematic reviews, with the rationale: ‘For this study, we assumed that a recall of 0.95 or greater was required for the system to identify an adequate fraction of the positive papers. Precision should be as high as possible, as long as recall is at least 0.95.’

In a followup study [2], Cohen et al. used the same dataset extended to 18 reviews. This extended data does not appear to have been published. Apart from the larger dataset, this study uses the same method as in the previous study [1].

<sup>1</sup> <https://dmice.ohsu.edu/cohenaa/systematic-drug-class-review-data.html>

3 (**Ma, 2007**): Ma, Y. (2007). Text classification on imbalanced data: Application to systematic reviews automation. Master's thesis, University of Ottawa (Canada)

The first known use of active learning for systematic reviews was done in the 2007 master thesis by Ma. Ma used 14,276 references from previous systematic reviews of nutrition and diet interventions for heart disease and stroke. The author used Complement Naive Bayes with Bi-Normal Separation for feature selection, and clustering based sample selection to deal with class imbalance, and reported a wss@100 of 53.4% on the dataset.

4 (**Cohen, 2008**): Cohen, A. M. (2008). Optimizing feature representation for automated systematic review work prioritization. *AMIA Annual Symposium proceedings*, pages 121–5

5 (**Cohen, 2011**): Cohen, A. M. (2011). Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association : JAMIA*, 18(1):author reply 104

In these studies, Cohen, evaluated the use of svm with  $n$ -grams ( $n < 5$ ), MeSH terms, and UMLS. All features were found to be useful for the classifier, except the UMLS features.

The authors also reported that intratopic classification yielded better results than intertopic classification, and the author conclude that topic specific training data is necessary for high performance.

The results from the paper was later used as comparisons by Matwin and Sazonova [19] and Khabsa et al. [48].

6 (**Martinez et al., 2008**): Martinez, D., Karimi, S., Cavedon, L., and Baldwin, T. (2008). Facilitating biomedical systematic reviews using ranked text retrieval and classification. In *Australasian Document Computing Symposium (ADCS)*, pages 53–60

In this study, the authors used SVMs (WEKA) to rank references from 17 unspecified AHRQ systematic reviews to screen for relevant and high quality studies. They report a wss@95 around 30%. The authors also reported that the searches in some systematic reviews were not reproducible, and either were not syntactically correct, or did not yield the same results as those reported in the systematic reviews.

7 (Yu et al., 2008): Yu, W., Clyne, M., Dolan, S. M., Yesupriya, A., Wulf, A., Liu, T., Khoury, M. J., and Gwinn, M. (2008). GAPscreener: An automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*, 9:205

In this study Yu et al. used SVMs (LibSVM) with standard bag-of-word features. They reported 97.5% recall, 98.3% specificity, and 31.9% precision. The training set was constructed artificially by selecting 10,000 known positive references from a bibliographic database (HuGE Navigator), and 10,000 random articles from PubMed. The test set was constructed prospectively from references indexed in PubMed over 5+4 consecutive weeks, and manually screened for inclusion in the database. Features were constructed using the 'z-score' method, which uses bag-of-word representations of reference data, with each term weighted according to its normalized z-score, calculated by comparing 10,000 positive references against 10,000 random articles in PubMed. The authors reported that the two-way method improves overall performance.

8 (Razavi et al., 2009): Razavi, A. H., Matwin, S., Inkpen, D., and Kouznetsov, A. (2009). Parameterized contrast in second order soft co-occurrences: a novel text representation technique in text mining and knowledge extraction. In 2009 *Ieee International Conference on Data Mining Workshops*, pages 471–476. IEEE

In this study, Razavi et al. constructed two models to rank candidate references in the TrialStat dataset. The first model used common bag-of-word representations, while the second used second-order soft co-occurrence. The method used to rank references is not clear from the paper.

They used a bag-of-word representation model to identify the 700 references most likely to be positive and find that this set included 590 true positives. They used the second-order soft co-occurrence model to exclude the 8000 least likely candidates, and found that this set included 54 false negatives. In practice only the second model would be relevant in an automated screening scenario to exclude references automatically.

Their results would hence have resulted in a 37.7% workload reduction for ~97.4% recall.

9 (Cohen et al., 2009): Cohen, A. M., Ambert, K., and McDonagh, M. (2009). Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. *Journal of the American Medical Informatics Association*, 16(5):690–704

In this study, Cohen et al. extended the dataset used previously [1] to 24 review topics. The authors reported AUC scores using random stratified intratopic cross-validation (table 2, diagonals) and intertopic cross-validation (table 2, non-diagonals).

The authors used SVMs and trained in two steps. First the classifier was trained using intertopic training data. Then the support vectors from the first step are used as training data in the second stage, with intratopical training data added. The method is reported to improve on a baseline SVM when intratopical training data is scarce, and seems to achieve similar performance as training on intratopical data.

**10 (Kouznetsov et al., 2009):** Kouznetsov, A., Matwin, S., Inkpen, D., Razavi, A. H., Frunza, O., Sehatkar, M., Seaward, L., and O’Blenis, P. (2009). Classifying biomedical abstracts using committees of classifiers and collective ranking techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5549 LNAI:224–228

**11 (Kouznetsov and Japkowicz, 2010):** Kouznetsov, A. and Japkowicz, N. (2010). Using classifier performance visualization to improve collective ranking techniques for biomedical abstracts classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6085 LNAI:299–303

In these studies, Kouznetsov et al. evaluated screening prioritization on the Trial-Stat dataset. The authors of this study ostensibly tried to rank in terms of query relevance, but actually seems to have done final selection of references for the systematic review.

The authors describe a method to fuse multiple rankers into a single ranking decision, i.e. sensor fusion. They used Complement Naive Bayes, Discriminative Multinomial Naive Bayes, Alternating Decision Trees, AdaBoost with Logistic Regression, and AdaBoost with J48.

The authors reported 91.6% and 84.3% versus a reported 90–95% recall and 80–85% precision by human screeners. The authors for some reason do not report the results of the individual classifiers, and it therefore not at all clear from the paper how much the proposed methods adds over simply using any single classifier on its own.

The author also state that the committee performs ‘with a confidence level similar to human experts.’ However, the goal of screening automation is generally not to beat any one human screener, but to perform similarly to the human ensemble, and the human ensemble in the study achieved over 95% recall.

**12** (Wallace et al., 2010b): Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. (2010b). Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 173–182. ACM

In this study, Wallace et al. (2010b) used active learning with SVM (LibSVM) and compared random sampling with uncertainty sampling to select candidate references for the review authors to screen on the COPD, Micro nutrients, and the Proton beam datasets. They also introduce labeled features to let review authors suggest features indicative of the target class in order to better leverage expert knowledge.

The authors demonstrate improved  $U_{10}$  curves for the labelled features over random sampling an uncertainty sampling, but do not measure the resulting performance using point metrics.

**13** (Frunza et al., 2010): Frunza, O., Inkpen, D., and Matwin, S. (2010). Building systematic reviews using automatic text classification techniques. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 303–311. Association for Computational Linguistics

Frunza et al. (2010) used complement naive bayes using WEKA to simulate replacing one of the human screeners with an automated system on the TrialStat dataset [10, 8]. The authors report an optimal 17.1% precision for 92.7% recall using this human-machine ensemble.

**14** (Wallace et al., 2010c): Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., and Schmid, C. H. (2010c). Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):55

In this study, Wallace et al. used active learning with aggressive undersampling to mitigate the class imbalance on three datasets: COPD, Proton Beam, and Micronutrients.

This study appears to contain the first mention of the subsequently often quoted estimate of the time it takes for a screener to screen a single abstract (30s on average).

The active learning uses something the authors call patient active learning as an initial sampling protocol, where the protocol initially tries to retrieve a representative selection of the reference space. Only when a representative sample has been labelled by the human reviewers will the system switch to uncertainty sampling.

The authors simulate the algorithm on three datasets, and report a decrease in burden of ~40%.

**15** (**Bekhuis and Demner-Fushman, 2010**): Bekhuis, T. and Demner-Fushman, D. (2010). Towards automating the initial screening phase of a systematic review. *Studies in Health Technology and Informatics*, 160(PART 1):146–150

In this study, Bekhuis and Demner-Fushman evaluated a number of machine learning approaches to automate the second screening. They evaluated in terms of recall, precision, and  $F_1$  measure with included studies as the positive class.

The authors use search filters for relevance to topic and study design, and the method described in the study therefore assumes that methodological search filter work for the research question of the review. The authors tried to use naive Bayes and SVMs but report that their implementations failed. They instead evaluated decision trees, evolutionary SVMs, and WAODE.

The evaluated dataset consists of 400 references, where 13% are positive. The data was selected prospectively for one review and results were cross-validated in that one review. The authors report lackluster scores for the Decision Trees and WAODE. For the evolutionary SVM, the authors report 100% recall for ~40–48% average precision for RBF kernels and 4th degree Epanechnikov kernels. The authors report ~67% recall and precision for lower degree polynomial kernels. The general trend in the results is that more complex kernels (i.e. with higher VC dimension) results in higher recall, and lower precision.

For unclear reasons, the recall dropped to 76.92% for both radial and Epanechnikov kernel when evaluated on a held-out test set.

**16** (**Fiszman et al., 2010**): Fiszman, M., Bray, B., Shin, D., Kilicoglu, H., Bennett, G., Bodenreider, O., and Rindflesch, T. (2010). Combining Relevance Assignment with Quality of the Evidence to Support Guideline Development. *Stud Health Technol Inform*, 160(1):709–713

Similarly to [26], which this paper predates, Fiszman et al. distinguished between query relevance and article quality. Labelings are consequently encoded as a tuple consisting of relevance and quality judgments. Final decisions for inclusion in the review are taken to be those both relevant and high-quality. When training the classifier to recognize both labelings, the authors report 56% recall and 91% precision. When only considering relevance, the results are reported to ‘drop’ to 62% recall, 79% precision.

Details of the method are given in an earlier paper (Rindflesch and Fiszman, 2003). The method uses Hearst patterns to construct an ontology of hypernymic relations,

and then using these to find relevant and high quality articles.

17 (**Matwin et al., 2010**): Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., and O’Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association : JAMIA*, 17(4):446–53

18 (**Matwin et al., 2011**): Matwin, S., Kouznetsov, A., Inkpen, D., and O’Blenis, P. (2011). Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association : JAMIA*, 1(18):author reply 105

19 (**Matwin and Sazonova, 2012**): Matwin, S. and Sazonova, V. (2012). Direct comparison between support vector machine and multinomial naive Bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, 19(5):917–917

In this study, Matwin et al. used Factorized Complement Naive Bayes (FCNB) on the DERP dataset, and compared against Cohen et al.’s previously reported results [1]. FCNB is a modification of CNB, that introduces a multiplication factor  $F_c$  that is multiplied to the non-constant term in the CNB expression. They also describe a process they call ‘weight engineering’ where the CNB expression for each feature is further multiplied by another multiplicative factor learned by the model. The reported wss@95 scores ranged from 8.5% to 62.2%, and the authors reported a 15% average absolute improvement over the baseline [1].

20 (**Wallace et al., 2010a**): Wallace, B. C., Small, K., Brodley, C. E., Lau, J., and Trikalinos, T. a. (2010a). Modeling Annotation Time to Reduce Workload in Comparative Effectiveness Reviews Categories and Subject Descriptors Active Learning to Mitigate Workload. *Proceedings of the 1st ACM International Health Informatics Symposium. ACM.*, pages 28–35

21 (**Wallace et al., 2012b**): Wallace, B. C., Small, K., Brodley, C. E., Lau, J., and Trikalinos, T. a. (2012b). Deploying an interactive machine learning system in an evidence-based practice center. *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI ’12*, page 819

22 (**Rathbone et al., 2015**): Rathbone, J., Hoffmann, T., and Glasziou, P. (2015). Faster title and abstract screening? evaluating abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic reviews*, 4(1):80

These studies introduce the publicly available Abstrackr screening tool. The authors focused on modeling how long it would take for humans to screen references, with the intent of using predictions about screening difficulty for assigning refer-

ences to screeners. The authors hypothesized that the total amount of time it would take to screen all references might be reduced if such assignments were done sensibly.

The authors found that screeners take longer to screen references early in the process, but found only a weak correlation between time to screen and distance to the SVM separating hyperplane. To put it more simply, the references the learned SVM model is uncertain about are not the same references a human screener will be uncertain about.

They argue that active learning can be improved by using the predicted time to screen as a further parameter for sampling the next references to screen. Intuitively, we are best served to select references that improve the decision boundary but are also easy for human screeners to judge. This also allows the system to assign easy cases to novice screeners, and ambiguous cases to the more experienced [21].

Rathbone et al. retrospectively evaluated the use of Abstrackr on three systematic reviews of interventions: *dietary fibre interventions for colorectal cancer*, *rituximab and adjunctive chemotherapy interventions for non-Hodgkin's lymphoma*, and *eculizumab for atypical hemolytic uremic syndrome*, as well as one DTA systematic review on *echocardiography for stroke*.

They reported workload reductions between 9%–80% while missing one reference in each of two reviews.

**23** (Frunza et al., 2011): Frunza, O., Inkpen, D., Matwin, S., Klement, W., and O'Brien, P. (2011). Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*, 51(1):17–25

In this study, the authors used the TrialStat dataset [10, 8]. The authors used complement naive Bayes, with development and evaluation on the same review topic. The authors divide the task into two distinct versions. In the first they simply apply the method on the whole dataset (using train/test splits), which they call the global method. In the second they partition the task into a separate subtask for each question in the inclusion criteria. They report a maximum recall of 67.8%, and a maximum precision of 37.9% on the global method, and a maximum recall of 99.7% and a maximum precision of 63.0% on the per-question method. The results seem to suggest that we can expect better performance if we treat individual inclusion criteria as distinct in the screening process.

**24** (Tomassetti et al., 2011): Tomassetti, F., Rizzo, G., Vetro, A., Ardito, L., Torchiano, M., and Morisio, M. (2011). Linked data approach for selection process automation in systematic reviews. *15th Annual Conference on Evaluation & Assess-*

*ment in Software Engineering (EASE 2011)*, pages 31–35

In this study, Tomassetti et al. used a fairly straightforward active learning process based on Naive Bayes to retrieve articles relevant to software cost estimation. The evaluation set consists of 106 articles retrieved by the search term “Software Cost Evaluation” from the journal the authors pre-specified as having the highest percentage of relevant articles. The authors note that generally some articles are known to the reviewers at the beginning of the review, which can be used as a seed set.

The authors reported a 20% workload decrease for 100% recall.

**25** (Wallace et al., 2011): Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. (2011). Who should label what? instance allocation in multiple expert active learning. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 176–187. SIAM

In this study Wallace et al. investigated the effects of assigning different references to different screeners, with easier tasks going to junior screeners, and more difficult tasks going to senior screeners. They used SVMs with standard NLP features on the COPD dataset, and report improvements over the baseline, but it is difficult to interpret the results intuitively.

**26** (Choi et al., 2012): Choi, S., Ryu, B., Yoo, S., and Choi, J. (2012). Combining relevancy and methodological quality into a single ranking for evidence-based medicine. *Information Sciences*, 214:76–90

In this study Choi et al. distinguished between query relevance and article quality and used sensor fusion to arbitrate between the two.

Article relevance was determined by Okapi BM25. Quality classification was done using Naive Bayes and SVMs. The authors reported that SVMs yielded the best results.

**27** (Shekelle et al., 2012): Shekelle, P. G., Dalal, S. R., and Shetty, K. D. (2012). A Pilot Study Using Machine Learning and Domain Knowledge To Facilitate Comparative Effectiveness Review Updating. *AHRQ*

In this study, Shekelle et al. used generalized linear models and gradient boosting machines for retrospective use in one systematic review of *interventions to prevent fractures in persons with low bone density*, and one systematic review of *atypical antipsychotic drugs*. The models were trained on 1) 46 binary features based on

whether the target intervention and outcome terms were present in the citations and linked to particular subheadings; and 2) 29 binary features related to broader characteristics from the MeSH terms and publication type terms (e.g. demographic group, treatment target, and publication type). The authors reported 99% and 100% recall, for 55.4% and 63.2% workload reduction respectively for the two reviews.

**28** (Sun et al., 2012): Sun, Y. B., Yang, Y., Zhang, H., Zhang, W., and Wang, Q. (2012). Towards evidence-based ontology for supporting systematic literature review. *Evaluation and Assessment in Software Engineering*, 2012(1):171–175

In this study, Sun et al. reportedly constructed an ontology of the target paper structure, and populate this ontology automatically for each reference based on the structured abstracts. Rather than semi-automation, the author appear to intend to replace the human reviewers entirely in systematic reviews in computer science. The authors do not report traditional metrics such as precision/recall of found references. They report that in their test, their system retrieves the same 11 references as four students, although the students seem to each have found different sets of papers. The results are overall not very convincing since the competence of the students doing the review does not appear to be representative of reviewers in an actual review, and the size of the evaluation is small. Details of the implementation are unclear from the paper.

**29** (Bekhuis and Demner-Fushman, 2012): Bekhuis, T. and Demner-Fushman, D. (2012). Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artificial intelligence in medicine*, 55(3):197–207

In this study, Bekhuis and Demner-Fushman used screening automation as a second screener on one systematic review of *surgical interventions for treating ameloblastomas of the jaws*, and one systematic review of *vaccines for preventing influenza in the elderly*. The objective of the study was to determine the relative performance of different machine learning methods.

The authors compared kNN, naive Bayes, complement naive Bayes, and evolutionary svm using *tf-idf* weighted bag-of-word features and MeSH/EMTREE terms. The authors evaluated in terms of recall, precision, and  $F_3$  using 10-fold cross-validation.

The authors report 46% second screener workload reduction for 95.5% recall using the evolutionary svm, and 35% reduction for 96.7% recall using complement naive Bayes. Evolutionary svm and complement naive Bayes were significantly better than kNN, but no single classifier setting was consistently optimal.

**30** (**Wallace et al., 2012a**): Wallace, B. C., Small, K., Brodley, C. E., Lau, J., Schmid, C. H., Bertram, L., Lill, C. M., Cohen, J. T., and Trikalinos, T. A. (2012a). Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in Medicine*, 14(7):663–669

In this study, Wallace et al. retrospectively evaluated the use of svms in updating four reviews, and reported recall (sensitivity) scores of 100% for three reviews and 99% for one, with associated specificity scores of 90, 93, 90, 73% respectively. The associated reductions in workload are reported as 89, 82, 87, and 67% respectively. The svms are modified to deal with class imbalance by reweighting and undersampling. The authors use the randomness caused by the undersampling to train an ensemble of 11 classifiers, of which the majority vote is used as the final decision.

**31** (**Kim and Choi, 2012**): Kim, S. and Choi, J. (2012). Improving the performance of text categorization models used for the selection of high quality articles. *Healthcare informatics research*, 18(1):18–28

**32** (**Kim and Choi, 2014**): Kim, S. and Choi, J. (2014). An svm-based high-quality article classifier for systematic reviews. *Journal of biomedical informatics*, 47:153–159

In this study, Kim and Choi (2012) used standard bag-of-word features and MeSH terms to train a svm (svm<sup>light</sup>) with default settings and a linear kernel. The authors used 19 procedural systematic reviews each including at least 10 studies, as well as four topics from the DERP drug class efficacy dataset. They reported maximum AUC scores of 0.95 on the first dataset, and 0.84 on the second when using intra-topic classification [31]. They reported 88.32% accuracy using inter-topic classification over 75.38% accuracy using intra-topic classification [32].<sup>1</sup>

**33** (**Jonnalagadda and Petitti, 2013**): Jonnalagadda, S. and Petitti, D. (2013). A new iterative method to reduce workload in systematic review process. *International journal of computational biology and drug design*, 6(1-2):5–17

In this study, Jonnalagadda and Petitti (2013) used active learning with random indexing. The authors simulated the system on the DERP dataset [1], and reported a wss@95 ranging from 6%–30%.

<sup>1</sup> Note that a baseline which excludes all references would achieve 95.48% accuracy on the DERP dataset

**34** (Shemilt et al., 2014): Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., Kelly, M. P., and Thomas, J. (2014). Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 5(1):31–49

In this study Shemilt et al. evaluated the use of machine learning to reduce screening effort in two extremely large scoping reviews (> 800k references and > 1M references).

The authors used svm (LibSVM) with radial basis kernels and undersampling, with bagging over three classifiers. The authors report 88% and 90% workload reductions, but due to the prospective nature of the trial, no recall values were reported.

**35** (Cormack and Grossman, 2014): Cormack, G. and Grossman, M. (2014). Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. *SIGIR*, pages 153–162

**36** (Cormack and Grossman, 2015): Cormack, G. V. and Grossman, M. R. (2015). Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv preprint arXiv:1504.06868*

**37** (Cormack and Grossman, 2017): Cormack, G. V. and Grossman, M. R. (2017). Technology-assisted review in empirical medicine: Waterloo participation in CLEF eHealth 2017. In *CLEF (Working Notes)*

**38** (Cormack and Grossman, 2018): Cormack, G. V. and Grossman, M. R. (2018). Technology-assisted review in empirical medicine: Waterloo participation in CLEF eHealth 2018. In *CLEF (Working Notes)*

Conventional active learning use uncertainty sampling to query new labels from the human trainer, where examples are chosen to learn the target concept as quickly as possible. Examples are thus chosen based on how uncertain the model is about them, or conversely, how informative they would be to the model.

In this series of studies, Cormack and Grossman introduce continuous active learning (CAL). In CAL, new candidates are chosen greedily, prioritizing the examples that are the most likely to be positive. CAL therefore show positive examples earlier in the process, and achieve better recall/effort curves and average precision than uncertainty sampling (which Cormack and Grossman call simple active learning).

The authors use learning-to-rank approaches using svm and logistic regression using standard unigram features. The method – while simple – achieved the best performance on most metrics in all iterations of the TREC Total Recall Track, and all iterations of CLEF eHealth task 2. In CLEF 2017, they achieved a 0.701 WSS@95

to retrieve references included by abstract and title, compared to 0.400 for the BM25 baseline. In CLEF 2018, they achieved a 0.841 wss@95 to retrieve references included by full-text.

**39** (Miwa et al., 2014): Miwa, M., Thomas, J., O'Mara-Eves, A., and Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, 51:242–253

In this study, Miwa et al. used an active learning approach, which starts by using LDA to sample initial candidates across different clusters before the user has classified any articles.

The authors compare uncertainty sampling with certainty sampling and shows that certainty sampling is useful for retrieving relevant articles, similarly to Cormack and Grossman [35].

They report that aggressive undersampling combined with uncertainty sampling, and patient active learning combined with uncertainty sampling yield better results for the first 20–30%, after which it yields worse results. The authors report their results in terms of utility and coverage, and the results are therefore difficult to interpret intuitively.

**40** (Bekhuis et al., 2014): Bekhuis, T., Tseytlin, E., Mitchell, K. J., and Demner-Fushman, D. (2014). Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PLoS ONE*, 9(1):1–10

**41** (Bekhuis et al., 2015): Bekhuis, T., Tseytlin, E., and Mitchell, K. J. (2015). A prototype for a hybrid system to support systematic review teams: A case study of organ transplantation. In 2015 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 940–947. IEEE

In a series of studies, Bekhuis et al. applied machine learning as a second screener. Bekhuis et al. constructed a dataset from 5 systematic reviews of interventions, where reviewers reported that nonrandomized or observational studies were eligible for inclusion. The dataset was constructed by replicating the original search queries and using the included references as positive labels. They then develop and test a CNB classifier on the data, using standard bag-of-word features, index terms, concept defined in a predefined ontology, and features extracted using LDA. The authors report a reduction in screening burden of 88–98% in the second screening iteration, and 38–48% overall. However, the recall reported ranges from 60–97%, making it difficult to assess the feasibility of the method in the context of a systematic review.

In a later study [41], the authors complemented the machine learning model with a rule-based approach (Jess Rule Engine), where the authors applied 9 domain specific rules to exclude negative references (restricting by publication type, study design, species, organ transplantation, cell transplantation, mycophenolic acid, blood, physiological monitoring, and outcome). However, since the authors report a 55% recall, it is unclear whether the method is feasible for use in a systematic review.

**42** ([García Adeva et al., 2014](#)): García Adeva, J. J., Pikatza Atxa, J. M., Ubeda Carrillo, M., and Ansuategi Zengotitabengoa, E. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, 41(4 PART 1):1498–1508

In this study, García Adeva et al. used NB, kNN, SVM, and Rocchio to automate second screening in one systematic review of internet-based RCTs. The data appears to have been collected retrospectively. Implementation details are unspecified, suggesting that the methods were used with default settings.

The authors do not appear to maximize recall, and achieves an optimal recall of 86%. The results may therefore be ill-suited for systematic reviews. The discussion in the paper focuses on the F1 measure, and the authors reported an ‘optimal’ F1 of 70%.

The machine learning methods appear to be well implemented, but the unfortunate disconnect with the problem domain mean the results may not be useful for screening automation.

**43** ([Timsina et al., 2015](#)): Timsina, P., El-Gayar, O. F., and Liu, J. (2015). Leveraging advanced analytics techniques for medical systematic review update. In *2015 48th Hawaii International Conference on System Sciences*, pages 976–985. IEEE

**44** ([Timsina et al., 2016b](#)): Timsina, P., Liu, J., El-Gayar, O., and Shang, Y. (2016b). Using semi-supervised learning for the creation of medical systematic review: An exploratory analysis. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1195–1203. IEEE

**45** ([Liu et al., 2018](#)): Liu, J., Timsina, P., and El-Gayar, O. (2018). A comparative analysis of semi-supervised learning: the case of article selection for medical systematic reviews. *Information Systems Frontiers*, 20(2):195–207

In this study Timsina et al. evaluated the use of a semi-supervised algorithms (the label spreading algorithm RBF kernels) on four topics in DERP (ACE Inhibitors, Atypical Antipsychotics, Estrogens, and NSAIDs).

They reported substantial workload reductions, but the reported recall values range from 82%–90%, and the results may therefore not be relevant for systematic reviews. The authors report that the recall values were better than supervised SVM with 50%/50% training splits, but the reported recall values are much lower than what has been reported by Cohen using the same method [5].

**46 (Timsina et al., 2016a):** Timsina, P., Liu, J., and El-Gayar, O. (2016a). Advanced analytics for the automation of medical systematic reviews. *Information Systems Frontiers*, 18(2):237–252

In this study, Timsina et al., compared linear SVMs, (soft-margin) polynomial SVMs, EvoSVMs, Voting Perceptrons, and Naive Bayes. The linear SVMs are reported to yield no results in several of the tests. It is unclear from the report why the authors used hard margins for the linear kernel SVMs and soft margins for the polynomial SVMs.

**47 (Howard et al., 2016):** Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., Holmgren, S., Pelch, K. E., Walker, V., Rooney, A. A., et al. (2016). Swift-review: a text-mining workbench for systematic review. *Systematic reviews*, 5(1):87

In this paper, Howard et al. present SWIFT-review, a publicly available screening prioritization tool. Their system uses bag-of-word representations of documents as well as MeSH terms and LDA clusters. They use the system to simulate an active learning approach on each systematic review, starting from a seed set of annotated articles from the dataset.

They evaluated the model on 5 systematic reviews of exposures, reporting a mean wss@95 of 0.766. They also evaluated on the DERP dataset, reporting a mean wss@95 of 0.488.

The authors compared the results on the DERP dataset against the previous baselines [5, 17]. However, due to the differences in approaches, this may not be a fair comparison.

**48 (Khabsa et al., 2016):** Khabsa, M., Elmagarmid, A., Ilyas, I., Hammady, H., and Ouzzani, M. (2016). Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, 102(3):465–482

**49 (Olofsson et al., 2017):** Olofsson, H., Brolund, A., Hellberg, C., Silverstein, R., Stenström, K., Österberg, M., and Dagerhamn, J. (2017). Can abstract screening workload be reduced using text mining? user experiences of the tool

Rayyan. *Research synthesis methods*, 8(3):275–280

In this study, Khabisa et al. used random forests on the DERP dataset. Besides bag-of-word features the authors also use co-citation information and Brown clusters. These additional features are shown to improve performance over the baseline. The system is shown to outperform CNB and Voting Perceptron, but the results appear similar to Cohen’s SVM results. It is not clear whether the performance gains are due to the extra features, or due the use of random forests, and whether the SVM results would also have been improved with the addition of these features.

This system is available for use with the Rayyan reference manager. The extrinsic performance of the system has subsequently been independently evaluated by Olofsson et al., who reported that the system achieved between 85%–99% recall for a 50% workload reduction on an unspecified set of systematic reviews.

**50** (Ji et al., 2017): Ji, X., Ritter, A., and Yen, P.-Y. (2017). Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. *Journal of biomedical informatics*, 69:33–42

In this study, Ji et al. [50] evaluated the use of screening reduction on the DERP dataset, and compare with Cohen et al.’s results using voting perceptrons [1], and Khabisa et al.’s results using random forests [48].

Ji et al. proposed to use ontology based features and evaluate three different models, using SNOMED-CT, MeSH terms, or both. They report an optimal average wss@95 of 0.409 using the combined model.

**51** (Shekelle et al., 2017): Shekelle, P. G., Shetty, K., Newberry, S., Maglione, M., and Motala, A. (2017). Machine learning versus standard techniques for updating searches for systematic reviews: a diagnostic accuracy study. *Annals of internal medicine*, 167(3):213–215

In this study, Shekelle et al. used SVMs on 3 intervention reviews on *low bone density*, *gout* and *osteoarthritis*. They reported a workload reduction of 67%–83% while missing 1 reference in each of 2 reviews.

**52** (Van Altena and Olabarriaga, 2017): Van Altena, A. and Olabarriaga, S. D. (2017). Predicting publication inclusion for diagnostic accuracy test reviews using random forests and topic modelling. In *CLEF (Working Notes)*

For their participation in CLEF, Van Altena and Olabarriaga used random forests over a 75-topic LDA representation, and achieved a 0.333 wss@95 compared to

0.400 for the BM25 baseline.

**53** (**Chen et al., 2017**): Chen, J., Chen, S., Song, Y., Liu, H., Wang, Y., Hu, Q., He, L., and Yang, Y. (2017). Ecnu at 2017 eHealth task 2: Technologically assisted reviews in empirical medicine. In *CLEF (Working Notes)*

For their participation in CLEF, Chen et al. used learning-to-rank with BM25, PL2, and BB2 as features. They combined their model with a vector space model. They achieved a 0.121 wss@95 compared to 0.400 for the BM25 baseline.

**54** (**Lee, 2017**): Lee, G. E. (2017). A study of convolutional neural networks for clinical document classification in systematic reviews: sysreview at CLEF eHealth 2017. In *CLEF (Working Notes)*

For their participation in CLEF, Lee used convolutional neural networks and achieved a 0.131 wss@95 compared to 0.400 for the BM25 baseline.

**55** (**Di Nunzio et al., 2017**): Di Nunzio, G. M., Beghini, F., Vezzani, F., and Henrot, G. (2017). An interactive two-dimensional approach to query aspects rewriting in systematic reviews. ims unipd at CLEF eHealth task 2. In *CLEF (Working Notes)*

**56** (**Di Nunzio et al., 2018**): Di Nunzio, G. M., Ciuffreda, G., and Vezzani, F. (2018). Interactive sampling for systematic reviews. ims unipd at CLEF 2018 eHealth task 2. In *CLEF (Working Notes)*

For their participation in CLEF, Di Nunzio et al. used a two-dimensional probabilistic version of BM25 to rank articles. The top abstract returned by BM25 was provided to two non-experts who generated one additional query each. The three queries were then used to re-rank articles. They achieved 0.517 wss@95 compared to 0.400 for the BM25 baseline. In 2018 [56] they complemented the approach with active learning using naive Bayes, achieving 0.792 wss@95 to retrieve relevant studies included by full-text.

**57** (**Scells et al., 2017**): Scells, H., Zuccon, G., Deacon, A., and Koopman, B. (2017). Qut ielab at CLEF 2017 technology assisted reviews track: Initial experiments with learning to rank. In *CLEF (Working Notes)*

For their participation in CLEF, Scells et al. trained a learning-to-rank model using PICO annotations as features (Population, Intervention, Control, Outcome). The

features were extracted automatically from articles and manually from the Boolean queries. They achieved 0.294 wss@95 compared to 0.400 for the BM25 baseline.

**58** (**Alharbi and Stevenson, 2017**): Alharbi, A. and Stevenson, M. (2017). Ranking abstracts to identify relevant evidence for systematic reviews: The university of sheffield's approach to CLEF eHealth 2017 task 2. In *CLEF (Working Notes)*

For their participation in CLEF, Alharbi and Stevenson automatically parsed the Boolean queries to extract terms and MeSH headings and used *tf·idf* cosine similarity to rank references. They achieved 0.493 wss@95 compared to 0.400 for the BM25 baseline.

**59** (**Kalphov et al., 2017**): Kalphov, V., Georgiadis, G., and Azzopardi, L. (2017). Sis at clef 2017 ehealth tar task. In *CEUR Workshop Proceedings*, volume 1866, pages 1–5

For their participation in CLEF, Kalphov et al. used 1) LDA clusters to identify the topics most likely relevant to the search queries, 2) active learning using Rocchio; and 3) a combination of the both approaches. They achieved 0.530 wss@95 compared to 0.400 for the BM25 baseline.

**60** (**Singh et al., 2017**): Singh, G., Marshall, I., Thomas, J., and Wallace, B. (2017). Identifying diagnostic test accuracy publications using a deep model. In *CEUR Workshop Proceedings*, volume 1866. CEUR Workshop Proceedings

For their participation in CLEF, Singh et al. trained a deep convolutional model achieved 0.076 wss@95 compared to 0.400 for the BM25 baseline.

**61** (**Anagnostou et al., 2017**): Anagnostou, A., Lagopoulos, A., Tsoumakas, G., and Vlahavas, I. P. (2017). Combining inter-review learning-to-rank and intra-review incremental training for title and abstract screening in systematic reviews. In *CLEF (Working Notes)*

**62** (**Tsoumakas, 2018**): Tsoumakas, G. (2018). Learning-to-rank and relevance feedback for literature appraisal in empirical medicine. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, volume 11018, page 52. Springer

**63** (**Minas et al., 2018**): Minas, A., Lagopoulos, A., and Tsoumakas, G. (2018). Aristotle university's approach to the technologically assisted reviews in

empirical medicine task of the 2018 CLEF eHealth lab. In *CLEF (Working Notes)*

For their participation in CLEF, Anagnostou et al. used a combination of inter-topic static learning trained on the development corpus, and active learning iteratively trained on the target topic. The intertopic model used XGBoost trained on 24 topic/document features computed over the title, abstract and mesh terms of the articles and the query. In CLEF 2017, they achieved a 0.697 wss@95 to retrieve references included by abstract and title, compared to 0.400 for the BM25 baseline [61]. In CLEF 2018, they achieved a 0.848 wss@95 to retrieve references included by full-text [63].

74

**64** (**Cheng et al., 2018**): Cheng, S., Augustin, C., Bethel, A., Gill, D., Anzaroot, S., Brun, J., DeWilde, B., Minnich, R., Garside, R., Masuda, Y., et al. (2018). Using machine learning to advance synthesis and use of conservation and environmental evidence. *Conservation biology: the journal of the Society for Conservation Biology*, 32(4):762

In this study, Cheng et al. reported using word2vec for article screening, and GloVe for automated data extraction, but no implementation details are given. The topic and nature of the systematic reviews are not specified. The risk of bias is likely high, since the evaluation scores appear to have been cherry-picked for each review.

**65** (**Donoso-Guzmán and Parra, 2018**): Donoso-Guzmán, I. and Parra, D. (2018). An interactive relevance feedback interface for evidence-based health care. In *23rd International Conference on Intelligent User Interfaces*, pages 103–114. ACM

In this study, Donoso-Guzmán and Parra (2018) construct an interactive interface for an active learning system based on Rocchio and BM25 with standard NLP features. The authors reported an optimal 23% recall on the dataset, but the performance and workload reduction of the system is unclear from the report.

The authors created the dataset from the Epistemonikos database of primary studies previously included in systematic reviews, and the data is therefore not representative of candidate references normally considered in a screening scenario. The test user were not experienced reviewers, and only 86% reported being able to read English without problems. Overall, since the study methodology is unrepresentative of the screening methodology normally encountered in systematic reviews, it is not clear whether the results are applicable to systematic review screening.

**66** (Tsafnat et al., 2018): Tsafnat, G., Glasziou, P., Karystianis, G., and Coiera, E. (2018). Automated screening of research studies for systematic reviews using study characteristics. *Systematic reviews*, 7(1):64

In this study, Tsafnat et al. [66] used mentions of exposure and outcome in abstracts for prospective screening reduction in three systematic reviews of *outdoor particulate matter exposure and lung cancer*, *PFOA effects on fetal growth*, and *Bisphenol A (BPA) exposure and obesity*. They reported a mean 93.7% workload reduction for 98% overall recall.

The risk of bias in the study is likely high, due to the small sample size and the use in systematic reviews with apparently consistent terminology.

**67** (Przybyła et al., 2018): Przybyła, P., Brockmeier, A. J., Kontonatsios, G., LePogam, M.-A., McNaught, J., von Elm, E., Nolan, K., and Ananiadou, S. (2018). Prioritising references for systematic reviews with robotanalyst: A user study. *Research synthesis methods*, 9(3):470–488

In this study, Przybyła et al. [67] used SVM with  $L_2$  regularization to simulate screening reduction in ‘several reference collections [...] related to public health topics.’ They reported wss@95 ranging from -3.62%–66.17%

The authors evaluated on 17 topics, but the largest is only 4,964 references. Most topics with good wss@95 are small, which may suggest chance is influencing the results.

The method also uses preprocessing, with identification of terms using the C-value algorithm, as well as clustering with LDA/spectral clustering.

**68** (Wu et al., 2018): Wu, H., Wang, T., Chen, J., Chen, S., Hu, Q., and He, L. (2018). Ecnv at 2018 ehealth task 2: Technologically assisted reviews in empirical medicine. *Methods*, 4(5):7

For their participation in CLEF, Wu et al. used vector similarity using Paragraph2-Vector and achieved 0.147 wss@95 to retrieve relevant studies included by full-text.

**69** (Cohen and Smalheiser, 2018): Cohen, A. M. and Smalheiser, N. R. (2018). Uic/OHSU CLEF 2018 task 2 diagnostic test accuracy ranking using publication type cluster similarity measures. In *CEUR Workshop Proceedings*, volume 2125

For their participation in CLEF, Cohen and Smalheiser used clustering on PubMed data to identify 6 publication types, including DTA studies. They then used a SVM to

classify candidate references by similarity to each of the 6 cluster centroids. They achieved 0.486 wss@95 to retrieve relevant studies included by full-text.

**70** (**Alharbi et al., 2018**): Alharbi, A., Briggs, W., and Stevenson, M. (2018). Retrieving and ranking studies for systematic reviews: The university of sheffield's approach to CLEF eHealth 2018 task 2. In *CEUR Workshop Proceedings*, volume 2125. CEUR Workshop Proceedings

For their participation in CLEF, Alharbi et al. attempted to enrich queries with terms designed to identify diagnostic test accuracy studies used active learning using Rocchio. They achieved 0.681 wss@95 to retrieve relevant studies included by full-text.

**71** (**Lerner et al., 2019**): Lerner, I., Créquit, P., Ravaud, P., and Atal, I. (2019). Automatic screening using word embeddings achieved high sensitivity and workload reduction for updating living network meta-analyses. *Journal of clinical epidemiology*, 108:86–94

In this study, Lerner et al. [71] used logistic regression (SGD) using L2 regularization with word embeddings and oversampling to simulate screening reduction in network meta-analyses in pneumonology, urology, oncology, and psychiatry. They reported 53% wss@100.

**72** (**Bannach-Brown et al., 2019**): Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., and Macleod, M. R. (2019). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic reviews*, 8(1):23

In this study, Bannach-Brown et al. [72] used SVM (SGD), with and without LDA clustering for prospective screening reduction in two systematic reviews of animal studies: *animal models of neuropathic pain* and *animal models of depression*. They reported 70.5% and 69.3% wss@95 for the two systematic reviews respectively. The author also used machine learning to identify 11 false positives and 36 false negatives in the training set.

**73** (**Olorisade et al., 2019**): Olorisade, B. K., Brereton, P., and Andras, P. (2019). The use of bibliography enriched features for automatic citation screening. *Journal of biomedical informatics*, 94:103202

In this study, Olorisade et al. [73] evaluated the use of screening reduction on the DERP dataset, and compare with Cohen et al.'s results using voting perceptrons [1], and Khabsa et al.'s results using random forests [48]. They cite Matwin and Sazonova [19], but do not compare against their results.

Olorisade et al. proposed to use MeSH terms, and a combination of MeSH terms and reference lists from articles to improve training. They report an optimal average wss@95 of 0.408 on average.

#### SUMMARY

SEVERAL AUTOMATION APPROACHES EXIST, and have been developed since 2006. Few have seen consistent or repeated use in real systematic reviews, and there are few prospective use-cases of the methods

EXISTING METHODS ARE OFTEN MISMATCHED to the existing workflow, or have not been evaluated under the same conditions as would be encountered in a systematic review.

FEW EXISTING METHODS HAVE BEEN COMPARED to each other. Except for the studies participating in CLEF, few have been evaluated on common datasets



## PART II

# SCREENING AUTOMATION SYSTEMS

This part of the thesis is based on the following conference papers:

**(Norman et al., 2018c):** Norman, C., Leeﬂang, M., Zweigenbaum, P., and Név  ol, A. (2018c). Automating document discovery in the systematic review process: How to use chaff to extract wheat. In Calzolari, N. et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA)

**(Norman et al., 2017b):** Norman, C., Leeﬂang, M., and N  v  ol, A. (2017b). LIMS  @CLEF eHealth 2017 task 2: Logistic regression for automatic article ranking. *Working Notes of CLEF*

**(Norman et al., 2018b):** Norman, C., Leeﬂang, M., and N  v  ol, A. (2018b). LIMS  @CLEF eHealth 2018 task 2: Technology assisted reviews by stacking active and static learning. *Working Notes of CLEF*, pages 10–14

The work has also been described in the following conference paper, omitted from this thesis:

**(Norman et al., 2017c):** Norman, C., Leeﬂang, M., Zweigenbaum, P., and N  v  ol, A. (2017c). Tri automatique de la litt  rature pour les revues syst  matiques. *24   Conf  rence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 234–41

The work has been presented as the following system demonstration:

**(Norman et al., 2017a):** Norman, C., Grouin, C., Lavergne, T., Zweigenbaum, P., and N  v  ol, A. (2017a). Traitement de la langue biom  dicale au LIMS  . *24   Conf  rence sur le Traitement Automatique des Langues Naturelles (TALN)*, page 33



QUESTIONS IN HEALTHCARE are increasingly resolved by systematic reviews, and the production of systematic reviews is consequently increasing at a rapid pace. PubMed indexed 17,254 new systematic reviews in 2018 alone, and this number has increased more than five-fold over the last decade (see chapter 2). While the demand for systematic reviews is growing, the number of publications that systematic reviews need to sift through is also increasing at a similarly break-neck pace. Despite the technological progresses seen over the last few decades, systematic reviews take longer to produce, and cost more than 35 years ago (Lau, 2019). We today spend more time and money producing new systematic reviews than we ever have, and the amount will surely increase further.

Search queries to identify diagnostic studies tend to have low accuracy (Beynon et al., 2013), and are discouraged for use in systematic reviews (Leeftang et al., 2006). Systematic reviews of diagnostic test accuracy may therefore be in particular need of alternative approaches to cope with the rapidly increasing workloads. Furthermore, since DTA systematic reviews are considered more difficult to automate than e.g. systematic reviews of interventions a breakthrough in this domain may lead to breakthroughs for several other types of systematic reviews (Kanoulas et al., 2017b, 2018; Petersen et al., 2014).

In this part of the thesis we will examine how screening automation methods can be used to reduce the workload. We will look at how these methods can be made to work, and how the performance of these methods compare with each other. The focus in this section will be technical, and all performance comparisons will concern intrinsic performance. In other words, we here seek to evaluate the performance of the component models in reproducible laboratory settings. We will examine the extrinsic performance of the methods – i.e. how the methods influence the systematic review process – in part III. We will however start thinking about how different approaches fit into different systematic review contexts and settings.

Readers not interested in the gritty technical details of the algorithms can safely skip to the chapter 9, which summarizes the main points of the papers.

## 5.1 SCREENING AUTOMATION METHODS

Screening automation is an umbrella term for several disparate approaches with the common goal of reducing the workload during the screening stage in systematic reviews (O'Mara-Eves et al., 2015). We will concern ourselves with the two main approaches that lend themselves to intrinsic evaluations: screening reduction and screening prioritization (see chapter 4).

SCREENING REDUCTION methods work by using automated methods to reduce the number of studies that need to be manually screened. The workload reduction in this approach comes from the reduction in number, not the order in which references are screened. Once the number is reduced, screening may proceed in any order.

The first screening reduction approach is to use a classification algorithm, where the algorithm is trained to explicitly model binary include/exclude decisions.

The second screening reduction approach is to use a ranking algorithm, where a regressor is trained to model the probability of inclusion/exclusion (Fuhr, 1992). All items falling below some threshold are then excluded from consideration (O'Mara-Eves et al., 2015).

The main difference between a classifier and a regressor with a cut-off is how the two are trained – the classifier is typically trained to minimize the number of misclassifications, the regressor is typically trained to minimize the number of inversions, i.e. cases where non-relevant items are ranked higher than relevant ones.

For the purposes of the systematic review, there is little difference between classification and probability regression with a threshold, and screening automation process can in either case simply be added as an additional stage between the database search (after records have been deduplicated, and meta-data have been retrieved) and the title and abstract screening. The remainder of the process can proceed entirely unaltered. This kind of screening automation therefore conceptually works as a second search filter to further exclude non-relevant references, with finer granularity than is possible with a boolean search filter alone.

SCREENING PRIORITIZATION similarly uses ranking to reduce the workload, but the primary intent is to change the order of screening, so that relevant records are screened before non-relevant ones. The number of records to screen can be reduced by combining screening prioritization with a cut-off threshold. The main difference compared to screening reduction is that screening prioritization does not add an extra filtering step before the screening commences, but rather modifies the screening process to screen in descending order of likelihood of relevance.

## 5.2 DESIGNING A SCREENING AUTOMATION MODEL

Screening automation can be performed using a number of different approaches. To some degree the choice of approach will be decided by their relative expected performance – some approaches are clearly better than others. However, the choice will also necessarily be constrained by the context of the systematic review. Different automation approaches make different assumptions regarding what training data is available, and how the screening process is to be performed. Some

approaches may be impossible or infeasible to use, depending on 1) what training data is available, 2) whether it is logistically feasible for the screeners to adopt the software and process mandated by the screening automation method, and importantly, 3) to what degree the altered screening process can be expected to ensure the integrity of the systematic review and its results (table 5.1).

The last two constraints are seldom mentioned or acknowledged by previous literature, but have been cited as one of the most frequent reasons to eschew screening automation (Van Altena et al., 2019). Cavalier attitudes to these concerns are unlikely to aid adoption of screening automation methods.

82

	Static		Active Learning
	Intratopic	Intertopic	
Can be used de novo?	No	Maybe	Yes
Unchanged workflow?	Yes	Yes	No
Unchanged results?	Yes	Yes	Unclear

Table 5.1 – The implications of different design decisions on a systematic review. The intertopic static approach can be used if and only if relevant data is available from similar systematic reviews. Active learning will necessarily result in a systematic review that ‘does not look like’ a systematic review. We take a closer look at what the implication are for the reviews results and conclusions in chapter 12.

Training data used to train screening automation models consist of examples of included and excluded references. What automation approaches are possible will depend on several factors. First, whether such training examples will address the same research question as the systematic review to be conducted and use the same inclusion criteria (so-called intratopic training), or from systematic reviews on similar questions (so-called intertopic training). Second, whether such training examples will be gathered from already existing sources (static learning), or gathered continuously as part of the screening effort (active learning)

*Static Approaches*

The static approach is the most straightforward. In this approach, training data is collected from some already existing source and is then used to train a ranking or classification model. This model is then applied prospectively in ongoing systematic reviews. The model is not updated using training examples gathered during screening.

THE INTRATOPIC STATIC APPROACH is likely to be the best choice if the screening automation is intended to be used in a systematic review update, and if there is an ample number of included references collected in previous review updates (including the original review). The trained model can be combined with a cut-off threshold for exclusion, and remaining references can be randomized. The subsequent screening process therefore does not need to differ from the conventional screening process, and is therefore fundamentally compatible with the conventional process provided the automation method is accurate and unbiased. Active learning could further refine the model, but such improvements are subject to diminishing returns, and if sufficient amounts of training data are available from the start, then the performance improvements gained from active learning may be limited. This straightforward approach can however only be used in systematic review updates. If the screening automation is intended to be used in a new systematic review (conducted de novo), then training data is unlikely to be available from the same topic.

THE INTERTOPIC STATIC APPROACH can be used in new systematic reviews (conducted de novo). In this case, training data is unlikely to be available from the same topic (i.e. screened using the same inclusion criteria), but there may be training data from similar topics (i.e. similar inclusion criteria for a different target condition).

If the model is trained on data from one such similar review the model will learn the inclusion criteria for this particular review, and the screening automation model may thus be of limited usefulness. However, if the model is trained on data from several similar reviews, the model may learn to generalize the common denominator of the inclusion criteria used across the different reviews. For instance, by training a model on the studies included in a range of systematic reviews of diagnostic test accuracy, the model may learn to recognize general diagnostic test accuracy studies of sufficient quality and clarity to be included in a systematic review, and learn to disregard the particular target condition addressed by different systematic reviews.

Unlike active learning, the intertopic static approach requires no targeted training data. Furthermore, the intertopic static approach may also be combined with a prespecified cut-off and randomization, and is therefore compatible with the conventional systematic review process.

### *Active Learning*

The ACTIVE LEARNING APPROACH has few technical constraints, and can be used in any systematic review where screening logistics and methodology constraints

allows its use. Since the training data is gathered as part of the ongoing screening effort, the active learning approach will by definition always use intratopic training. Active learning does not require training data when screening is started. The process may then be bootstrapped ('seeded') by sampling the references randomly or by using unsupervised models such as clustering or topic modelling to find an initial sample of relevant items for training. However, if some form of training data is already available, this can be used to kick-start the process. It may for instance be possible to use information retrieval methods with the database query or review protocol as input (Cormack and Grossman, 2017).

The main constraints of the active learning approach is that it requires the screening process to be fundamentally altered.

First, screening must be performed in specialized software that can continuously retrain the model, and present candidate references to the screener. This requires the systematic review authors to acquire and install such software, a hurdle that frequently make systematic reviewers eschew automation (Van Altena et al., 2019). Second, and importantly, active learning necessarily changes the order in which references are screened, with more likely candidates presented first. This could introduce so called rank-order bias, where screeners are influenced by the relevance scores given by the model. Screeners may in such cases be more likely to include top-ranked references, and less likely to include lower ranked references.

Third, with active learning the order of the references will change and adapt according to the screening decisions taken during screening. This may make it difficult to reproduce the screening process, and systematic review authors therefore need to confirm that active learning remains compatible with the purpose of the review.

### 5.3 USING TRAINING DATA FROM DIFFERENT STAGES OF SCREENING

Screening in systematic reviews is performed in two stages, first by title and abstract, then by full-text. This separation into two stages is largely a product of the difficulty in retrieving full-text of large number of records, a process that is hindered by restrictions imposed by publishers, as well as limited archival of and electronic access to articles (see chapter 3). In many cases, full-texts can only be retrieved by contacting the authors. Many records can however be excluded based only on titles and abstracts, and it usually makes sense to reduce the number of full-texts that need to be retrieved.

It is not clear whether screening automation systems should attempt to learn to recognize studies included in the first or second stage of screening. Reasonable arguments can be made either way. On the one hand, trying to mimic the human screening process would use decisions from the first stage. At the same time, studies included in the first stage but excluded in the second ultimately do not matter

to the review, and there is little value in the screening automation to identify these. Either way, unless full-texts are available, there will necessarily be some number of records that cannot be judged based on the available information.

As a practical example of the varying opinion on this matter, only the studies included in the second stage were considered relevant in the first iteration of the CLEF shared task (Kanoulas et al., 2017a). In the second iteration this decision was inverted, and instead all studies included in the first stage were considered relevant (Kanoulas et al., 2018).

Similarly, it is not clear whether studies in both stages of screening are equally representative of the studies the screening automation should try to identify. Judgments made in the first screening stage are preliminary, and often over-inclusive to avoid missing relevant studies. These may therefore be less representative and therefore inferior examples for training.

#### 5.4 EVALUATION OF PERFORMANCE

Comparing the relative performance of different methods is difficult since most previous work have been evaluated on different datasets, under different settings, and often using different performance measures. A number of public datasets are available to compare performance, but usage is fragmented and relative performance comparisons therefore difficult. Comparisons are often hindered by science's search for novelty. The 'pseudo-innovative masquerade in the quest for making a questionable case for novelty' (Ioannidis, 2016) frequently compel authors to make small incremental changes to existing methods – not enough to make the work meaningfully novel, but enough to preclude comparisons to previous work (Olorisade et al., 2016; O'Mara-Eves et al., 2015).

There have been attempts to compare previous methods by replicating reported methods on the same datasets, but the replication of published methods is often difficult or impossible due to insufficient reporting (Olorisade et al., 2016).

Another way to compare the relative performance of methods is through the use of a *shared task*, a community challenge where participating systems are trained on the same training data, and evaluated blindly using pre-decided metrics (Chapman et al., 2011; Huang and Lu, 2015). The shared task model removes many of the problems of replication studies, and also safe-guards against cheating, mistakes, the cherry-picking of metrics or data, as well as publication bias.

The only shared task for screening prioritization we are aware of is the CLEF Shared Task on Technology Assisted Reviews in Empirical Medicine, focusing on diagnostic test accuracy reviews (Kanoulas et al., 2017a, 2018). The CLEF shared task diverged from standard practice for shared tasks by not blinding participants to the gold standard data used for evaluation (see section 7.5.5). This may have been a

cause of biases in the reported results, and the results should be interpreted accordingly. Even so, the results of the CLEF campaign serves as an important venue to compare results across different methods and implementations

The CLEF shared task assumed the context of a new systematic review. Thus, all models evaluated in CLEF used either intertopic training or active learning. To date, and to our knowledge, there are no shared tasks addressing evaluation of intratopic static approaches.

## 5.5 OBJECTIVE

86

In this section we present three conference papers published during 2017-2018 where we attempt to address the following research questions:

- RQ 1 *What kind of data should we use to train screening automation methods?*
- RQ 2 *How do different screening automation approaches compare with each other for DTA screening?*
- RQ 3 *Are the screening automation methods we develop competitive with the current state-of-the-art?*



The material in chapter 6 has been published as:

(Norman et al., 2018c): Norman, C., Leeﬂang, M., Zweigenbaum, P., and N  v  ol, A. (2018c). Automating document discovery in the systematic review process: How to use chaff to extract wheat. In Calzolari, N. et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA)

This is a computer science conference article. Conferences are the main venue for publication in computer science. The article has undergone peer review and has been published in its entirety in the conference proceedings.

In this article we tried to address the following question:

RQ 1 *What kind of data should we use to train screening automation methods?*

We set out to do this by breaking the question into the following sub-questions:

- RQ 1 a) *Can we separate screening into two stages?*  
b) *Do we need examples from all stages of screening?*  
c) *Should the positive labels match the decisions in the first or second stage of screening?*

The results of this study was used as the basis for our participation in the CLEF eHealth task 2 described in chapter 7 and 8. This study was published after the study presented in chapter 7, but the two studies were conducted concurrently.

#### AUTHOR'S CONTRIBUTIONS

CN wrote the first draft and conducted the experiments. All authors conceived and designed the study. All authors read and approved the final manuscript.

## Automating Document Discovery in the Systematic Review Process: How to Use Chaff to Extract Wheat

Christopher R. Norman, Mariska M.G. Leeftang,  
Pierre Zweigenbaum, & Aurélie Névéol

Language Resources and Evaluation Conference (LREC), 2018

### Abstract

Systematic reviews in e.g. empirical medicine address research questions by comprehensively examining the entire published literature. Conventionally, manual literature surveys decide inclusion in two steps, first based on abstracts and title, then by full text, yet current methods to automate the process make no distinction between gold data from these two stages. In this work we compare the impact different schemes for choosing positive and negative examples from the different screening stages have on the training of automated systems. We train a ranker using logistic regression and evaluate it on a new gold standard dataset for clinical NLP, and on an existing gold standard dataset for drug class efficacy. The classification and ranking achieves an average AUC of 0.803 and 0.768 when relying on gold standard decisions based on title and abstracts of articles, and an AUC of 0.625 and 0.839 when relying on gold standard decisions based on full text. Our results suggest that it makes little difference which screening stage the gold standard decisions are drawn from, and that the decisions need not be based on the full text. The results further suggest that common-off-the-shelf algorithms can reduce the amount of work required to retrieve relevant literature.

### 6.1 INTRODUCTION

Systematic reviews seek to systematically gather all published evidence addressing a given research question and analyze the aggregate results. Systematic reviews constitute some of the strongest forms of scientific evidence, are an integral part of evidence based medicine, and serve a key role in informing and guiding public and institutional decision-making (Wright et al., 2007).

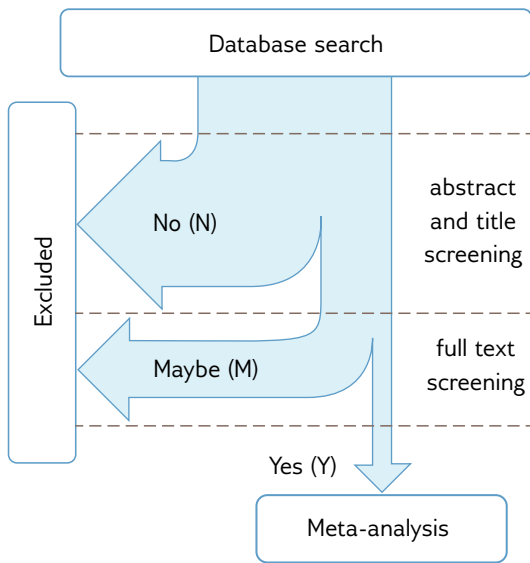


Figure 6.1 – Overview of the data flow during the screening process in systematic reviews.

using boolean queries handcrafted by experts. From this initial set of references, reviewers first screen for inclusion based on titles and abstracts, and then based on the full text (O'Mara-Eves et al., 2015) as illustrated in figure 6.1. In this paper we will call the references excluded in the first screening stage No ('N'), references excluded in the second screening stage Maybe ('M'), and references included in the final analysis Yes ('Y').

This selection is divided into two stages because while final decisions can only be based on the full text of articles, many references can be rejected based only on title and abstract. Retrieving the full text articles, which often needs to be done manually, is generally only feasible for a fraction of the articles in large systematic reviews (Tsafnat et al., 2014). However, even though humans approach screening as a two-step process, automation methods to date have generally approached the problem as a one-step process to find the relevant articles.

In this paper we ask if there is value in recognizing the distinction between each successive stage of the process. Our contribution is two-fold: First, we conduct experiments to inform methodology choices for automating the literature screening, and to find ways to improve the quality of constructing datasets used to train

One limiting factor of systematic reviews is that they tend to be prohibitively costly to produce.<sup>1</sup> The number of references needed to be manually screened in order to satisfy the requirement that virtually all relevant articles have been identified can number in the tens of thousands. Often only some dozens of these references are selected for the final meta-analysis, and the selection process may require months of work for several reviewers (O'Mara-Eves et al., 2015).

The screening process starts with identifying an initial set of candidate references, typically by searching databases

<sup>1</sup> Although primary clinical research is often more expensive.

such retrieval methods. Second, we experiment on an existing reference dataset and introduce a new, complementary dataset.

### 6.1.1 *Related Work*

Methods for automation have been attempted with varying degrees of success in technology assisted review in several topics in biomedicine (O'Mara-Eves et al., 2015). Technology assisted review has also been implemented in other fields with similarly stringent recall requirements, such as patent search (Stein et al., 2012), and electronic discovery (Cormack and Grossman, 2014). Automated document discovery is typically cast as a ranking or classification problem (O'Mara-Eves et al., 2015).

Common methods for automation include Support Vector Machines and variants of Naive Bayes, including Complement Naive Bayes (Matwin et al., 2010), and Multinomial Naive Bayes (Matwin and Sazonova, 2012). Other methods have been tried, including Voting Perceptrons (Cohen et al., 2006), Decision Trees (Bekhuis and Demner-Fushman, 2010), Evolutional SVM (Bekhuis and Demner-Fushman, 2010), WAODE (Bekhuis and Demner-Fushman, 2010), kNN (García Adeva et al., 2014), Rocchia (García Adeva et al., 2014), hypernym relations (Sun et al., 2012), Generalized Linear Models (Shekelle et al., 2012), Gradient Boosting Machines (Shekelle et al., 2012), Random Indexing (Jonnalagadda and Petitti, 2014), and Random Forests (Khabsa et al., 2016). Few of the methods proposed have been evaluated on common datasets however, and it is therefore difficult to draw conclusions about relative performance (O'Mara-Eves et al., 2015).

Recently, Khabsa et al. (2016) proposed using random forests, and compared the performance of their system with the reported performance of earlier systems on Cohen's 15 reviews (see section 6.2). Other methods have also been evaluated on the same dataset (Jonnalagadda and Petitti, 2014). For these reasons, and because the dataset is publicly available we will use this dataset as our baseline.

However, even though humans approach screening as a series of filters of increasingly fine granularity, all methods we have reviewed in previous literature approach the problem as a one stage process.

### 6.1.2 *Objective*

We construct an automatic screening system using a standard, off-the-shelf classifier. We describe our implementation and compare it with the state of the art to show that it functions as intended. We then apply our implementation on two datasets for systematic reviews, one of which is novel, in order to answer the following questions:

Dataset	Topic	Y	M	N
Yearbook	ClinicalNLP (2017)	11	70	244 (177)
	ClinicalNLP (2016)	23	60	267 (191)
Cohen	CalciumChannelBlockers	100	180	938
	ACEInhibitors	41	142	2361
	BetaBlockers	42	260	1770
	Opioids	15	33	1867
	OralHypoglycemics	136	3	364
	Statins	85	88	3292
	SkeletalMuscleRelaxants	9	25	1609
	Antihistamines	16	76	218
	ProtonPumpInhibitors	51	187	1095
	Triptans	24	194	453
	NSAIDS	41	47	305
	ADHD	20	64	767
	AtypicalAntipsychotics	146	218	756
	UrinaryIncontinence	40	38	249
	Estrogens	80	0	288

Table 6.1 – The distribution of class labels in each dataset. The Yearbook makes an additional separation of N into references that are off-topic and those that are on-topic but does not fit the research question of the review. The number of off-topic references is given in parentheses.

1. Can we separate the screening into two stages?
2. Do we need examples from all stages of screening (Y, M, N)?
3. Should the positive labels match the decisions in the first or second stage of the screening?

To our knowledge, these questions have not yet been considered by existing literature.

Note that the aim of this study is not to improve upon the state of the art, but to investigate how different labeling schemes affect datasets for literature screening.

## 6.2 DATASETS

To address our research questions, we use two datasets that label not only Y and N judgments, but explicitly mark the M subset.

Topic \ Measure	Intertopic			Intratopic		
	wss@95	AUC		wss@95		AUC
		(Cohen)		(Khabsa)		(Khabsa)
CalciumChannelBlockers	.129	<b>.759</b>	.712	<b>.398</b>	.287 (RF)	.825 <b>.873</b> (SVM)
ACEInhibitors	.566	<b>.817</b>	.806	<b>.629</b>	.523 (CNB)	.917 <b>.951</b> (RF)
BetaBlockers	.400	<b>.837</b>	.801	<b>.511</b>	.367 (CNB)	.863 <b>.893</b> (RF)
Opiods	.301	<b>.885</b>	.856	<b>.590</b>	.554 (CNB)	.905 <b>.913</b> (RF)
OralHypoglycemics	.072	<b>.657</b>	.573	<b>.111</b>	.080 (CNB)	.568 <b>.781</b> (SVM)
Statins	.266	<b>.826</b>	.773	<b>.436</b>	.400 (RF)	.873 <b>.915</b> (RF)
SkeletalMuscleRelaxants	.241	.828 <b>.836</b>		<b>.429</b>	.371 (RF)	.740 <b>.794</b> (RF)
Antihistamines	.073	<b>.652</b>	.620	<b>.149</b>	.148 (CNB)	.650 <b>.722</b> (SVM)
ProtonPumpInhibitors	.377	<b>.823</b>	.793	<b>.307</b>	.288 (RF)	.826 <b>.880</b> (RF)
Triptans	.464	.819 <b>.823</b>		.303 <b>.312</b> (RF)		.792 <b>.909</b> (SVM)
NSAIDS	.671	<b>.912</b>	.899	.537 <b>.528</b> (CNB)		.861 <b>.951</b> (SVM)
ADHD	.128	<b>.591</b>	.469	.616 <b>.668</b> (VP)		.908 <b>.951</b> (RF)
AtypicalAntipsychotics	.162	<b>.759</b>	.653	<b>.210</b>	.206 (CNB)	.779 <b>.835</b> (RF)
UrinaryIncontinence	.374	<b>.887</b>	.851	<b>.422</b>	.411 (RF)	.784 <b>.890</b> (SVM)
Estrogens	.176	<b>.693</b>	.588	.292 <b>.375</b> (CNB)		.689 <b>.887</b> (SVM)

Table 6.2 – Results comparing our implementation to the state of the art. Intertopic results report the average over 5 runs. Intratopic results report the average over 10 runs ( $5 \times 2$  cross validation). Both cases use  $(Y || MN)$ . Intertopic state of the art results are taken from Cohen (2008). Intratopic state of the art results are taken from Khabsa et al. (2016), who also report results on Complement Naive Bayes (CNB) by Matwin et al. (2010), Voting Perceptrons (VP) by Cohen et al. (2006), and Support Vector Machines (SVM) by Cohen (2008). Exact intertopic AUC scores are not explicitly reported by Cohen (2008) and have instead been extracted from Figure 1 in his paper.

The datasets each consist of references in the form of PubMed identifiers (PMID) with corresponding inclusion labels (i.e.  $y$ ,  $m$ , or  $n$ ) and topic labels. Article meta-data, as well as titles and abstracts, are not included in either dataset, but can be downloaded from Medline using the Entrez API.<sup>1</sup> The distribution of references from each review stage is reported in Table 6.1. Like in the majority of previous literature, we assume that labeled training data is available, which is generally not true for new reviews. Training data might however exist from past reviews on the same or similar topics. We call such cases where the training data is drawn from similar, but not exactly the same topic, *inter-topic* training.

It may also be possible to have reviewers label small batches of references, and use these as training data for the remainder of the process. Furthermore, systematic reviews sometimes need to be updated, in which case we can use the data from

<sup>1</sup> <https://www.ncbi.nlm.nih.gov/home/develop/api/>

previous iterations for training. We call such cases where the training data is drawn from exactly the same topic *intra-topic* training.

#### 6.2.1 *The Yearbook Dataset*

We construct this dataset by using the references that were considered on topic in the review on clinical NLP done by Névéal and Zweigenbaum (2017); Névéal et al. (2016) for the IMIA Yearbook of Medical Informatics.

This review is updated annually, and the resulting dataset illustrates systematic reviews updates. In each iteration, previous data can be leveraged to train an *intra-topic* classifier.

This dataset is made available in csv and json format,<sup>1</sup> and is planned to be updated to incorporate future iterations of the review.

#### 6.2.2 *The Cohen Dataset*

In one of the early papers on screening automation, Cohen et al. (2006) constructed a dataset from 15 systematic reviews on drug efficacy. This dataset was later extended to 18 (Cohen et al., 2010), then to 24 reviews (Cohen et al., 2009). The smaller dataset comprising 15 reviews has been made available (Cohen et al., 2006).<sup>2</sup> Several methods, including Voting Perceptrons (Cohen et al., 2006), Complement Naive Bayes (Matwin and Sazonova, 2012), SVM (Cohen, 2006, 2008; Cohen et al., 2009), Random Indexing (Jonnalagadda and Petitti, 2014), and Random Forests (Khabisa et al., 2016) have been tested on this dataset, and we can therefore use this dataset to compare our performance against previous work.

This dataset illustrates leveraging training data from similar topics. For each subtopic, data from the other subtopics may be leveraged to build an *inter-topic* classifier.

### 6.3 DOCUMENT RANKING METHOD

We construct a ranker by extracting bag-of- $n$ -grams ( $n \leq 3$ ) over words in the titles and abstracts. We use both *tf-idf* scores and binary features, and both stemmed and unstemmed versions. The  $n$ -grams from the background, method, results, and conclusion of the abstract are also each considered in separation. We also extract article metadata, namely author-assigned keywords, journal names, and publication types. For Cohen we also extract MeSH terms, but omit these for Yearbook since MeSH terms are generally not yet available when reviews are updated.

---

<sup>1</sup> Available from DOI: [doi:10.5281/zenodo.1173076](https://doi.org/10.5281/zenodo.1173076)

<sup>2</sup> The old link has however expired. The data can now be found at <https://dmice.ohsu.edu/cohenaa/systematic-drug-class-review-data.html>

We use a ranking approach only. In practice we ignore the decision boundary used by the logistic regression, and instead leave the decision as to where to stop the search entirely to the reviewer(s). Point measures, such as recall, can therefore only be computed as a function of the position in the ranked list.

We use the implementation of logistic regression in `sklearn` (Pedregosa et al., 2011) trained using stochastic gradient descent, i.e. the `SGDClassifier` trained using log loss. We train the ranker for a maximum of 100,000 iterations.

We generally follow the setup of Cohen et al. (2006), and Khabsa et al. (2016). For intra-topic cross validation we use 2-fold cross validation on each topic and repeat this 5 times. For intertopic training we report the average of 5 repetitions. In each experiment we report the average and standard deviation over all folds and repetitions. All hyperparameters remain constant throughout each experiment.

Unless otherwise stated, we use the default settings for all parameters. We train the ranker and calculate the AUC similarly to Cohen (2008); Cohen et al. (2009). Cross validation was done both inter-topic and intra-topic similarly to the later work of Cohen et al. (2009), and results are reported for each case. We also report the `wss@95` scores (Cohen et al., 2006) in order to compare our results against the naive bayes methods of Matwin et al. (2011). We handle class imbalance by (pseudo)randomly undersampling the majority class to have the same number of instances as the minority class. We however observe that this yields poor results when the number of examples in the majority class is low, and therefore include a minimum of 500 majority class examples.

We increase the weights on the relevant references to 80 to emulate differing costs of misclassification. We also chose  $\alpha = 10^{-4}$  as a reasonable value for the regularization term for the Cohen dataset, and  $\alpha = 0.05$  for Yearbook. We selected these values through experimentation on one of the topics in Cohen (CalciumChannelBlockers), and the first iteration of the Yearbook dataset (2016).

### 6.3.1 *Experimental Setup*

We perform two types of experiments; First, we run our implementation on the Cohen dataset and compare it with the reported performance of previous work. We do this in order to verify the correctness of our algorithm. Second, we perform experiments where we enumerate different ways to treat  $\gamma$ ,  $m$ , and  $N$  labels as positive and negative examples.

We test if it is feasible to emulate the way humans conduct systematic reviews by considering a two-stage approach where we first separate  $\gamma m$  from  $N$ , and then  $\gamma$  from  $m$ .

We test whether treating the  $m$  subset as positive or negative labels impacts the performance by comparing the performance when separating  $\gamma m$  from  $N$  with the

	$(Y  MN)$		$(YM  N)$		$(Y M N)$		$(Y  M)$		$(Y  N)$		$(M  N)$	
	WSS	AUC	WSS	AUC	WSS	AUC	WSS	AUC	WSS	AUC	WSS	AUC
Yearbook	.003	.625	.229	.803	.189	.808	.012	.481	.020	.738	.256	.785
Cohen	.449	.839	.265	.768	.472	.814	.163	.557	.423	.832	.239	.714

(a) Intra-topic results averaged over 10 runs ( $5 \times 2$  cross validation) for different dataset compositions. The averages were computed using weights proportional to the number of articles in each topic ( $Y+M+N$ ,  $Y+M$ ,  $Y+N$ , or  $M+N$ ).

Topic	$(Y  MN)$				$(YM  N)$				$(Y M N)$			
	WSS@95		AUC		WSS@95		AUC		WSS@95		AUC	
	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
ClinicalNLP (Yearbook)	.003	.000	.625	.005	.229	.011	.803	.001	.189	.008	.808	.002
CalciumChannelBlockers	.398	.098	.825	.024	.218	.056	.764	.030	.338	.073	.790	.012
ACEInhibitors	.629	.158	.917	.020	.277	.050	.800	.021	.598	.126	.879	.027
BetaBlockers	.511	.157	.863	.030	.187	.047	.730	.025	.476	.210	.831	.021
Opioids	.590	.193	.905	.052	.366	.096	.817	.033	.705	.063	.881	.035
OralHypoglycemics	.111	.048	.568	.026	.138	.068	.579	.036	.089	.020	.583	.026
Statins	.436	.176	.873	.021	.254	.094	.779	.025	.421	.101	.864	.015
SkeletalMuscleRelaxants	.429	.221	.740	.113	.264	.180	.826	.064	.445	.116	.746	.057
Antihistamines	.149	.089	.650	.089	.126	.038	.566	.026	.239	.092	.596	.013
ProtonPumpInhibitors	.307	.191	.826	.044	.167	.043	.731	.023	.378	.058	.770	.037
Triptans	.303	.237	.792	.075	.300	.039	.746	.030	.412	.067	.691	.026
NSAIDs	.537	.184	.861	.022	.402	.072	.755	.042	.458	.057	.727	.024
ADHD	.616	.148	.908	.026	.697	.096	.910	.017	.828	.057	.906	.011
AtypicalAntipsychotics	.210	.044	.779	.012	.123	.024	.714	.027	.284	.057	.803	.022
UrinaryIncontinence	.422	.144	.784	.032	.207	.089	.660	.040	.475	.072	.750	.038
Estrogens	.292	.089	.689	.026	.266	.093	.715	.040	.319	.056	.693	.026

(b) Intratopic results averaged over 10 runs ( $5 \times 2$  cross validation) for different dataset compositions.

Topic	$(Y  M)$				$(Y  N)$				$(M  N)$			
	WSS@95		AUC		WSS@95		AUC		WSS@95		AUC	
	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
ClinicalNLP (Yearbook)	.012	.000	.481	.005	.020	.002	.738	.003	.256	.004	.785	.001
CalciumChannelBlockers	.141	.039	.590	.030	.421	.106	.852	.024	.208	.069	.743	.032
ACEInhibitors	.165	.083	.631	.059	.410	.370	.918	.032	.256	.063	.771	.020
BetaBlockers	.383	.096	.737	.021	.515	.135	.870	.034	.190	.031	.713	.018
Opioids	.131	.096	.526	.006	.592	.205	.906	.064	.249	.177	.762	.045
OralHypoglycemics	.058	.000	.387	.167	.105	.039	.579	.030	.754	.194	.826	.112
Statins	.125	.052	.560	.037	.439	.184	.879	.047	.240	.086	.708	.028
SkeletalMuscleRelaxants	.240	.143	.547	.017	.297	.149	.668	.078	.226	.163	.800	.067
Antihistamines	.204	.165	.554	.062	.161	.090	.700	.033	.128	.073	.583	.036
ProtonPumpInhibitors	.159	.052	.584	.022	.421	.168	.852	.026	.122	.046	.694	.032
Triptans	.199	.130	.695	.072	.437	.244	.880	.042	.272	.064	.746	.028
NSAIDs	.129	.050	.576	.056	.479	.185	.851	.017	.316	.094	.723	.027
ADHD	.193	.138	.588	.093	.707	.169	.938	.021	.639	.170	.916	.013
AtypicalAntipsychotics	.112	.023	.548	.017	.259	.114	.792	.030	.113	.025	.629	.031
UrinaryIncontinence	.090	.038	.550	.024	.433	.159	.792	.033	.121	.103	.591	.046
Estrogens	-	-	-	-	.233	.034	.686	.038	-	-	-	-

(c) Intratopic results averaged over 10 runs ( $5 \times 2$  cross validation) for different dataset compositions.

Table 6.3 –  $(Y||MN)$  denotes results using  $Y$  as the positive class.  $(YM||N)$  denotes results using  $Y$  and  $M$  as the positive class.  $(Y|M|N)$  denotes results using  $Y$  and  $M$  as the positive class in training, and  $Y$  as the positive class in evaluation.  $(Y||M)$  denotes results using  $Y$  as the positive, and  $M$  as the negative class.  $(Y||N)$  denotes results using  $Y$  as the positive, and  $N$  as the negative class.  $(M||N)$  denotes results using  $M$  as the positive, and  $N$  as the negative class. Estrogens has no  $M$ , and is consequently excluded from the calculations of the results for  $(Y||M)$  and  $(M||N)$ .

performance when separating  $Y$  from  $MN$ .

And finally, we evaluate models where we treat the  $M$  subset as positive examples during training but negative during testing in order to test whether classification in earlier stages generalize to classification in later stages.

We report the work saved over sampling at 95% recall ( $wss@95$ ) (Cohen et al., 2006) and the area under the receiver operator characteristic curve (AUC) (Cohen, 2008) in order to bring our results in line with previous literature (Khabisa et al., 2016). The  $wss@95$  metric measures the theoretical work saved when using the model to retrieve 95% of the relevant articles.

## 6.4 RESULTS

We present our comparison with the state of the art in Table 6.2. In Tables 6.3a–6.3c we present the results of our experiments using data with different compositions of examples in terms of  $Y$ ,  $M$ , and  $N$ .

## 6.5 DISCUSSION

In this section we discuss the results, in order to verify that our system works as intended, and to address the questions we set out in Section 6.1.2 Objective.

### 6.5.1 Performance of Our System

Intuitively: based on the  $wss@95$  scores (Tables 6.3a, 6.3b), our method could save the reviewers from having to look at 46 (Antihistamines) to 1058 (BetaBlockers) references depending on the topic, or about 605 references on average.

The results of our implementation are comparable to state of the art results across the board (Table 6.2). Our implementation exhibits equal or better results for intertopic training (Table 6.2). For intratopic training, our implementation exhibit worse results in terms of AUC, but better scores in terms of  $wss@95$ . Our implementation seems to perform worse than the state of the art mainly on the topics where there are no or very few  $M$  (OralHypoGlycemics, Estrogens). It is also possible that the additional features used by Khabisa et al. (references cited) can explain some of the difference in results.

### 6.5.2 Can We Separate the Screening into Two Stages?

Separating the screening into two stages would entail first screening in terms of  $(Y|M||N)$  followed by  $(Y||M)$ . However, from Tables 6.3a–6.3c it is clear that while  $(Y||MN)$  is feasible,  $(Y||M)$  is considerably more difficult than  $(Y||N)$  or  $(Y||MN)$  (Ta-

bles 6.3a–6.3c). The ranker is however doing a slightly better job on BetaBlockers and Triptans (Tables 6.3b and 6.3c).

In particular, when separating  $\gamma$  from  $M$ , the ranker is not performing much better than chance on many of the topics. This is to be expected, since  $M$  represent those references the human annotators required the full text to judge, and it would be unreasonable to expect the ranker to be able to judge these based only on title and abstract.

Consequently, we can certainly perform  $(\gamma M || N)$  as an initial step, but  $(\gamma || M)$  would at the very least require ranking the full text articles.

98

### 6.5.3 *Do We Need Examples from All Stages of Screening ( $\gamma$ , $M$ , $N$ )?*

We observe similar results for  $(\gamma M || N)$  and  $(\gamma || MN)$  on Cohen, i.e. we can train a ranker using positive examples that were included based on title and abstract  $(\gamma + M)$ , even if these were to turn out to be non-relevant upon inspecting the full text ( $M$ ). On the Yearbook dataset we observe better scores for  $(\gamma M || N)$  than  $(\gamma || MN)$ , likely due to the number of  $\gamma$  available for training (23) being much smaller. In Table 6.3b we can generally observe similar results for  $(\gamma M || N)$  and  $(\gamma || MN)$ , the exceptions being Triptans and NSAIDs where we observe better results for  $(\gamma || MN)$ . We also observe similar results for  $(\gamma M || N)$  and  $(\gamma || MN)$  on the Yearbook data. On some topics we observe better results for  $(\gamma M || N)$ , but the difference is small.

Furthermore, both  $(\gamma || N)$  and  $(M || N)$  seem to give reasonable results, although these results are not directly comparable to the results for  $(\gamma M || N)$ . We can also observe that  $(\gamma || N)$  is generally easier than  $(M || N)$ . This could be due to  $\gamma$  containing fewer borderline cases.

Consequently, we do need positive examples drawn from  $\gamma$  or  $M$ , as well as negative examples drawn from  $N$ . It seems to make less difference whether we consider  $M$  to be positive or negative examples and we may be able to exclude either  $\gamma$  or  $M$  in training.

Interestingly it seems from Table 6.3a that it is more difficult to classify in terms of  $(\gamma M || N)$  than  $(\gamma || MN)$  on Cohen, but the inverse is true on Yearbook. This might be explained by the small number of  $\gamma$  on Yearbook (11), and we can observe the same on the topics in Cohen with few  $\gamma$  (SkeletalMuscleRelaxants, ADHD). Oral-Hypoglycemics have only 3  $M$  and Estrogens no  $M$  at all, and we therefore exclude these topics from the results.

### 6.5.4 *Can We Use $M$ as Positive Examples for Training?*

Cohen et al. previously discovered that while intratopic data is generally better than intertopic data (Cohen et al., 2006), the less targeted intertopic data can com-

plement the intratopic data if the intratopic data is scarce (Cohen et al., 2009, 2006). Our results suggest the same (Table 6.2), but also that we can generally use  $M$  as training examples to complement the  $Y$ . The intuition behind these ideas is similar: while it is generally important to have training data targeted for the particular problem, it is also important to have sufficient amounts of data, and less targeted training data can provide a supplement if only scarce amounts of data is available. We can further compare the results for intratopic ( $Y|M|N$ ) versus the results for intertopic ( $Y||MN$ ) in Tables 6.3b and 6.2 to get a sense of whether complementing our training data by using  $M$  as positive examples works better than complementing our training data with less targeted data from similar topics.

We observe better results for intertopic ( $Y||MN$ ) for OralHypoglycemics, SkeletalMuscleRelaxants, Antihistamines, Triptans, NSAIDs, and UrinaryIncontinence. This might in part be explained by OralHypoglycemics, SkeletalMuscleRelaxants and Antihistamines having few  $Y$ . We observe better results for intratopic ( $Y|M|N$ ) on ACEInhibitors, ProtonPumpInhibitors, and ADHD. It is not clear why we observe this difference on these topics.

#### 6.5.5 Strategies for Ranking Articles

From Tables 6.3a–6.3c it seems that there is no single approach that is clearly better for any kind of data. Which approach works best depends on the number of articles in each class, as well as the exact nature of articles in each stage. What parts of the data to e.g. use for training must therefore be decided based on the characteristics of the dataset, or by testing multiple approaches.

The results and conclusions of this study guided the strategic choices we made for the system submitted to the CLEF eHealth shared task *Technology Assisted Reviews in Empirical Medicine* (Kanoulas et al., 2017a; Norman et al., 2017b). We submitted four runs using different machine learning methods: 1) the ( $YM||N$ ) approach described here 2) an ( $YM||N$ ) approach using standard logistic regression (i.e. not trained using SGD), and 3) two variations of logistic regression with active learning, where the system starts using the ( $Y|M|N$ ) approach and later switches to using the ( $Y||MN$ ) approach once a sufficient number of  $Y$  have been discovered.

On the Cohen dataset approach 2 worked better than approach 1 for intratopic training and vice-versa, and we could reliably see improvements over either of these by using active learning. On the CLEF data however, approach 1 achieved much better results than either approach 2 or 3. We believe that this was at least partly due to the small number of relevant articles per topic in the CLEF dataset (Norman et al., 2017b).

Our participation placed third to fifth in the evaluation overall, depending on metric used, and placed first among the systems not using active learning.

### 6.5.6 *Limitations*

This work relied on two datasets and a ranker developed in-house. It is not clear how the results generalize to other domains and datasets, or to other machine learning methods.

We observe fairly large variance for many of the runs (Tables 6.3b, 6.3c), and on many topics. This is particularly problematic for the wss metric, but it also affects the AUC metric even averaged over ten repetitions. For instance, Estrogens has no  $M$ , and we should therefore expect the same results for  $(Y||MN)$  and  $(YM||N)$ , yet we observe differences roughly equal to the standard deviation for the AUC. Previous literature generally do not report their variance, which complicates the comparison with previous results.

### 6.5.7 *Future work*

We are working on extending the system to use additional machine learning methods, including deep artificial neural networks, and to complement the system with information retrieval methods.

## 6.6 CONCLUSION

We find that in order to train rankers to automate the screening process we need to use 1) examples of excluded references ( $N$ ), and 2) references included in either the first ( $M$ ) or second stage of the screening ( $Y$ ). In the systematic reviews, the  $M$  are those articles that were excluded after reading the full text, and so are in reality negative examples. However, our results suggest that these can still be used as positive examples for training. It may well be possible to construct an accurate ranker using only the  $M$  as the positive examples, without any real positive examples (i.e.  $Y$ ) at all.

Our best results are achieved with  $(Y||MN)$  on the Cohen dataset, whereas our best results are achieved with  $(Y|M|N)$  on the Yearbook dataset. Given that the distribution of the labels is similar in both datasets it is likely that greater contribution of the  $M$  on the Yearbook dataset is due to its smaller size. For any new systematic review we only have whatever training data we label ourselves, and data scarcity is therefore one of the major issues we need to overcome. Even for systematic review updates the amount of positive training data available is typically modest since the number of included articles in any systematic review tends to be small (the  $Y$  column in Table 6.1).

Since the number of references that are provisionally included based on title and abstract ( $Y+M$ ) can outnumber the final includes ( $Y$ ) by almost ten to one (Table 6.1),

using examples of  $M$  in addition to  $Y$  suggests a straightforward way to increase the amount of training data available (i.e. the  $Y|M|N$  approach), and thus potentially overcome the data scarcity problem, particularly if we do not have access to inter-topic training data. This does not seem to have been considered in previous work. Our results also agree with the state of the art and suggest that common-off-the-shelf machine learning algorithms can accurately predict topical relevance of candidate articles for inclusion in systematic reviews.

In light of the results, we recommend that future datasets intended to be used either for training or for evaluation of document screening should include a tripartite labeling reflecting the two filtering stages in manual systematic reviews. Strictly, only the distinction between  $Y_M$  and  $N$  is necessary for training, but we still likely want to only treat  $Y$  as positive during evaluation, since only these would be considered relevant for the purposes of the systematic review.

#### ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676207.

The material in chapter 7 has been published as:

(Norman et al., 2017b): Norman, C., Leeﬂang, M., and N  v  ol, A. (2017b). LIMSI@CLEF eHealth 2017 task 2: Logistic regression for automatic article ranking. *Working Notes of CLEF*

This is a shared task participation. A shared task provides a common platform to compare performance, and is a common way to assess the relative performance of competing approaches in NLP or IR. However, the CLEF eHealth task diverged from accepted shared task practices by not blinding participants to the gold standard, by using self-reported results, and by not providing pre-specified evaluation metrics (section 7.5.5). This may have introduced bias, and the relative results between systems should be interpreted with this in mind.

The article has undergone peer review and has been published in its entirety in the conference proceedings.

With this article we have tried to address the following research questions:

- RQ 2 *How do different screening automation approaches compare with each other for DTA screening?*
- RQ 3 *Are the screening automation methods we develop competitive with the current state-of-the-art?*

#### AUTHOR’S CONTRIBUTIONS

CN wrote the first draft and conducted the experiments. All authors conceived and designed the study. All authors read and approved the final manuscript.

## LIMSI@CLEF eHealth 2017 Task 2: Logistic Regression for Automatic Article Ranking

Christopher R. Norman, Mariska M.G. Leeflang, & Aurélie Névéol

103

### Abstract

This paper describes the participation of the LIMSI-MIROR team at CLEF eHealth 2017, task 2. The task addresses the automatic ranking of articles in order to assist with the screening process of Diagnostic Test Accuracy (DTA) Systematic Reviews. We used a logistic regression classifier and handled class imbalance using a combination of class reweighting and undersampling. We also experimented with two strategies for relevance feedback. Our best run obtained an overall Average Precision of 0.179 and Work Saved over Sampling @95% Recall of 0.650. This run uses stochastic gradient descent for training but no feature selection or relevance feedback. We observe high performance variation within the queries in the test set. Nonetheless, our results suggest that automatic assistance is promising for ranking the DTA literature as it could reduce the screening workload for review writer by 65% on average.

### 7.1 INTRODUCTION

Systematic reviews seek to gather all available published evidence for a given topic and provide an informed analysis of the results. This work constitutes some of the strongest forms of scientific evidence. Systematic reviews are an integral part of evidence based medicine in particular, and serve a key role in informing and guiding public and institutional decision-making. Systematic reviews for Diagnostic Test Accuracy (DTA) studies have been shown particularly challenging compared to other types of reviews because of the difficulty in defining search strategies offering adequate levels of sensitivity and specificity (Petersen et al., 2014). For this reason, there is a need to particularly investigate automation strategies to assist DTA systematic review writers in the time-consuming screening process.

Methods for automating the screening process in systematic reviews have been actively researched over the years (O'Mara-Eves et al., 2015), with promising results

obtained using a range of machine learning methods. However, previous work has not addressed DTA studies.

This paper describes the work underlying our participation in the CLEF 2017 eHealth Task 2 (Goeuriot et al., 2017; Kanoulas et al., 2017a). This work is part of an ongoing effort on providing automatic assistance for the screening process in systematic reviews addressing a variety of topics, including DTA studies.

The remainder of this paper is organized as follows; subsection 2 presents the datasets used for system development. subsection 3 provides an overview of our system and describes each component. Finally, subsection 4 reports our results and subsection 5 provides an analysis of our methods and participation in the task.

## 7.2 DATASETS

The task relied on a corpus comprising 50 DTA systematic review topics associated with the full list of articles retrieved by an expert query and assessed for inclusion based on title and abstract or full text. The corpus was split into a development dataset comprising 20 topics and a test set comprising the remaining 30 topics. Our classifier was trained on the development dataset and evaluated on the test dataset. We have also used a dataset of systematic reviews on drug class efficacy due to Cohen et al. (Cohen et al., 2006) to develop the methods applied in this task. Several groups have been using this dataset in the past (Cohen et al., 2006; Khabsa et al., 2016), which gives us a way to compare our results with previous work, although we can of course only do by using the same evaluation metrics and training modes as previous work.

For both the CLEF and Cohen datasets we know the inclusion decisions based on the abstracts, as well as the inclusion decisions based on the full text. We thus have two definitions of positive examples, depending on whether we use the abstract decisions or full text decisions as the gold standard.

We use a tripartite labeling to reflect this:

- ◆ NO (N) is the set of articles that were excluded based on the abstract
- ◆ MAYBE (M) is the set of articles that were preliminarily included based on the abstract, but later excluded based on the full text
- ◆ YES (Y) is the set of articles that were included based on both the abstract and the full text, and later used in the meta-analysis

Table 7.1 shows a breakdown of the distribution of examples for each class the CLEF and Cohen datasets used in our work.

Following the work of Cohen (2008), we also distinguish between two modes of training:

Dataset	Topic	Absolute number			Relative number		
		Y	M	N	Y	M	N
Cohen	CalciumChannelBlockers	100	180	938	8.21%	14.78%	77.01%
	ACEInhibitors	41	142	2361	1.61%	5.58%	92.81%
	BetaBlockers	42	260	1770	2.03%	12.55%	85.42%
	Opioids	15	33	1867	0.78%	1.72%	97.49%
	OralHypoglycemics	136	3	364	27.04%	0.60%	72.37%
	Statins	85	88	3292	2.45%	2.54%	95.01%
	SkeletalMuscleRelaxants	9	25	1609	0.55%	1.52%	97.93%
	Antihistamines	16	76	218	5.16%	24.52%	70.32%
	ProtonPumpInhibitors	51	187	1095	3.83%	14.03%	82.15%
	Triptans	24	194	453	3.58%	28.91%	67.51%
	NSAIDs	41	47	305	10.43%	11.96%	77.61%
	ADHD	20	64	767	2.35%	7.52%	90.13%
	AtypicalAntipsychotics	146	218	756	13.04%	19.46%	67.50%
	UrinaryIncontinence	40	38	249	12.23%	11.63%	77.61%
	Estrogens	80	0	288	21.74%	0.00%	78.26%
	Total	846	1555	16333	4.52%	8.30%	87.18%
CLEF (train)	CD007394	47	48	2450	1.85%	1.89%	96.27%
	CD007427	17	106	1398	1.12%	6.97%	91.91%
	CD008054	41	233	2940	1.28%	7.25%	91.47%
	CD008643	7	4	15065	0.05%	0.03%	99.93%
	CD008686	5	2	3946	0.13%	0.05%	99.82%
	CD008691	20	53	1243	1.52%	4.03%	94.45%
	CD009020	12	150	1422	0.76%	9.47%	89.77%
	CD009323	9	113	3757	0.23%	2.91%	96.85%
	CD009591	41	103	7847	0.51%	1.29%	98.20%
	CD009593	24	54	14844	0.16%	0.36%	99.48%
	CD009944	64	53	1064	5.42%	4.49%	90.09%
	CD010409	41	35	43287	0.09%	0.08%	99.82%
	CD010438	3	36	3211	0.09%	1.11%	98.80%
	CD010632	14	18	1472	0.93%	1.20%	97.87%
	CD010771	1	47	274	0.31%	14.60%	85.09%
	CD011134	49	166	1738	2.51%	8.50%	88.99%
	CD011548	1	108	12591	0.01%	0.85%	99.14%
	CD011549	1	1	12699	0.01%	0.01%	99.98%
	CD011975	60	559	7582	0.73%	6.82%	92.45%
	CD011984	28	426	7738	0.34%	5.20%	94.46%
	Total	485	2315	146568	0.32%	1.55%	98.13%
CLEF (test)	CD007431	47	9	2050	2.23%	0.43%	97.34%
	CD008081	10	16	944	1.03%	1.65%	97.32%
	CD008760	9	3	52	14.06%	4.69%	81.25%
	CD008782	34	11	10460	0.32%	0.10%	99.57%
	CD008803	99	0	5121	1.90%	0.00%	98.10%
	CD009135	19	58	714	2.40%	7.33%	90.27%
	CD009185	23	69	1523	1.42%	4.27%	94.30%
	CD009372	10	15	2223	0.44%	0.67%	98.89%
	CD009519	46	58	5867	0.77%	0.97%	98.26%
	CD009551	16	30	1865	0.84%	1.57%	97.59%
	CD009579	79	59	6317	1.22%	0.91%	97.86%
	CD009647	17	39	2729	0.61%	1.40%	97.99%
	CD009786	6	4	2055	0.29%	0.19%	99.52%
	CD009925	55	405	6071	0.84%	6.20%	92.96%
	CD010023	14	38	929	1.43%	3.87%	84.70%
	CD010173	10	13	5472	0.18%	0.24%	99.58%
	CD010276	24	30	5441	0.44%	0.55%	99.02%
	CD010339	9	105	12689	0.07%	0.82%	99.11%
	CD010386	1	1	623	0.16%	0.16%	99.68%
	CD010542	8	12	328	2.30%	3.45%	94.25%
	CD010633	3	1	1569	0.19%	0.06%	99.75%
	CD010653	0	45	7957	0.00%	0.56%	99.44%
	CD010705	18	5	91	15.79%	4.39%	79.82%
	CD010772	11	36	269	3.48%	11.39%	85.13%
	CD010775	4	7	230	1.66%	2.90%	95.44%
	CD010783	11	19	10875	0.10%	0.17%	99.72%
	CD010860	4	3	87	4.26%	3.19%	92.55%
	CD010896	3	3	163	1.78%	1.78%	96.45%
	CD011145	48	154	10670	0.44%	1.42%	98.14%
	CD012019	1	2	10314	0.01%	0.02%	99.97%
	Total	639	1250	115698	0.54%	1.06%	98.39%
	Total (train + test)	1124	3565	262266	0.42%	1.34%	98.24%

Table 7.1 – The distribution of class labels in each dataset.

- ❖ **INTERTOPIC** training uses articles from a different topic (systematic review) for training
- ❖ **INTRATOPIC** training uses articles from the current topic (systematic review) for training

### 7.3 METHOD

We first give an overview of our system, which relies on logistic regression, in subsection 7.3.1. Further details about the system are given in sections 7.3.2–7.3.5, including features, strategies to handle class imbalance and implement relevance feedback.

#### 7.3.1 Overview

We have tried the following two classifiers:

- ❖ **CLASSIFIER 1** uses logistic regression trained using stochastic gradient descent on all features
- ❖ **CLASSIFIER 2** uses standard logistic regression trained using standard methods on a subset of the features, and with additional preprocessing to improve the throughput

We have tried three approaches to relevance feedback:

- ❖ **NO RELEVANCE FEEDBACK**
- ❖ **ABRUPT** uses intertopic ranking until a sufficient number of relevant and non-relevant articles have been identified, and then switches to using intratopic ranking based on the identified articles
- ❖ **GRADUAL** initially uses intertopic ranking, and gradually improves the model using both **Y** and **M** identified through relevance feedback

In total, we have submitted the following four runs to the **CLEF** evaluation:

- ❖ **NO · AF · FULL** uses classifier 1 with no relevance feedback
- ❖ **NO · AF** uses classifier 2 with no relevance feedback
- ❖ **ABRUPT** uses classifier 2 with **ABRUPT** relevance feedback
- ❖ **GRADUAL** uses classifier 2 with **GRADUAL** relevance feedback

### 7.3.2 *Classification Approach*

We are currently using two classification systems. Both use logistic regression but differ in how the model is optimized and the amounts and types of pre- and postprocessing that is performed. Both methods use implementations provided by sklearn (Pedregosa et al., 2011).

Our first method, which is used in `NO · AF · FULL` tends to work well for intertopic classification on previous datasets (see table 7.3), presumably because it generalizes better. This system uses logistic regression trained using stochastic gradient descent. The only preprocessing done is the normalization of numerals.

Our second method, which is used in `NO · AF`, `ABRUPT`, and `GRADUAL` uses standard methods for training (liblinear). This version tends to work well on intratopic classification on previous datasets (see table 7.3), but does not scale as well with data volume. We therefore need to do additional preprocessing to reduce the number of features and keep running times down. We thus remove features with variance less than a predefined threshold, we only consider  $n$ -grams with high mutual information with the target class in the training set, we normalize numerals, and we extract the principal components from the resulting data.

Principal component analysis tends to reduce overfitting in our experiments, and it also drastically reduces the time it takes to train and apply the classifier, which is mostly important when we use relevance feedback.

### 7.3.3 *Features*

For all classifiers we extract  $n$ -grams ( $n \leq 5$ ) from the titles and abstracts. We also extract publication type, journal names, author assigned keywords, MeSH terms, and backward references, where these are available. The backward references are only available for references pointing to articles available in PubMed Central, and this feature set is therefore fairly sparse.

Not all feature sets are useful for identifying DTA studies, but the current model has been constructed such that irrelevant features should not adversely affect the performance. All the feature sets have been shown to be useful on some domain. For instance MeSH terms might not be useful for DTA studies, but we have previously found them to be useful in identifying topics related to drug efficacy.

### 7.3.4 *Class Imbalance*

Class imbalance can be handled using undersampling, or by class reweighting. We are currently using a combination of both these approaches.

Topic	NO · AF · FULL	NO · AF	VP	CNB	RF
CalciumChannelBlockers	.398	<b>.408</b>	<.100	.234	.287
ACEInhibitors	<b>.629</b>	.517	.318	.523	.447
BetaBlockers	<b>.511</b>	.427	.284	.367	.361
Opioids	.590	<b>.641</b>	<.190	.554	.455
OralHypoglycemics	.111	<b>.153</b>	<.050	.080	.074
Statins	.436	<b>.573</b>	.242	.315	.400
SkeletalMuscleRelaxants	<b>.429</b>	.179	-.050	.265	.371
Antihistamines	.149	<b>.157</b>	.080	.148	.030
ProtonPumpInhibitors	.307	<b>.320</b>	<.180	.229	.288
Triptans	.303	<b>.312</b>	.030	.279	<b>.312</b>
NSAIDS	.537	<b>.600</b>	.352	.528	.404
ADHD	.616	.530	<b>.668</b>	.622	.447
AtypicalAntipsychotics	.210	<b>.234</b>	.140	.206	.199
UrinaryIncontinence	<b>.422</b>	.365	.260	.290	.411
Estrogens	.292	<b>.475</b>	.140	.375	.180

Table 7.2 – Comparison in terms of wss@95% with previous literature using Voting Perceptrons, Complement Naive Bayes, and Random Forests, as reported by Khabsa et al. (2016). We here only have state of the art metrics for the intratopic case.

**CLASS WEIGHTS** We set the weight for the positive class to 80 for the initial inter-topic classifier. We have determined this to be a reasonable weight experimentally using the Cohen dataset.

For the GRADUAL relevance feedback we also attached higher weights to the intratopic training examples identified through relevance feedback.

**UNDERSAMPLING** In order to reduce the effects of the class imbalance we under-sample the training set to include an equal number of Y, M, and N. However, by doing so we end up with only around 1500 training samples. PCA yields at most the same number of principal components as we have input samples, and 1500 is generally too few principal components to build an accurate classifier. For the second model we therefore perform undersampling in two steps; We first select a maximum of 500 Y, 1000 M, and 1500 N that we feed into the feature extraction pipeline, which thus determines the number of features in our model. We then select a smaller undersample to use for training.

We take a new undersample in each iteration of relevance feedback.

Topic	Intertopic			RF	Intratopic		
	NO · AF				NO · AF		
	FULL	Cohen	GRADUAL		FULL	Cohen	Khabsa
CalciumChannelBlockers	.759	.773	.712	.862	.825	.868	.873
ACEInhibitors	.817	.782	.806	.899	.917	.925	.946
BetaBlockers	.837	.832	.801	.860	.863	.871	.891
Opioids	.885	.902	.856	.936	.905	.893	.897
OralHypoglycemics	.657	.581	.573	.753	.568	.768	.781
Statins	.826	.798	.773	.797	.873	.922	.900
SkeletalMuscleRelaxants	.826	.823	.836	.812	.740	.527	.738
Antihistamines	.652	.600	.620	.752	.650	.655	.722
ProtonPumpInhibitors	.823	.790	.793	.886	.826	.860	.860
Triptans	.819	.796	.823	.804	.792	.808	.909
NSAIDS	.912	.828	.899	.922	.861	.935	.951
ADHD	.591	.606	.469	.740	.908	.897	.924
AtypicalAntipsychotics	.759	.645	.653	.855	.779	.803	.835
UrinaryIncontinence	.887	.875	.851	.888	.784	.885	.890
Estrogens	.693	.649	.588	.879	.689	.912	.887

Table 7.3 – Comparison in terms of AUC with previous literature using Support Vector Machines (Cohen) and Random Forests (Khabsa), as reported by Khabsa et al. (Khabsa et al., 2016), and Cohen et al. (Cohen, 2008). Exact intertopic AUC scores are not explicitly reported by Cohen et al. and have instead been extracted from Figure 1 in their paper.

### 7.3.5 Relevance Feedback

We use two schemes for relevance feedback. For both schemes we retrain the classifier each time we retrieve relevance feedback.

ABRUPT trains an initial intertopic classifier on the training dataset and ranks the test dataset in descending order of confidence. The system then iteratively asks for feedback for the top ranked results. When enough positive and negative examples have been identified, the system switches to using a classifier trained on the examples identified from relevance feedback. Additional examples are added to the intratopic classifier as they are discovered.

The idea behind this system is that on some topics in Cohen we can train highly performing intratopic classifiers using very small amounts of data, and we have observed that even trained on small amounts of data these sometimes outperform intertopic classifiers by a large margin. In these cases it might make sense to switch to intratopic classification as soon as we can.

We set the minimum number of positive examples to 4, and the minimum number of negative examples to 10.

Topic	Y				YM				N			
	w/o RF		w/ RF		w/o RF		w/ RF		w/o RF		w/ RF	
	NO · AF		GRADUAL		NO · AF		GRADUAL		ABRUPT		GRADUAL	
	FULL		ABRUPT		FULL							baseline
CD007431	0.047	0.016	0.026	0.013	0.010 ± 0.005	0.065	0.026	0.044	0.019	0.015 ± 0.005		
CD008081	0.146	0.099	0.087	0.046	0.016 ± 0.010	0.114	0.097	0.060	0.041	0.032 ± 0.009		
CD008760	0.790	0.516	0.569	0.835	0.169 ± 0.052	0.886	0.644	0.734	0.807	0.210 ± 0.050		
CD008782	0.057	0.231	0.032	0.042	0.004 ± 0.002	0.060	0.242	0.040	0.050	0.005 ± 0.002		
CD008803	0.181	0.131	0.147	0.120	0.020 ± 0.003	0.181	0.131	0.147	0.120	0.020 ± 0.002		
CD009135	0.382	0.217	0.149	0.324	0.030 ± 0.009	0.485	0.349	0.266	0.493	0.102 ± 0.012		
CD009185	0.041	0.049	0.080	0.027	0.018 ± 0.006	0.139	0.096	0.135	0.085	0.060 ± 0.007		
CD009372	0.122	0.189	0.078	0.081	0.007 ± 0.006	0.080	0.107	0.056	0.060	0.014 ± 0.004		
CD009519	0.031	0.022	0.034	0.020	0.009 ± 0.002	0.059	0.051	0.067	0.038	0.019 ± 0.002		
CD009551	0.199	0.140	0.222	0.157	0.011 ± 0.006	0.287	0.259	0.259	0.284	0.027 ± 0.006		
CD009579	0.172	0.105	0.257	0.286	0.013 ± 0.002	0.253	0.157	0.259	0.286	0.022 ± 0.002		
CD009647	0.038	0.024	0.019	0.026	0.008 ± 0.004	0.052	0.040	0.034	0.068	0.022 ± 0.004		
CD009786	0.028	0.024	0.012	0.008	0.006 ± 0.007	0.034	0.055	0.190	0.014	0.008 ± 0.007		
CD009925	0.114	0.080	0.044	0.077	0.010 ± 0.002	0.334	0.168	0.151	0.285	0.071 ± 0.003		
CD010023	0.089	0.058	0.051	0.085	0.020 ± 0.008	0.303	0.273	0.168	0.222	0.058 ± 0.009		
CD010173	0.014	0.008	0.010	0.001	0.003 ± 0.004	0.025	0.014	0.015	0.003	0.006 ± 0.003		
CD010276	0.072	0.055	0.032	0.003	0.006 ± 0.003	0.108	0.100	0.057	0.007	0.011 ± 0.003		
CD010339	0.018	0.067	0.021	0.035	0.001 ± 0.002	0.043	0.046	0.020	0.040	0.010 ± 0.001		
CD010386	0.053	0.083	0.091	0.167	0.009 ± 0.023	0.031	0.044	0.050	0.085	0.010 ± 0.017		
CD010542	0.082	0.145	0.190	0.038	0.036 ± 0.021	0.131	0.158	0.188	0.110	0.068 ± 0.018		
CD010633	0.015	0.010	0.010	0.002	0.006 ± 0.014	0.071	0.028	0.023	0.003	0.006 ± 0.010		
CD010653	-	-	-	-	-	0.011	0.016	0.012	0.005	0.006 ± 0.002		
CD010705	0.240	0.220	0.389	0.312	0.174 ± 0.037	0.250	0.247	0.444	0.380	0.214 ± 0.036		
CD010772	0.117	0.035	0.069	0.086	0.048 ± 0.020	0.211	0.155	0.214	0.343	0.158 ± 0.021		
CD010775	0.187	0.101	0.170	0.069	0.034 ± 0.031	0.623	0.462	0.433	0.258	0.062 ± 0.023		
CD010783	0.071	0.037	0.020	0.009	0.002 ± 0.003	0.044	0.103	0.051	0.026	0.004 ± 0.002		
CD010860	0.188	0.139	0.135	0.032	0.070 ± 0.042	0.168	0.126	0.134	0.047	0.104 ± 0.042		
CD010896	0.347	0.093	0.248	0.239	0.037 ± 0.033	0.213	0.100	0.154	0.163	0.054 ± 0.028		
CD011145	0.027	0.009	0.023	0.011	0.005 ± 0.001	0.108	0.044	0.058	0.038	0.019 ± 0.002		
CD012019	0.003	0.002	0.003	0.002	0.001 ± 0.008	0.003	0.002	0.002	0.001	0.001 ± 0.001		
Average	0.179	0.145	0.143	0.146	0.027 ± 0.003	0.133	0.100	0.111	0.109	0.047 ± 0.003		

Table 7.4 – Average precision score for all topics in the CLEF dataset.

Topic	Y			YM			N		
	W/O RF		baseline	W/ RF		baseline	W/ RF		baseline
	NO · AF			W/O RF			W/ RF		
	FULL			ABRUPT	GRADUAL		ABRUPT	GRADUAL	
CD007431	0.825	0.673	0.762	0.704	0.773	0.684	0.769	0.700	0.501 ± 0.060
CD008081	0.907	0.872	0.801	0.695	0.801	0.751	0.653	0.603	0.506 ± 0.057
CD008760	0.963	0.895	0.933	0.955	0.976	0.917	0.927	0.920	0.536 ± 0.082
CD008782	0.942	0.983	0.351	0.876	0.939	0.977	0.360	0.888	0.500 ± 0.044
CD008803	0.944	0.889	0.898	0.915	0.944	0.889	0.898	0.915	0.504 ± 0.028
CD009135	0.962	0.875	0.856	0.959	0.875	0.841	0.744	0.897	0.524 ± 0.033
CD009185	0.790	0.676	0.677	0.744	0.779	0.693	0.618	0.687	0.514 ± 0.031
CD009372	0.947	0.970	0.817	0.853	0.815	0.839	0.714	0.749	0.501 ± 0.059
CD009519	0.865	0.802	0.862	0.807	0.851	0.792	0.838	0.766	0.503 ± 0.028
CD009551	0.960	0.961	0.892	0.946	0.945	0.953	0.862	0.930	0.506 ± 0.043
CD009579	0.902	0.784	0.871	0.913	0.902	0.827	0.821	0.875	0.505 ± 0.025
CD009647	0.747	0.774	0.674	0.830	0.720	0.706	0.628	0.835	0.504 ± 0.038
CD009786	0.918	0.854	0.756	0.736	0.895	0.858	0.743	0.762	0.501 ± 0.093
CD009925	0.947	0.822	0.695	0.839	0.883	0.674	0.616	0.753	0.518 ± 0.013
CD010023	0.890	0.806	0.780	0.879	0.872	0.864	0.780	0.889	0.513 ± 0.039
CD010173	0.929	0.805	0.882	0.379	0.901	0.770	0.766	0.383	0.502 ± 0.062
CD010276	0.956	0.938	0.882	0.279	0.940	0.904	0.801	0.345	0.503 ± 0.038
CD010339	0.887	0.860	0.777	0.873	0.816	0.760	0.580	0.764	0.504 ± 0.028
CD010386	0.971	0.982	0.984	0.992	0.820	0.686	0.804	0.531	0.510 ± 0.202
CD010542	0.794	0.748	0.793	0.650	0.692	0.606	0.696	0.676	0.512 ± 0.066
CD010633	0.846	0.873	0.830	0.315	0.884	0.903	0.869	0.414	0.500 ± 0.145
CD010653	-	-	-	-	0.688	0.753	0.721	0.497	0.500 ± 0.043
CD010705	0.713	0.621	0.867	0.802	0.655	0.611	0.868	0.817	0.544 ± 0.059
CD010772	0.805	0.455	0.581	0.679	0.661	0.485	0.551	0.719	0.537 ± 0.041
CD010775	0.956	0.893	0.601	0.847	0.982	0.947	0.762	0.914	0.508 ± 0.084
CD010783	0.935	0.935	0.926	0.916	0.918	0.941	0.848	0.870	0.502 ± 0.052
CD010860	0.832	0.840	0.837	0.217	0.697	0.656	0.667	0.152	0.514 ± 0.111
CD010896	0.855	0.648	0.721	0.904	0.756	0.691	0.605	0.733	0.503 ± 0.115
CD011145	0.860	0.736	0.792	0.750	0.868	0.751	0.724	0.723	0.504 ± 0.020
CD012019	0.964	0.960	0.962	0.946	0.917	0.767	0.806	0.542	0.497 ± 0.160
Average	0.890	0.825	0.795	0.766	0.839	0.780	0.735	0.708	0.509 ± 0.014

Table 7.5 – Normalized average precision score for all topics in the CLEF dataset.

Topic	Y				Y   MN				Y   N			
	W/O RF		W/ RF		W/O RF		W/ RF		W/O RF		W/ RF	
	NO · AF		ABRUPT		baseline		NO · AF		ABRUPT		GRADUAL	
	FULL						FULL					baseline
CD007431	0.621	0.298	0.356	0.415	0.071 ± 0.078	0.297	0.079	0.356	0.415	0.079	0.323	0.030 ± 0.052
CD008081	0.452	0.260	0.056	0.391	0.042 ± 0.082	0.430	0.138	0.283	0.365	0.283	0.365	0.023 ± 0.048
CD008760	0.731	0.512	0.591	0.575	0.023 ± 0.075	0.731	0.575	0.591	0.575	0.591	0.575	0.075 ± 0.087
CD008782	0.767	0.873	-0.037	0.476	0.036 ± 0.046	0.706	0.857	-0.039	0.476	0.857	0.476	0.013 ± 0.035
CD008803	0.787	0.584	0.528	0.612	0.009 ± 0.023	0.787	0.584	0.528	0.612	0.584	0.312	0.009 ± 0.023
CD009135	0.759	0.457	0.739	0.783	0.048 ± 0.067	0.439	0.493	0.935	0.580	0.493	0.580	0.012 ± 0.026
CD009185	0.377	0.073	0.096	0.500	0.031 ± 0.057	0.377	0.026	0.024	0.114	0.026	0.114	0.013 ± 0.025
CD009372	0.654	0.844	0.051	0.170	0.040 ± 0.083	0.353	0.461	0.139	0.170	0.139	0.170	0.035 ± 0.050
CD009519	0.597	0.294	0.442	0.624	0.011 ± 0.034	0.483	0.219	0.336	0.291	0.336	0.291	0.006 ± 0.022
CD009551	0.856	0.866	0.584	0.834	0.070 ± 0.075	0.757	0.838	0.368	0.667	0.368	0.667	0.014 ± 0.037
CD009579	0.531	0.153	0.327	0.522	0.012 ± 0.027	0.580	0.275	0.203	0.351	0.203	0.351	0.008 ± 0.021
CD009647	0.321	0.469	0.124	0.577	0.058 ± 0.073	0.240	0.243	0.028	0.499	0.243	0.499	0.018 ± 0.033
CD009786	0.799	0.656	0.134	0.248	0.098 ± 0.124	0.621	0.656	0.234	0.248	0.234	0.248	0.041 ± 0.085
CD009925	0.810	0.346	0.050	0.277	0.022 ± 0.033	0.469	0.0	-0.040	0.002 ± 0.010	0.515	-0.037	0.002 ± 0.010
CD010023	0.714	0.649	0.572	0.693	0.085 ± 0.085	0.492	0.474	0.515	0.024 ± 0.034	0.474	0.662	0.024 ± 0.034
CD010173	0.777	0.476	0.555	0.078	0.039 ± 0.083	0.671	0.271	0.174	0.034 ± 0.057	0.271	0.078	0.034 ± 0.057
CD010276	0.811	0.803	0.486	-0.031	0.031 ± 0.050	0.719	0.511	0.217	0.024 ± 0.034	0.511	-0.031	0.024 ± 0.034
CD010339	0.193	0.114	0.077	0.330	0.052 ± 0.095	0.346	0.192	0.026	0.011 ± 0.023	0.192	0.219	0.011 ± 0.023
CD010386	0.920	0.931	0.932	0.940	0.461 ± 0.289	0.616	0.337	0.571	0.286 ± 0.237	0.337	0.019	0.286 ± 0.237
CD010542	0.464	0.171	0.099	0.191	0.056 ± 0.097	0.232	0.065	0.099	0.041 ± 0.061	0.065	0.191	0.041 ± 0.061
CD010633	0.637	0.741	0.584	0.050	0.203 ± 0.197	0.637	0.741	0.584	0.143 ± 0.164	0.741	0.050	0.143 ± 0.164
CD010653	-	-	-	-	-	0.218	0.272	0.227	0.014 ± 0.035	0.272	0.050	0.014 ± 0.035
CD010705	0.213	0.187	0.625	0.503	0.040 ± 0.064	0.064	0.161	0.564	0.014 ± 0.048	0.161	0.494	0.014 ± 0.048
CD010772	0.532	0.070	0.247	0.153	0.108 ± 0.101	0.077	0.001	-0.009	0.006 ± 0.031	0.001	-0.009	0.006 ± 0.031
CD010775	0.867	0.726	0.170	0.701	0.140 ± 0.163	0.867	0.813	0.174	0.109 ± 0.098	0.813	0.718	0.109 ± 0.098
CD010783	0.856	0.906	0.819	0.842	0.124 ± 0.106	0.701	0.381	0.340	0.015 ± 0.043	0.381	0.072	0.015 ± 0.043
CD010860	0.578	0.695	0.716	0.046	0.134 ± 0.155	0.237	0.067	-0.050	0.065 ± 0.106	0.067	-0.039	0.065 ± 0.106
CD010896	0.517	0.098	0.151	0.684	0.192 ± 0.196	0.518	0.098	0.151	0.086 ± 0.121	0.098	0.051	0.086 ± 0.121
CD011145	0.497	0.342	0.228	0.175	0.011 ± 0.034	0.446	0.327	0.108	0.004 ± 0.015	0.327	0.103	0.004 ± 0.015
CD012019	0.914	0.909	0.912	0.896	0.455 ± 0.286	0.797	0.369	0.534	0.195 ± 0.190	0.369	0.276	0.195 ± 0.190
Average	0.640	0.500	0.390	0.457	0.093 ± 0.023	0.497	0.348	0.241	0.015 ± 0.016	0.348	0.271	0.015 ± 0.016

Table 7.6 – Work saved over sampling at 95% recall for all topics in the CLEF dataset.

GRADUAL trains an initial intertopic classifier using the training set and ranks the test set in descending order of confidence. The system then iteratively asks for feedback for the top ranked result. Articles queried for relevance feedback are then added to the model as they are queried, but with higher weights than the intertopic examples. The model thus starts out as an intertopic classifier, but gradually turns into an intratopic classifier as more targeted data is added to the model. Since the intratopic examples identified through relevance feedback are given higher weights, these will eventually drown out the original classifier, provided enough examples exist to be discovered.

Besides using  $\gamma$  and  $N$ , we also use intratopic  $M$  as positive examples, with lower weights than intratopic  $\gamma$ , but higher than intertopic  $\gamma$ . The reasoning behind this is that we often encounter  $M$  earlier than  $\gamma$ , and in greater numbers, in particular on topics with very few  $\gamma$ . We have observed on other datasets that we can sometimes improve performance by using both  $\gamma$  and  $M$  as positive examples, when the number of  $\gamma$  is very low.

after the number of  $\gamma$  found is larger than 40, we stop using  $M$  as positive examples. Reasonable parameter settings were identified experimentally on the Cohen dataset.

#### 7.3.6 Use of the CLEF Development Dataset

We do not split the training data into separate training and validation splits, since we do not have the necessary number of  $\gamma$  to do this without hurting the performance of the classifier. We do however use a small set of samples that overlaps with the training set for validation. The performance we observe on this validation suffers from severe overfitting, but we can observe when the model fails to build a classifier on the current undersample. In such cases we can observe an AUROC  $< 0.5$  even on the training set. In these cases we simply discard the classifier and try again with a new undersample. We observe that this improves performance dramatically when we have a very small amount of training data (approximately four or less positive examples).

## 7.4 RESULTS

We present a comparison with previous work on the Cohen dataset for wss@95 in table 7.2 and for AUC in table 7.3. Results from previous literature are taken from Khabsa et al. (Khabsa et al., 2016), and Cohen et al. (Cohen, 2008). Exact intertopic AUC scores are not explicitly reported by Cohen et al. and have instead been extracted from Figure 1 in their paper. The majority of these results, with the exception of one result by Cohen et al. (Cohen, 2008) use intratopic classification. We present our results on the CLEF dataset for average precision in table 7.4, nor-

malized average precision in table 7-5, wss@95 in table 7-6, and in aggregate in table 7-7. The results in these tables correspond to those submitted as official runs. For comparison, we also calculate a baseline by evaluating each metric on the data ordered randomly. This has been repeated 1000 times and we report the average and standard deviation.

We also report the mean, standard deviation, minimum and maximum wss@95 and AUC over ten runs for a selection of topics in the CLEF dataset in table 7-8.

## 7-5 DISCUSSION

114

### 7-5-1 *Datasets*

One of the topics in the CLEF dataset, CD010653, has no  $\gamma$ . While we can still calculate performance scores relative to  $M$ , this topic might arguably have been omitted from the test data. One of the topics, CD008803, similarly has no  $M$ . This also happens to be the topic with the largest number of  $\gamma$ .

As a general tendency, we can observe that the relative number of  $\gamma / M / N$  in the CLEF dataset varies dramatically across topics. At the one end we have one topic consisting of 14.06%  $\gamma$  (CD008760), and one topic consisting of 15.79%  $\gamma$  (CD010705). At the other end we have three topics with a mere 0.01%  $\gamma$  (CD011548, CD011549, and CD012019). Most topics in the CLEF dataset have a very small number of  $\gamma$  compared to Cohen, both in terms of relative and absolute numbers. Several topics have a large number of  $M$  however (CD007427, CD008054, CD009020, CD009323, CD009591, 011134, CD011548, CD0011975, CD011984, CD009925, CD10339, CD011145). Curiously, more topics in the training set have a large number of  $M$  than in the test set, despite this comprising a smaller number of topics.

The number of  $N$  also varies wildly, from 52 up to 43287. Compared to the Cohen dataset we also have a smaller minimum number of  $N$ , as well as much larger maximum number.

If we compare the training and test sets, the training set contains almost double the absolute number of  $M$ , many more  $N$ , but fewer  $\gamma$ .

### 7-5-2 *Performance*

While relevance feedback sometimes gives an improvement in performance, relevance feedback often seems to only confuse the system (tables 7-4–7-7). This should be contrasted with our experiments on the Cohen dataset, where the same implementation reliably yields an improvement (table 7-3), and generally yields performance intermediate between intertopic and intratopic classification, as one would expect. There are perhaps better approaches to relevance feedback than

ours, which can reliably improve upon the baseline, but it might also be that there is simply little to gain from relevance feedback on several of the topics. Of particular note, we should not expect any improvements by using RF on topics such as CDO10386, CDO10633, CDO10860, CDO10896, and CDO12019, that have a low absolute number of  $\gamma$  and  $M$ . It is also worth pointing out that our ABRUPT scheme requires at least 4  $\gamma$  before switching to the intratopic model, and any differences between NO · AF and ABRUPT on these topics can thus only be due to chance.

We can see an improvement on the topic CDO10705 when using relevance feedback (tables 7·4–7·7). This topic is also the topic with the highest percentage of  $\gamma$  at 15.79%. We do not see any improvement for CDO08760, the other topic with a high percentage of  $\gamma$  (14.06%), but this may be due to the initial classifier having much higher performance.

We can observe that GRADUAL outperforms ABRUPT on topic CDO08760, despite this topic having only 3  $M$ , which is probably too low a number for GRADUAL to have an advantage. The simplest explanation for this is likely random chance.

It is however easy to see that relevance feedback does not appear to lead to an improvement for our system. For instance ABRUPT outperforms NO · AF 15 times out of 30, and GRADUAL outperforms NO · AF only 10 times out of 30 (tables 7·4). Of course, it seems unlikely for relevance feedback to be useful for those topics where the number of positives is extremely low, even in theory. In particular, if there is only one relevant article, as is the case for CDO12019 and CDO10386, then relevance feedback cannot really add any value to the classification. Any successful use of relevance feedback on such topics would necessarily have to use the negative examples.

We get better performance for NO · AF · FULL than NO · AF. We have however generally observed that this difference is generally reversed for intratopic classification, which is what we should end up with when we after relevance feedback, but it is possible that we would get better performance if we were to use NO · AF · FULL as a base for our relevance feedback experiments, since we would start with a much better initial classifier.

Ordinarily, screeners would be free to choose the order in which they screen each article, and may proceed for instance in alphabetical or chronological order. For the purposes of our baseline, we assume that any such order ordinarily available to screeners would be indistinguishable from random order on average.

Topic	Y			YM			N		
	W/o RF			W/ RF			W/ RF		
	NO · AF			ABRUPT			GRADUAL		
	FULL	std	min	FULL	std	min	FULL	std	min
WSS@95	0.640	0.500	0.390	0.457	0.093 ± 0.023	0.497	0.348	0.241	0.045 ± 0.016
WSS@100	0.591	0.420	0.350	0.407	0.112 ± 0.022	0.412	0.261	0.173	0.056 ± 0.015
last_rel	1678	2263	2619	2384	3393.7 ± 118.1	2250	2993	3414	3749.7 ± 68.8
NCG@10	0.517	0.407	0.357	0.346	0.081 ± 0.010	0.475	0.367	0.316	0.092 ± 0.006
NCG@20	0.802	0.639	0.644	0.685	0.180 ± 0.015	0.717	0.554	0.518	0.192 ± 0.008
NCG@30	0.908	0.783	0.753	0.789	0.280 ± 0.018	0.825	0.674	0.609	0.291 ± 0.010
NCG@40	0.946	0.843	0.814	0.832	0.379 ± 0.020	0.887	0.746	0.678	0.391 ± 0.011
NCG@50	0.972	0.890	0.842	0.881	0.479 ± 0.020	0.929	0.800	0.727	0.491 ± 0.011
NCG@60	0.984	0.921	0.886	0.911	0.579 ± 0.020	0.955	0.851	0.789	0.591 ± 0.011
NCG@70	0.990	0.942	0.911	0.937	0.679 ± 0.018	0.976	0.903	0.834	0.691 ± 0.011
NCG@80	0.997	0.960	0.939	0.959	0.778 ± 0.016	0.987	0.930	0.878	0.791 ± 0.007
NCG@90	0.998	0.987	0.965	0.980	0.878 ± 0.013	0.996	0.964	0.920	0.890 ± 0.007
NCG@100	1.000	0.998	1.000	1.000	0.977 ± 0.006	1.000	0.999	0.998	0.990 ± 0.002
norm_area	0.890	0.825	0.795	0.766	0.504 ± 0.022	0.839	0.780	0.735	0.509 ± 0.014
AP	0.133	0.100	0.111	0.109	0.027 ± 0.003	0.179	0.145	0.143	0.047 ± 0.003

Table 7.7 – Aggregate performance for each ranking metric.

Topic	WSS@95										AUC					
	NO · AF · FULL					NO · AF					NO · AF · FULL					
	mean	std	min	max	mean	std	min	max	mean	std	min	max	mean	std	min	max
	CD008760	.723	.034	.653	.762	.666	.080	.481	.764	.949	.012	.933	.971	.937	.020	.899
CD010386	.899	.012	.883	.921	.932	.006	.923	.942	.942	.949	.012	.933	.971	.982	.006	.992
CD010705	.085	.036	.025	.143	.143	.047	.027	.011	.099	.696	.013	.683	.705	.595	.028	.572
CD012019	.920	.006	.907	.929	.929	.923	.007	.903	.927	.962	.010	.949	.982	.973	.007	.979
CD010339	.250	.085	.084	.415	.415	.438	.139	.265	.607	.884	.013	.864	.903	.039	.798	.923

Table 7.8 – The average, standard deviation, minimum, and maximum WSS@95 and AUC over ten iterations on a subset of the topic in CLEF for our systems NO · AF and NO · AF · FULL.

7.5.3 *Metrics*

Average Precision has been selected as the main metric for this task as it was previously found particularly adapted to evaluate retrieval performance for highly imbalanced datasets (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015). However, these studies rely on common assumptions that we value high precision at the top of the ranking, whereas for systematic review screening we value recall almost exclusively. Of particular note, average precision heavily penalizes rankings where the top few results are non-relevant, even if the ranking manages to place all relevant articles in the upper percentiles of the ranking.

Furthermore, average precision is strongly correlated with the number of positives in the topic, with most of the cases where we achieve  $AP > 0.2$  are for topics with high prevalence. While this is to be expected, it means that average precision makes it difficult to compare performance across topics, since we can see a strong correlation with the prevalence of relevant articles in the topic (tables 7.1, 7.4–7.7). Similarly, Mean Average Precision will likely be dominated by the results on the topics with many relevant articles and a small number of total candidates, i.e. arguably the topics which are the least representative systematic reviews of DTA studies, and where automated methods are likely the least useful.

7.5.4 *Reliability of the Experiments*

Our classification method is stochastic, and thus does not produce deterministic results that are always the same every time we run on the same input data. To gauge the reliability of the experiment we repeat it ten times for a subset of the topics and calculate the standard deviations, as well as examine the minimum and maximum values (table 7.8).

We can generally observe a fairly large variability for topics with a small total number of candidates, such as CDO08760 and CDO10705, and for topics with a comparably smaller proportion of  $\gamma$ , such as CDO10339. When we consider topics with a large number of candidates we can observe a large variability for the CDO12019, but small variability for CDO10386. We might speculate that small topic size and a small relative number of  $\gamma$  is correlated with larger variability, but it is clear that the variability for some topics is quite large, regardless of the underlying causes and mechanisms. The standard deviation can be as large as .139, which is large enough that it casts doubts about the reliability of the results. Furthermore, the minimum and maximum values are much more skewed towards extreme values than we should expect from the standard deviations were the values normally distributed, suggesting that the distribution is heavy-tailed and skewed towards outliers.

Considering the above, we might suspect that the differences in performance in tables 7.4–7.7 are not significant. For instance ABRUPT outperforms GRADUAL 17 times out of 30, but we do not know whether this means that ABRUPT is a better method, or if this is simply due to random chance. We might speculate that our GRADUAL implementation works better for the cases where we have a sufficient number of  $M$ , but the experiment is ultimately too low-powered to draw conclusions. Future iterations of the campaign could consider whether performance should be computed as an average over multiple runs, in order to get more precise results for stochastic systems such as ours.

We can however see smaller variability in the mean performance across all topics, which might suggest that these are more reliable estimates. However, these give little indication as to how the performance depends on topic composition.

118

#### 7.5.5 *General Remarks on the Shared Task Model*

The Shared Task Model is typically implemented in evaluation campaigns that seek to perform a community-wide technical evaluation of systems addressing a particular task. A Shared Task thus offers an evaluation paradigm that includes: 1) a specific definition of the task and evaluation metrics 2) an implementation through the dissemination of datasets and evaluation tools and 3) the execution of the evaluation in a controlled setting where participants have access to data at the same time and are evaluated blindly by an independent third party. As outlined below, this year the TAR task was not conducted according to the Shared Task Model.

In this iteration of the evaluation campaign, the final set of evaluation metrics was decided only shortly before participants were required to freeze their systems. One of the expected outcomes of evaluation campaigns such as this is indeed the discussion of the relative merits of the various metrics to be used. However, changing the target metric close to the submission deadline means that some participants may have optimized for different metrics than those ultimately used for evaluation.

The gold standard labeled test data was distributed directly to the participants at the beginning of the test phase. This is explained by the lack of an assessor through which participants could receive relevance feedback as has been the case in e.g. TREC Total Recall. While common labeled test collections are routinely used for research, this procedure is unusual in a shared task setting where participants are typically asked to process a test dataset while being blind to the gold standard associated with the dataset. This could alternatively have been accomplished in part by requiring the submission of runs without relevance feedback before the distribution of the gold standard labels.

Another feature of the shared task model is the computation of performance metrics for all participants by a common, independent party which ensures that all par-

ticipations are evaluated using the exact same conditions. This confers a stronger reliability in the comparability and reproducibility of results. At the time of writing, while a common evaluation tool has been released, the performance reported by participants has been self-computed without validation from the task organizers. In addition to result validation, it would also have been useful to receive an indication of the overall performance of the participants prior to the deadline for the submission of the working notes. This would have enabled a discussion about the relative performance of the system that is currently difficult to do without comparing with previous literature using external datasets.

## 7.6 CONCLUSIONS

Our best system is the one using logistic regression trained using stochastic gradient descent, using a minimum of preprocessing, and no relevance feedback. This system achieves a workload reduction of 64.0% on average, with a minimum workload reduction of 19.3%, and a maximum workload reduction of 92.0%. On average, we would have to screen 1,678 articles per topic to retrieve all relevant articles. Overall there is a large variation in performance across topics however. We do not generally see an improvement when using relevance feedback. For the topics where relevance feedback is hypothetically feasible we sometimes see an improvement, although the effect does not appear very reliable, and the low power of the experiment means that the results are unlikely to be significant.

## ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676207.

The material in chapter 8 has been published as:

(Norman et al., 2018b): Norman, C., Leeﬂang, M., and N  v  ol, A. (2018b). LIMSI@CLEF eHealth 2018 task 2: Technology assisted reviews by stacking active and static learning. *Working Notes of CLEF*, pages 10–14

This is a shared task participation. A shared task provides a common platform to compare performance, and is a common way to assess the relative performance of competing approaches in NLP or IR. However, the CLEF eHealth task diverged from accepted shared task practices by not blinding participants to the gold standard, by using self-reported results, and by not providing pre-specified evaluation metrics (section 7.5.5 and 8.6). This may have introduced bias, and the relative results between systems should be interpreted with this in mind.

The article has undergone peer review and has been published in its entirety in the conference proceedings.

With this article we have tried to address the following research questions:

- RQ 2 *How do different screening automation approaches compare with each other for DTA screening?*
- RQ 3 *Are the screening automation methods we develop competitive with the current state-of-the-art?*

#### AUTHOR’S CONTRIBUTIONS

CN wrote the first draft and conducted the experiments. All authors conceived and designed the study. All authors read and approved the final manuscript.

LIMSI@CLEF eHEALTH 2018 TASK 2: TECHNOLOGY ASSISTED  
REVIEWS BY STACKING ACTIVE AND STATIC LEARNING

Christopher R. Norman, Mariska M.G. Leeflang, &amp; Aurélie Névéol

121

**Abstract**

This paper describes the participation of the LIMSI-MIRROR team at CLEF eHealth 2018, task 2. The task addresses the automatic ranking of articles in order to assist with the screening process of Diagnostic Test Accuracy (DTA) Systematic Reviews. We ranked articles by stacking two models, one linear regressor trained on untargeted training data, and one model using active learning. The workload reduction to retrieve 95% of the relevant articles was estimated at 82.4%, and we observe a workload reduction less than 70% in only two topics. The results suggest that automatic assistance is promising for ranking the DTA literature. *Keywords:* Evidence Based Medicine, Information Storage and Retrieval, Review Literature as Topic, Supervised Machine Learning

## 8.1 INTRODUCTION

Systematic reviews seek to gather all available published evidence for a given topic and provide an informed analysis of the results. This work constitutes some of the strongest forms of scientific evidence. Systematic reviews are an integral part of evidence based medicine in particular, and serve a key role in informing and guiding public and institutional decision-making. Systematic reviews for Diagnostic Test Accuracy (DTA) studies have been shown particularly challenging compared to other types of reviews because of the difficulty in defining search strategies offering acceptable recall (Petersen et al., 2014). For this reason, there is a need to investigate automation strategies to assist DTA systematic review writers, particularly in the time-consuming screening process.

Methods for automating the screening process in systematic reviews have been actively researched over the years (O'Mara-Eves et al., 2015), with promising results obtained using a range of machine learning methods. However, previous work has not addressed DTA studies.

This paper describes the work underlying our participation in the CLEF 2018 eHealth Task 2 (Kanoulas et al., 2018; Suominen et al., 2018). This work is part of an ongoing effort to provide automated assistance in the screening process in systematic reviews addressing a variety of topics, including DTA studies.

The remainder of this paper is organized as follows; Section 2 presents the dataset used for system development. Section 3 provides an overview of our system and describes each component. Finally, section 4 reports our results and section 5 provides an analysis of our methods and participation in the task.

## 8.2 MATERIAL

In this work we have used the CLEF DATASET (Kanoulas et al., 2017a) as the gold standard for evaluation. The first iteration (2017) of the CLEF dataset (Kanoulas et al., 2017a) comprised 50 DTA systematic review topics (20 for training, 30 for testing) associated with the full list of articles retrieved by an expert query and assessed for inclusion based on title and abstract or full text. The second iteration (2018) uses the previous 50 topics for training, and supplies an additional 30 topics for testing.

For each of the datasets we know the inclusion decisions based on the abstracts, as well as the inclusion decisions based on the full text. We thus have two definitions of positive examples, depending on whether we use the abstract decisions or full text decisions as the gold standard.

We use a tripartite labeling to reflect this:

- ❖ No (N) is the set of articles that were excluded based on the abstract
- ❖ MAYBE (M) is the set of articles that were preliminarily included based on the abstract, but later excluded based on the full text
- ❖ YES (Y) is the set of articles that were included based on both the abstract and the full text, and later used in the meta-analysis

Table 8.1 shows a breakdown of the distribution of examples for each class in the CLEF dataset.

## 8.3 METHODS

To rank candidate articles we construct three machine learning models:

Split	Topic	Absolute number			Relative number		
		Y	M	N	Y	M	N
train split 1 (2017 train split)	CD008643	4	7	15065	0.0%	0.0%	99.9%
	CD009593	24	54	14844	0.2%	0.4%	99.5%
	CD011549	1	1	12699	0.0%	0.0%	100.0%
	CD010771	1	47	274	0.3%	14.6%	85.1%
	CD010438	3	36	3211	0.1%	1.1%	98.8%
	CD007427	17	106	1398	1.1%	7.0%	91.9%
	CD008686	5	2	3946	0.1%	0.1%	99.8%
	CD011548	5	108	12591	0.0%	0.9%	99.1%
	CD007394	47	48	2450	1.8%	1.9%	96.3%
	CD009323	9	113	3757	0.2%	2.9%	96.9%
	CD010632	14	18	1472	0.9%	1.2%	97.9%
	CD011975	60	559	7582	0.7%	6.8%	92.5%
	CD009944	64	53	1064	5.4%	4.5%	90.1%
	CD009591	41	103	7847	0.5%	1.3%	98.2%
	CD011134	49	166	1738	2.5%	8.5%	89.0%
	CD009020	12	150	1422	0.8%	9.5%	89.8%
	CD010409	41	35	43287	0.1%	0.1%	99.8%
	CD008691	20	53	1243	1.5%	4.0%	94.5%
	CD011984	28	426	7738	0.3%	5.2%	94.5%
	CD008054	41	233	2940	1.3%	7.2%	91.5%
train split 2 (2017 test split)	CD010783	11	19	10875	0.1%	0.2%	99.7%
	CD009135	19	58	714	2.4%	7.3%	90.3%
	CD009185	23	69	1523	1.4%	4.3%	94.3%
	CD010023	14	38	929	1.4%	3.9%	94.7%
	CD010653	0	45	7957	0.0%	0.6%	99.4%
	CD009647	17	39	2729	0.6%	1.4%	98.0%
	CD011145	48	154	10670	0.4%	1.4%	98.1%
	CD008760	9	3	52	14.1%	4.7%	81.2%
	CD010775	4	7	230	1.7%	2.9%	95.4%
	CD009925	55	405	6071	0.8%	6.2%	93.0%
	CD009372	10	15	2223	0.4%	0.7%	98.9%
	CD010896	3	3	163	1.8%	1.8%	96.4%
	CD010542	8	12	328	2.3%	3.4%	94.3%
	CD008803	99	0	5121	1.9%	0.0%	98.1%
	CD009519	46	58	5867	0.8%	1.0%	98.3%
	CD010386	1	1	623	0.2%	0.2%	99.7%
	CD008782	34	11	10462	0.3%	0.1%	99.6%
	CD009579	79	59	6317	1.2%	0.9%	97.9%
	CD010772	11	36	269	3.5%	11.4%	85.1%
	CD009551	16	30	1865	0.8%	1.6%	97.6%
	CD010173	10	13	5472	0.2%	0.2%	99.6%
	CD010339	9	105	12689	0.1%	0.8%	99.1%
	CD010633	3	1	1569	0.2%	0.1%	99.7%
	CD010705	18	5	91	15.8%	4.4%	79.8%
	CD012019	1	2	10314	0.0%	0.0%	100.0%
	CD007431	15	9	2050	0.7%	0.4%	98.8%
	CD010276	24	30	5441	0.4%	0.5%	99.0%
	CD009786	6	4	2055	0.3%	0.2%	99.5%
	CD008081	10	16	944	1.0%	1.6%	97.3%
	CD010860	4	3	87	4.3%	3.2%	92.6%
test split (2018 test split)	CD011602	1	7	6149	0.0%	0.1%	99.9%
	CD011515	1	126	7117	0.0%	1.7%	98.2%
	CD010864	3	41	2461	0.1%	1.6%	98.2%
	CD012083	5	6	311	1.6%	1.9%	96.6%
	CD010680	0	26	8379	0.0%	0.3%	99.7%
	CD011431	26	271	885	2.2%	22.9%	74.9%
	CD012216	1	10	206	0.5%	4.6%	94.9%
	CD012281	9	14	9853	0.1%	0.1%	99.8%
	CD011686	2	53	9388	0.0%	0.6%	99.4%
	CD009175	7	58	5579	0.1%	1.0%	98.8%
	CD010213	33	566	14599	0.2%	3.7%	96.1%
	CD010657	35	104	1720	1.9%	5.6%	92.5%
	CD012599	19	556	7473	0.2%	6.9%	92.9%
	CD011420	5	37	209	2.0%	14.7%	83.3%
	CD012009	4	33	499	0.7%	6.2%	93.1%
	CD009263	10	114	78679	0.0%	0.1%	99.8%
	CD011926	29	11	4010	0.7%	0.3%	99.0%
	CD008122	57	215	1639	3.0%	11.3%	85.8%
	CD008587	35	44	9073	0.4%	0.5%	99.1%
	CD011912	18	18	1370	1.3%	1.3%	97.4%
	CD009694	9	7	145	5.6%	4.3%	90.1%
	CD010296	38	15	4549	0.8%	0.3%	98.8%
	CD012165	47	261	9914	0.5%	2.6%	97.0%
	CD008759	42	18	872	4.5%	1.9%	93.6%
	CD012179	117	187	9528	1.2%	1.9%	96.9%
	CD010502	71	158	2756	2.4%	5.3%	92.3%
	CD008892	30	39	1430	2.0%	2.6%	95.4%
	CD012010	8	282	6540	0.1%	4.1%	95.8%
	CD011053	7	5	2223	0.3%	0.2%	99.5%
	CD011126	9	4	5987	0.1%	0.1%	99.8%

Table 8.1 – The distribution of class labels in the dataset.

### 8.3.1 Overview

**CNRS STATIC** Our **STATIC RANKER** uses logistic regression trained on a large number ( $> 500,000$ ) of features. This model is trained once on train split 1 (Table 8-1), and can then be used to rank candidate articles in any unseen DTA systematic review, without a provided search query or topic description. This model is intended to capture diagnostic test accuracy studies without considering whether the articles are topically relevant.

124 **CNRS RF (UNI-/BIGRAM)** We construct two **RELEVANCE FEEDBACK** (active learning) models uses logistic regression on a smaller number ( $\sim 2,000$ ) of features. These models are trained using relevance feedback on the target topic, starting with the topic description as an artificial seed document. The unigram model is a reimplementa-tion of the **CAL** model by Cormack and Grossman (Cormack and Grossman, 2015, 2017). We also experiment on a model which uses bigrams in addition to unigrams. These models are intended to capture topicality, and to incrementally improve per-formance through the screening process.

**CNRS COMBINED** Our **STACKED METAClassifier** uses a three-layer feedforward dense neural net-work to estimate the optimal ranking based on the output of the **STATIC** model and the **RF · BIGRAM** model.

We describe each system in detail in the remainder of this section.

### 8.3.2 Static Ranking Model

We here use a machine learning approach and train a classifier on the training split, largely identical to the implementation of our static model submitted in 2017 (Norman et al., 2017b). The decision function of the classifier can then be used to calculate probability scores for unseen candidate articles. This is a static model, intended to capture diagnostic test accuracy studies without considering whether the articles are topically relevant.

We use logistic regression trained using stochastic gradient descent (sklearn) on a sparse feature matrix consisting of a large number ( $> 500,000$ ) of features. We have tried using other classifiers, including **SVMs**, random forests, feed-forward neural networks, convolution networks and **LSTMs**, but logistic regression yields consistently better performance in our experiments with a fraction of the training time.

We handle class imbalance by class reweighting. We have implemented undersam-pling mechanisms, but these tend to decrease performance. We set the weight for the positive class to 80 for the initial intertopic classifier. We have determined

this to be a reasonable weight experimentally in previous experiments on another dataset (Norman et al., 2017b).

This model was trained on the 2017 training split.

### 8.3.3 *Active Learning*

We here use an active learning approach, where we at each timestep train a classifier (ranker) on the relevant articles screened so far. We start the process using the topic description as an artificial seed document. The model is intended to capture topical relevance, and to use the data collected through the screening process, which is generally more targeted than the data we have available in the training split.

The model largely follows the continuous active learning approach of Cormack and Grossman (Cormack and Grossman, 2015, 2017), except for using bigrams in addition to unigrams. We repeat the procedure for clarity.

At each timestep we rank the candidate articles and show the top  $B$  articles to the oracle, and the oracle labels these as  $\gamma$ ,  $m$ , or  $N$ . The number of articles  $B$  is initially set to 1 and is incremented by  $\lfloor B \rfloor$  at each timestep.

We use the following process to construct positive training data:

**if  $\gamma$  have been encountered:**

Then we use all encountered  $\gamma$  as positive training data. The synthetic seed document and any encountered  $m$  are discarded.

**else if  $m$  have been encountered, but no  $\gamma$ :**

Then we use all encountered  $m$  as positive training data. The synthetic seed document is discarded.

**else (no  $\gamma$  or  $m$  have been encountered):**

We use the synthetic seed document as positive training data.

To construct negative training data we sample 100 articles (or as many as remains) from the unseen candidates and temporarily label these  $N$ , irrespective of their true labels. Any articles already shown to the oracle are not considered for use as negative data.

We train our model using the above positive and negative data to re-rank the candidate articles and repeat the process until all articles have been shown to the oracle. This model only uses the candidate articles and the topic description as training data, and thus do not depend on other training data, such as the topics in the training split.

Topic	Y  MN					YM  N				
	STATIC	RF		COMB.	baseline	STATIC	RF		COMB.	baseline
		UNI.	BI.				UNI.	BI.		
ALL	.169	.176	.124	.203	.014 ± 0	.313	.314	.218	.337	.053 ± 0
CD008122	.331	.274	.327	.344	.042 ± .013	.744	.706	.652	.748	.146 ± .001
CD008587	.045	.033	.043	.094	.004 ± 0	.076	.063	.062	.109	.009 ± .001
CD008759	.477	.543	.283	.549	.047 ± .001	.562	.620	.326	.609	.101 ± .010
CD008892	.278	.342	.329	.511	.022 ± .001	.323	.376	.361	.462	.043 ± .002
CD009175	.085	.095	.003	.059	.002 ± .001	.206	.156	.025	.130	.013 ± .002
CD009263	.060	.022	.000	.103	.000 ± .000	.116	.104	.003	.038	.002 ± .001
CD009694	.435	.447	.494	.843	.084 ± .014	.734	.774	.411	.694	.102 ± .018
CD010213	.040	.061	.018	.053	.002 ± .000	.260	.250	.195	.226	.042 ± .003
CD010296	.450	.535	.074	.541	.011 ± .002	.512	.563	.082	.568	.017 ± .005
CD010502	.209	.254	.186	.334	.028 ± .003	.339	.409	.323	.467	.080 ± .007
CD010657	.176	.206	.070	.196	.028 ± .001	.386	.406	.213	.421	.079 ± .003
CD010864	.079	.054	.013	.020	.002 ± .001	.084	.082	.113	.133	.023 ± .000
CD011053	.065	.063	.019	.048	.007 ± .005	.105	.105	.035	.080	.011 ± .005
CD011126	.111	.107	.018	.042	.003 ± .001	.145	.141	.027	.070	.003 ± .001
CD011420	.062	.056	.263	.215	.021 ± .000	.341	.336	.644	.742	.178 ± .000
CD011431	.216	.166	.167	.231	.026 ± .004	.649	.626	.662	.669	.262 ± .018
CD011515	.050	.028	.071	.042	.001 ± .001	.298	.369	.302	.360	.017 ± .001
CD011602	.002	.002	.002	.003	.001 ± .000	.018	.014	.021	.037	.004 ± .002
CD011686	.015	.012	.005	.047	.002 ± .001	.289	.201	.111	.162	.005 ± .001
CD011912	.212	.195	.453	.266	.013 ± .001	.374	.365	.447	.481	.031 ± .007
CD011926	.428	.540	.028	.129	.008 ± .000	.479	.569	.037	.165	.013 ± .002
CD012009	.051	.149	.027	.041	.009 ± .002	.387	.317	.192	.455	.085 ± .010
CD012010	.090	.125	.102	.106	.002 ± .001	.253	.295	.272	.354	.050 ± .001
CD012083	.612	.436	.335	.602	.022 ± .003	.373	.313	.243	.378	.040 ± .004
CD012165	.072	.075	.013	.073	.005 ± .001	.347	.348	.046	.291	.031 ± .002
CD012179	.183	.193	.075	.201	.015 ± .002	.374	.343	.123	.356	.033 ± .002
CD012216	.016	.016	.013	.012	.014 ± .008	.268	.246	.222	.285	.089 ± .023
CD012281	.012	.024	.091	.155	.001 ± .000	.026	.027	.080	.210	.003 ± .001
CD012599	.054	.059	.080	.042	.002 ± .000	.266	.266	.260	.253	.074 ± .004

Table 8.2 – Average precision score for each topic, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The COMBINED model uses the STATIC and RF · BIGRAM models as subcomponents.

8.3.4 Stacked Model

We use a three-layer dense neural network as a function approximator to estimate the joint score for a candidate document given the scores from our static and active models. We use 16 nodes in each layer, apply 30% dropout after each layer and use softmax activation on the final layer to simulate two-class logistic regression. The model is trained by sampling training data uniformly from recorded active learning output. We have tried using uncertainty sampling, but this has yielded inferior results.

## 8. Ranking Performance for DTA Reviews (2018)

Topic	Y  MN					YM  N				
	STATIC	RF		COMB.	baseline	STATIC	RF		COMB.	baseline
		UNI.	BI.				UNI.	BI.		
ALL	.741	.815	.668	.824	.104 ± .024	.513	.617	.519	.657	.028 ± .009
CD008122	.800	.794	.772	.788	.018 ± .033	.403	.455	.415	.453	.005 ± .013
CD008587	.839	.838	.836	.896	.034 ± .047	.772	.746	.696	.759	.012 ± .026
CD008759	.746	.764	.612	.736	.019 ± .037	.685	.703	.612	.668	.015 ± .030
CD008892	.891	.884	.788	.883	.048 ± .052	.040	.534	.694	.486	.006 ± .027
CD009175	.936	.916	.546	.915	.073 ± .111	.027	.532	.285	.532	.011 ± .029
CD009263	.465	.920	.117	.861	.041 ± .084	.418	.408	.122	.557	.006 ± .020
CD009694	.826	.832	.678	.813	.045 ± .091	.521	.795	.320	.683	.061 ± .073
CD010213	.278	.834	.647	.825	.038 ± .049	.065	.590	.556	.341	.002 ± .009
CD010296	.928	.924	.723	.924	.028 ± .042	.906	.909	.588	.918	.022 ± .034
CD010502	.346	.617	.757	.646	.019 ± .030	.298	.587	.405	.609	.002 ± .014
CD010657	.739	.741	.345	.757	.034 ± .044	.473	.453	.404	.503	.006 ± .018
CD010864	.914	.885	.837	.854	.197 ± .193	.215	.506	.571	.619	.017 ± .036
CD011053	.909	.913	.537	.903	.076 ± .112	.913	.913	.766	.906	.105 ± .095
CD011126	.921	.929	.819	.910	.048 ± .091	.933	.935	.860	.917	.096 ± .094
CD011420	.719	.715	.831	.823	.114 ± .144	.572	.575	.585	.627	.015 ± .034
CD011431	.763	.733	.696	.703	.025 ± .048	.017	.162	.275	.173	.003 ± .011
CD011515	.947	.945	.948	.947	.459 ± .290	.398	.178	.679	.721	.005 ± .020
CD011602	.879	.864	.877	.890	.448 ± .283	.750	.786	.806	.870	.059 ± .098
CD011686	.937	.910	.844	.875	.284 ± .231	.584	.285	.457	.811	.022 ± .034
CD011912	.871	.874	.854	.883	.053 ± .067	.843	.850	.654	.841	.032 ± .044
CD011926	.933	.933	.483	.916	.017 ± .047	.928	.926	.483	.909	.024 ± .041
CD012009	.713	.734	.362	.476	.150 ± .158	.584	.592	.362	.476	.026 ± .042
CD012010	.020	.671	.579	.744	.064 ± .105	.004	.534	.261	.581	.001 ± .013
CD012083	.925	.900	.835	.897	.122 ± .144	.180	.512	.727	.605	.117 ± .102
CD012165	.818	.824	.308	.828	.013 ± .035	.779	.769	.234	.774	.002 ± .012
CD012179	.804	.790	.403	.819	.010 ± .022	.750	.723	.363	.769	.002 ± .012
CD012216	.669	.655	.597	.577	.444 ± .289	.669	.655	.583	.581	.112 ± .103
CD012281	.880	.886	.931	.923	.054 ± .095	.716	.745	.622	.762	.031 ± .054
CD012599	.080	.413	.807	.877	.953 ± .067	.154	.384	.422	.476	.001 ± .009

Table 8.3 – wss@95 score for all topics in the CLEF dataset, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The COMBINED model uses the STATIC and RF · BIGRAM models as subcomponents.

As input to the model we use the score values we get from the static and active learning models, along with meta-level features. The full set of features is as follows:

1. Static model document score (STATIC)
2. Active model document score (RF · BIGRAM)
3. Number of Y found
4. Amount of relevance feedback (absolute number)

Topic	Y  MN					YM  N				
	STATIC	RF		COMB.	baseline	STATIC	RF		COMB.	baseline
		UNI.	BI.				UNI.	BI.		
ALL	.640	.762	.633	.779	.130 ± .024	.349	.460	.339	.510	.027 ± .007
CD008122	.459	.496	.378	.481	.016 ± .015	.289	.320	.040	.332	.003 ± .003
CD008587	.782	.848	.769	.845	.029 ± .028	.419	.475	.393	.412	.012 ± .012
CD008759	.031	.276	.325	.368	.021 ± .020	.031	.276	.325	.368	.016 ± .016
CD008892	.828	.887	.576	.875	.031 ± .031	.072	.358	.576	.390	.014 ± .014
CD009175	.986	.966	.596	.965	.123 ± .111	.010	.381	.264	.269	.015 ± .015
CD009263	.515	.970	.167	.911	.091 ± .084	.018	.061	.047	.218	.008 ± .008
CD009694	.876	.882	.728	.863	.095 ± .091	.565	.720	.228	.708	.051 ± .051
CD010213	.019	.520	.582	.727	.029 ± .029	.001	.043	.274	.061	.001 ± .002
CD010296	.918	.914	.638	.917	.026 ± .026	.918	.914	.418	.917	.019 ± .018
CD010502	.335	.629	.626	.684	.014 ± .014	.324	.581	.163	.585	.004 ± .004
CD010657	.550	.526	.331	.553	.028 ± .028	.057	.058	.103	.047	.007 ± .007
CD010864	.964	.935	.887	.904	.247 ± .193	.254	.423	.383	.351	.021 ± .022
CD011053	.959	.963	.587	.953	.126 ± .112	.959	.957	.587	.953	.078 ± .070
CD011126	.971	.979	.869	.960	.098 ± .091	.971	.979	.869	.960	.073 ± .070
CD011420	.769	.765	.881	.873	.164 ± .144	.343	.530	.575	.534	.020 ± .020
CD011431	.707	.665	.724	.695	.036 ± .034	.019	.029	.033	.064	.003 ± .003
CD011515	.997	.995	.998	.997	.509 ± .290	.171	.012	.386	.575	.007 ± .008
CD011602	.929	.914	.927	.940	.498 ± .283	.800	.836	.856	.920	.109 ± .098
CD011686	.987	.960	.894	.925	.334 ± .231	.069	.051	.198	.798	.018 ± .017
CD011912	.886	.902	.704	.897	.051 ± .048	.877	.866	.460	.877	.027 ± .026
CD011926	.302	.871	.383	.867	.033 ± .034	.302	.871	.383	.867	.025 ± .024
CD012009	.763	.784	.412	.526	.200 ± .158	.437	.457	.270	.285	.024 ± .024
CD012010	.070	.721	.629	.794	.114 ± .105	.027	.180	.067	.226	.003 ± .003
CD012083	.975	.950	.885	.947	.172 ± .144	.168	.540	.294	.618	.084 ± .075
CD012165	.072	.362	.179	.442	.020 ± .019	.039	.347	.087	.367	.003 ± .003
CD012179	.141	.482	.205	.464	.008 ± .008	.141	.367	.200	.401	.003 ± .003
CD012216	.719	.705	.647	.627	.494 ± .289	.576	.599	.303	.627	.078 ± .075
CD012281	.930	.936	.981	.973	.104 ± .095	.724	.726	.453	.730	.040 ± .038
CD012599	.129	.301	.857	.619	.051 ± .048	.092	.089	.109	.067	.001 ± .002

Table 8.4 – wss@100 score for all topics in the CLEF dataset, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The COMBINED model uses the STATIC and RF · BIGRAM models as subcomponents.

5 Amount of relevance feedback (percentage)

6 Relevance feedback stage (whether using seed, m or y as positive training data)

Features 3 and 4 are normalized using the following log transform

$$\text{sgn}(x) \times \frac{\log_2(1 + |x|)}{8}$$

to keep numbers in mainly in the range [0, 1]. We do not truncate large numbers. Feature 6 take discrete values in {−1, 0, 1}

However, we observe that features 5 and 6 decrease model performance and we therefore excluded these in the model used in our officially submitted runs.

This model is trained on data generated from training split 2 (Table 8-1) to avoid overfitting. We generate the training data for the stacked model by letting the active model run on the training data, and at each step in the process we record the score generated by the active learning model, as well as the above features. We do this 100 times for each topic. One data point thus consists of the score from the static model (feature 1), and features 2–6 from this pre-generated data.

We train the stacked model on data sampled randomly from this pool of data points, by sampling 50 runs in each iteration, and sampling an equal number of positive and negative training examples from each run (with a minimum of 20 total). The model is trained on a batch of size 32. The training data is resampled every training iteration.

## 8.4 RESULTS

We present our results for average precision in table 8-2, wss@95 in table 8-3, wss@100 in table 8-4, Last Rel in table 8-6, as well as the aggregate scores in table 8-5. For comparison, we also calculate a baseline by evaluating each metric on the data ordered randomly. The baseline values are calculated using the average and the standard deviation of 1000 repetitions. The RF · UNIGRAM, and RF · BIGRAM, and the COMBINED model were submitted as our official runs. The results omit one topic with no  $\gamma$  (CDO10680).

## 8.5 DISCUSSION

### 8.5.1 Datasets

One of the topics in the CLEF dataset, CDO10653, has no  $\gamma$ . While we can still calculate performance scores relative to  $M$ , this topic might arguably have been omitted from the test data. One of the topics, CDO08803, similarly has no  $M$ . This also happens to be the topic with the second largest number of  $\gamma$ .

As a general tendency, we can observe that the relative number of  $\gamma / M / N$  in the CLEF dataset varies dramatically across topics. At the one end we have one topic consisting of 14.06%  $\gamma$  (CDO08760), and one topic consisting of 15.79%  $\gamma$  (CDO10705). At the other end we have five topics with less than 0.1%  $\gamma$  (CDO11548, CDO11549, CDO12019, CDO11515, and CDO09263). The number of  $N$  also varies wildly, from 52 up to 78,679.

Topic	Y    MN			YM    N			
	STATIC	RF		STATIC	RF		COMB.
		UNI.	BI.		UNI.	BI.	
ALL	3349.448	1305.034	3798.000	6405.696 ± 27.238	5173.467	550.600	4378.900
CD008122	1034	964	1189	188.775 ± 29.665	1358	1835	1276
CD008587	1998	1390	2113	889.107 ± 252.363	4803	5559	5378
CD008759	903	675	630	912.361 ± 18.900	675	630	589
CD008892	258	170	636	1452.336 ± 46.459	962	636	914
CD009175	80	190	2282	4947.315 ± 626.643	3492	4156	4125
CD009263	38214	2340	65642	71659.995 ± 665.289	73961	75061	61604
CD009694	20	19	44	145.670 ± 14.589	45	125	47
CD010213	14915	7297	6348	14753.984 ± 445.766	14543	11039	14269
CD010296	379	394	1665	4481.412 ± 121.595	394	2677	382
CD010502	1986	1108	1116	2942.072 ± 42.574	1252	2500	1238
CD010657	836	882	1244	1806.271 ± 52.839	1753	1668	1772
CD010864	90	164	283	1886.456 ± 482.878	1445	1546	1625
CD011053	92	83	923	1952.562 ± 25.296	92	923	106
CD011126	174	128	784	5414.703 ± 543.169	128	784	238
CD0111420	58	59	30	209.813 ± 36.150	118	107	117
CD0111431	346	396	326	1139.302 ± 4.689	1148	1144	1106
CD0111515	20	36	14	3553.976 ± 2097.615	6003	7160	3079
CD0111602	435	529	448	3088.216 ± 174.224	1229	4452	495
CD0111686	123	382	997	6291.365 ± 2182.568	8787	886	1903
CD0111912	160	138	417	1334.026 ± 67.988	188	7573	173
CD0111926	2827	524	2501	3915.805 ± 135.943	524	2501	537
CD012009	127	116	316	428.898 ± 84.684	291	392	383
CD012010	6352	1907	2537	6049.525 ± 719.498	5601	6374	5284
CD012083	8	16	37	266.458 ± 46.415	148	228	123
CD012165	9488	6521	8394	10013.510 ± 193.570	6673	9337	6468
CD012179	8446	5097	7813	975.778 ± 8.874	6225	7863	5893
CD012216	61	64	77	109.840 ± 62.640	87	152	81
CD012281	695	631	183	8851.610 ± 939.890	2706	5400	2669
CD012599	7009	5626	1153	7636.046 ± 387.458	7328	7171	7512
							8035.456 ± 13.165

Table 8.6 – Last rel score for all topics in the CLEF dataset, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The COMBINED model uses the STATIC and RF · BIGRAM models as subcomponents.

Metric	Y  MN				
	STATIC	RF		COMB.	baseline
		UNI.	BI.		
AP	.169	.176	.124	.203	.014 ± .000
WSS@95	.741	.815	.668	.824	.104 ± .024
WSS@100	.640	.762	.633	.779	.130 ± .024
Last Rel	3349.448	1305.034	3798.000	1224.655	6405.696 ± 272.238
Metric	YM  N				
	STATIC	RF		COMB.	baseline
		UNI.	BI.		
AP	.313	.314	.218	.337	.053 ± .000
WSS@95	.513	.617	.519	.657	.028 ± .009
WSS@100	.349	.460	.339	.510	.027 ± .007
Last Rel	5708.400	5173.467	550.600	4378.900	7131.769 ± 36.629

Table 8.5 – Aggregate scores, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The COMBINED model uses the STATIC and RF · BIGRAM models as subcomponents.

### 8.5.2 Performance

No single model performs best on all topics. Generally however, RF · UNIGRAM consistently outperforms the static model, and the COMBINED model (STATIC + RF · BIGRAM) outperforms the other three models.

Surprisingly, the RF · UNIGRAM model consistently outperforms the RF · BIGRAM model, despite using a subset of the features of the RF · BIGRAM model. For this reason it seems likely that a stacked model consisting of the STATIC model and the RF · UNIGRAM model would have achieved better performance than the stacked model submitted as our official run.

The RF · UNIGRAM model is particularly adept at finding all relevant articles, resulting in better last rel score than the STATIC model for 19 topics out of 29, and a better last rel score than the RF · BIGRAM model for 24 out of 29. This also results in a WSS@100 score of 76.2% for the RF · UNIGRAM, versus 64.0% for the STATIC model, and 63.3% for RF · BIGRAM.

Note however that last rel generates scores of wildly varying scale, and the large last rel scores for STATIC and RF · BIGRAM are therefore almost entirely due to a few large outliers. In particular, 59% of the information contained in the last rel score for RF · BIGRAM is due to a single topic with a large number of candidate articles (CDO09263). The metric may thus be useful when interpreted on individual topics, but not when averaged. The WSS@100 metric, which is equivalent to last

rel on individual topics, produces scores on the same scale and therefore makes sense also when averaged.

## 8.6 CONCLUSIONS

Our best system combines a static model and a relevance feedback model using stacking. The workload reduction to retrieve 95% of relevant articles is estimated at 82.4% on average, with a minimum workload reduction of 47.6%, and a maximum workload reduction of 94.7%. The workload reduction is consistent across topics, and we note a workload reduction less than 70% in only two topics. Due to the highly variable number of candidate articles in different topics, however, we may still need to screen several thousands of articles to find all relevant articles in any given systematic review.

Our remarks on the implementation of the shared task model and task organization from last year (Norman et al., 2017b) remain valid for this edition of the TAR task.

## ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676207.



AS WE HAVE SEEN in this part, screening automation methods need to cope with several technical constraints, many of which are well attested for general systematic reviews (O'Mara-Eves et al., 2015). For instance, systematic review screening is subject to extreme class imbalance – one included study for every thousand excluded is not uncommon in DTA systematic reviews (Norman et al., 2017b, 2018b,c). As a comparison, the Cohen dataset of drug class efficacy includes 4.52% of the candidate references by full-text and 8.30% studies provisionally by abstract, whereas the same numbers for the CLEF dataset were 0.42% and 1.34% respectively (Norman et al., 2018b). This confirms findings from previous methodological studies that the relative screening burden is higher in systematic reviews of diagnostic test accuracy compared to systematic reviews of interventions (Petersen et al., 2014). However, while class imbalance is an obstacle for training, there are a number of solutions to overcome this challenge, including under- and oversampling methods (He and Garcia, 2009).

Among the 50 systematic reviews used in the CLEF shared task, the median number of included studies was 14 (range: 0 to 99) (Norman et al., 2018b). Eleven of the reviews included four studies or less. The training data that we can expect from many systematic reviews are therefore far below the numbers necessary to reach data-saturation in machine learning models. It is unclear how to effectively train screening automation model on such limited data, particularly to reach the close to perfect recall commonly required by systematic reviews. This may be a problem particular to screening automation in diagnostic systematic reviews. The reviews in the Cohen dataset included a median of 41 studies (range: 9 to 146) and thus do not present the same problem (Norman et al., 2018b).

Coping with extremely small amounts of training data is therefore likely to be the main concerns for screening automation in systematic reviews of DTA.

#### 9.1.1 *Differences Between Studies Included by Abstract and Full-Text*

As we have seen in the first study in this section, references included in the systematic review ( $\gamma$ ) and references provisionally included in the systematic review but excluded based on full-text ( $\mathcal{M}$ ) are not sufficiently different in terms of language or word choice that general machine learning algorithms such as logistic regression can distinguish the two based only on title and abstract. This is unsurprising, since by definition the  $\mathcal{M}$  are precisely those references where human screeners were unable to assess inclusion by title and abstract, and it would be unreasonable to expect automated methods to fare better.

However, there are observable differences between  $\gamma$  and  $\mathcal{M}$ . It is generally easier to distinguish  $\gamma$  from  $\mathcal{N}$  than  $\mathcal{M}$  from  $\mathcal{N}$ . This may be due to  $\mathcal{M}$  including more

borderline cases than  $\gamma$ . In particular, human screeners tend to be overinclusive during the initial screening – preferring to reserve judgement until the full-text is available, and as a consequence, the  $M$  may include references that are little different from those that should be excluded (Bekhuis et al., 2014; Frunza et al., 2011; O’Mara-Eves et al., 2015). Overall, the performance may degrade when only  $M$  are used as positive examples for training, although the differences appear to be minor.

#### 9.1.2 *Using Training Data from Both Screening Stages*

134

When training an automation model, we preferably want high quality data. As we have seen in this part, the  $\gamma$  are usually more representative of included studies than the  $M$ , although the differences are often minor. Similarly, intratopic training data will – by definition – always be better examples than intertopic training data. Thus the optimal training scheme uses only intratopic  $\gamma$  as positive examples. However, we also want as much data as possible to use for training, and the best quality positive examples is seldom available in large quantities from previous systematic reviews. Out of the 50 systematic reviews distributed in CLEF, only 19 included at least 20 studies. Furthermore, five reviews only included a single study, and one review included no studies at all. It is not clear how it would be possible to train screening automation systems in these cases. In practice, the want for quality and the want for quantity will frequently be at loggerheads – if we increase the quality by being more restrictive we will necessarily reduce the quantity and vice versa. To use automation effectively in many systematic reviews, we thus need to develop strategies for complementing the best training examples – intratopic  $\gamma$  – with sub-optimal training examples – intertopic  $\gamma$  or intratopic  $M$ , or possibly intertopic  $M$  as a last resort. Other approaches may be possible, including e.g. pre-trained models or few-shot learning, but these too work similarly by leveraging untargeted training data. Combining and weighting these different kinds of variously untargeted training data is not straightforward. The stacked model introduced in chapter 8 for our participation in CLEF in 2018 was our attempt to use meta-regression to learn to balance inter- and intratopic training data, and to balance  $\gamma$  and  $M$  using active learning.

#### 9.1.3 *Screening Approaches*

During this project we have investigated three different models, for slightly different systematic review contexts.

THE STATIC MODEL is trained on the inclusion/exclusion decisions of references screened in previous systematic reviews. This model thus requires training data to be available at the time the screening is started. This typically limits the applicability of this model to systematic review updates, or to use intertopic training to train general models, e.g. to identify general DTA studies.

THE ACTIVE MODEL uses active learning to improve its performance throughout screening. This model does not require training data at the time the screening is started, and can therefore be used also in systematic reviews conducted de novo, where training data is not available. This process can be started from scratch, with no training data at all, but if some quantity of training data is available – some relevant studies are often known when a systematic review is started – or if training data can be constructed artificially, such data can be used as a starting point.

THE STACKED MODEL combines the static and active models to achieve the best of both. It uses a static (intertopic) model as a base and then uses the more targeted intratopic data collected through the screening process to improve the model further, using active learning.

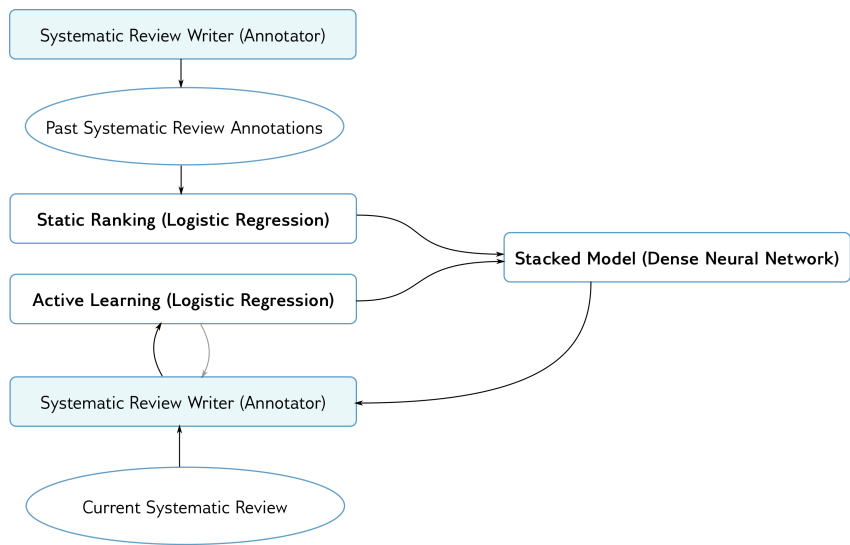


Figure 9.1 – Overview of the stacked model, integrating judgments from an intertopic model trained on previous systematic reviews, and an intratopic active learning model.

## 9.2 SCREENING PERFORMANCE

Evaluating screening performance relative to previous literature is difficult, since most previous work is evaluated using different measures, with different settings, and usually on different datasets. Comparisons are therefore only possible relative to a small subset of competing approaches.

In this project we have relied on two datasets for state-of-the-art comparisons: the Cohen dataset of drug class efficacy to assess the performance of the static model with inter- and intratopic training, and the CLEF dataset to assess the performance of the active learning models, as well as the static model with intertopic training. The CLEF shared task has been conducted three times, in 2017, 2018, and 2019. We participated in the first two iterations. In each iteration of the shared task, the best performing method has been the Waterloo Continuous Active Learning approach by Cormack and Grossman (2016), which has remained unchanged through the CLEF campaign. The same approach (using SVM rather than logistic regression for training) was previously used in the TREC Total Recall shared task, where it has also remained undefeated in all iterations of the shared task.

### 9.2.1 *Static Model (Intratopic)*

In one of the earlier papers on the subject, Cohen et al. (Cohen et al., 2006) constructed a dataset from 15 reviews on drug efficacy. This dataset was later extended to 18 (Cohen et al., 2010), and then to 24 reviews (Cohen et al., 2009). The smaller dataset comprising 15 reviews has been made available (Cohen et al., 2006)<sup>1</sup>. When the first study in this section was conducted, several approaches had been evaluated on the Cohen dataset, including Voting Perceptrons (Cohen et al., 2006), Complement Naive Bayes (Matwin and Sazonova, 2012), SVM (Cohen, 2006, 2008; Cohen et al., 2009), Random Indexing (Jonnalagadda and Petitti, 2014), and Random Forests (Khabsa et al., 2016), and this dataset therefore constitutes the closest thing we had to a standard dataset for comparison of performance for intratopic static approaches. At the time, there was no clearly superior state-of-the-art, and the best previous results was complement Naive Bayes (Matwin and Sazonova, 2012) on 8 topics, random forests (Khabsa et al., 2016) on 6 topics, and voting perceptrons (Cohen et al., 2006) on one topic. Our intratopic results were better than the then state-of-the-art in terms of wss@95 (0.392 on average), but worse in terms of AUC. This suggests that our model works well for finding all relevant studies, whereas the competing approaches are better at finding the first relevant studies, but struggles to find the last ones.

---

<sup>1</sup> The data can be found at  
<https://dmice.ohsu.edu/cohenaa/systematic-drug-class-review-data.html>

After we performed this work, two more studies have been published evaluated on this dataset, which have since pushed the state-of-the-art further (Ji et al., 2017; Olorisade et al., 2019). Ji et al. (2017) do not cite any previous studies other than Cohen’s original work on voting perceptrons (Cohen et al., 2006), and the work on random forests by Khabisa et al. (2016). Olorisade et al. (2019) also cite Matwin and Sazonova (2012) but do not compare against their results. The two studies thus give an obsolete and incomplete view of the state-of-the-art, and underestimates the best previous performance.

Ji et al. (2017) proposed to use ontology based features and evaluate three different models. Their first model uses SNOMED-CT and achieves a wss@95 of 0.392 on average, with better results than our model on 8/15 topics. Their second model uses MeSH terms and achieves a wss@95 of 0.347 on average, with better results than our model on 6/15 topics. Their third model uses both SNOMED-CT and MeSH terms and achieves a wss@95 of 0.409 on average, with better results than our model on 9/15 topics.

Olorisade et al. (2019) proposed to use MeSH terms and reference lists from articles to improve training. Their first model uses MeSH terms only and achieves a wss@95 of 0.408 on average, with better results than our model on 10/15 topics. Their second model uses MeSH terms and reference lists and achieves a wss@95 of 0.301 on average, with better results than our model on 5/15 topics.

Thus, new methods have improved the state-of-the-art further since we first evaluated our method. Still, the performance of our model is still almost the same as the new state-of-the-art, and consistently performs better on several topics. Furthermore, our model does not require UMLS-based ontological features, which are often comparatively slow to use and require large amounts of memory.

### 9.2.2 Static Model (Intertopic)

For intertopic static learning, our results on the Cohen dataset were better than the state-of-the-art across the board. Subsequent studies evaluating on the Cohen dataset only evaluated in terms of intratopic training, and thus do not allow any comparisons (Ji et al., 2017; Olorisade et al., 2019).

Despite the simplicity of the approach, the intertopic static approach frequently gives performance comparable to the current state-of-the-art active learning models in the CLEF evaluation. There may be two reasons. First, the static intertopic model used in CLEF was trained on approximately 250,000 titles and abstracts, whereas the Waterloo active learning approach starts training on 100, and never uses more than a few thousand training examples. Second, the candidate references in the CLEF dataset – like in a conventional systematic review – are based on a database query that is typically already restricted by condition. Search queries

are typically more accurate when limiting to specific diseases than when limiting to DTA studies, and it is likely more important that the screening automation is accurate where the search query is inaccurate, rather than where the query is accurate.

Metric	Static	Active Learning		Combined static + bigram	baseline
		unigram	bigram		
Last Rel	3349.448	1305.034	3798.000	<b>1224.655</b>	6405.696 ± 272.238
wss@100	0.640	0.762	0.633	<b>0.779</b>	0.130 ± 0.024
wss@95	0.741	0.815	0.668	<b>0.824</b>	0.104 ± 0.024
AP	0.169	0.176	0.124	<b>0.203</b>	0.014 ± 0.000

Table 9.1 – Results of each ranking model on the 2018 CLEF gold standard dataset.

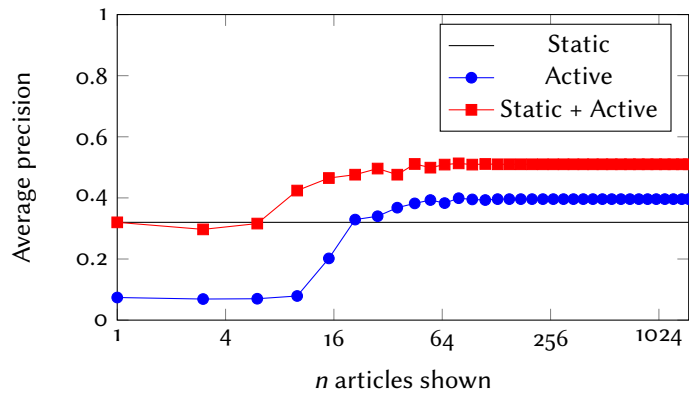


Figure 9.2 – Comparison between the learning modes of 1) intertopic static, 2) intratopic active, and 3) a stacked model combining intertopic static and intratopic active. The points on the curve correspond to the timesteps when the active learning is retrained using the Waterloo continuous active learning method (Cormack and Grossman, 2016; Norman et al., 2018b).

9.2.3 Active Learning

We submitted two variations of the active learning approach in the 2018 CLEF iteration, one based on unigrams, and one based on bigrams. Surprisingly, the unigram model consistently outperformed the bigram model (table 9.1). The active learning approach using unigrams consistently outperformed the intertopic static approach in both iterations of CLEF (table 9.1).

9.2.4 *Stacked Model*

While the active learning approach outperformed the static intertopic approach in the final results in CLEF, the static model tends to do better initially (figure 9.2). This is to be expected, since the active learning has to kickstart the process from nothing, and it is not going to do very well until starts finding relevant records to train the model. The active learning tend to overtake the static model after about four relevant records have been identified.

The stacked model is a hybrid approach that attempts to combine the benefit of the intertopic approach – the high initial performance – with the benefit of the intratopic active learning approach – improvement over time. Using the stacked model, both approaches are run concurrently, and the relevance score produced by both are combined in a neural network, along with data describing the current progress of screening (Norman et al., 2018b). The purpose of the neural network is to intelligently combine the output of the models, to produce a relevance score that is better than either. The stacking model should put more weight on the score from the static model initially – when the static model performs better – and then gradually switch over to put more weight on the active learning model – as the active learning model becomes more reliable.

The result of the stacking is illustrated in figure 9.2. The active learning approach again improves the performance of the model once a few relevant records have been identified. The stacked model starts with an initially better ranking, allowing it to get a headstart compared to the active learning approach, and consequently also a better final performance.

While the output of the static intertopic submodel may improve the performance by supplying additional information to the combined ranking algorithm, the main reason for the improved performance is likely that the initially better ranking given by the static intertopic model allows the active learning approach to gain a foothold and start improving earlier.

## 9.3 CONCLUSIONS

We have presented a screening automation system that can be used in a variety of systematic review contexts – ranging from review updates to reviews conducted de novo. The system is general in purpose, and performs well on reference screening datasets on clinical NLP, drug class efficacy (intervention studies), DTA studies, and core outcome set development. The system is furthermore highly customizable, and the underlying preprocessing pipeline and classification or ranking algorithms can be changed to finetune the system for specific systematic review topics or contexts.



# PART III

## THE IMPACT OF SCREENING AUTOMATION

This part of the thesis is based on the following publications:

([Norman et al., 2019f](#)): Norman, C. R., Gargon, E., Leeflang, M. M. G., Névéal, A., and Williamson, P. R. (2019f). Evaluation of an automatic article selection method for timelier updates of the COMET core outcome set database. *Database*. Oxford University Press

([Norman et al., 2019c](#)): Norman, C., Leeflang, M., Porcher, R., and Névéal, A. (2019c). Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. *BMC Systematic Reviews*. Springer Nature

The contents in this section is also based on material presented at the following conference:

([Norman et al., 2019d](#)): Norman, C., Leeflang, M., Porcher, R., and Névéal, A. (2019d). Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. In *AMIA Annual Symposium Proceedings*

The contents in this section is also based on material accepted and originally planned to be presented at the following conference:

([Norman et al., 2019b](#)): Norman, C., Leeflang, M., Porcher, R., and Névéal, A. (2019b). Does screening automation negatively impact meta-analyses in systematic reviews of diagnostic test accuracy? In *Cochrane Colloquium*

Unfortunately, this 2019 Cochrane Colloquium was cancelled due to the October 2019 civil unrest in Santiago, Chile. The presentation was instead presented virtually.

**F**OR THE PERFORMANCE EVALUATIONS in part II, we used screening prioritization to order candidate references during screening in order of likelihood of being relevant. In other words, we simulated screeners being presented with references more likely to be relevant before being presented with references less likely to be relevant. Using this approach, we only change the order in which we screen references, not which references will be screened. The screening process still examine each and every reference, but will accumulate relevant studies earlier and quicker than when screened in random or quasi-random order.

The advantage of this approach is that we do not run the risk of missing relevant studies. This makes this approach ‘safe’ to use in live systematic reviews (O’Mara-Eves et al., 2015). Presenting the most likely candidates for inclusion as early as possible could theoretically improve the screening process. It may for instance be possible to start the analysis process earlier if relevant studies are identified earlier (Cohen et al., 2009; Thomas, 2013). There are however no published evaluations demonstrating reductions in overall workload (Thomas, 2013). Crucially, it may be difficult to justify the added complexity and risk of bias in exchange for workload reductions that have yet to be measured or demonstrated. To truly take advantage of screening automation methods, it may therefore be necessary to forego screening prioritization in favor of screening reduction methods.

The main purpose of this part of the thesis is to establish criteria to automatically exclude records with screening reduction methods, while still resulting in the ‘same’ systematic review. In order to do this however, we also need to formalize what it means for two reviews to be the ‘same.’

Such sameness involves two aspects. The first aspect is procedural: an automated review process is compatible with the established process if it does not fundamentally alter it, and if steps are taken to reduce any sources of bias that may arise from the parts of the process that are altered. The second aspect is observational: an automated review process may be compatible with the conventional process even if it is different, provided we can show convincing evidence that the new process does not increase the risk of bias.

#### 10.1 HOW CAN WE MEASURE REVIEW INTEGRITY?

A number of performance metrics for screening prioritization have been used in previous literature, including recall (sensitivity), precision, specificity, accuracy, F measure (F1), area under the receiver operator curve (AUROC/AUC), work saved under sampling (WSS@R), average precision (AP), cumulative gain (CG), discounted cumulative gain (DCG), normalized discounted cumulative gain (NDCG), reciprocal rank, loss<sub>R</sub>, loss<sub>E</sub>, and reliability (Kanoulas et al., 2017a; O’Mara-Eves et al., 2015).

With the exception of sensitivity, specificity, and AUC, many of these metrics originate from machine learning or information retrieval. These are metrics expected in machine learning and information retrieval conferences and journals, and many are thus likely chosen because they are familiar to reviewers and readers in these venues.

Unfortunately, many of the metrics used to evaluate automation methods are seldom used in fields outside of computer science, and are unlikely to be familiar to systematic reviewers. Furthermore, they are almost invariably difficult or impossible to translate into systematic review integrity or impact. As an example, the AUC score can be interpreted as the probability that an arbitrary relevant item is ranked before an arbitrary non-relevant item in the list. There is no obvious way to set a threshold for this probability that distinguishes between the systematic review drawing correct or wrong conclusions.

These metrics may still be useful in comparing different algorithms with each other, or between different choices of training features (i.e. what data is fed into the model). This is how they are often used (e.g. see part II). When comparing algorithms or the impact of different technical choices, metrics do not need to be interpretable, they only need to be consistent in distinguishing better performing algorithms from worse performing ones.

Unfortunately, many metrics make assumptions about utility that are wrong or detrimental when used for systematic review screening. Systematic reviews almost always require disproportionately high recall (search sensitivity), whereas commonly used information retrieval metrics assume that both recall and precision are equally important.

In a systematic review, false positives and false negatives are not associated with the same cost (Wallace et al., 2010b). Including an extra study typically leads to 30–60 seconds higher workload for the screeners. In contrast, missing a relevant study runs the risk of invalidating the results of the review and potentially recommending diagnostic tests or treatments that are poor or harmful to patients.

Several commonly used metrics, including the  $F_1$  measure and AUC, give equal weight to false positives and false negatives, and are therefore poorly suited to evaluate screening automation methods. Alternatives have been proposed to counter this, including the  $F_3$  score which gives recall three times the weights of precision (Bekhuis and Demner-Fushman, 2010), and the  $U_{19}$  score which gives recall 19 times the weight of the workload (Wallace et al., 2010b,c).

Similarly, the majority of conventional information retrieval metrics (AP, CDG with variants, reciprocal rank, loss, reliability) are predominantly ‘top-heavy.’ That is to say, they try to measure the ability of the retrieval methods to fill the top of the list with relevant items, rather than its ability to find all relevant items. This typically works well for web searches, where users are unlikely to examine the

retrieved results beyond the first few pages, but may be disastrous in the context of screening automation, where perfect recall is expected.

Among the metrics commonly used, only the recall and the work saved under sampling (wss) can readily be interpreted in terms of the impact the screening automation has on the systematic review. Even so, recall may be a poor measure of systematic review integrity on its own. On the one hand, a systematic review could potentially miss only a single study, but this could still be unacceptable if the missed study is disproportionally large and would change the conclusions of the review. Conversely, it may be acceptable even if the screening missed several studies, if the missed studies are not important for the review. Conventional screening metrics make no such distinction. Furthermore, there may be diminishing returns for identifying a large number of studies – once a sufficient number of studies have been found, finding more studies may not yield any new information, and only limited improvements in confidence. We may see substantial further workload reductions by only retrieving as much evidence as is necessary to perform the systematic review.

We need metrics that allow us to know how screening reduction methods would impact the systematic review. Such metrics should at a minimum give some guarantee that the screening reduction will not influence the results or conclusions of the review.

#### 10.2 MEASURING INTEGRITY PROSPECTIVELY

Previous work on screening automation often report substantial workload reductions – ranging between 30–70% – but much of this is evaluated retrospectively, and more importantly, speculatively. For instance, the commonly used wss@95% metric means intuitively that there was a point during (the retrospectively simulated) screening procedure where at least 95% of all relevant references had been identified, and the screening could have been interrupted at this point for a substantial workload reduction, *if this had somehow been known during screening*. Since we cannot know the number of outstanding references in a prospective systematic review, there is no way to know when 95% of references have been identified.

This is not just the case for the wss metric, in fact all metrics mentioned above are retrospective – these are all evaluated based on gold standard inclusion and exclusion decisions from previous data.

#### 10.3 REDUCING BIAS

Conventionally, the screening order in systematic reviews is randomized, or quasi-randomized by ordering e.g. by author name. This ensures that the decisions to

include or exclude studies are not affected by the order in which references are screened, and serves to reduce potential bias.

The prospective application of the cut-off is particularly important, because this allows the stopping criterion to be decided and specified at the time the protocol is written, which in turn ensures that the decision to stop screening is not made ad-hoc, and do not depend on the results of the screening.

Randomization is fundamentally incompatible with screening prioritization, which work by ordering the studies by their likelihood of being relevant. Screening prioritization could thus potentially introduce so-called rank-order bias, where screeners are influenced by the ranking, and be more likely to include non-relevant studies at the top of the list, and more likely to exclude relevant studies at the bottom (Gargon et al., 2019). While the ranked approach thus produces shorter turnaround for the review, the process may no longer be compatible with the expected qualities of a systematic review (Thomas, 2013).

Screening prioritization is still rarely used in systematic reviews however, and there are consequently no trials estimating biases due to automation. To what extent rank order bias affect systematic reviews thus remain to be seen. Proponents of screening prioritization point to the beneficial effects (O'Mara-Eves et al., 2015). For instance, screeners may get 'up to speed on the current developments' more quickly (Cohen, 2008), screeners may gain a better understanding of the inclusion criteria earlier (O'Mara-Eves et al., 2015), or lower ranked studies may be screened by less experienced screeners (Cohen et al., 2009).

At the same time, changing inclusion criteria after seeing the evidence are a known cause of bias in systematic reviews (Lasserson et al., 2019, in Higgins et al., 2019) and there are concerns that screening prioritization could introduce bias (Gargon et al., 2019). Letting screeners with less expertise screen lower ranked studies may only serve to acerbate such biases.

For the time being, the only way to blind screeners to the ranking and avoid the potential for bias is by using predetermined thresholds and subsequently randomize the remaining references, foregoing screening prioritization altogether. Screening reduction, coupled with subsequent randomization is conceptually similar to a boolean search filter, which is already used for the database search, and is therefore largely compatible with the established systematic review process. All other requirements of the systematic review process can – with proper care – be fulfilled by a static model, including reproducibility. Machine learning methods are not normally transparent however, and in a systematic review where transparency is important additional work may be necessary to ensure transparency.

#### 10·4 DETERMINING SAFE SCREENING THRESHOLDS

To use screening reduction methods in a live systematic review, we must be able to determine thresholds for the automated methods that ensures that the systematic review is not adversely affected, and which ensures that the process still fulfil all requirements of transparency and reproducibility. In this part we will look at two studies in different review contexts, and with consequently different approaches to determining such a threshold.

146

#### 10·5 OBJECTIVES

In this part we present two journal papers, both published in 2019, where we attempt to address the following research questions:

- RQ 4 *Can we use screening automation in live systematic review while keeping the same rigorous methodology?*
- RQ 5 *What are the minimum conditions for a systematic review to guarantee the same results and conclusions using screening prioritization as with the conventional process?*



The material in chapter 11 has been published as:

(Norman et al., 2019f): Norman, C. R., Gargon, E., Leeflang, M. M. G., Névéol, A., and Williamson, P. R. (2019f). Evaluation of an automatic article selection method for timelier updates of the COMET core outcome set database. *Database*. Oxford University Press

This article documents the use of the screening automation for the COMET core outcome set systematic review in 2019. The study was conducted to confirm that the screening automation method used for the database update performs within acceptable parameters for the COMET review (Gargon et al., 2019).

The underlying research question in this study was:

RQ 4 *Can we use screening automation in live systematic review while keeping the same rigorous methodology?*

#### AUTHOR'S CONTRIBUTIONS

CN wrote the first draft and conducted the experiments. All authors conceived and designed the study. All authors read and approved the final manuscript.

## Evaluation of an automatic article selection method for timelier updates of the COMET Core Outcome Set database

Christopher R. Norman, Mariska M.G. Leeflang, Elizabeth Gargon,  
Aurélié Névéol, & Paula R. Williamson

149

DATABASE, 2019

### Abstract

Curated databases of scientific literature play an important role in helping researchers find relevant literature, but populating such databases is a labour intensive and time-consuming process. One such database is the freely accessible COMET Core Outcome Set database, which was originally populated using manual screening in an annually updated systematic review. In order to reduce the workload and facilitate more timely updates we are evaluating machine learning methods to reduce the number of references needed to screen. In this study we have evaluated a machine learning approach based on logistic regression to automatically rank the candidate articles. Data from the original systematic review and its four first review updates were used to train the model and evaluate performance. We estimated that using automatic screening would yield a workload reduction of at least 75% while keeping the number of missed references around 2%. The results suggest that machine learning methods can reduce the workload required to select Core Outcome Set articles, and the method is now being used for the next round of the COMET database update.

Database: <http://www.comet-initiative.org>

### 11.1 INTRODUCTION

A wealth of biomedical information is buried in the free text of scientific publications. Curated databases play a major role in helping researchers and clinicians access this data, by selecting articles and specific facts of interest in the subfield of biomedicine they address (Dowell et al., 2009; Krallinger et al., 2011).

One such database is maintained by the Core Outcome Measures in Effectiveness Trials (COMET) Initiative, which aims to improve the usefulness of outcomes in research and help tackle problems such as outcome reporting bias, inconsistency and

lack of importance or relevance of outcomes to patients. These problems are being addressed through the development and use of core outcome sets (COS). A COS is an agreed standardised set of outcomes that should be measured and reported, as a minimum, in all trials for a specific clinical area (Williamson et al., 2017). COMET facilitates the development and application of COS, by bringing relevant material together and thus making it more accessible. Since 2011, COMET has maintained a public repository of studies relevant to the development of COS (The COMET database, <http://www.comet-initiative.org/studies/search>). The database was originally populated through completion of a systematic review (Gargon et al., 2014), which is annually updated to include all published COS, currently up to and including December 2017 (Davis et al., 2018; Gargon et al., 2018; Gorst et al., 2016a,b). The database is an integral resource not only to the development of COS, but also to the uptake of COS in research and in the avoidance of unnecessary duplication and waste of scarce resources (Gargon et al., 2018). Relevant studies are added to the database as they are found, but the annual update to the systematic review is necessary to ensure completeness. We encounter challenges in undertaking this comprehensive approach, such as the variability in free text terms and index terms used for COS development, further confounded by the absence of a specific index term or Medical Subject Heading (MeSH) main heading for this study type (Gargon et al., 2015). A direct consequence of these challenges is the work involved in manually screening a large number of records on an annual basis. The latest update (Gargon et al., 2018) took seven months from running the searches in March 2018 to submission of the manuscript in early October 2018, and involved five reviewers. It is a labour intensive review and therefore costly to keep this up to date. With the need to update this annually, a balance needs to be struck between managing this workload and the likelihood that all eligible studies will be identified. The addition of a new index term or MeSH heading to identify COS is unlikely at this time, so it is imperative that we explore alternative routes in an attempt to streamline this process.

#### 11.1.1 *Screening Automation in Systematic Reviews*

Automation has great potential to make systematic reviews quicker and cheaper (Beller et al., 2018; Tsafnat et al., 2014). Recent advances in text mining, natural language processing and machine learning have demonstrated that tasks within the systematic review process can be automated or assisted by automation. Possible tasks include screening of titles and abstracts, sourcing full texts and data extraction. Automation to assist the screening process is of particular interest in these systematic review updates due to the high number of hits retrieved in the annual searches.
















Using automated methods to prioritise the order in which references are screened is considered safe for use in prospective systematic reviews, but using cut-off values to eliminate studies automatically is not recommended practice (O'Mara-Eves et al., 2015). A wide range of methods have been proposed for this kind of screening prioritisation, including Support Vector Machines, Naive Bayes, Voting Perceptrons, LAMBDA-Mart, Decision Trees, EvolutionalSVM, WAODE, kNN, Rocchia, hypernym relations, ontologies, Generalized Linear Models, Convolutional Neural Networks, Gradient Boosting Machines, Random Indexing, and Random Forests (Kanoulas et al., 2017a; Khabsa et al., 2016; O'Mara-Eves et al., 2015; Suominen et al., 2018). Several screening prioritisation systems are publicly available, including EPPI-Reviewer, Abstrackr, SWIFT-Review, Rayyan, Colandr, and RobotAnalyst (Howard et al., 2016; Khabsa et al., 2016; Przybyła et al., 2018; Thomas and Brunton, 2007; Wallace et al., 2012b).

Comparing the relative performance of different methods is difficult since most methods have been evaluated on different datasets, under different settings, and with different metrics. There have been attempts to compare previous methods by replicating reported methods on the same datasets, but the replication of published methods is often difficult or impossible due to insufficient reporting Olorisade et al. (2016). Performance varies depending on included study type (e.g. randomized control trial, diagnostic study), clinical setting, research question, number of candidate references, et c, and it is therefore seldom possible to extrapolate performance on new, untested systematic reviews from previous experiments.

Conventional screening automation is based on learning-to-rank, an information retrieval approach that uses machine learning or statistics to learn a ranking model from existing training data (Fuhr, 1992). In the original formulation, a model is trained to estimate the relevance of each candidate reference (pointwise learning), and the references can then be presented to the screeners in descending order of estimated relevance. This is a form of probability regression and has been implemented using a multitude of methods from machine learning and statistics (O'Mara-Eves et al., 2015). However, in a ranking scenario it may be better to minimise the number of inversions, the number of pairs such that a relevant reference occurs after a non-relevant one, rather than the estimated probability score. This is known as ordinal regression, and can be done using machine learning methods by training on pairs of references (pairwise training) or on an entire list of references (listwise training) (Burgess, 2010).

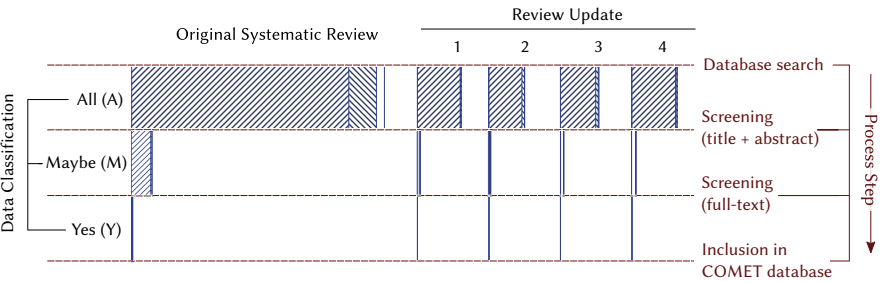
## 11.2 MATERIAL AND METHODS

To develop and evaluate our method, we used the results from the manual screening conducted in the systematic review, and its four annual updates (fig. 11.1) (Davis et al., 2018; Gargon et al., 2018, 2014; Gorst et al., 2016a,b).

	Original Systematic Review			Review Update											
															
All (A)	24,384	27,375	28,371	4,587	4,226	4,980	3,785	3,984	4,090	4,043	4,226	4,406	4,963	5,140	5,140
Maybe (M)	2,220	2,290	2,346	297	414	429	187	238	248	370	492	519	455	514	514
Yes (Y)	195	217	220	29	30	31	22	24	24	12	15	16	68	70	70

(a) The number of references considered in each stage of the screening process.

152



(b) Visual diagram of the flow of references during the manual screening process in the systematic review and the four review updates, The width of each bar corresponds to their respective numbers in (a).

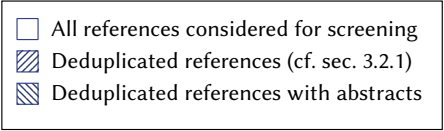


Figure 11.1 – Description of the data used in this study, resulting from the original systematic review (Gargon et al., 2014), and its review updates (Davis et al., 2018; Gargon et al., 2018; Gorst et al., 2016a,b). ‘All references considered for screening’ refers to the references retrieved from the database search in the systematic review, and includes references without abstracts, as well as duplicates across review updates. ‘Deduplicated references’ and ‘Deduplicated references with abstracts’ refer to the data after preprocessing (cf. sec. 11.2.1). We use the following shorthand for the different stages of the screening process: All (A): References initially identified through the database search. Maybe (M): References provisionally included based on title and abstract, but not yet screened based on full-text. Yes (Y): References judged relevant based on full-text and included in the COMET database.

11.2.1 *Data Preprocessing*

Before experimenting on the data we preprocessed it to ensure that it conforms to a few standard constraints necessary for the experiments to work as intended. In particular, training and evaluating on the same data points would overestimate the performance, and we therefore preprocess the data so that the training and evaluation sets do not overlap. Furthermore, removing duplicate data points means that each data point is counted only once in the evaluation of the results.

References may have two publication dates in their bibliographic records, once when they are published online (ahead of print), and once when they appear in the printed journal. When duplicate publication dates span review iterations, references may therefore occasionally be considered in two consecutive review iterations. In the manual screening for the COMET systematic review such duplicate references were screened in both updates they appeared in. Removing these would have required more work than simply screening them, and screening the same references twice will only provide an extra check and will not be detrimental to the review.

For this reason, 1,026 references in the original systematic review were re-screened in update 1, 103 references in update 1 were re-screened in update 2, 95 references in update 2 were re-screened in update 3, and 180 references in update 3 were re-screened in update 4. In total, 5 out of 354 included references were considered in at least two review updates (see set  $\gamma$ , fig. 11.1).

We opted to remove these duplicate references from the training set, rather than the test set, to mirror how these were handled in the systematic review. In practical terms, this means the model will always judge re-examined references without being biased by (or simply repeating) the judgment shown in the previous review update.

11.2.2 *Document Ranking Method*

To rank references, we used a static ranking method that we have described previously (Norman et al., 2018c) and which performed in the top tier of methods evaluated in the CLEF eHealth international challenge dedicated to Technologically Assisted Reviews in Empirical Medicine (Kanoulas et al., 2017a, 2018), and which compared particularly favourably to other models not relying on active learning (similarly to the setup used in this study).

We evaluated the model on each review update by examining how early it would have ranked the included references (set  $\gamma$  deduplicated in fig. 11.1).

We performed two sets of experiments. First, we performed a simulated prospec-

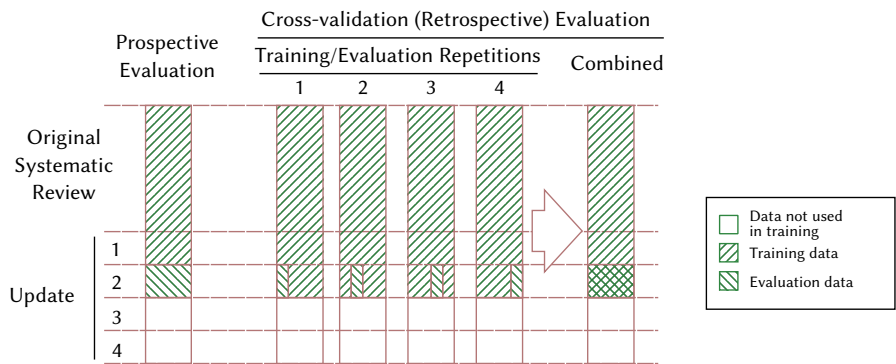


Figure 11.2 – Illustration of our prospective and retrospective (cross-validation) experimental setups when evaluating the performance of the model on update 2. For simplicity we illustrate using four folds instead of ten. This setup allows us to also use update 2 as training data when evaluating on update 2, while avoid training and evaluating on the the same individual references.

tive evaluation<sup>1</sup> on each of the four review updates. In each of these four experiments, we trained a model on the deduplicated data from the prior review iterations. Thus, we for instance trained the model on the data from the original systematic review and updates 1 and 2 when we evaluated the model on the update 3. Second, we evaluated on the original systematic review and on each of the four review updates by adding cross-validated data to the data from previous review data (see fig. 11-2). For instance, for update 2 we split the references into ten random sets. For each of these ten sets we trained a model on the data from other nine sets in addition to the original review and update 1, and let the model calculate scores for each reference in the set held out from training. We then constructed a single ranking by merging the ten sets and ranking this set by the score assigned to each reference. We performed these experiments because we suspected that we might get better performance when adding data from the same review update, either because of conceptual drift (Cohen et al., 2004) or simply because of the increase in the amount of training data. This setup also allowed us to evaluate the performance on the original systematic review update, which contains more data than the four review updates combined.

<sup>1</sup> The evaluation is prospective for the model, since it is not allowed to see the future data in the experiments. This study as a whole is still retrospective since the ‘future’ data already existed when we performed the experiments.

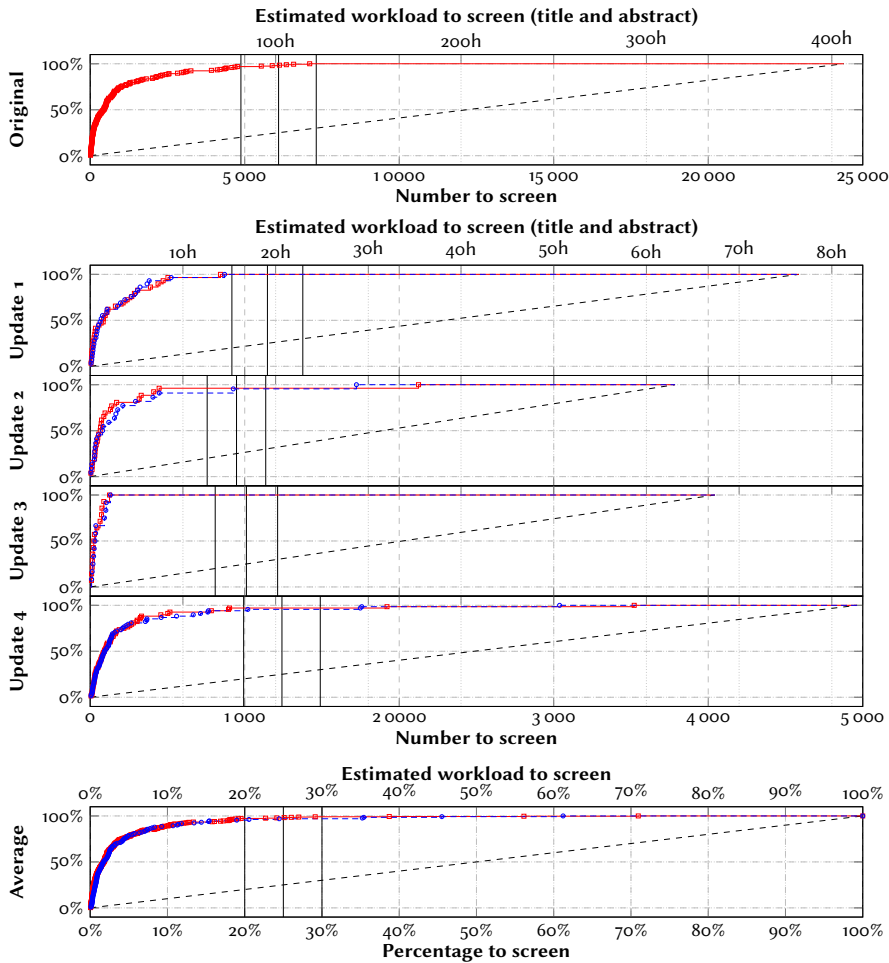


Figure 11.3 – EFFORT-RECALL CURVES EVALUATING THE SYSTEM PERFORMANCE ON THE REFERENCES WITH ABSTRACTS IN EACH REVIEW UPDATE. The total number of data points are given in the leftmost column for each update in fig. 11.1. The marks denote the positions in the ranking at which the included references would have been identified with screening prioritisation, evaluated prospectively (blue circles) or retrospectively using cross-validation (red squares). The y-axes denote the percentage of identified included references (recall) throughout the screening process. The dashed lines denote the mean expected curve when screening in random order (equivalent to standard practice). We mark three hypothetical cut-offs at 20%, 25%, and 30%. For scale, we give an estimate of the workload required by an experienced screener (1 abstract in 1 minute). Inexperienced screeners may take longer, and we estimate fulltext screening to take approximately ten times longer than screening titles and abstracts.

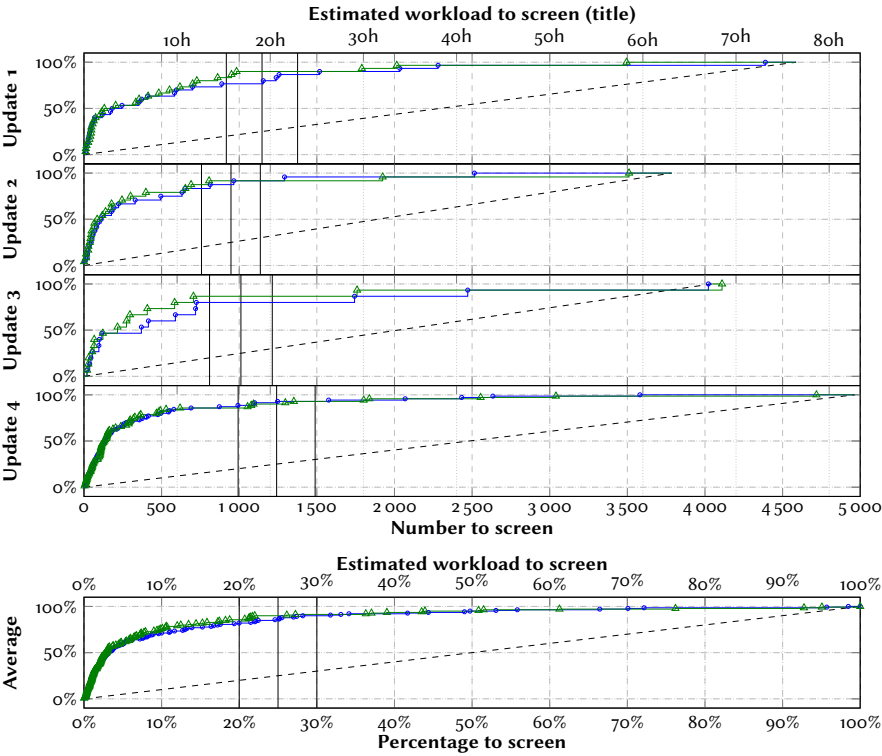


Figure 11.4 – EFFORT-RECALL CURVES EVALUATING THE SYSTEM PERFORMANCE ON TITLES ONLY. The marks denote the positions in the ranking at which the included references would have been identified with screening prioritisation, trained and evaluated only on titles (blue circles) or trained on titles and abstract and evaluated on titles (green triangles). The y-axes denote the percentage of identified included references (recall) throughout the screening process. The dashed lines denote the mean expected curve when screening in random order (equivalent to standard practice). We mark three hypothetical cut-offs at 20%, 25%, and 30%. For scale, we give an estimate of the workload required by an experienced screener (1 title in 1 minute). We estimate that the time required to screen titles is on average the same as screening abstracts.

Abstracts were not available for all references, and we therefore performed two sets of experiments to determine to what extent abstracts are necessary for judgment. First, we performed one set of experiments where we trained and evaluated a model using information from the titles and abstracts. In this setup we excluded references for which abstracts were missing. Second, we performed one set of experiments where we evaluated the model using information only from the titles. We used the same model as in previously, trained on titles and abstracts from all references in the training sets, as well as a model trained only on titles.

### 11.3 IMPLEMENTATION

We constructed a ranker by extracting bag-of- $n$ -grams ( $n \leq 5$ ) over words in the titles and abstracts. We used both tf-idf scores and binary features, in both stemmed and unstemmed form. In previous experiments, 4-grams and 5-grams have yielded consistent but very minor performance improvements, and could have been omitted without substantially decreasing performance. However, the stochastic gradient descent training does not take substantially longer to train on higher order  $n$ -grams, and we prefer that unhelpful features be discarded by the training, based on the data. We did not use feature selection, or dimension reduction.

We used the implementation of logistic regression in `sklearn` (Pedregosa et al., 2011) using version 0.20.2 trained using stochastic gradient descent, i.e. the `SGDClassifier` trained using log loss. We trained the ranker for 50 iterations.

We have also tested logistic regression optimised using `liblinear`, Long Short-Term Memories, Neural Networks, Passive Aggressive classifiers, Random Forests, as well as Support Vector Machines with linear, polynomial, and Radial Basis Function kernels. Logistic regression trained using Stochastic Gradient Descent is fast to train, does not require feature selection or dimension reduction, and performs as well as or better than all other methods we have tested. We have observed no performance gains by using pairwise training over pointwise training.

To compensate for the imbalance between the number of positive and negative references we increased the training weight for the positive examples to 80. Furthermore, we performed logistic regression with  $L_2$  regularization using  $\alpha = 10^{-4}$ . Each of these parameter settings was chosen as good default values in experiments on systematic reviews of drug class efficacy, and has proved to generalize well to systematic reviews of diagnostic test accuracy. Unlike in our previous work, we did not use under- or oversampling to compensate for the imbalance, because our previous experiments suggest this has limited benefit when used in addition to adjusting the training weights, and that the amount of under- or oversampling is often difficult to tune. We used default settings for all other parameters.

11.3.1 *Evaluation*

We evaluate in terms of observed trade-off between effort and recall (sensitivity). We define effort as the absolute number of articles screened manually by the human screeners:

$$\text{effort} = \text{TP} + \text{FP}$$

where TP denotes the true positives, and FP denotes the false positives.

We define recall as the proportion of positives (relevant articles) that are correctly identified:

$$\text{recall} = \text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP denotes the true positives, and FN denotes the false negatives.

The effort and recall are positively correlated, and vary as the cut-off value is varied. Similarly to e.g. ROC curves, we will plot pairs of effort/recall value pairs over all possible cut-offs to simplify the selection of an appropriate trade-off between effort and recall.

## 11.4 RESULTS

We report the results of our experiments as effort-recall curves in fig. 11.3 and fig. 11.4.

In the simulated prospective evaluation, we would have found the last included reference at position 870/4,587 in update 1 (19.0%), position 1,723/3,785 in update 2 (45.5%), position 131/4,043 in update 3 (3.4%), and position 3,038/4,963 in update 4 (61.2%) (fig. 11.3). Accepting some losses in update 4, we could have identified 67/68 references (98.5%) at position 1,758 (35.4%), 66/68 references (97%) at position 1,748 (35.2%), or 65/68 references (95.6%) at position 1,020 (20.6%). The last two references in update 2 appear to be outliers, and we would have identified 21/22 references (95.5%) at position 926 (18.7%), or 20/22 references (90.1%) at position 447 (9.0%).

If we had used this system and had stopped after screening 25% of the candidate references, we would have identified 126 out of the 129 deduplicated references with abstracts in the four review updates (97.7%) (fig. 11.3).

In the simulated retrospective evaluation (using cross-validation), we would have found the last included reference at position 7,102/24,384 in the original review (29.1%), position 843/4,587 in update 1 (18.4%), position 2,125/3,785 in update 2 (56.1%), position 125/4,043 in update 3 (3.1%), and position 3,521/4,963 in update 4 (70.9%) (fig. 11.3). Accepting some losses in update 4, we could have identified 67/68 references

(98.5%) at position 1,921 (38.7%), 66/68 references (97%) at position 902 (18.2%). Similarly to the prospective evaluation results, the last reference in update 2 appears to be an outlier, and we would have identified 21/22 references (95.5%) at position 446 (9.0%).

Overall, there was only a small difference between the prospective and the retrospective results, and the retrospective results were consistently better only in update 2 (fig. 11.3). Stopping after screening 25% of the candidate references in the retrospective evaluation would have identified 317 out of the 324 deduplicated references with abstracts in the four review updates (97.7%) (fig. 11.3).

The model performed substantially worse when evaluating on only titles (fig. 11.4). A model trained using all prior references, but trained and evaluated only on titles would on average have identified 86% of the relevant references after screening 25% of the candidates (fig. 11.4, bottom). A model trained on titles and abstracts from all prior references, but evaluated only on titles would on average have identified 90% of the relevant references after screening 25% of the candidates (fig. 11.4, bottom). Using a more conservative threshold would not have helped – several of the relevant references were identified only at the end of the simulated screening. However, only 3,840 out of 45,602 references in the dataset lacked abstracts (8.4%). These references constitute less than 300 references in each review, corresponding to a workload of less than 5 hours of screening per reviewer.

## 11.5 DISCUSSION

We used a logistic regression model for automatic article ranking to assess the suitability of automated screening for future updates to an annual systematic review of COS. We estimate that this model of automatic ranking can decrease the number of references that needs to be screened by 75% while identifying approximately 98% of all relevant references on average.

The results of this study are encouraging, and suggest that automated screening can be used to reduce the workload and therefore time and cost associated with this annual update. While we anticipate a reduction of workload by 75% and 62.5 hours per screener in the abstract screening stage, a balance needs to be struck with the prospect of identifying all eligible studies. With the last included reference identified at position 3,038 in the previous update, it is realistic to accept that all studies might not be identified using this ranking method if a reduction in time and workload is desired. However, 97.8% of articles (317/324) could still be identified retrospectively and 97.7% of articles (126/129) could be identified prospectively if a different position was selected for the cut-off point for screening. Other methods of identifying relevant studies are employed in the update of the systematic review of COS, such as hand-searching, reference checking, relevant database alerts for

key words and references, as well as checking with known experts. These other methods of identifying relevant papers increase the likelihood that all eligible studies will continue to be identified, and mean that a balance can be struck between managing the workload and identifying all eligible studies.

The results of this study showed that the screening automation can be reliable, provided both titles and abstracts are available, but that the automated ranking cannot reliably identify included references based only on titles. However, the number of references without abstracts is relatively low and we estimate that screening these manually would only take 2–4 hours per screener. We therefore recommend these be screened manually also in future updates of the systematic review of COS.

160

## 11.6 CONCLUSIONS

Based on the results in this study we determined that stopping after screening the first 25% of the candidate studies would result in a loss of roughly 2% of the relevant studies, which we deemed an acceptable trade-off in this systematic review. However, the same stopping criterion would have resulted in a loss of over 10% of the relevant studies without abstracts. Balancing the risk of missing relevant references against the limited number of such references, we opted to screen all references without abstracts manually.

We are currently using this system based on logistic regression to identify Core Outcome Sets published in 2018 for the fifth update of the COMET database. The database searches were performed in March 2019 and the screening is currently ongoing. The prospective use of these methods will further validate these results and this model of automated screening. This study has demonstrated that automation has great potential to make the annual updates of this systematic review quicker and cheaper.

## ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676207.



The material in chapter 12 has been published as:

(Norman et al., 2019c): Norman, C., Leeflang, M., Porcher, R., and Névéol, A. (2019c). Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. *BMC Systematic Reviews*. Springer Nature

This study was performed to address long-standing unresolved questions regarding the minimum requirement for a systematic review to be complete. In particular, systematic reviews require the identification of *all* relevant studies, in contrast to the majority of primary research, which requires unbiased samples of sufficient power. We suspected that for sufficiently large meta-analyses a sufficiently large and unbiased sample would yield the same results.

Furthermore, there have long been a disconnect between the way screening automation methods have been evaluated, and the goals and purposes of the systematic review. Fundamentally, no previous study had stopped to consider what screening automation means for the systematic review. In particular, will the use of screening automation change the results and conclusions of the review?

Conversely, the answers to these question would also arm us with better methods to interrupt screening. After all, if we have determined that the meta-analyses will not change further, then we can be content to stop added more studies to them. We envisioned this as a Bayesian approach, which would stop once the value of additional studies would be less than the cost involved in identifying them.

The underlying research question in this study was:

RQ 5 *What are the minimum conditions for a systematic review to guarantee the same results and conclusions using screening prioritization as with the conventional process?*

We set out to do this by breaking the question into the following sub-questions:

- RQ 5 a) *Can we measure the impact of the screening on the meta-analyses prospectively?*  
b) *Does perfect (or 95%) recall make sense as a target?*

#### AUTHOR'S CONTRIBUTIONS

CN wrote the first draft and conducted the experiments. All authors conceived and designed the study. All authors read and approved the final manuscript.

## Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy

Christopher R. Norman, Mariska M.G. Leeflang, Raphaël Porcher,  
✉ Aurélie Névéol

163

BMC Systematic Reviews

### Abstract

**BACKGROUND** The large and increasing number of new studies published each year is making literature identification in systematic reviews ever more time-consuming and costly. Technological assistance has been suggested as an alternative to the conventional, manual study identification to mitigate the cost, but previous literature has mainly evaluated methods in terms of recall (search sensitivity) and workload reduction. There is a need to also evaluate whether screening prioritization methods leads to the same results and conclusions as exhaustive manual screening. In this study we examined the impact of one screening prioritization method based on active learning on sensitivity and specificity estimates in systematic reviews of diagnostic test accuracy.

**METHODS** We simulated the screening process in 48 Cochrane reviews of diagnostic test accuracy, and re-run 400 meta-analyses based on a least 3 studies. We compared screening prioritization (with technological assistance) and screening in randomized order (standard practice without technology assistance). We examined if the screening could have been stopped before identifying all relevant studies while still producing reliable summary estimates. For all meta-analyses, we also examined the relationship between the number of relevant studies and the reliability of the final estimates.

**RESULTS** The main meta-analysis in each systematic review could have been performed after screening an average of 30% of the candidate articles (range: 0.07% to 100%). No systematic review would have required screening more than 2,308 studies, whereas manual screening would have required screening up to 43,363 studies. Despite an average 70% recall, the estimation error would have been 1.3% on average, compared to an average 2% estimation error expected when replicating summary estimate calculations.

**CONCLUSION** Screening prioritization coupled with stopping criteria in diagnostic test accuracy reviews can reliably detect when the screening process

has identified a sufficient number of studies to perform the main meta-analysis with an accuracy within pre-specified tolerance limits. However, many of the systematic reviews did not identify a sufficient number of studies that the meta-analyses were accurate within a 2% limit even with exhaustive manual screening, i.e. using current practice.

## 12.1 BACKGROUND

The increasing reliance on evidence provided by systematic reviews, coupled with rapidly increasing publishing rates is leading to an increasing need to automate the more labor-intensive parts of the systematic review process (Elliott et al., 2014). Beyond simply reducing the cost involved in producing systematic reviews, automation technologies, used judiciously, could also help produce more timely systematic reviews. For systematic reviews of diagnostic test accuracy (DTA), no sensitive and specific methodological search filters are known, and their use is therefore discouraged (Beynon et al., 2013; De Vet et al., 2008; Leeflang et al., 2006). Consequently, the number of citations to screen in a systematic review of diagnostic test accuracy is often several times higher than for systematic reviews of interventions, and the need for automation may therefore be particularly urgent (Kanoulas et al., 2017a, 2018; Petersen et al., 2014).

Methods for automating the screening process have been developed since at least 2006 (Cohen et al., 2006; O'Mara-Eves et al., 2015), but have so far seen limited adoption by the systematic review community. While there are examples of past and ongoing systematic reviews using automation, many more use manual screening. Thomas noted in 2013 that in order for widespread adoption to occur screening automation must confer a *relative advantage* (time saved), but must also ensure *compatibility* with the old paradigm, i.e. ensuring that screening automation is equivalent to manual screening (Thomas, 2013). There has been a large number of studies measuring the amount of time saved by automated screening, which may suggest that automation methods are maturing in terms of relative advantage. We are however not aware of any studies focusing on the compatibility aspect: whether automated screening results in the 'same' systematic review, and much of the literature to date have implicitly assumed that recall values over 95% are both necessary and sufficient to ensure an unchanged systematic review (O'Mara-Eves et al., 2015). In this study we aim to revisit this hypothesis, which to our knowledge has never been tested.

Among possible automation approaches, only screening prioritization is currently considered safe for use in systematic reviews (O'Mara-Eves et al., 2015). In this approach, systematic review authors screen all candidate studies, but in descending order of likelihood of being relevant. It is often assumed that we can achieve some

amount of reduction in workload by using screening prioritization (O'Mara-Eves et al., 2015), but the extent to which this is true has not been evaluated (Thomas, 2013). Screening prioritization can be combined with a cut-off (stopping criterion) to reduce the workload, for example by stopping screening when the priority scores assigned to remainder of the retrieved studies falls below some threshold. Using cut-offs is generally discouraged since it is not possible to guarantee that no relevant studies remain after the cut-off point and would thus be falsely discarded (O'Mara-Eves et al., 2015). However, using cut-offs would likely reduce the workload down to a fraction compared to using screening prioritization alone, and may therefore be necessary to fully benefit from screening prioritization.

### 12.1.1 *Meta-analyses of Diagnostic Test Accuracy*

Systematic reviews of diagnostic test accuracy may yield estimates of diagnostic performance with higher accuracy and stronger generalizability than individual studies, and are also useful for establishing whether and how the results vary by subgroup (Leeflang et al., 2008). Systematic reviews of diagnostic test accuracy are critical for establishing what tests to recommend in guidelines, as well as for establishing how to interpret test results.

Unlike randomized control trials, which typically report results as a single measure of effect (e.g. as a relative risk ratio), diagnostic test accuracy necessarily involves a trade-off between sensitivity and specificity depending on the threshold for positivity for the test (Leeflang et al., 2008; Macaskill et al., 2010). Diagnostic test accuracy studies therefore usually report results as two or more statistics: e.g. sensitivity and specificity, negative and positive predictive value, or the Receiver Operating Characteristic (ROC) curve. The raw data underlying these statistics is called a  $2 \times 2$  table, consisting of the true positives, the false positives, the true negatives, and the false negatives for a diagnostic test evaluation.

Meta-analyses of diagnostic test accuracy pool the  $2 \times 2$  tables reported in multiple DTA studies together to form a summary estimate of the diagnostic test performance. The results of DTA studies are expected to be heterogeneous, and the meta-analysis thus needs to account for both inter- and intra-study variance (Macaskill et al., 2010). This is commonly accomplished using hierarchical random effects models, such as the bivariate method, or the hierarchical summary ROC model (Reitsma et al., 2005; Rutter and Gatsonis, 2001). Pooling sensitivity and specificity separately to calculate separate summary values is discouraged, as it may give an erroneous estimate, e.g. a sensitivity/specificity pair not lying on the ROC curve (Leeflang et al., 2008).

12.1.2 *Systematic Reviews Require Perfect Recall*

Systematic reviews are typically expected to identify *all* relevant literature. In the Cochrane Handbook for DTA Reviews (De Vet et al., 2008) we can read:

‘Identifying as many relevant studies as possible and documenting the search for studies with sufficient detail so that it can be reproduced is largely what distinguishes a systematic review from a traditional narrative review and should help to minimize bias and assist in achieving more reliable estimates of diagnostic accuracy’.

166

Thus, the requirement to retrieve all relevant literature may just be a means to achieve unbiased and reliable estimates in the face of e.g. publication bias, rather than an end in itself. In this context, ‘as many relevant studies as possible’ may be better understood as searching multiple sources, including gray literature, in order to mitigate biases in different databases (De Vet et al., 2008). Missing a single study in a systematic review could result in the systematic review drawing different conclusions, and recall can therefore, in general, only guarantee an unchanged systematic review if it is 100%. For some systematic reviews, finding all relevant literature may be the purpose of the review, i.e. when the review is conducted to populate literature databases (Gargon et al., 2014). On the other hand, for systematic reviews addressing diagnostic accuracy or treatment effects, the review may be better helped by identifying an unbiased sample of the literature, sufficiently large to answer the review question (Booth, 2010). In systematic reviews of interventions, such a sample is often substantially larger than can be identified with the systematic review process (Wetterslev et al., 2017), but we hypothesize that it can also be substantially smaller.

Of course, many systematic reviews aim not just to produce an accurate estimate of the mean and confidence intervals, but also estimate prevalence, as well as identify and produce estimates for subgroups. Thus, to ensure an unchanged systematic review we would really need to ensure that the unbiased sample is sufficient to properly answer all aspects of the research question of the review. For instance, an unchanged systematic review of diagnostic test accuracy could require unchanged estimates of summary values, confidence intervals, the identification of all subgroups, and unchanged estimates of prevalence. We will in this study restrict ourselves to measuring the accuracy of the meta-analyses in systematic reviews of diagnostic test accuracy, i.e. the means and confidence intervals of the sensitivity and specificity.

There are multiple potential sources of bias that can affect a systematic review, including publication bias, language bias, citation bias, multiple publication bias,

database bias, and inclusion bias (Egger and Smith, 1998; Kung et al., 2010; Shea et al., 2009). While some sources of bias, such as publication bias, mainly occur across databases, others, such as language bias or citation bias may be present within a single database.

However, bias (i.e. only finding studies of a certain kind) is often conflated with the exhaustiveness of the search (i.e. finding all studies). While an exhaustive search implies no bias, a non-exhaustive search may be just as unbiased, provided the sample of the existing literature it identifies is essentially random. If the goal of the systematic review is to estimate the summary diagnostic accuracy of a test, the recall (the sensitivity of the screening procedure) may therefore be less important than the number of studies or total number of participants identified, provided the search process does not systematically find e.g. English language literature over literature in other languages. However, previous evaluations of automation technologies usually measure only recall, or use metrics developed primarily for web searches (Kanoulas et al., 2017a, 2018) while side-stepping the (harder to measure) reproducibility, bias, and reliability of the parameter estimation process.

### 12.1.3 *The Impact of Rapid Reviews on Meta-Analysis Accuracy*

Screening prioritization aims to decrease the workload in systematic reviews, while incurring some (presumably acceptable) decrease in accuracy. Similarly to screening prioritization, rapid reviews also seek to reduce the workload in systematic reviews and produce timelier reviews by taking shortcuts during the review process, and is sometimes used as an alternative to a full systematic review when a review needs to be completed on a tight schedule (Tricco et al., 2015). Examples of rapid approaches include limiting the literature search by database, publication date, or language (Marshall et al., 2019).

Unlike screening prioritization, the impact of some rapid review approaches on meta-analyses have been evaluated (Egger et al., 2003; Halladay et al., 2015; Hartling et al., 2017; Marshall et al., 2019; Nussbaumer-Streit et al., 2018; Sampson et al., 2003). However, a 2015 review identified 50 unique rapid review approaches, and only a few of these have been rigorously evaluated or used consistently (Tricco et al., 2015). Limiting inclusion by publication date, excluding smaller trials, or only using the largest found trial have been reported to increase risk of changing meta-analysis results (Marshall et al., 2019). By contrast, removing non-English language literature, unpublished studies, or grey literature rarely change meta-analysis results (Egger et al., 2003; Hartling et al., 2017).

The percentage of included studies in systematic reviews that are indexed in PubMed has been estimated between 84–90%, and restricting the literature search to PubMed has been reported to be relatively safer than other rapid review ap-

proaches (Booth, 2016b; Halladay et al., 2015; Marshall et al., 2019). However, Nussbaumer-Streit et al. have reported 36% changed conclusions for randomly sampled reviews, and 11% changed conclusions for review with at least 10 included studies (Nussbaumer-Streit et al., 2018). The most common change was a decrease in confidence. Marshall et al. also evaluated a PubMed only search for meta-analyses of interventions, and demonstrated changes in result estimates of 5% or more in 19% of meta-analyses, but observed changes were equally likely to favor controls as interventions (Marshall et al., 2019). Thus, a PubMed only search appears to be associated with lower confidence, but not with consistent bias. Halladay et al. have reported significant differences between PubMed indexed studies and non-PubMed indexed studies in 1 out of 50 meta-analyses including at least 10 studies (Halladay et al., 2015). While pooled estimates from different database searches may not be biased to favor either interventions or controls, Sampson et al. have reported that studies indexed in Embase but not in PubMed exhibit consistently smaller effect sizes, but also reasoned that the prevalence of such studies is low enough that this source of bias is unlikely to be observable in meta-analyses (Sampson et al., 2003).

#### 12.1.4 *Related Methods for Screening Prioritization*

The earliest known screening prioritization methods were published in 2006, and a number of methods have been developed since then (Cohen et al., 2006). Similar work on screening the literature for database curation have been published since 2005 (Aphinyanaphongs et al., 2005; Dobrokhotov et al., 2005). A wide range of methods (generally from machine learning) have been used to prioritize references for screening, including Support Vector Machines, Naive Bayes, Voting Perceptrons, LAMBDA-Mart, Decision Trees, EvolutionalSVM, WAODE, kNN, Rocchia, hypernym relations, ontologies, Generalized Linear Models, Convolutional Neural Networks, Gradient Boosting Machines, Random Indexing, and Random Forests (Kanoulas et al., 2017a, 2018; Khabisa et al., 2016; O'Mara-Eves et al., 2015). Several screening prioritization systems are publicly available, including EPP1-Reviewer, Abstrackr, SWIFT-Review, Rayyan, Colandr, and RobotAnalyst (Howard et al., 2016; Khabisa et al., 2016; Przybyła et al., 2018; Thomas and Brunton, 2007; Wallace et al., 2012b).

The most straightforward screening prioritization approach trains a machine learning model on the included and excluded references from previous iterations of the systematic review, and then uses this model to reduce the workload in future review updates (O'Mara-Eves et al., 2015). For natural reasons, this approach can only be used in review updates, and not in new systematic reviews. By contrast, in the *active learning* approach the model is continuously retrained as more and more references are screened. In a new systematic review, active learning starts

with no training data, and the process is typically bootstrapped ("seeded") by sampling the references randomly, by using unsupervised models such as clustering or topic modelling, or by using information retrieval methods with the database query or review protocol as the query (Cormack and Grossman, 2017).

Comparing the relative performance of different methods is difficult since most are evaluated on different datasets, under different settings, and often report different measures. There have been attempts to compare previous methods by replicating reported methods on the same datasets, but the replication of published methods is often difficult or impossible due to insufficient reporting (Olorisade et al., 2016). Another way to compare the relative performance of methods is through the use of a *shared task*, a community challenge where participating systems are trained on the same training data, and evaluated blindly using pre-decided metrics (Chapman et al., 2011; Huang and Lu, 2015). The shared task model removes many of the problems of replication studies, and also safe-guards against cheating, mistakes, the cherry-picking of metrics or data, as well as publication bias. The only shared task for screening prioritization we are aware of is the CLEF Shared Task on Technology Assisted Reviews in Empirical Medicine, focusing on diagnostic test accuracy reviews (Kanoulas et al., 2017a, 2018).

The purpose of this study is not to compare the relative performance of different methods, and we will focus on a single method (Waterloo CAL) that ranked highest on most metrics in the CLEF shared task both 2017 and 2018 (Kanoulas et al., 2017a, 2018). As far as we can determine, Waterloo CAL represents the state-of-the-art for new systematic reviews of diagnostic test accuracy (i.e. performed de novo). The training done in Waterloo CAL is also similar to methods currently used prospectively in recent systematic reviews, and mainly differs in terms of preprocessing (Bannach-Brown et al., 2019; Lerner et al., 2019; Przybyła et al., 2018).

### 12.1.5 Objectives

Our objectives in this study are twofold:

- ✦ We aim to retrospectively and prospectively measure the impact of screening automation on meta-analysis accuracy. We will use one single method for analysis in this study, but the criteria should be usable with any screening automation method. We will pay special attention to prospective criteria, since these can also be calculated while the screening is ongoing, and we will examine cut-offs for the prospective criteria that could be used in a prospective setting to bound the loss in accuracy within prespecified tolerance limits.
- ✦ We aim to evaluate the (retrospective) 95% recall criterion, which has long been the target to strive for in screening automation, and will test whether this criterion is

necessary and sufficient to guarantee an unchanged systematic review. In the case the criterion is not necessary or sufficient, we aim to develop criteria that could be used instead.

## 12·2 METHODS

### 12·2·1 *Data Used in the Study*

**THE LIMSI-COCHRANE DATASET** This dataset consists of 1,939 meta-analyses from 63 systematic reviews of diagnostic test accuracy from the Cochrane Library (the full dataset is available online: DOI: 10.5281/zenodo.1303259) (Norman et al., 2018a). The dataset comprises all studies that were included in the systematic reviews, from any database or from gray literature, as well as the 2×2 tables (the number of true positives, false positives, true negatives, and false negatives) extracted from each included study by the systematic review authors, grouped by meta-analysis. This dataset can be used to replicate the meta-analyses in these systematic reviews, in full or over subsets of the data, for instance to evaluate heterogeneity or bias of subgroups.

**THE CLEF DATASET** This dataset consists of all references from PubMed considered for inclusion – both those included in the systematic review and those ultimately judged not relevant to the systematic review – in 80 systematic reviews of diagnostic test accuracy also from the Cochrane Library (Kanoulas et al., 2017a, 2018).<sup>1</sup> Due to the way the data was collected, this dataset only contains references from PubMed, but not from other databases or gray literature. The dataset only includes the PubMed identifiers for each reference, and whether the studies were included in the reviews.

**COMBINED DATASET** For our experiments we combined the two datasets by collecting the reviews, meta-analyses and references common to both. In total, this intersection comprises 48 systematic reviews and 1,354 meta-analyses of diagnostic tests. All analyses in this study will be based on this intersection unless otherwise specified.

Since the CLEF dataset only includes references from PubMed, the meta-analyses performed in this study will only be based on studies from PubMed. Some meta-analyses may therefore be smaller than they were in the original reviews. The exclusion of studies from other sources than PubMed has been demonstrated to have moderate impact, and no bias on meta-analyses of interventions, and we will

---

1 The full dataset is available online: <https://github.com/CLEF-TAR/tar>

make the explicit assumption that the same is true for systematic reviews of diagnostic test accuracy (we are not aware of studies measuring this directly) (Halladay et al., 2015; Marshall et al., 2019).

Cochrane guidelines for systematic reviews of diagnostic test accuracy discourage drawing conclusions from small meta-analyses, but do not offer a specific minimum number of studies required for a meta-analysis (Macaskill et al., 2010). In this study, we will only consider meta-analyses based on three or more studies, because the R package we use (mada) issues a warning when users attempt to calculate summary estimates based on fewer studies (Doebler and Holling, 2015). This minimum likely errs on the side of leniency.

We considered as meta-analysis any summary estimate reported individually in the ‘summary of findings’ section of the systematic reviews, regardless of how the estimates were calculated. Thus we considered meta-analyses of subgroups to constitute distinct meta-analyses, in addition to any meta-analyses of the entire groups of participants. We further considered meta-analyses distinct for the same diagnostic test evaluated with e.g. multiple cut-off values, whenever these are reported separately in the systematic reviews.

171

#### 12.2.2 Automated Screening Method

We used a previously developed active learning approach to rank all candidate references for each systematic review in descending order of likelihood of being relevant (Norman et al., 2018b). The method was selected since it was the best performing method for new systematic reviews (performed de novo) in the 2017–2018 Shared Task on Technology Assisted Reviews of Empirical Medicine (Kanoulas et al., 2017a, 2018).

We used this ranking to simulate the literature screening process in each systematic review, and for those meta-analyses where at least 3 diagnostic studies were included we simulated the meta-analysis continuously throughout the screening process. As a control, we performed the same simulation with references screened in randomized order. We assumed that screeners will only stop if prompted to do so by the system. If not prompted to stop, the screeners will continue screening until all candidate studies have been screened.

We use a variant of active learning that has demonstrated good performance in systematic reviews of diagnostic test accuracy as well as in article discovery in the legal domain (Cormack and Grossman, 2016). In this method we start with an artificial training set, where we use the protocol of the review as a single initial positive training example (seed document). This artificial seed document is discarded as soon as real positive examples are found. We select 100 references randomly from the evaluation set and use these as negative examples, regardless

of whether they are really positive or negative. In each iteration, new “negative” examples are randomly selected in this way such that the total number of negative examples is always at least 100. Following Cormack and Grossman, we show  $B$  references to the screener in each iteration, where  $B$  is initially set to 1, then increased by  $\lceil B/10 \rceil$  in each subsequent iteration (Cormack and Grossman, 2015). To train, we use logistic regression with stochastic gradient descent on bigrams and unigrams extracted from the text in titles and abstracts.

172

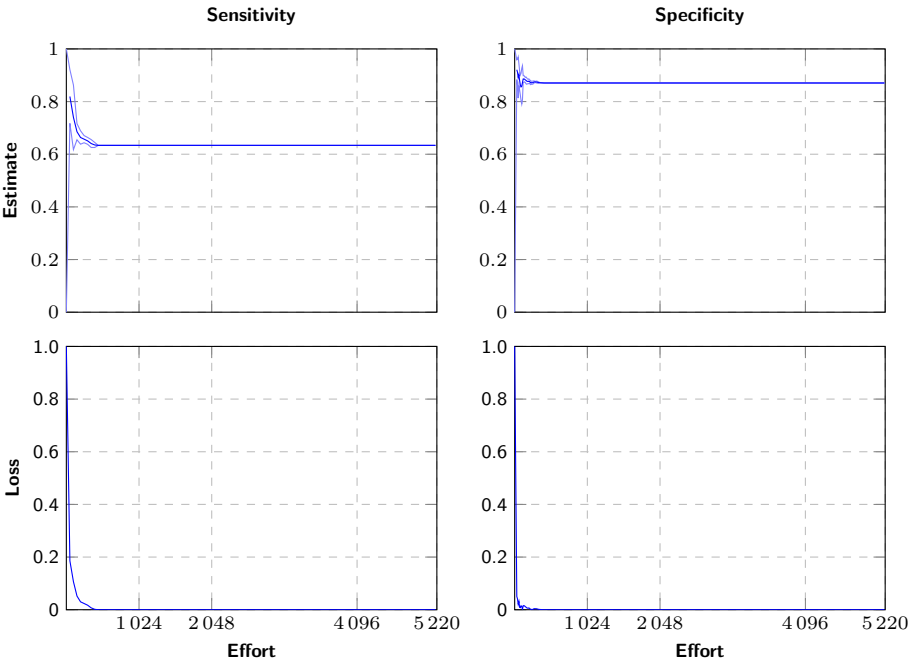


Figure 12.1 – EXAMPLE OF EFFORT/LOSS CURVE FOR A SINGLE META-ANALYSIS USING SCREENING PRIORITIZATION. The evolution of the sensitivity and specificity estimates for one diagnostic test 'CD008803 1 GDX: Inferior average' ( $n = 48$ ), where the candidate studies are screened using screening prioritization. The x axis measures the number of screened studies (effort) and the the y axis measures the summary estimates at the 25%, 50%, and 75% percentiles over 20 simulated screenings using screening prioritization. We also plot the difference to the ‘true’ values (bottom).

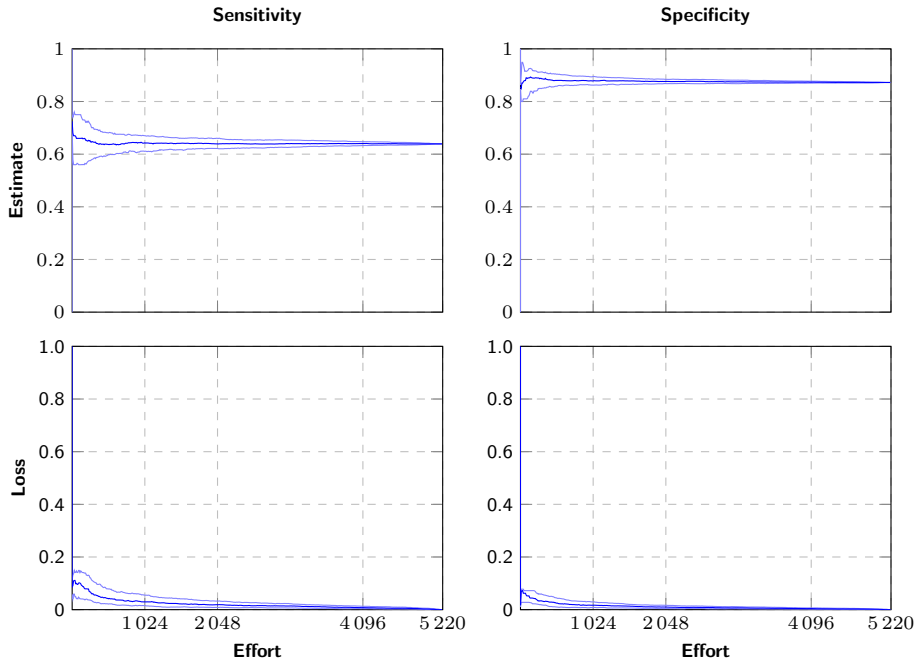


Figure 12.2 – EXAMPLE OF EFFORT/LOSS CURVE FOR A SINGLE META-ANALYSIS USING RANDOMIZED ORDER. The evolution of the sensitivity and specificity estimates for one diagnostic test 'CD008803 1 GDx: Inferior average' ( $n = 48$ ), where the candidate studies are screened in arbitrary order. The x axis measures the number of screened studies (effort) and the y axis measures the summary estimates at the 25%, 50%, and 75% percentiles over 400 simulated screenings using arbitrary (pseudorandom) order. We also plot the difference to the 'true' values (bottom).

### 12.2.3 Evolution of a Summary Estimate

We define the **EFFORT** in a screening process as the number of candidate studies screened so far. Thus we will for simplicity assume that screening a single article will always incur the same cost.

To measure the reliability of a summary estimate, we define the **LOSS** at each timestep as the absolute distance to its 'true' value, similarly to previous work on

the evolution of heterogeneity estimates by Thorlund et al. (Thorlund et al., 2012). To obtain a scalar loss score for a sensitivity/specificity pair we use the euclidean  $L_2$  distance to the true value. That is, given a true sensitivity/specificity  $(\mu, \nu)$ , then for any estimate  $(\hat{\mu}, \hat{\nu})$  we define its  $L_2$  loss as

$$L_2(\hat{\mu}, \hat{\nu}) = \sqrt{(\mu - \hat{\mu})^2 + (\nu - \hat{\nu})^2}$$

Similarly to Thorlund et al., we used the final estimate over all relevant studies as a good approximation of the ‘truth’ (Thorlund et al., 2012). This however assumes that the number of relevant studies is sufficiently large that the final summary estimates have converged and are stable.

Conventionally, the screening process first identifies all relevant studies, and the summary estimates are only estimated after the screening process has finished. However, nothing prevents systematic review authors from calculating an estimate as soon as some minimum number of studies have been identified, and then recalculate this estimate every time a relevant article is discovered (see Figures 12.1–12.2). Continuously updated, we should expect the estimate to be unreliable at first, but converge to its true value, and equivalently, the loss to approach zero.

#### 12.2.4 *Finding a Balance Between Loss and Effort*

To search for an optimal balance between loss and effort we consider two types of stopping criteria, retrospective and prospective.

**RETROSPECTIVE STOPPING CRITERIA** (cut-offs) are evaluated on the effort/loss curve (Figures 12.1–12.2) or using other information only available after screening has finished, and these criteria can therefore only be applied retrospectively. While we cannot use these criteria to decide when to stop the screening, we can use them for evaluation, i.e. to retrospectively see where we could theoretically have interrupted screening without impacting the accuracy of the summary estimate.

**PROSPECTIVE STOPPING CRITERIA** can be evaluated without knowing the final estimates or the total number of relevant studies among the candidates, and can therefore be used for decision support in a live systematic review.

##### *Retrospective stopping criteria*

**RECALL (R)** The recall, or the sensitivity of the screening procedure, measures what fraction of the relevant studies were identified by the screening procedure.

Commonly, only very high values are considered acceptable ( $R=95\%$ , and  $R=100\%$ ), but values as low as  $R=55\%$  have been considered (Cohen et al., 2012).

This is one of the only measures commonly used in previous literature (O'Mara-Eves et al., 2015), and forms the basis for evaluation measures such as  $wss@95$  (Cohen et al., 2006). Common performance metrics such as  $wss@95$  evaluates the theoretical workload reduction if screening were somehow to

be interrupted after identifying 95% of all relevant studies. However, it is not possible to know when this point has been reached during a systematic review, since it is not possible to know the number of relevant studies before screening all references.

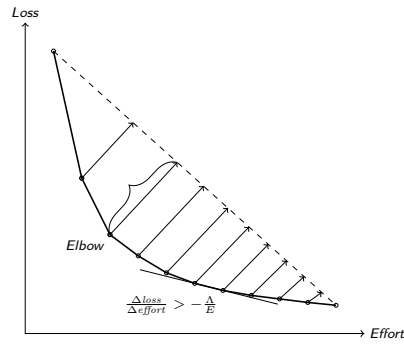


Figure 12.3 – THE ELBOW ALGORITHM AND THE SLOPE CRITERION.

**KNEE/ELBOW METHOD** We here stop at the ‘elbow’ point on the effort/loss curve (Figure 12.3). This is a point on the curve corresponding to the optimal point in terms of balance between effort and estimated precision.

Multiple definitions of the elbow point exist. We here use the definition due to Satopää et al. (Satopää et al., 2011), which is easy to implement and robust against noise. Under this definition, the knee point on the effort/gain curve is the one furthest from a straight line drawn from the first and last points on the curve.

**LOSS/EFFORT** We here stop at the point on the effort/loss curve where we would have needed to screen at least  $E$  references to further reduce the  $L_2$  loss by at least  $\Lambda$  (Figure 12.3).

This corresponds to the first consecutive pair of points  $(e_{t-1}, \lambda_{t-1})$ ,  $(e_t, \lambda_t)$  on the convex hull of the effort/loss curve such that

$$\frac{\Delta \text{loss}_t}{\Delta \text{effort}_t} = \frac{\Delta \lambda_t}{\Delta e_t} = \frac{\lambda_t - \lambda_{t-1}}{e_t - e_{t-1}} > -\frac{\Lambda}{E}$$

Since we can only calculate the loss after the screening has finished we can only apply the criterion retrospectively in this study.

The same stopping criterion has been used in similar applications, for instance for determining when all themes have been identified in ecological surveys (Tran et al., 2017). However, the effort/loss curve does not move in only one direction, since adding a single study frequently shifts the estimate away from the truth. When-

ever this happens  $\frac{\Delta\lambda_t}{\Delta e_t}$  will change signs and immediately trigger the condition. To prevent this from happening, we take the convex hull of the curve, which makes the curve monotonously decreasing.

#### *Prospective stopping criteria*

**NUMBER OF RELEVANT STUDIES RETRIEVED** We here stop as soon as we have identified  $n$  relevant studies.

176

**FOUND/EFFORT** This criterion is conceptually similar to the loss/effort criterion, except that we use the number of relevant studies found instead of the loss. We here stop at the point where we have to screen at least  $E$  references to find  $F$  additional relevant studies (Tran et al., 2017).

This corresponds to the first consecutive pair of points  $(e_{t-1}, f_{t-1})$ ,  $(e_t, f_t)$  on the the found/effort curve such that

$$\frac{\Delta found_t}{\Delta effort_t} = \frac{\Delta f_t}{\Delta e_t} = \frac{f_t - f_{t-1}}{e_t - e_{t-1}} < \frac{F}{E}$$

Unlike the loss/effort, the number of found relevant studies is monotonously increasing and we therefore do not need to take the convex hull of the found/effort curve.

This criterion is equivalent to stopping when we have not encountered a new relevant study among the last  $E/F$  candidate studies screened, and the criterion will therefore always incur a constant effort penalty equal to  $E/F$ .

**DISPLACEMENT** Every time we identify an additional relevant study we calculate how much the sensitivity and specificity estimates change when the study is included in the meta-analysis. That is, if two consecutively identified relevant studies were identified at time steps  $t$  and  $t - 1$ , and  $s_t = (\mu_t, \nu_t)$  and  $s_{t-1} = (\mu_{t-1}, \nu_{t-1})$  are the summary estimates of sensitivity and specificity at these time points, then we define the displacement at time  $t$  as

$$\Delta\lambda_t = \sqrt{(\mu_t - \mu_{t-1})^2 + (\nu_t - \nu_{t-1})^2}$$

To make the results less sensitive to noise, we will mainly consider the moving average (MA) of the displacement with window size 2 (abbreviated MA2).

This criterion can only be calculated if a summary estimate can be calculated, and is therefore undefined until at least three relevant studies have been found.

**DISPLACEMENT (LOOCV)** For any set of references, we calculate the *Leave-One-Out Cross-Validated* (LOOCV) (Molinaro et al., 2005) displacement as the median displacement when excluding each reference from the summary estimate calculations.

That is, consider that a set of studies  $S$  has been identified at some point in the screening process, where  $(\mu, \nu)$  is the summary estimate that would result when calculated based on all studies in  $S$ . Further, let  $(\mu_s, \nu_s)$  be the summary estimate that would result from excluding a single study  $s \in S$ . Then we define the LOOCV displacement as

$$\Delta\lambda_S = \text{median}_{s \in S} \left[ \sqrt{(\mu - \mu_s)^2 + (\nu - \nu_s)^2} \right]$$

This criterion can only be calculated if a summary estimate can be calculated, and is therefore undefined until at least three relevant studies have been found.

### 12.2.5 Calculation of Summary Statistics

To calculate the summary estimates we used the `REITSMA` function from the `MADA R` package (Doebler and Holling, 2015), which implements the Reitsma bivariate random effects model (Reitsma et al., 2005).

## 12.3 RESULTS

### 12.3.1 Characteristics of the Systematic Reviews

In the 63 systematic reviews in the Limsi-Cochrane dataset, the minimum number of meta-analyses was 1 (3 reviews), the mode was 2 (11 reviews), the median was 6, and the maximum was 170.

We used the combined dataset for all analyses. This dataset comprises 48 systematic reviews and 1,354 meta-analyses of diagnostic test accuracy, but only 400 of the meta-analyses were based on at least 3 primary studies in PubMed, and thus included in our analysis. Ninety-six of the meta-analyses were based on ten or more studies in PubMed. While we only consider studies from PubMed in this study, which decreases the number of studies per meta-analysis, the large majority of meta-analyses in the original systematic reviews were based on only one or two studies collected from multiple databases (Norman et al., 2018a).

The small size of the meta-analyses were reflected in the number of times the stopping criteria triggered. With cutoff set to 1 relevant per 500 screened, the found/effort criterion would have triggered for 277/400 meta-analyses, and for all meta-analyses in 30/48 systematic reviews (Ranked, found/effort (1/500) in Table

Criterion Type	Criterion	Ranked				$L^2$ loss	Sensitivity loss		Specificity loss	
		Triggered MA	SR	Effort abs	perc		mean	lb	mean	ub
Retrospective	Recall (95%)	41	1	5494.122	0.489	0.960	0.001	0.001	0.000	0.000
	Knee/elbow	314	30	120.752	0.047	0.299	0.056	0.040	0.074	0.086
	Loss/effort (0.02/1,000)	314	30	259.570	0.104	0.527	0.023	0.018	0.140	0.053
	Loss/effort (0.015/1,000)	314	30	271.589	0.108	0.539	0.022	0.018	0.136	0.052
	Loss/effort (0.01/1,000)	314	30	297.064	0.113	0.546	0.021	0.016	0.136	0.049
Prospective	Found/effort (1/500)	277	30	811.220	0.268	0.770	0.012	0.008	0.011	0.014
	Found/effort (1/1,000)	254	22	1338.465	0.386	0.783	0.007	0.005	0.006	0.005
	Found/effort (1/2,000)	174	17	2330.448	0.424	0.815	0.002	0.002	0.002	0.002
	Relevant found ( $n = 20$ )	41	1	162.732	0.050	0.565	0.023	0.020	0.026	0.022
	Relevant found ( $n = 15$ )	57	3	140.000	0.039	0.534	0.029	0.024	0.029	0.026
	Relevant found ( $n = 10$ )	96	4	109.042	0.044	0.529	0.031	0.024	0.039	0.028
	Displacement MA2 (0.005)	35	0	344.714	0.086	0.667	0.014	0.012	0.015	0.012
	Displacement MA2 (0.010)	57	2	280.000	0.065	0.578	0.018	0.016	0.021	0.018
	Displacement MA2 (0.015)	74	3	205.189	0.059	0.538	0.024	0.019	0.030	0.021
	Displacement MA2 (0.020)	91	4	124.099	0.044	0.511	0.029	0.024	0.034	0.025
	Displacement LOOCV (0.005)	42	0	131.667	0.050	0.538	0.021	0.019	0.030	0.016
	Displacement LOOCV (0.010)	90	3	192.356	0.054	0.461	0.030	0.026	0.038	0.025
	Displacement LOOCV (0.015)	130	5	161.923	0.063	0.427	0.035	0.028	0.042	0.030
	Displacement LOOCV (0.020)	152	5	172.947	0.053	0.370	0.042	0.034	0.054	0.034
Criterion Type	Criterion	Randomized				$L^2$ loss	Sensitivity loss		Specificity loss	
		Triggered MA	SR	Effort abs	perc		mean	lb	mean	ub
Retrospective	Recall (95%)	36	1	6956.583	0.961	0.959	0.002	0.002	0.000	0.001
	Knee/elbow	314	30	1658.322	0.299	0.203	0.064	0.042	0.208	0.103
	Loss/effort (0.02/1,000)	309	27	1315.816	0.351	0.264	0.047	0.036	0.269	0.101
	Loss/effort (0.015/1,000)	309	27	1429.359	0.364	0.285	0.043	0.033	0.249	0.097
	Loss/effort (0.01/1,000)	309	27	1552.871	0.390	0.313	0.040	0.030	0.237	0.093
Prospective	Found/effort (1/500)	218	16	1854.106	0.374	0.169	0.112	0.073	0.329	0.143
	Found/effort (1/1,000)	151	10	2942.583	0.465	0.237	0.097	0.062	0.268	0.113
	Found/effort (1/2,000)	68	6	5097.559	0.493	0.272	0.082	0.052	0.239	0.083
	Relevant found ( $n = 20$ )	41	1	5106.000	0.574	0.565	0.022	0.018	0.023	0.021
	Relevant found ( $n = 15$ )	57	3	4288.070	0.550	0.534	0.024	0.020	0.026	0.025
	Relevant found ( $n = 10$ )	93	4	3275.194	0.537	0.517	0.032	0.024	0.039	0.028
	Displacement MA2 (0.005)	33	0	6478.970	0.663	0.644	0.016	0.014	0.017	0.014
	Displacement MA2 (0.010)	54	2	4395.889	0.537	0.519	0.023	0.019	0.027	0.021
	Displacement MA2 (0.015)	74	3	4103.730	0.533	0.505	0.025	0.021	0.030	0.023
	Displacement MA2 (0.020)	87	4	3478.276	0.509	0.485	0.026	0.022	0.034	0.024
	Displacement LOOCV (0.005)	40	1	5525.555	0.557	0.537	0.017	0.014	0.020	0.014
	Displacement LOOCV (0.010)	87	3	3928.483	0.498	0.454	0.027	0.022	0.037	0.022
	Displacement LOOCV (0.015)	130	4	2902.831	0.475	0.415	0.038	0.031	0.048	0.030
	Displacement LOOCV (0.020)	154	5	2808.455	0.453	0.396	0.041	0.035	0.051	0.034

Table 12.1 – AVERAGE MEASURED LOSS FOR EACH CRITERION, MEASURED FOR ALL TESTS WHERE THE CRITERIA TRIGGERED. Triggered signifies the number of meta-analyses (MA, maximum 400) for which the criterion triggered, and the number of systematic reviews (SR, maximum 48) where the criterion triggered for all meta-analyses. Effort signifies the absolute and relative number of references needed to be screened before triggering the stopping criteria. Recall signifies the percentage of relevant studies identified when the stopping criterion triggered. The loss in sensitivity and specificity are measured as the difference to the final estimates at the criterion threshold. We also include the the difference between the measured lower and upper bounds of the 95% confidence intervals and their final estimated values (lb, ub).

12·1). With cutoff set to 1 relevant per 2,000 screened, it would have triggered for 174/400 meta-analyses, and for all meta-analyses in 17/48 systematic reviews (Ranked, found/effort (1/2,000) in Table 12·1). With cutoff set to 0.02, the displacement criterion would have triggered for 91/400 meta-analyses, or for all meta-analyses in 4/48 systematic reviews (Ranked, displacement MA2 (0.02) in Table 12·1). With cutoff set to 0.005, it would have triggered for 35/400 meta-analyses, and for all meta-analyses in no systematic review (Ranked, displacement MA2 (0.005) in Table 12·1).

### 12·3·2 *How Many Studies Does it Take to Make a Meta-analysis?*

179

The displacement when including the last relevant study in the meta-analyses decreases with the total number  $n$  of studies included in the meta-analysis (Figure 12·4). The last primary study added to the summary estimate calculations displace the estimates by 16 percentage points or less for  $n \geq 5$ , by 4 points or less for  $n \geq 10$ , by 2 points or less for  $n \geq 20$ , by 1 point or less for  $n \geq 50$ .

There is a moderately strong correlation (Pearson  $r = 0.54$ ) between the last displacement and the  $L_2$  loss at each summary estimate update. The correlation can be made somewhat stronger by taking the moving average over the last few successive summary estimate updates to cancel out some of the spurious values (MA2:  $r = 0.59$ , MA3:  $r = 0.60$ , MA4:  $r = 0.59$ , MA5:  $r = 0.58$ ). Averaging the displacement using leave-one-out cross-validation (Molinaro et al., 2005) gives similar correlation to MA3 ( $r = 0.60$ ).

### 12·3·3 *Contribution of Screening Prioritization*

12

Screening prioritization requires screening a much smaller number of candidate references to reach the cut-off point for all criteria, particularly for prospective criteria. For instance, identification of at least 10 relevant primary studies for each applicable meta-analysis would be reached after screening an average of 4.4% of the candidate studies, while we would have needed to screen an average of 53.7% of the candidate studies in randomized order to achieve the same (Relevant Found ( $n = 10$ ) in Table 12·1). To identify 20 relevant studies for each meta-analysis, it would have been necessary to screen an average of 57.4% of the references in random order, but only 5.0% using screening prioritization (Relevant Found ( $n = 20$ ) in Table 12·1). For all criteria except the found/effort the estimation error is similar at the cut-off point for prioritized screening and screening in random order.

On average, the displacement threshold criterion and the number of relevant found exhibit roughly similar behavior in terms of accuracy and efficiency. In Table 12·1 we see that if we stop after finding 10 relevant studies (Criterion: 'Relevant Found

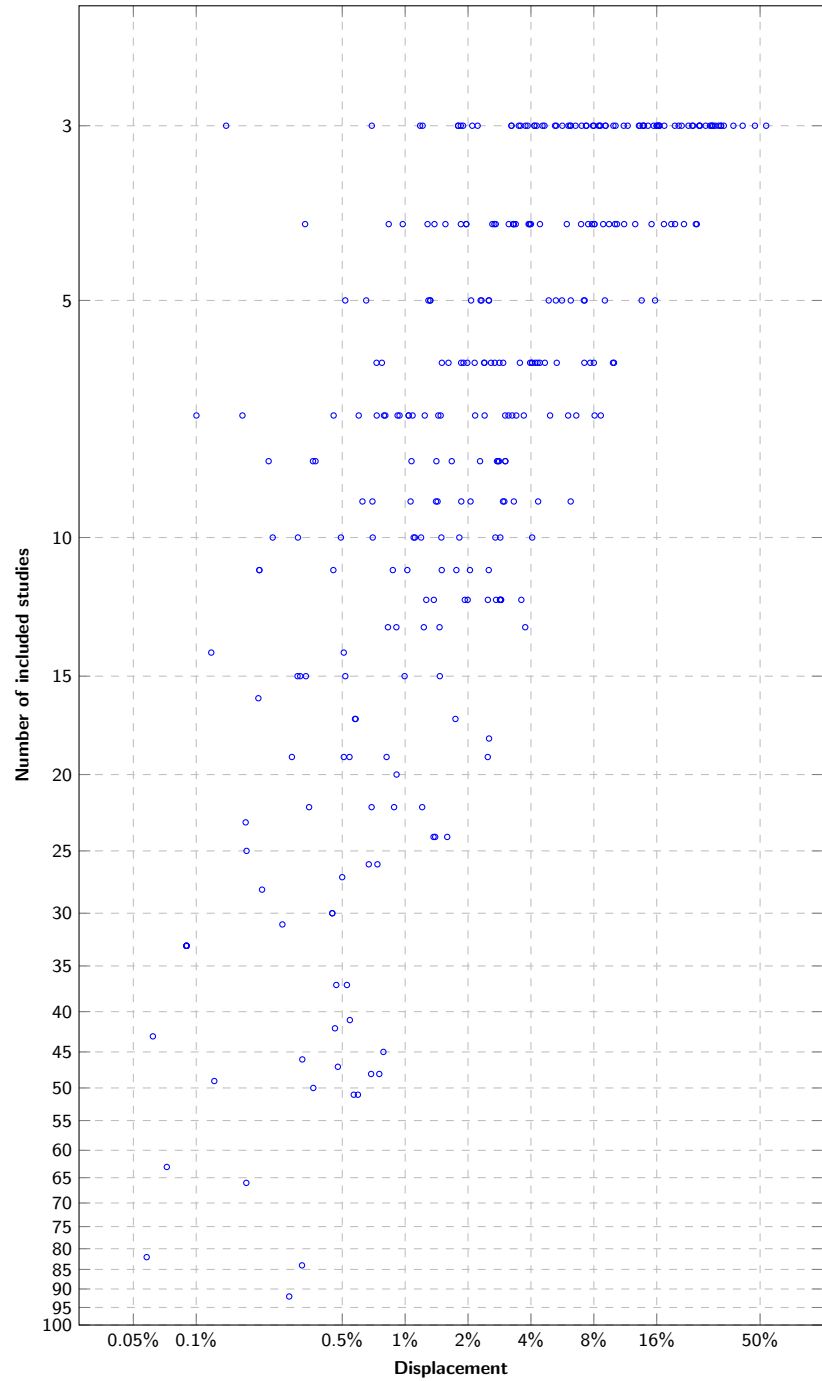


Figure 12.4 – DISPLACEMENT VERSUS NUMBER OF RELEVANT PRIMARY STUDIES The  $x$ -axis denotes how much the estimate changed when the last relevant primary study was included ( $L_2$  distance between successive sensitivity/specificity pairs). The  $y$ -axis denotes the total number of relevant primary studies found for the diagnostic test.

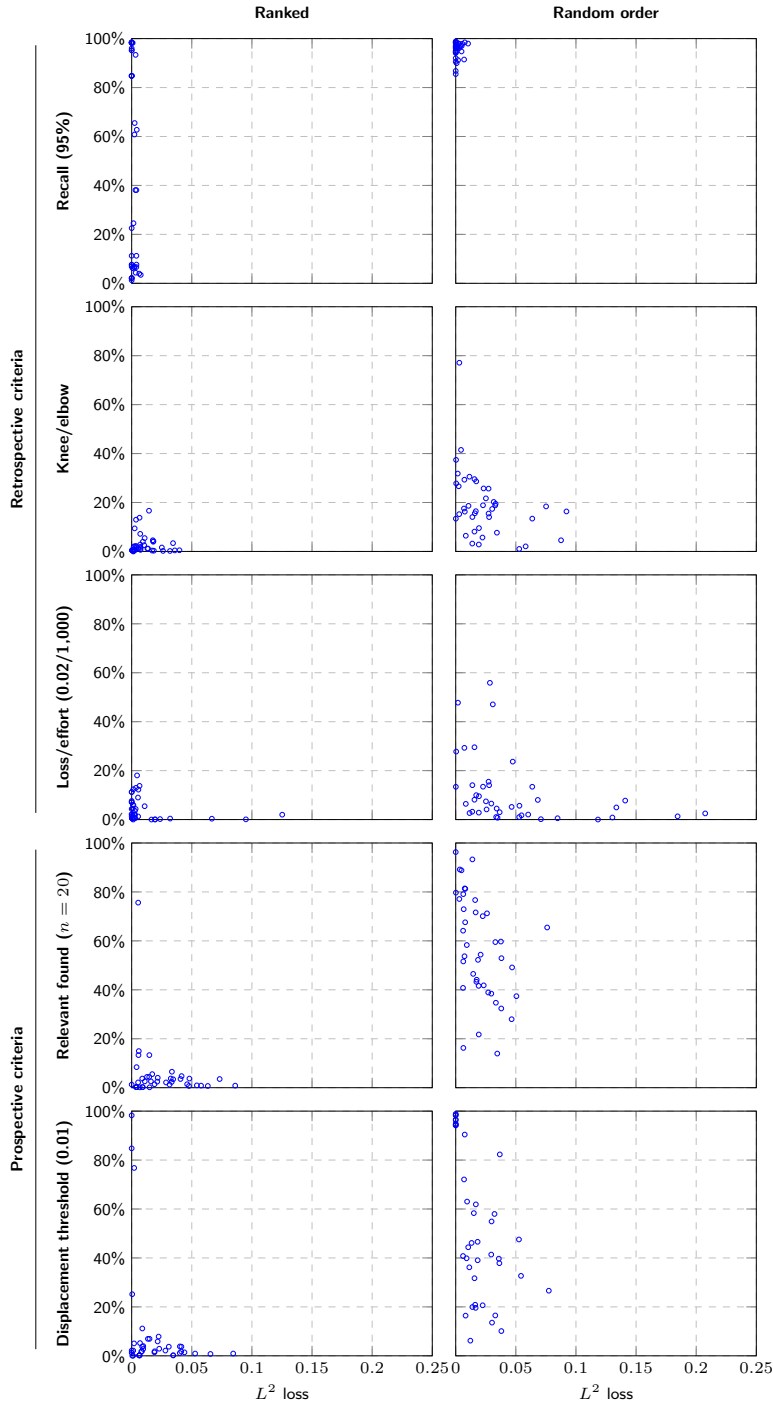


Figure 12.5 – COMPARISON BETWEEN STOPPING CRITERIA Effort (y-axis) versus  $L^2$  distance to final summary estimate (x-axis) for each stopping criteria in the meta-analyses. We only included meta-analyses based on at least 20 studies, so that the criteria were applicable to all meta-analyses, and consequently that all data points occur in all scatterplots. This is limited by the relevant found criterion, which only makes sense for meta-analyses based on at least 20 studies.

( $n = 10$ )') we would mis-estimate the mean sensitivity by approximately 2.4 percentage point and the mean specificity by approximately 1.3 percentage point. If we stop after observing a mean 0.02 displacement over the last two updates (Criterion: 'Displacement MA2 (0.02)') we would also have needed to screen 4.4% of the candidate studies on average, and we would have mis-estimated the mean sensitivity by approximately 2.4 percentage point and the mean specificity by 1.0 percentage point.

Stricter thresholds allow trading a higher screening workload for lower estimation error. For instance, stopping after finding 20 relevant studies (Criterion: 'Relevant Found ( $n = 20$ )') leads to screening 5.0% of the candidate studies on average, and mis-estimates the mean sensitivity by approximately 2.0 percentage point and the mean specificity by approximately 0.7 percentage point. Similarly, stopping after observing a mean 0.005 displacement over the last two updates (Criterion: 'Displacement MA2 (0.005)') leads to screening 8.6% of the candidate studies on average, and mis-estimates the mean sensitivity by approximately 1.2 percentage point and the mean specificity by approximately 0.7 percentage point.

However, while the average discrepancy is only 2 percentage point, the results vary greatly between meta-analyses, and the discrepancy for a given meta-analysis may be as high as 8 percentage point, even with a conservative threshold (Figure 12.5).

## 12.4 DISCUSSION

By monitoring the moving average of the displacement we were able to estimate the current precision of the diagnostic test accuracy estimates through the screening process. However, the meta-analyses of diagnostic test accuracy were accurate within 2% only for meta-analyses including at least 20 studies (Figure 12.4). A criterion to interrupt screening once the displacement falls below 2% would consequently have triggered in 91/400 meta-analyses (Table 12.1). Many meta-analyses had poor accuracy even when based on all relevant studies (Figure 12.4).

### 12.4.1 *Estimates Converge Faster Using Screening Prioritization*

Screening prioritization identifies most or all relevant primary studies much earlier in the screening process compared to randomized order (Figure 12.5). The rate of identification of relevant studies will generally be high initially, before dropping down to a trickle. This rate can be used either to estimate how many relevant studies exist among the candidates (Cormack and Grossman, 2016), or directly as a stopping criteria (cf. found/effort in Table 12.1). When screening in randomized order the gaps between successive relevant studies is likely to be large, with highly variable size, which makes it more difficult to estimate the identification rate, or

the total number of relevant studies. Consequently the found/effort criterion interrupts too prematurely in randomized order leading to higher loss for sensitivity, specificity, and their associated confidence intervals, for all evaluated cut-offs (Table 12-1, bottom section).

We can also observe that the summary estimates converge to their final values much more quickly and reliably than when screening in arbitrary order (Figures 12-1 and 12-2). In other words, screening prioritization allows producing almost the same estimates with reduced effort – the problem is knowing whether it is safe to interrupt the screening prematurely. However, screening prioritization may allow meta-analyses to be started after screening a few percent of the candidate references. Even if the authors of the systematic review decide that all references need to be screened to ensure that nothing is missed, the meta-analysis may be conducted in parallel with screening the remaining references, and can later be updated to account for any additional studies found.

183

#### 12.4.2 *Sufficiently Large Meta-analyses Can be Stopped Prematurely*

For any individual summary estimate, we can have two outcomes:

1. The systematic review fails to identify sufficient evidence, and the estimates produced by the published systematic review may in fact be biased or unreliable due to the insufficient amount of evidence.
2. The estimate is unbiased and reliable at some point in the screening process. Continuing the screening process is unlikely to change the precision of the estimate (cf. Figure 12-1), and the effort could arguably be spent elsewhere.

The systematic review process implicitly assumes the borderline case between these two, where the estimate becomes unbiased and reliable only and exactly at the end of the screening. Our results suggest this may not be an unreasonable assumption when screening in random order – the displacement fell below a tolerance of 0.1 only during the last 10% of the screened references for 10 out of 41 meta-analyses based on at least 10 studies (Random, Displacement threshold (0.01) in Figure 12-5). However, the same was only true for 1 out of 41 meta-analyses when using ranked order (Ranked, Displacement threshold (0.01) in Figure 12-5).

In case 1, we could arguably stop screening (and possibly refine the database search) as soon as it becomes clear that a sufficient number of relevant studies cannot be retrieved. We cannot know with absolute certainty how many remaining studies exist for us to find. However, the found/effort curve will typically be convex when the candidate list is ranked, and extrapolating from its current slope therefore pro-

vides a probabilistic upper bound of the number of remaining studies (Cormack and Grossman, 2016).

Case 2 assumes a sufficiently large amount of evidence to base the summary estimates upon. Then, as additional evidence is accumulated, the summary estimate will converge to its true value. The value of additional evidence will drop accordingly as the estimate becomes increasingly stable.

We previously estimated the average discrepancy when replicating summary estimates in the systematic reviews at approximately 2 percentage point (Norman et al., 2018a), and we can take this as a minimum requirement for estimation accuracy. On average, we can achieve the same or better estimation accuracy with the displacement criterion with a cut-off of 0.01 or lower, or with the found/effort criterion with a cut-off of 1/500 or lower.

#### 12.4.3 *Data Saturation is Seldom Reached in DTA Systematic Reviews*

We observe a consistent positive relationship between meta-analysis size and the accuracy of the estimates (Figure 12.4). The least accurate diagnostic test accuracy estimates occurred for meta-analyses of three included studies and were accurate only within roughly 50% of their final values (Figure 12.4). The vast majority of estimates were not accurate within 2% at the end of the screening process. These results mirror the work of Wetterslev et al, who have previously observed that most Cochrane systematic reviews of interventions are insufficiently powered to even detect or reject large intervention effects (Wetterslev et al., 2017)

Our stopping criteria based on displacement will only interrupt the screening process once the estimates have stabilized due to data saturation. If data saturation fails to occur because too few studies exist to find the screening will not be interrupted. We can however also interrupt screening if it becomes clear that no further studies will be uncovered by the screening process, i.e. by using a stopping criterion like found vs effort.

For instance, using a combination of stopping criteria (Displacement (0.01) OR Relevant ( $n = 15$ ) OR Found/Effort (1/1,000)) would have reduced the screening effort by 21.5–99.9285% (mean: 81.7%, median: 90.56%) for the main meta-analysis in 33 out of 38 systematic reviews with an average 1.2% estimation error (Figure 12.6). The 5 systematic reviews where the effort would not have been reduced were among the smallest with a total number of candidate studies ranging from 64 to 981. Ten systematic reviews performed no meta-analysis with at least three studies in PubMed and were therefore excluded from this analysis.

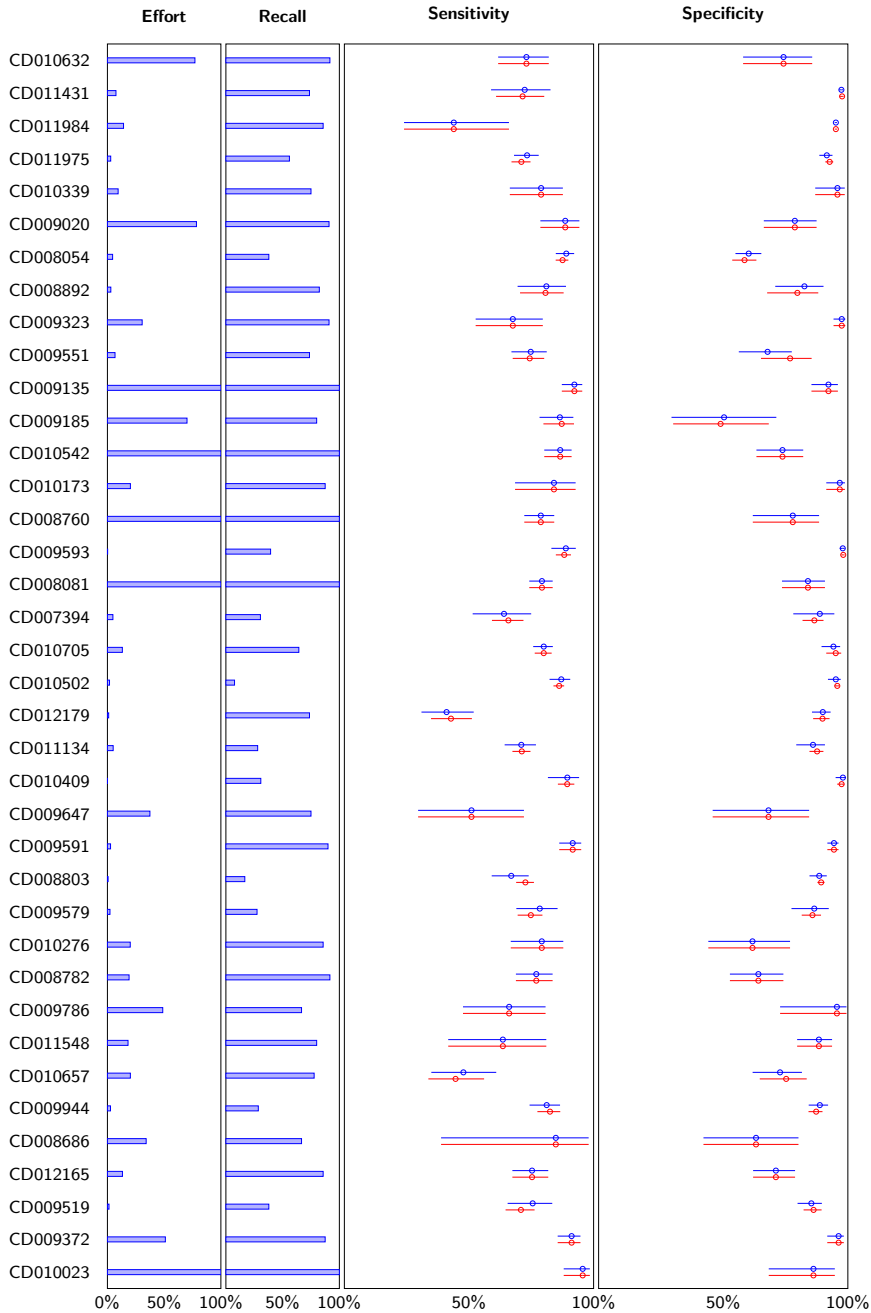


Figure 12.6 – THE IMPACT OF SCREENING PRIORITIZATION AND STOPPING CRITERIA ON META-ANALYSES. Difference in meta-analysis results for the largest meta-analysis in each systematic review using a combination of stopping criteria (Displacement (0.01) OR Relevant ( $n = 15$ ) OR Found/Effort (1/1,000)). Ten systematic reviews did not include any meta-analysis based on three or more studies (in PubMed) and were therefore excluded from the results. Effort denotes the fraction of candidate references screened. Recall denotes the fraction of identified relevant studies. Blue data points correspond to the simulated results using early stopping. Red data points correspond to results without early stopping, i.e. equivalent to current practice (which would have 100% effort and 100% recall).

12.4.4 *External validity*

We have presented 7 criteria and have evaluated how these perform when using logistic regression for ranking, and when using random order. We expect these criteria to generalize differently if used with other methods.

The  $L_2$  loss guarantees presented for the recall, the relevant found, and the displacement (either with MA2 or LOOCV) only depend on the relative order of the relevant studies, and is otherwise independent of where in the ranking the relevant studies occur. In other words, whether our results for these criteria extend to other methods only depends on how the method orders the relevant studies. In this study we demonstrate that using these criteria with logistic regression results in the same  $L_2$  loss compared to random order, and thus that logistic regression does not bias the meta-analyses compared to random order. In light of this, we expect these criteria to yield similar  $L_2$  loss for any ranking method that is similarly unbiased.

The knee/elbow criterion, the loss/effort criterion, and the found/effort criterion all depend on the relative order of all studies, both relevant and non-relevant, and can therefore be expected to give different results depending on the strength of the ranking method. We can observe this in Figure 12.5, where the knee/elbow criterion and the loss/effort criterion result in larger and more frequent  $L_2$  losses for random order than for ranked order. The found/effort criterion breaks down entirely for random order and yields unacceptably large  $L_2$  losses (see Randomized: Found/effort in Table 12.6). In light of this, the parameters we use for these criteria thus cannot be assumed to yield the same  $L_2$  losses for other ranking methods, and would need to be recalibrated when used with other methods.

In this study we have only considered meta-analyses with at least three included studies. However, the prospective criteria are conservative and will simply not trigger when used in a systematic review where there are only two or less studies to find. The only exception is the found/effort criterion, but this criterion can easily be modified so that it is ignored before at least three relevant studies have been found.

12.4.5 *Recommendations*

We explicitly refrain from recommending specific stopping criteria or specific cut-off values, since there is no one size that fits all systematic reviews – the criteria and their parameters need to be decided to suit the purposes of the review. If automation is adopted in a systematic review, acceptable tolerances should be decided as part of the protocol, and the protocol should include a strategy to ensure that the tolerance criteria will be satisfied.

We recommend that several stopping criteria be monitored in parallel, and that

screening is interrupted only once criteria for all necessary aspects of the systematic review are satisfied. In this study we focus on the accuracy of the main meta-analysis – similar criteria should also be specified for all other aspects deemed necessary for the review, such as the identification of all subgroups, or estimates of prevalence of the diagnosed condition.

Specifically, to monitor the accuracy of the sensitivity and specificity estimates, we recommend the use of:

- ✦ The displacement  $MA_2$  criterion, set to half the required tolerance
- ✦ The displacement LOOCV criterion, set to half the required tolerance
- ✦ The relevant found criterion, set conservatively (15 at a minimum)

The  $MA_2$  and LOOCV displacement yield similar information and do not need to be monitored simultaneously. The LOOCV variant underestimated the loss in our experiments more than the  $MA_2$ , and triggered more often with larger average  $L_2$  loss, and we therefore recommend the  $MA_2$  variant over LOOCV. On average, both variants overestimated the final  $L_2$  loss and we recommend the displacement be interpreted with this in mind.

These criteria triggering mean that the current estimate is accurate within a given tolerance, and that further studies are unlikely to change the estimates, even if a large number of relevant studies still exist to find. These criteria can also be used with randomized screening, and likely also for any screening prioritization method that does not bias the order of the relevant studies. If the displacement criterion is infeasible to calculate, the relevant found criterion can be used alone, but it may be difficult to infer meta-analysis accuracy from the number of relevant studies included.

- ✦ The loss/effort criterion with a conservative parameter setting (1/1,000 or stricter)

This criterion triggering is an indication that no further studies exist to find. This criterion should be treated with more caution than the other criteria. In particular, the criterion depends on the strength of the screening prioritization method, and can trigger prematurely e.g. if the method struggles to find some subset of the relevant studies, or if the screening prioritization method is generally poor.

The found/effort criterion is also more likely to trigger prematurely if the total number of relevant studies is low. Therefore we also recommend not using this criterion until some minimum number of studies have been identified (three appear to be a safe choice for the current setting and the current method).

#### 12.4.6 *Limitations of this Study*

This study focused on systematic reviews of diagnostic test accuracy studies. Therefore, we do not know what the implications are for other types of systematic reviews. However, the methods in this study are applicable to systematic reviews estimating numerical values, and our results may therefore be applicable also to systematic reviews of interventions.

Due to the nature of the datasets we could only recalculate meta-analyses using data from studies indexed in PubMed. Previous studies examining the impact of only searching PubMed on meta-analyses of interventions demonstrated moderate changes in estimates, and observed changes were equally likely to favor controls as interventions (Halladay et al., 2015; Marshall et al., 2019). In this study we assume that searching only PubMed is similarly unbiased for diagnostic test accuracy, but we are aware of no studies examining this directly. Limiting the meta-analyses to PubMed does however reduce the number of studies available for analysis, and may therefore mean that we are underestimating the applicability of these stopping criteria, and that we may be observing greater variance than we would in a prospective setting.

This study focused on Cochrane systematic reviews, which are known to have higher consistency and lower bias than other systematic reviews (Jadad et al., 1998). It is not clear what the implications are for systematic reviews conducted with less stringency than Cochrane systematic reviews.

The definition of loss we use for evaluation ( $L_2$ ) makes the simplifying assumption that sensitivity and specificity are equally important. Specificity values of diagnostic tests tend towards values close to one, and thus often exhibit smaller variance than the sensitivity. As a result, the  $L_2$  loss is often dominated by the sensitivity loss (Table 12.1). We also report loss separately for sensitivity and specificity in our analysis.

#### 12.4.7 *Future Work*

Future work will evaluate the validity of these results in prospective settings. We also plan to use Bayesian methods to estimate final meta-analysis accuracy from the study data accumulated through the screening process. Furthermore, we will also aim to extend this approach to other study types beyond diagnostic test accuracy, such as intervention studies.

## 12.5 CONCLUSIONS

Our results suggest that diagnostic summary sensitivity and specificity can be estimated within an accuracy of 2 percentage points while deliberately missing over 40% of the relevant studies within a single database. This is contrary to current guidelines which assume that an exhaustive search is necessary to produce reliable estimates with low bias. On the other hand, we find a clear relationship between the absolute size of the meta-analysis and the reliability and precision of the estimates. In other words, a reliable meta-analysis requires identifying a sufficient number of studies, but how large a fraction of relevant studies is identified is less important.

In the simulations, a combination of stopping criteria reduced the screening effort by 71.2% on average (median: 86.8%, range: 0% to 99.93%) for the main meta-analysis in each systematic review, and triggered in every systematic review with more than 1,000 candidate studies. No systematic review required screening more than 2,308 studies, whereas exhaustive manual screening required screening up to 43,363 studies. Despite an average 70% recall the estimation error was 1.3% on average, much less than the estimation error expected when replicating summary estimate calculations.

The (retrospective) 95% recall criterion yielded an average 0.1% error when ranking with logistic regression, and an average 0.2% error when using random order. Thus, we confirm the hypothesis that 95% recall is sufficient to accurately estimate the main meta-analysis in systematic reviews of diagnostic test accuracy, provided the ranking method is unbiased. On the other hand, we observe almost unchanged estimates (within 2% tolerance) for recall as low as 30%, and 95% recall is thus not necessary to reach accurate estimates.

## LIST OF ABBREVIATIONS

- CLEF Conference and Labs of the Evaluation Forum, formerly known as the Cross-Language Evaluation Forum
- DTA Diagnostic test accuracy
- LOOCV Leave-one-out cross-validation
- MA Moving average. We further denote moving average with different window sizes by MA2, MA3, MA4, et c.

## DECLARATIONS

ETHICS APPROVAL AND CONSENT TO PARTICIPATE Not applicable.

CONSENT FOR PUBLICATION Not applicable.

AVAILABILITY OF SUPPORTING DATA The first dataset supporting the conclusions of this article is available in the Zenodo repository (DOI: [10.5281/zenodo.1303259](https://doi.org/10.5281/zenodo.1303259), or <https://zenodo.org/record/1303259#.XBpPGMZ7kUE>), the second is available from github (<https://github.com/CLEF-TAR/tar>).

COMPETING INTERESTS The authors declare that they have no competing interests.

FUNDING This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676207.

AUTHOR'S CONTRIBUTIONS CN wrote the first draft and conducted the experiments. All authors conceived and designed the study. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS Not applicable.

**T**HE TWO STUDIES presented in this section have similar aims. They both seek to ensure that using screening automation in the systematic review does not fundamentally change the integrity of the systematic review. A systematic review that uses screening automation should not be methodologically inferior to one conducted according to the established systematic review process. The systematic review should remain reproducible, transparent, and free of bias.

However, the approach used to ensure this integrity are markedly different in the two papers. In the first paper we attempted to adhere as closely as possible to the conventional process, down to screening in randomized order in EndNote. Since the process is fundamentally unaltered we argue that is as unbiased as the conventional process. In the second paper we replace the conventional process with one using screening prioritization, and demonstrate that this results in almost exactly the same results and conclusions for the meta-analyses.

191

### 13.1 SCREENING REDUCTION COMPATIBLE WITH CURRENT PRACTICE

The first study in this part is prospective, and details the use of the static intratopic method (described in part II) in the 2019 update of the COMET database. In this study the cut-off was determined retrospectively on previous review updates. In other words, we identified a threshold that would have resulted in an acceptable balance between workload reduction and screening exhaustiveness in previous review updates, and applied this criterion in the screening for the 2019 review update. Consequently, the approaches work by assuming that the results would be practically similar for each update, and that we can therefore extrapolate from historical data to future updates.

We judged that missing 2% of the references was an acceptable trade-off for a 75% workload reduction. This particular review is intended to populate a literature database, and recall is therefore a direct measure of the impact the screening automation has on the review.

References where meta-data did not include any abstract could not be ranked with acceptable performance guarantees, and were therefore ineligible for screening automation. There were however only a small number of such references, corresponding to a workload of approximately 2–4 hours per screener.

Recall can only be calculated retrospectively, which means that the 2% loss in recall is just an estimate. Relevant studies for inclusion in the COMET database are identified from multiple sources, and time will tell whether the estimated number of missed articles match reality. Screening was done on a small sample (1%) of the excluded references to verify the results however, and all of these references were found to be correctly excluded.

13

The application of screening automation was done to adhere to the established process as closely as possible. First we use a screening reduction approach where the inclusion threshold could be determined as part of the protocol. We applied the model before starting the screening, and the set of remaining references were exported in EndNote format. The records were randomized to avoid rank order bias prior to screening. The screening was then performed as normal using EndNote. Unlike previous years, the 2019 update only involved two screeners, but the remainder of the process was unchanged – apart for the use of screening reduction. No specialized software was required by the screeners.

Neither transparency or reproducibility were hard requirements for this review. The screening reduction method used (logistics regression with stochastic gradient descent) is stochastic, and repeated applications will result in slightly different output. Even so, the screening model and the resulting ranked list have been kept, and the systematic review screening process is therefore traceable.

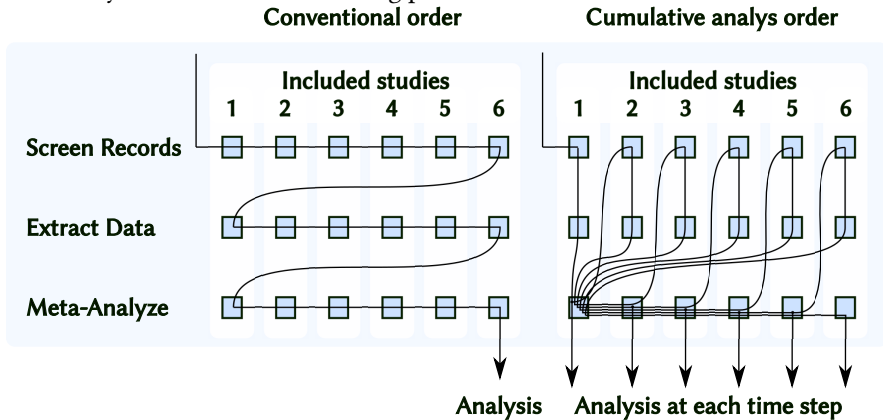


Figure 13.1 – The order in which records are screened in the conventional screening process and with the cumulative meta-analysis process.

## 13.2 BETTER METRICS FOR SCREENING AUTOMATION

In the second study in this section we have tried to measure ‘information loss’ directly for systematic reviews of diagnostic test accuracy. In simple terms, we have tried to ask the question: what does it mean for an abridged method to yield the ‘same’ systematic review as with exhaustive screening?

Such a measure should optimally satisfy three criteria:

- ❖ The measure should be possible to calculate cumulatively through the screening process
- ❖ It should be possible to stop screening once we are confident that further screening will not change the conclusions of the review; and
- ❖ It should be possible to determine criteria for stopping as part of the review protocol to avoid bias

If screening automation methods are to be used in systematic reviews, reviewers need to judge what amount of loss is acceptable for the current review. But more importantly, reviewers need to judge what *kind* of loss is acceptable, likely across multiple dimensions. A loss in rigor or bias is unlikely to be acceptable. On the other hand, a loss in recall or exhaustiveness may yield a review that does not ‘look like’ a systematic review, but may not meaningfully impact the results and conclusions of the review – provided a sufficient selection of studies are identified to address all review questions, and provided the selection of studies is essentially random. To avoid reviewer bias and ad-hoc decisions during the screening process there should be a clear, pre-specified protocol for judging when the screening is complete.

In the second study in this section, we have looked at using the meta-analysis accuracy as a performance metric during screening by performing cumulative meta-analyses through the screening process. This accuracy can be estimated prospectively and thresholds can be decided as part of the protocol. This however requires the screening process to be performed in parallel with the data extraction, synthesis and meta-analysis stages of the systematic review process, and would thus result in an unconventional systematic review process (see figure 13-1).

The benefit of this measure is that it is conservative and reliable, and interrupting screening once the accuracy falls within prespecified limits is unlikely to lead to wrong results or conclusions in the systematic review. Furthermore, this allows the screening to be interrupted much earlier in the process and reduce the workload by orders of magnitude more than with conventional stopping criteria.

Performing full-text retrieval, data extraction, and cumulative meta-analyses on each additional identified relevant study could introduce bias by influencing what studies will be including in the remainder of the screening. However, this kind of bias is only relevant to humans seeing cumulative results – and if the full-text retrieval and data extraction stages could be sufficiently automated, the screeners could be blinded from the full-texts and extracted data.

In practice, a simpler way to achieve the same blinding – and one that is possible today – would be to let different authors perform screening and data extraction. This may not be an option for small reviews involving a small number of authors, but

may be practical for reviews with very large numbers of candidate records, where additional authors would otherwise have been set to screen references. Generally, it is these large reviews that are in the greatest need of automation.

### 13.3 CONCLUSIONS

Systematic review automation method can be used in systematic reviews without fundamentally altering the process. Screening reduction method can be used as an extra search filter, leaving the remainder of the review process identical to the conventional process, including screening in random order, and the use of standard reference managers like EndNote.

The accuracy of the screening process, and the impact it has on the results and conclusions of the review can be measured prospectively through the screening process using cumulative meta-analyses. This does however require modifying the systematic review process to perform data extraction and meta-analyses concurrently.

## PART IV

# DATA EXTRACTION & SYNTHESIS

This part of the thesis is based on the following publications:

([Norman et al., 2018a](#)): Norman, C., Leeflang, M., and Névéol, A. (2018a). Data extraction and synthesis in systematic reviews of diagnostic test accuracy: A corpus for automating and evaluating the process. In *AMIA Annual Symposium Proceedings*, volume 2018, page 817. American Medical Informatics Association

([Norman et al., 2019e](#)): Norman, C., Spijker, R., Kanoulas, E., Leeflang, M., and Névéol, A. (2019e). A distantly supervised dataset for automated data extraction from diagnostic studies. *ACL BioNLP*)


The contents in this section has also been presented at the following venue:

([Norman et al., 2019d](#)): Norman, C., Leeflang, M., Porcher, R., and Névéol, A. (2019d). Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. In *AMIA Annual Symposium Proceedings*

The contents in this section is also based on material accepted and originally planned to be presented at the following conference:

([Norman et al., 2019a](#)): Norman, C., Leeflang, M., and Névéol, A. (2019a). Automated checking for human errors in meta-analyses of diagnostic test accuracy. In *Cochrane Colloquium*

Unfortunately, this 2019 Cochrane Colloquium was cancelled due to the October 2019 civil unrest in Santiago, Chile. The presentation was instead presented virtually.

NCE THE SELECTION of included studies is finished, the results of the studies will be synthesized and analyzed, and the review will be written up for publication. This involves several steps. Some of these steps – such as drawing conclusions based on the data – require human judgement and are not time-consuming. These are therefore difficult to automate, and would result in only limited benefits. However, other steps, such as extracting data from articles and calculating statistical analyses, are repetitive, mechanical, and time-consuming, and thus prime candidates for automation.

In this part of the thesis we will look at the later stages of the systematic review process, after relevant studies have been identified, until the results of the review are written up for publication, with a particular focus on what parts may be possible to automate using current methods.

We will focus on the *data extraction*, *data synthesis* and the *analysis* stages. In the conventional systematic review process, systematic reviewers may perform other actions besides these, such as snowballing and re-checking the literature, but these actions conceptually belong to the article selection process, rather than the synthesis process. Furthermore, the main reason these action are done during the last stages of the review is largely an artifact of the manual nature of the review. Snowballing is performed after full-text screening because this requires the reference lists from the included studies. Re-checking the literature is only necessary because of the delay between starting and finishing the review, and would therefore be unnecessary if the remainder of the review process could be performed quicker. Most previous work have focused on screening automation, and almost exclusively on automating the title and abstract screening. There is comparatively less previous work on automating the later stages of the review process, including the article retrieval, article screening, data extraction, data synthesis, and the analysis stages. Furthermore, the work that does exist have largely focused on systematic reviews of interventions, with no previous work on diagnostic systematic reviews.

Previous work on systematic review automation in DTA systematic reviews have focused exclusively on screening automation (Kanoulas et al., 2017b, 2018). Datasets are thus available for training screening automation methods, but no such datasets are available describing any other review stage. The purpose of this part of the thesis is on the one hand to partially fill some of these gaps, by creating a dataset describing the data that passes through the data extraction and synthesis stages in DTA systematic reviews taken from Cochrane Library. We will also experiment with automating those parts of the process that are amenable to automation.

Data Item	Published Method
Total number of participants	[33; 82; 87; 140; 141; 144; 153; 154; 169; 174; 191; 312; 330; 332]
Age	[169; 191; 332; 333]
Sex	[169; 332; 333]
Country	[191; 332]
Co-morbidity	[255]
Spectrum of presenting symptoms, current treatments, recruitment centers	[87; 169; 255; 330; 332; 333]
Ethnicity	[333]
Date of study	[191]
Total number of intervention groups	[287; 288]
Specific intervention	[33; 34; 51; 82; 87; 144; 153; 154; 174; 191; 287; 312; 332]
Intervention description and details	[175]
Outcomes and time points (collected and reported)	[33; 34; 82; 87; 144; 151; 153; 154; 174; 175; 287; 288; 312; 332]
Comparison	[33; 51; 82; 141]
Sample size	[34; 175]
Overall evidence	[81; 280]
Generalizability/external validity	[151]
Research questions and hypotheses	[151; 332]
Study design	[144; 174; 312; 332]
Total study duration	[34; 82; 169]
Sequence generation	[201]
Allocation sequence concealment	[201]
Blinding	[201]
Random sequence allocation	[151]
Participant flow	[34; 175; 252]
Key conclusions of the study authors	[280]

Table 14.1 – List of previous methods for automated or semi-automated data extraction. Adapted from Jonnalagadda et al. (2015)

#### 14.1 AUTOMATED DATA EXTRACTION

Data extraction in the context of a systematic review refers to the identification of key characteristics of included primary studies, such as the methods used to perform the study, and the condition or population targeted (Li et al., 2019, in Higgins et al., 2019), but also involves producing assessments of the methodological quality of the included studies (Reitsma et al., 2008, in Deeks et al., 2013a). ‘Data extraction’ may be a misnomer – identifying these characteristics may be a matter of judgment, and may require domain knowledge and data from external sources. The data extraction stage is one of the more time-consuming stages of the systematic review process (Pham et al., 2018).

Prior to our publication of the first paper in this section (Norman et al., 2019e), there had been no previous work on data extraction systems applicable to diagnostic test accuracy studies or systematic reviews, and no published datasets.

There are a number of systems for automated data extraction, but these have focused almost exclusively on systematic reviews of interventions. Unfortunately,

Sentence
'A total of 59 patients were included in the study. [...] THIRTY-SEVEN patients underwent staging laparoscopy while 22 proceeded directly to laparotomy.'
'89 patients with primary solid abdominal tumors were eligible for evaluation; of those 49 patients had a gastric cancer, 33 patients a pancreatic cancer and seven an adenocarcinoma of the esophagus.'
'127 patients with primary solid abdominal tumors were eligible for evaluation; of those 66 patients had a gastric cancer and 61 a pancreatic cancer.'
'The inclusion criteria were met by 205 patients. Of these 131 patients underwent a staging laparoscopy detecting metastases in 21 patients.'
'ONE HUNDRED FORTY-FOUR patients with radiologically resectable nonpancreatic adenocarcinoma, periampullary tumors were identified from a prospective database between August 1993 and December 2000.'
'Over a 4-year period, 25 patients with potentially resectable tumors and 33 patients with LAPC were staged with laparoscopy, with an equivalent prevalence of occult metastases found at laparoscopy (28% potentially resectable vs. 33% LAPC, P=0.8).'
'A total of 114 patients with pancreatic cancer and no evidence of metastatic disease by computed tomography underwent laparoscopy.'
'A cohort of 40 consecutive patients referred to a tertiary referral center and with a diagnosis of potentially resectable pancreatic or periampullary cancer underwent staging laparoscopy with laparoscopic ultrasonography.'
'Staging laparoscopy was performed in 16 of the patients, and 2-ICA was used to treat three of 16 because they were found to have small liver metastases during staging laparoscopy.'

Table 14.2 – Example sentences describing the number of participants in nine studies examining the diagnostic accuracy of laparoscopy following computed tomography scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer (Allen et al., 2013).

data extraction methods trained on RCTs are unlikely to be relevant for DTA studies. First, RCTs and DTA studies typically need to extract different data items. Second, even where the same data is ostensibly extracted in both, the ‘same’ data may mean different things, or there may be marked differences in the language used to describe the data in the primary studies.

A small number of the items extracted in a systematic review of diagnostic test accuracy are not domain specific.

Determining article language can be done using standard language identification methods, and does not require methods specialized for diagnostic test accuracy studies. Standard language identification methods achieve close to perfect accuracy for large text samples and are available for several hundred languages (Jauhi-

'A total of 82 patients (19–23/group) were recruited.'
'Between Jan 2, 2013, and Nov 27, 2014, we enrolled 81 participants.'
'104 individuals who had low density parasitaemia at screening were randomized and treated during the dry season.'
'We randomly allocated 468 participants to receive artemether-lumefantrine combined with placebo (119 children) or with 0.1 mg/kg (116), 0.4 mg/kg (116), or 0.75 mg/kg (117) primaquine base.'
'In a randomized, partial blind study, 90 hospitalized adults with Plasmodium falciparum malaria that was blood schizonticide-responsive and a gametocytemia of > 55/μl within 3 days of diagnosis were randomized to receive single doses of either PQ 45 mg or BQ 75 mg on day 4.'
'A total of 93 male patients were enrolled.'
'In this randomised, double-blind, placebo-controlled trial, 360 asymptomatic parasitaemic children aged 2–15 years were enrolled and assigned to receive: artemether-lumefantrine (AL) and a dose of placebo; AL and a 0.25 mg/kg primaquine dose; or AL and a 0.40 mg/kg primaquine dose.'
'SIXTY-NINE of these G6PD-deficient patients were randomly allocated to one of three treatment regimes with (a) chloroquine, (b) chloroquine and primaquine or (c) sulfadoxine-pyrimethamine (Fansidar).'
'Among 124 parasitaemic persons identified during mass blood screening and passive case detection from outpatient clinics, 117 were enrolled and randomized to one of the 4 treatment regimens.'
'Prior to trial halt for poor DHP treatment efficacy, 101 participants were randomized and 50 received primaquine.'

Table 14.3 – Example sentences describing the number of participants in ten studies examining the effectiveness of different 8-aminoquinolines for reducing Plasmodium falciparum transmission (Graves et al., 2018)

ainen et al., 2017).

The key conclusions of the primary studies are commonly extracted during Cochrane DTA systematic reviews. This has been addressed by Song et al. (2013), who automatically extracted study-type agnostic data including the key conclusions by the authors. Since they appear to have used diagnostic studies among its training data, their system may be relevant in DTA systematic reviews.

A number of data items are ostensibly extracted both for intervention reviews and for diagnostics, i.e. *blinding, number of participants, country, age, gender, and selection criteria*. These have been addressed by previous literature for systematic reviews of interventions. Unfortunately, the 'same' data items in interventions and diagnostics often refer to different things. Furthermore, even where the data may mean the same thing conceptually, the data may be expressed using different

language.

For instance, both intervention studies and a diagnostic studies have target conditions, but while the intervention study seeks to *treat* the condition, the diagnostic study seeks to *diagnose* it. Thus, the two are semantically different, and they will be expressed using different language. In an intervention study the target condition may be the name of the disease, while in a diagnostic study the target condition may be a description of the symptoms or the presence of pathogens rather than the name of the disease.

Similarly, in an RCT, the participants are often described as being ‘recruited’, ‘enrolled’, or ‘randomized’ to ‘receive’ the relevant interventions. In a diagnostic test accuracy study the participants are instead described as ‘undergoing’ the relevant index test or on whom the index test was ‘performed’. Furthermore, the patient flow is in the vast majority of cases different in RCTs and DTA studies. An RCT has (at least) two treatment arms, one being the control group. In a DTA study, all participants (optimally) undergo the same reference standard, and there is no equivalent to the control group. Consequently, the structure of the language used to describe the populations in RCTs and DTA studies will be markedly different (tables 14.2 and 14.3).

Some data items, such as study country, patient age, and patient gender may use more similar language in different study types. In the absence of evaluations performed on data from different systematic review types however, the extent to which this is true remains unclear.

#### 14.2 AUTOMATED DATA HOMOGENIZATION

In practice, even if individual studies report the same data items, they are likely to report these using different metrics, units, or formats. For instance, the ratio of male and female patients enrolled in a study can either be stated in absolute numbers (e.g. ‘14 female; 11 male’) or as a percentage (e.g. ‘56% female’).

In order to compare and analyze multiple studies, the extracted data needs to be homogenized into a consistent format. There are multiple hurdles. One is that it may not be possible to use canonical formats for all data items. For instance, a  $2 \times 2$  table can always be translated into sensitivity and specificity, but the reverse may not be possible.

However, systematic reviews typically report only the data after being synthesized, not the raw data that was reported in each primary study. Thus, by looking at systematic reviews, we can only learn what the preferred formats for data collection are, not how the data items are actually reported in primary studies.

We will not attempt to automate on this step in this thesis, since it is not a natural language processing problem. This step must however be adequately automated be-

fore a fully automated data extraction system can be integrated with an automated meta analysis system.

### 14.3 AUTOMATED DATA SYNTHESIS

Once data have been converted to a common format, the data from the different studies will be pooled together. However, first the studies will be compared to determine which are similar enough to be grouped into a separate comparison. It has been suggested that this grouping could be performed using clustering. However, determining the subgroup for meta analyses should be decided at the protocol stage of the review, rather than after data have been extracted (McKenzie et al., 2019a,b; Thomas et al., 2019, in Higgins et al., 2019) Furthermore, determining which 2x2 table is included in which analysis is neither associated with a substantial workload nor is it repetitive and mechanical work amenable to automation. Thus there are likely limited gains from automating the procedure, even if clustering were to perform similar to human screeners and such post-hoc decisions would not introduce additional bias.

This does not mean that software tools cannot assist the process. Such tools could help tabulated and compare data across studies, or partition the data based on criteria determined by the reviewers. However, the actual decision making is likely better left solely to the review authors for the foreseeable future.

### 14.4 AUTOMATED META-ANALYSIS

Using statistical methods to quantitatively combine data from multiple studies is known as a meta-analysis. A meta-analysis gives the ability to improve the precision, answer questions not posed by individual studies, and to settle conflicting findings in different studies (Deeks et al., 2019, in Higgins et al., 2019). In systematic reviews where a meta-analysis is not appropriate, the meta-analysis is typically replaced by a qualitative analysis of the included studies.

There are a number of statistical methods to perform meta-analysis, both for systematic reviews of interventions (Deeks et al., 2019, in Higgins et al., 2019), as well as for diagnostics (Macaskill et al., 2010, in Deeks et al., 2013a). The design and implementation of statistical methods are complicated – and sometimes a matter of debate – beyond the scope of this thesis (Deeks et al., 2019, in Higgins et al., 2019). Either the hierarchical summary ROC (HSROC) method (Rutter and Gatsonis, 2001) or the Bivariate method (Reitsma et al., 2005) are recommended for DTA systematic reviews (Deeks et al., 2019, in Higgins et al., 2019; Leeftang et al., 2008). These methods model both variation due to random effects within the studies, as well as heterogeneity between different studies (Leeftang et al., 2008).

Specialized software to calculate meta-analysis results are available in a number of software packages, several of which are in widespread use (Tsafnat et al., 2014). For systematic reviews of interventions, meta-analyses can be performed within RevMan using a number of different methods (Deeks et al., 2019, in Higgins et al., 2019). For DTA systematic reviews, both the HSROC and the Bivariate method must be performed using external software, since fitting these models requires methods too complex to implement within RevMan (Macaskill et al., 2010, in Deeks et al., 2013a). Several software packages are available, including the SAS NLMixed procedure,<sup>1</sup> the Stata xtmelogit or meqrlogit routines,<sup>2</sup> or the reitsma function from the mada R package (Doebler and Holling, 2015). The meta-analysis process thus requires some amount of manual work – exporting the data from RevMan to this external software, and then importing the meta-analysis results back into RevMan. Transferring data between software packages is generally not the most time-consuming parts of the systematic review process. However, the process may still involve considerable work in some reviews, and process may be error prone (Tsafnat et al., 2014). Better integration between review tools and statistical software therefore have to potential to save time and reduce errors (Tsafnat et al., 2014).

#### 14.5 OBJECTIVES

In this section we present two conference papers, published 2018–2019, where we attempt to address the following research questions:

- RQ 6 *How are the current data extraction, data synthesis, and meta-analysis stages of DTA systematic reviews performed by human authors?*
- RQ 7 *Can we extract important study characteristics automatically from primary DTA studies?*

---

<sup>1</sup> <https://support.sas.com/documentation/onlinedoc/stat/141/nlmixed.pdf>

<sup>2</sup> <https://www.stata.com/help.cgi?xtmelogit>



The material in chapter 15 has been published as:

([Norman et al., 2018a](#)): Norman, C., Leeflang, M., and Névéol, A. (2018a). Data extraction and synthesis in systematic reviews of diagnostic test accuracy: A corpus for automating and evaluating the process. In *AMIA Annual Symposium Proceedings*, volume 2018, page 817. American Medical Informatics Association

In this study we attempted to document the current process in the later stages of DTA systematic reviews. Our previous aim was to collect training data for us to be able to perform the (later) studies presented in chapters 12 and 16. We were also interested in formalizing the human process, so that our meta-analysis and extraction methods would conform to the accepted practice.

Identifying errors in the human process was not one of the aims of the study.

The underlying research question in this study was:

RQ 6 *How are the current data extraction, data synthesis, and meta-analysis stages of DTA systematic reviews performed by human authors?*

#### AUTHOR'S CONTRIBUTIONS

CN wrote the first draft annotated the data and conducted the experiments.  
All authors conceived and designed the study. All authors read and approved the final manuscript.

DATA EXTRACTION AND SYNTHESIS SYSTEMATIC REVIEWS OF  
DIAGNOSTIC TEST ACCURACY: A CORPUS FOR AUTOMATING  
AND EVALUATING THE PROCESS

Christopher R. Norman, Mariska M.G. Leeflang, &amp; Aurélie Névéol

Proceedings of the American Medical Informatics Association's Annual  
Symposium 2018

205

**Abstract**

**BACKGROUND:** Systematic reviews are critical for obtaining accurate estimates of diagnostic test accuracy, yet these require extracting information buried in free text articles, an often laborious process.

**OBJECTIVE:** We create a dataset describing the data extraction and synthesis processes in 63 DTA systematic reviews, and demonstrate its utility by using it to replicate the data synthesis in the original reviews.

**METHOD:** We construct our dataset using a custom automated extraction pipeline complemented with manual extraction, verification, and post-editing. We evaluate using manual assessment by two annotators and by comparing against data extracted from source files.

**RESULTS:** The constructed dataset contains 5,848 test results for 1,354 diagnostic tests from 1,738 diagnostic studies. We observe an extraction error rate of 0.06–0.3%.

**CONCLUSIONS:** This constitutes the first dataset describing the later stages of the DTA systematic review process, and is intended to be useful for automating or evaluating the process.

## 15.1 INTRODUCTION

Accurate estimates of diagnostic test accuracy (DTA) are critical for deciding what tests to recommend or use, as well as for interpreting test results, and is therefore important to clinicians and policy makers, as well as to individual patients. Diagnostic test accuracy results are usually reported independently in several small studies. In order to achieve accurate and generalizable estimates, we typically need

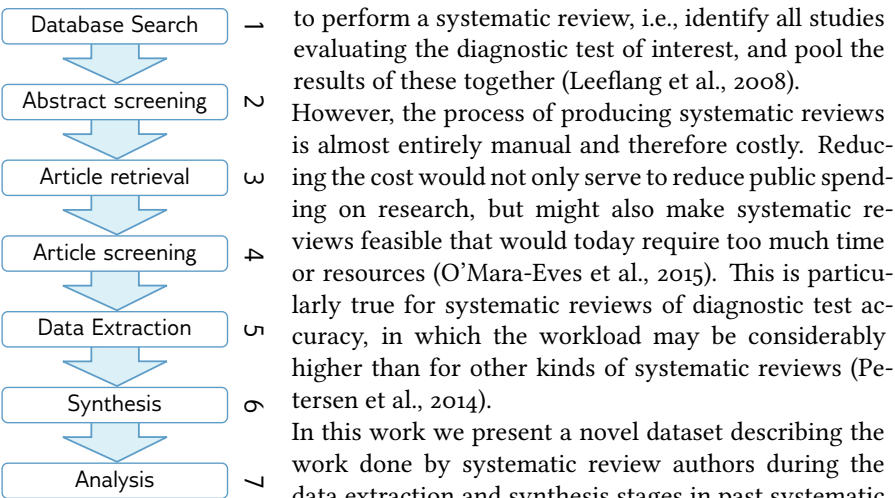


Figure 15.1 – Overview of the systematic review process. Simplified from Tsafnat et al. (Tsafnat et al., 2014).

15.1.1 Automated Data Extraction and Synthesis

Systematic reviews are conducted in a multistep process as illustrated in Figure 15.1, each step following a systematic and highly controlled procedure to ensure close to perfect recall, and a minimum of mistakes. The data extracted from the identified studies is pooled and synthesized, and the conclusions of the review are based on this synthesis. In a DTA systematic review, these results typically come in the form of a summary score, the mean sensitivity and specificity<sup>1</sup> estimated from the synthesized data. The high recall requirements, as well the high stakes associated with errors means that all of the stages are conventionally performed manually.

Most steps in the process are costly and could benefit from assistance through technical means, (Tsafnat et al., 2014) but previous work has focused on reducing the workload mainly in the article selection process (O'Mara-Eves et al., 2015), i.e.

<sup>1</sup> Defined as the number of true positives divided by the total number of positives and the number of true negatives divided by the total number of negatives respectively.

in stage 2 in Fig. 15-1. Less work has been done towards reducing the workload in the article retrieval, article screening, data extraction, data synthesis, and the analysis steps (3–7 in figure 15-1), even though these too are laborious and still entirely manual processes.

Consequently, datasets exist describing the abstract screening stage in DTA systematic review, or describing the data extraction in other domains, and these have been used for work towards automating these processes. We are aware of no datasets describing stages 3–7 for DTA systematic reviews. In this paper we aim to partially fill this gap, by presenting a corpus describing the processes performed by human authors in the data extraction and synthesis stages in DTA systematic reviews. This is intended to be useful for eventually automating the process, but also for reasoning about the work done by human systematic review authors.

207

### 15.1.2 *Systematic Review Reproducibility*

Multiple levels of reproducibility of research have been proposed (Cohen et al., 2018; Goodman et al., 2016), and exact definitions may differ. Here we use the following definitions: we *reproduce* research by redoing experiments in the same setting, we *replicate* research by redoing the analysis on the reported data, and we *repeat* research by retracing exactly the steps of published results.

Research reproducibility has been receiving increasing amount of attention by the research community in the last few decades (Baker, 2016; Collberg and Proebsting, 2016). In practice however, it is often difficult to even replicate the results of a paper, that is to say, to use the data presented to redo the analysis. In the setting of a systematic review on diagnostic test accuracy, provided the data used to calculate the summary scores are reported, we should be able to redo the calculations to yield the same results. To demonstrate the usefulness of our newly created dataset, we will test to what extent this is possible for the systematic reviews in the Cochrane Library, by replicating the calculation of the key summary scores reported in each systematic review.

### 15.1.3 *Related Work*

#### *Datasets Describing the Systematic Review Process*

Datasets have been published describing the database search and abstract screening steps in the systematic review process, both for DTA systematic reviews and for other topics. For instance, the included studies as well as the database search queries from 50 of the systematic reviews on DTA in the Cochrane Library have previously been published in one of the CLEF eHealth shared tasks (Kanoulas et al.,

2017b, 2018), thus addressing stages 1 and 2 in Figure 15-1. Similarly, Cohen has previously published a dataset describing the included and excluded studies in 15 systematic reviews on drug class efficacy (Cohen et al., 2006), thus addressing stage 2 in Figure 15-1, albeit in a different domain.

Datasets addressing the data extraction stage do not exist for diagnostic test accuracy, but exist for other domains, such as the PIBOSO corpus (Kim et al., 2011). Work has also been done on automatically extracting PICO<sup>1</sup> statements (Kiritchenko et al., 2010; Wallace et al., 2016), as well as other clinical trial information from article full text (Kiritchenko et al., 2010).

In order to extract data from DTA studies automatically using supervised machine learning, we need labeled gold standard datasets describing what data was extracted from each primary study, i.e. the data extraction forms in each systematic review. Such a dataset targeting systematic reviews of diagnostic test accuracy should include data extraction forms for the data necessary to perform the systematic review analysis, such as the index test, reference standard, target condition, and the 2 × 2 tables,<sup>2</sup> preferable with an emphasis on those items most difficult to extract manually. We are aware of no such datasets in current literature.

### *Systematic Review Replication*

Replication in science has been the focus of an increasing amount of discussion recently (Baker, 2016). However, we are not aware of work on replicating systematic reviews.

## 15-2 OBJECTIVES

We extract and reconstruct the reported data from each open-access or free systematic review in the diagnostic test accuracy section in the Cochrane Library,<sup>3</sup> with the following goals in mind:

1. **DATA EXTRACTION FORM CORPUS FOR DTA SYSTEMATIC REVIEWS:** We create a dataset by collecting the data extraction forms, the summary scores, and the included and excluded articles from each systematic review in the Cochrane Library. The dataset is intended to describe the work being done by human screeners in the data extraction and synthesis stages of a DTA systematic review, by documenting the input and output of these stages in past reviews.

<sup>1</sup> Population, intervention, control group, and outcome.

<sup>2</sup> The true/false positives and the true/false negatives for the test results, roughly equivalent to a confusion matrix in computer science.

<sup>3</sup> <https://www.cochranelibrary.com/cdsr/reviews/topics>

2. SUMMARY SCORE REPLICATION: We demonstrate the usefulness of our corpus by replicating the summary scores reported for the main tests in each systematic review. Our aim is to identify obstacles to replication, and to measure the discrepancies between our calculated summary scores and those reported.

### 15.3 MATERIAL AND METHODS

Our raw data consists of the systematic reviews on diagnostic test accuracy published in the Cochrane Library.<sup>1</sup> The Cochrane Library is the repository for the systematic reviews conducted under the auspices of Cochrane, one of the leading organizations for systematic reviews worldwide, which imposes more rigorous standards on systematic reviews than do paper-based journals (Jadad et al., 1998). We downloaded a snapshot of the systematic reviews on DTA in October 2017 to keep the data consistent during processing. For four of the reviews, we were able to obtain the XML source files used to compile the published systematic reviews from the authors, and we use these to evaluate the extraction quality. We do not have access to the source files for the other 59 systematic reviews, and for these we need to extract the data from the published articles.

#### 15.3.1 *The Cochrane DTA Data Form Corpus*

We construct our dataset from the published review articles, which come in a mixture of free text, formatted as HTML, and data tables, formatted as PNG images. Our contribution consists not only of extracting these elements piecemeal, but also in linking the elements together. We use automated extraction methods where possible, and complement these with manual extraction, verification and post-editing. Figure 15.2a presents the results of text extraction for a sample systematic review (Wijedoru et al., 2017): the list of included studies (left column), the list of diagnostic tests (right column) and which study evaluated which tests (links). Figure 15.2b presents a sample  $2 \times 2$  image table from the same review. Each row in the table describe the results on the same test (test 5) performed independently in six studies, and so each row corresponds to a link in Figure 15.2a.

Two annotators (CN and AN) manually post-edited the data from one systematic review sampled randomly. Based on this evaluation, we decided which parts of the automated extraction requires manual verification and post-editing, and which parts can be extracted automatically.

To assess the output quality, we also compare the post-edited data against the source files where available.

---

<sup>1</sup> [www.cochranelibrary.com](http://www.cochranelibrary.com)

**HTML PROCESSING** The HTML contents are processed using the LXML Python package.<sup>1</sup> Our system parses the HTML articles, and locates each section in the article using HTML ID and class attributes, as well as headers, and extracts the text contents.

**IMAGE PROCESSING** The diagnostic test results are only presented in images, and therefore requires optical character recognition (OCR) for extraction. In our extraction system, the images are first passed through a preprocessing stage where the images are scaled to roughly double the original size and antialiased. The data is then extracted using Tesseract.<sup>2</sup> We also use domain knowledge to correct mistakes, and flag possible errors for manual verification.

For each systematic review, we collect references to the *included and excluded studies*, i.e. the output of stage 4 in the systematic review process (Fig. 15.1). We collect these automatically using HTML processing.

For each primary study we collect the *data extraction forms* reported in the systematic review, i.e. the data extracted from each included primary study. We also collect the *data tables*, documenting the test results for each test, i.e. the  $2 \times 2$  tables, sensitivity and specificity along with their 95% confidence intervals (see Fig. 15.2b). Together, these constitute the output of stage 5 in the systematic review process (Fig. 15.1) for diagnostic test accuracy. We do this using HTML processing to locate the data tables, and image processing to extract the table contents.

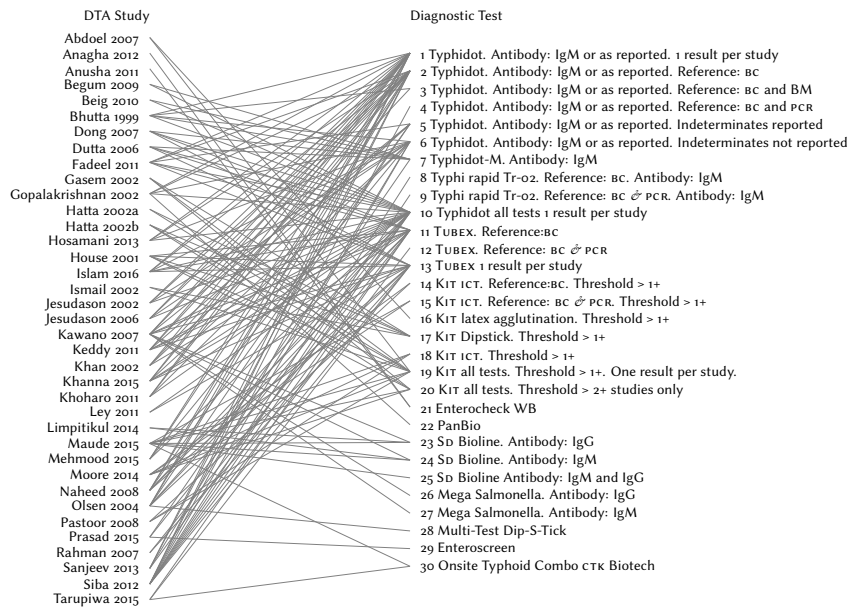
For each systematic review we collect the reported *summary scores* for each diagnostic test, i.e. the means estimated from the pooled test results. This constitutes the output of stage 6 in the systematic review process (Fig. 15.1) for diagnostic test accuracy. We do this manually by reading the summary of findings and locating the relevant data table matching the description in the text.

We also keep track of which test was performed by which study, which studies were included in which systematic review and which test results were used in which summary score calculation. These relations are not simple one-to-one relations, and in practice systematic reviews include sets of studies which may overlap with the included or excluded studies in other systematic reviews. The diagnostic tests performed by the studies within a systematic review generally overlap<sup>3</sup> (see Fig. 15.2a and 15.2b). A summary score pair should normally be connected to a single diagnostic test (but several test results), or two tests if it measures the relative performance of contrasted pairs of tests, but the summary scores are usually not reported for all diagnostic tests.

<sup>1</sup> <https://pypi.python.org/pypi/lxml>

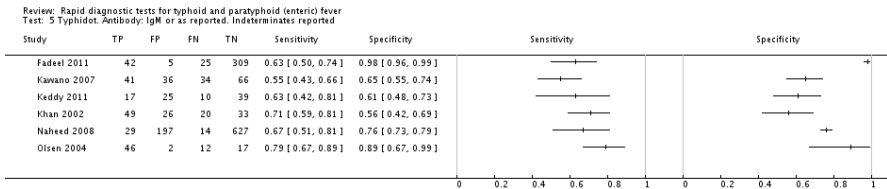
<sup>2</sup> <https://github.com/tesseract-ocr/tesseract>

<sup>3</sup> One of the primary objectives of a DTA systematic review is to pool the results of diagnostic tests performed by multiple primary studies together so this overlap is expected in a successful systematic review.



211

(a) The relations between the diagnostic tests and the studies included in the systematic review. Most of the tests are evaluated by several studies, and so are connected by several edges in the graph.



(b) The diagnostic test results, as reported in the systematic review, for test 5 in the graph above. Each row describes the test results reported in one study corresponds to a single edge in the graph above.

Figure 15.2 – Example of parts of the data in a systematic review on diagnostic test accuracy of Salmonella infection(Wijedoru et al., 2017), showing the relations between the data elements (a), and the source data from which the relations were extracted (b).

15.3.2 *Summary Score Replication*

We attempt to replicate the summary scores reported in each systematic review. We take note of which summary scores were reported with sufficient clarity that it would be possible to exactly repeat the original summary score calculations. However, in this work we do not attempt exact repetition, only replication using equivalent methods following Cochrane guidelines.

Summary scores should be calculated to account for the inter- and intrastudy variance (Macaskill et al., 2010, in (Deeks et al., 2013a)). There are multiple software packages available to perform these calculations, such as the SAS NLMixed procedure,<sup>1</sup> the Stata `xtmelogit` or `meqrlogit` routines,<sup>2</sup> or the `reitsma` function from the `mada` R package (Doeblér and Holling, 2015). Cochrane guidelines give no recommendation as to which software package to use (Macaskill et al., 2010, in (Deeks et al., 2013a)), and the choice consequently differs from review to review, as does the exact software version. Any of these choices is however considered valid, and should produce equivalent, albeit not necessarily identical results.

While the choice of software package and version is often reported in systematic reviews, it would be infeasible to repeat all systematic reviews using the exact same software package and version. Our intent is not however exactly repeating the original calculations, but replicating them using equivalent software, and we therefore use the same software package (`mada`) for all our trials.

## 15.4 RESULTS

From the 63 systematic reviews we extracted 5,848 test results together evaluating 1,354 unique diagnostic tests in 1,738 DTA studies (Table 15.1). We also extracted 589 reported summary score pairs, of which we replicate 103 and compare with the values reported in the systematic reviews.

15.4.1 *Dataset Construction*

Table 15.1 presents statistics of the dataset contents, which are publicly available.<sup>3</sup>

**EVALUATION BY MANUAL ANNOTATION BY TWO ANNOTATORS** Two independent annotators (AN and CN) manually verified and post-edited the included and excluded studies, the data tables, and the data forms automatically extracted from one systematic review. During the annotation, we highlighted the data elements

<sup>1</sup> <https://support.sas.com/documentation/onlinedoc/stat/141/nlmixed.pdf>

<sup>2</sup> <https://www.stata.com/help.cgi?xtmelogit>

<sup>3</sup> DOI: [10.5281/zenodo.1303259](https://doi.org/10.5281/zenodo.1303259)

Data	Extracted	Auto-corrected perc.		Evaluated			
				Manually perc.		Against source perc.	
Systematic reviews	63	—	—	1	1.6%	4	6.3%
Included studies	1,738	—	—	49	2.8%	203	11.7%
Data forms	1,356	—	—	49	3.6%	145	10.7%
Text entries	29,201	—	—	1,176	4.0%	3,281	11.2%
Excluded studies	6,699	—	—	132	2.0%	337	5.0%
Diagnostic tests	1,354	796	58.8%	43	3.2%	28	2.1%
Test results	5,848	1,981	33.9%	144	2.5%	330	5.6%
Study IDs	5,848	1,706	29.2%	144	2.5%	330	5.6%
Numerical	58,480	1,018	1.7%	1,440	2.5%	3,300	5.6%
Summary scores	589	—	—	—	—	—	—

Table 15.1 – The nature and amount of extracted data of the each type, along with the portions of auto-corrected and evaluated elements. We consider test results to be auto-corrected if they contain at least one auto-corrected value. We consider diagnostic tests to be auto-corrected if they contain at least one auto-corrected test result.

flagged by the automatic validation. We observed a 100% inter-annotator agreement on the corrected data. No errors were found in the extraction of the included and excluded studies, the data forms, or the numerical values from the tables. We did however find 3 errors among the study IDs extracted from the data tables. As these errors were all flagged by the automatic validation process, we decided that validation in subsequent reviews would be performed by one annotator (CN) focusing on the flagged data.

**EVALUATION BY COMPARING WITH THE SOURCE FILES** We compared our extracted data after post-editing against ground truth data from four source files. We observed 4 errors out of 330 study IDs (1.2%), of which 3 could be spotted by either checking whether the study IDs were in the list of included references, or by checking that the study IDs for each table appeared in alphabetical order. We thus observed 1 error out of 330 (0.3%) after sanity checking. We observed 2 errors out of 3,300 numerical values (0.06%), both of which could be spotted by sanity checking that the 2 by 2 table matches the sensitivity and specificity for each table row.

15.4.2 Summary Score Replication

Figure 15.3 presents the flow of reviews in the dataset according to replicability status. For the 103 of the 589 summary score pairs that could be meaningfully

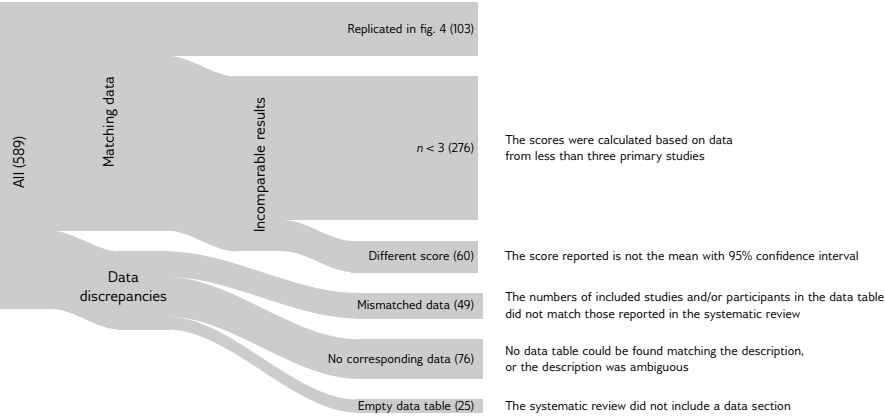


Figure 15.3 – Flow of the summary scores in the replication, describing how many summary scores were excluded for each reason in our replication attempt.

replicated, we plot the distribution of discrepancies in Figure 15.4, and we list the larger discrepancies in Table 15.2 (top section).

Twenty-five of the summary score pairs occurred in reviews with no data section. A further 76 of the summary scores descriptions did not clearly match any of the data tables in the data section. For 49 of the summary scores the number of studies or participants differed between the data tables and what was reported in the summary score description.

We excluded 60 summary scores from our replication attempts because they used measures other than the mean, such as the median, or the range. We also excluded 276 scores because they were based on less than three primary studies, and therefore can not be used to calculate reliable estimates (Macaskill et al., 2010, in (Deeks et al., 2013a)).

We replicated the remaining 103 summary scores pairs and plotted the difference in Fig 15.4. Of the 603 scalar values in the 103 summary scores,<sup>1</sup> we observed a 5 point difference in 79 / 603 (13%) of the scalars, and a 10 point difference in 25 / 603 (4%) scalars. In Table 15.2 (top section), we list the summary scores with at least one 10 point difference, roughly corresponding to the outliers in Fig. 15.4. We also list all the replicated summary scores for the summary scores where the number of studies or participants differ and had at least a 10 point difference in Table 15.2 (bottom section).

<sup>1</sup> A summary score is composed of 3 or 6 scalar values depending on whether both sensitivity and specificity are reported.

Test ID	Description	Replicated summary scores with 95% CI			Reported summary scores with 95% CI		
		n	Sensitivity (%)	Specificity (%)	n	Sensitivity (%)	Specificity (%)
Reported scores with matching number of studies and participants, at least one 10 point difference							
CD010705 37	Direct; SLID; cul...	8	71.6 [35.6, 92.0]	97.6 [90.0, 99.4]	8	87.0 [38.1, 98.6]	99.5 [93.6, 100]
CD012179 26	Anti-endometrial ...	4	80.0 [64.2, 90.0]	80.5 [60.0, 91.9]	4	81.0 [76.0, 87.0]	75.0 [46.0, 100]
CD012179 82	42.5. CA-19.9 (ca...	3	36.5 [29.7, 44.0]	90.2 [61.7, 98.1]	3	36.0 [26.0, 45.0]	87.0 [75.0, 99.0]
CD009591 21	Pelvic MRI	7	76.6 [64.0, 85.8]	69.2 [53.9, 81.2]	7	79.0 [70.0, 88.0]	72.0 [51.0, 92.0]
CD009591 7	Rvs TVUS	10	66.9 [40.3, 85.8]	97.4 [93.3, 99.0]	10	88.0 [82.0, 94.0]	100 [98.0, 100]
CD009591 30	Rvs MRI	3	76.7 [52.7, 90.7]	91.5 [56.9, 98.9]	3	81.0 [70.0, 93.0]	86.0 [78.0, 95.0]
CD009591 9	Vaginal TVUS	6	56.5 [31.6, 78.5]	97.3 [92.8, 99.0]	6	57.0 [21.0, 94.0]	99.0 [96.0, 100]
CD009591 31	Vaginal MRI	4	74.3 [61.1, 84.2]	93.2 [81.9, 97.6]	4	77.0 [67.0, 88.0]	97.0 [92.0, 100]
CD009591 33	POD MRI	5	86.2 [74.2, 93.1]	89.8 [70.5, 97.0]	5	90.0 [76.0, 100]	98.0 [89.0, 100]
CD009591 19	Rectosigmoid TRUS	4	89.5 [82.9, 93.8]	91.4 [78.5, 96.9]	4	91.0 [85.0, 98.0]	96.0 [91.0, 100]
CD009591 38	Rectosigmoid MDCT-e	3	95.6 [80.3, 99.1]	97.9 [92.7, 99.4]	3	98.0 [94.0, 100]	99.0 [97.0, 100]
CD010023 1	CT	4	72.7 [59.4, 82.9]	97.8 [94.3, 99.2]	4	72.0 [36.0, 92.0]	99.0 [71.0, 100]
CD010023 2	MRI	5	78.4 [61.8, 89.1]	96.2 [87.3, 98.9]	5	88.0 [64.0, 97.0]	100 [38.0, 100]
CD010023 3	Bs	6	95.3 [88.1, 98.2]	84.7 [69.8, 93.0]	6	99.0 [69.0, 100]	86.0 [73.0, 94.0]
CD009579 8	CCA poc haematobium	4	39.9 [12.3, 75.9]	77.5 [43.7, 93.9]	4	39.0 [6.0, 73.0]	78.0 [55.0, 100]
CD010653 2	Diagnosis of schi...	16	57.9 [50.1, 65.3]	73.8 [64.0, 81.7]	16	58.0 [50.3, 65.3]	74.7 [85.2, 82.3]
CD010079 7	IqCODE cut-off 3,6	3	74.4 [64.2, 82.4]	91.3 [83.9, 95.5]	3	78.0 [68.0, 86.0]	87.0 [71.0, 95.0]
CD008054 14	Triage of LSIL wi...	4	95.9 [90.6, 98.3]	22.4 [16.7, 29.5]	4	97.5 [69.6, 99.8]	24.8 [7.3, 58.1]
CD008054 18	Triage of LSIL wi...	4	80.4 [63.5, 90.6]	48.9 [22.8, 75.7]	4	84.6 [48.6, 97.0]	44.4 [16.0, 76.9]
Reported scores with mismatched number of studies and participants, at least one 10 point difference							
CD012165 26	PGP 9.5 (protein ...	8	88.7 [67.4, 96.7]	76.9 [69.2, 83.1]	7	96.0 [91.0, 100]	86.0 [70.0, 100]
CD009591 28	USL MRI	4	85.4 [78.2, 90.5]	81.6 [51.7, 94.8]	4	86.0 [80.0, 92.0]	84.0 [68.0, 100]
CD007394 2	Children	7	44.2 [15.1, 77.8]	96.4 [92.7, 98.3]	6	84.0 [66.0, 93.0]	88.0 [60.0, 97.0]
CD011975 5	Total hCG	2	7.4 [2.7, 18.3]	95.0 [94.1, 95.8]	3	19.0 [4.0, 58.0]	Assumed 95 %
CD008803 40	OCT: ONH horizont...	12	53.7 [40.8, 66.1]	88.9 [84.0, 92.4]	6	41.0 [26.0, 58.0]	94.0 [90.0, 96.0]
CD008803 34	OCT: ONH Cup area	15	52.6 [38.9, 65.9]	88.3 [84.0, 91.5]	9	45.0 [26.0, 67.0]	92.0 [87.0, 95.0]
CD008803 38	OCT: ONH Cup volume	17	43.6 [29.8, 58.3]	89.3 [85.3, 92.3]	9	30.0 [16.0, 49.0]	94.0 [92.0, 96.0]
CD008803 37	OCT: ONH Nerve he...	8	52.8 [40.8, 64.5]	88.3 [81.9, 92.6]	4	44.0 [28.0, 62.0]	93.0 [87.0, 96.0]
CD008803 36	OCT: ONH Rim volume	11	56.6 [45.7, 66.9]	89.7 [84.7, 93.2]	6	49.0 [35.0, 62.0]	95.0 [92.0, 96.0]
CD008081 2	OCT for detection...	3	80.3 [72.3, 86.4]	82.6 [71.9, 89.8]	3	74.0 [68.0, 86.0]	92.0 [87.0, 97.0]



Table 15.2 – Replicated vs reported summary score pairs differing from the reported summary scores by at least 10 point on one of the six scalar values (one cell per row in the table). Larger differences are highlighted.

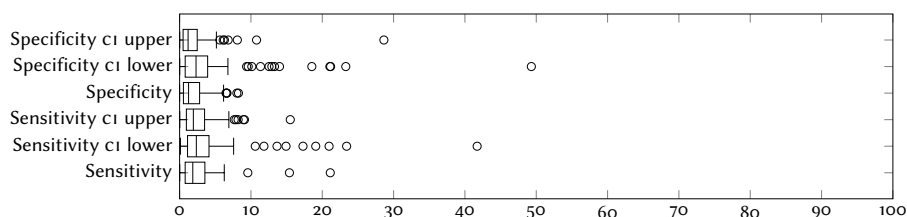


Figure 15.4 – Distribution of differences between reported and replicated summary scores for each of the six scalar values in the summary scores. All differences are in absolute point difference.

216

## 15.5 DISCUSSION

In this section, we discuss the data extraction and provide insight on the growth and possible uses of the dataset. We then discuss the findings and implications of our replication study in Section 15.5.2

### 15.5.1 Dataset Construction

Our manual extraction by two annotators on the automatically extracted parts of the data had a 100% inter-annotator agreement. Furthermore, we only observe errors for data extracted from the data tables, using OCR, and all of the errors were flagged for inspection by the automated extraction. In light of this, we content ourselves with letting a single annotator check and post-edit the extracted data tables for the remainder of the dataset, and do not verify the extraction of the lists of included and excluded studies or the data extraction forms.

Manual post-editing still lets the occasional error through, as we can see from the results in Section 15.4.1, but these can generally be spotted by automatic sanity checking.

The amount of manual effort required for to process a single review varies, from a few minutes to several hours for a single review. The effort required for verification and post-editing chiefly depends on the success rate of the automated extraction of the data tables, and the effort required to manually extract the summary scores chiefly depends the clarity of presentation in the systematic review. The amount of data in the review also plays a role, but only in the presence of automatic extraction failures, and unclear presentation.

### 15.5.2 Summary Score Replication

As we can see in Figure 15.4, our replicated summary scores are generally close to the summary scores reported in the original systematic reviews, but we also

observe a large number of discrepancies. The discrepancies tend to be larger and more common for the lower bound of the confidence intervals (Figure 15.4).

When there is a mismatch in terms of number of studies or participants between the summary scores and the data tables, this is typically deliberate, and the authors often state the reason why some of the studies or participants were excluded from the calculations. In order to replicate such summary scores, it may be necessary to modify the data tables manually. Simply using the original unmodified tables can obviously give different summary scores, although this varies (Table 15.2, bottom section).

In some cases the reason for the mismatch is not stated, but may be due to different definitions of number of studies or participants in different parts of the systematic review.<sup>1</sup>

The mismatch for ‘CD011975 Total hCG’ in Figure 15.2 (bottom section) appears to be due to a copy-paste error however, and the results presented is identical to the results for ‘CD011975 Inhibin.’ The summary scores are apparently calculated from data table 11 rather than from data table 5. This summary score is not mentioned in the systematic review body however, and so does not appear to have influenced the findings of the review.

A large part of the manual work required to connect summary scores to data tables were due to their order often being different in the two sections. We therefore recommend that data tables and summary scores presentation be aligned in future systematic reviews to make it easier to verify the results, as well as catch errors. We also note that this would go far towards automating the synthesis step in the systematic review process for diagnostic test accuracy.

The frequency of genuine errors in the systematic reviews is low (1 in 63 reviews). However, the one error we did find could be spotted simply by verifying the numbers of studies and participants. Such verification could potentially be performed automatically, provided the summary of findings and the data tables are organized consistently, and a standard definition of number of included studies and participants are used throughout the systematic review.

### 15.5.3 Dataset Applications and Future Work

This data is intended to be used to 1) train and evaluate methods for automating the data extraction and data synthesis stages in DTA systematic reviews, 2) perform replication studies, like the one we describe here, and 3) perform robustness studies by e.g. evaluating how the results of the analysis would differ on different subsets of the data (subset analysis or ablation studies).

<sup>1</sup> A paper can contain multiple studies, and for instance a diagnostic test for glaucoma may count individual eyes as participants.

In particular, this dataset contains all the information that needs to be extracted from the included studies in a systematic review on diagnostic test accuracy, and can therefore be used towards training supervised machine learning models to automate this process in future reviews.

## 15.6 CONCLUSION

In this paper, we presented a dataset describing the input and output of the data extraction and synthesis stages in systematic reviews on diagnostic test accuracy. The data extraction was manually validated and found successful with an error rate of 6 in 10,000 for the numerical values, 0.3% for the study ids, and no observed errors on the other data types. This is the first dataset to provide material for addressing automation of the later stages of the DTA systematic review process, including the data extraction and synthesis.

We demonstrate the value of this dataset by conducting a replication study of 103 summary scores from the data synthesis in 27 of the systematic reviews. Overall, we were able to replicate the results reported in the systematic reviews with less than 5 point difference for 87% of the values. We did not try to replicate the interpretation of the results to test whether these differences would have affected the general conclusions reached in the reviews. Our findings mirror insights gained in other fields: it is often not straightforward to replicate reported results, even when these are reported clearly.

We believe that the availability of material presented here (data and tools) can be helpful for the community to leverage the information contained in systematic reviews to a fuller extent, for instance to make it easier to replicate or update the data synthesis and analysis in systematic reviews.

## ACKNOWLEDGMENTS

We are grateful to Clive Adams from the Cochrane Schizophrenia Group, who kindly provided us with the source RevMan file for one of the systematic reviews we used for evaluation.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676207.



The material in chapter 16 has been published as:

([Norman et al., 2019e](#)): Norman, C., Spijker, R., Kanoulas, E., Leeﬂang, M., and Névéol, A. (2019e). A distantly supervised dataset for automated data extraction from diagnostic studies. *ACL BioNLP*)

In this study we attempted to explore methods to extract data automatically from DTA reports. The target condition, index test and reference standard were identified early on as the primary data elements to extract, since these were felt to be those that would lead to the largest work savings.

The underlying research question in this study was:

RQ 7 *Can we extract important study characteristics automatically from primary DTA studies?*

#### AUTHOR’S CONTRIBUTIONS

CN wrote the first draft and conducted the experiments. Rs and ML annotated the gold standard data. CN performed the data cleaning. All authors conceived and designed the study. All authors read and approved the final manuscript.

**A distantly supervised dataset for automated data extraction  
from diagnostic studies**

Christopher R. Norman, Mariska M.G. Leeftang,  
René Spijker, Evangelos Kanoulas, & Aurélie Névéol

221

ACL BioNLP, 2019

**Abstract**

Systematic reviews are important in evidence based medicine, but are expensive to produce. Automating or semi-automating the data extraction of index test, target condition, and reference standard from articles has the potential to decrease the cost of conducting systematic reviews of diagnostic test accuracy, but relevant training data is not available. We create a distantly supervised dataset of approximately 90,000 sentences, and let two experts manually annotate a small subset of around 1,000 sentences for evaluation. We evaluate the performance of BioBERT and logistic regression for ranking the sentences, and compare the performance for distant and direct supervision. Our results suggest that distant supervision can work as well as, or better than direct supervision on this problem, and that distantly trained models can perform as well as, or better than human annotators.

## 16.1 BACKGROUND

Evidence based medicine is founded on systematic reviews, which synthesize all published evidence addressing a given research question. By examining multiple studies, a systematic review can examine the variation between different studies, the discrepancies between them, as well as look at the quality of evidence across studies in a way that is difficult in a single trial. Since a systematic review needs to consider the entire body of published literature, producing a systematic review is an expensive and labor-intensive process, often requiring months of manual work (O'Mara-Eves et al., 2015).

To ensure that the results of a systematic review are as comprehensive and unbiased as possible, their production follows a strict and systematic procedure. To

catch and resolve disagreements, all steps of the process are performed in duplicate by at least two reviewers. There have recently been examples of systematic reviews using automation in a limited capacity (Bannach-Brown et al., 2019; Lerner et al., 2019; Przybyła et al., 2018), but the impact of automation on the reliability of systematic reviews is not yet fully understood. Automation is not part of accepted practice in current guidelines (De Vet et al., 2008).

After a set of potentially included studies have been identified, systematic reviewers complete a so-called *data extraction form* for each study. These forms comprise a semi-structured summary of the studies, identifying and extracting a consistent, pre-specified set of data items from abstracts or full-text articles in a coherent format (see the left part of Table 16.1 for sample excerpts). The coherent format allows the data from the studies to be synthesized qualitatively or quantitatively to address the research question of the review.

In this study we will focus on systematic reviews of diagnostic test accuracy (DTA), which examine the accuracy of tests and procedures for diagnosing medical conditions, and which have seen little attention in previous literature on automated data extraction. To compare and synthesize results across studies, reviewers extract diagnostic accuracy from each study, but also determine the *index test* (the specific diagnostic test or procedure that is being tested), what *target condition* the test seeks to diagnose, and the *reference standard* (the diagnostic test or procedure that is being used as the gold standard) (see Fig 16.1 for an example). These data must be determined for each study to know if the diagnostic accuracy in different studies can be compared.

#### 16.1.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model that is unsupervisedly pretrained on a large general language corpus, then supervisedly fine-tuned on natural language processing tasks (Devlin et al., 2018). Despite being a general approach, with almost no task-specific modifications, BERT achieves state-of-the-art performance across a number of natural language processing tasks, including text classification, question answering, inference, and named entity recognition.

Pretrained models like BERT can be used directly for screening automation or automated data extraction. However, by default BERT is trained on a general language corpus, which differs radically in word choice and grammar from the special language found in biomedicine and related fields (Sager et al., 1980). Pretraining on biomedical corpora, rather than general corpora, has been demonstrated to improve performance on several biomedical natural language processing tasks (Beltagy et al., 2019; Lee et al., 2019; Si et al., 2019).

16. Automated data extraction for systematic reviews of diagnostic accuracy

Original		Cleaned
Review: CD008892, study: Dutta 2006		
Index tests:	TUBEX Typhidot	Index test: TUBEX Typhidot
Target condition and reference standard(s):	Target condition Salmonella Typhi Reference standard: peripheral blood culture	Target condition: Salmonella Typhi Target condition: Typhoid fever Reference standard: Peripheral blood culture
Note: These are the data items corresponding to the example text in Fig. 16-1		
Review: CD010502, study: Schwartz 1997b		
Index tests:	Throat swab: not reported Commercial name of the RADT: QuickVue In-Line Strep A (Quidel) Type of RADT: EIA	Index test: QuickVue In-Line Strep A Index test: EIA Index test: ELISA Immunoassays
Target condition and reference standard(s):	See Schwartz 1997a	Target condition: Group A streptococcus  Target condition: Group A streptococcal infection Reference standard: Microbial culture Reference standard: Bacterial culture
Note: Neither the target condition nor the reference standard were mentioned in the table for Schwartz 1997a, but assumed the same for all studies included in this systematic review (they were presumably considered obvious by the authors).		
Review: CD011145, study: ADAMS Study 2007		
Index tests:	MMSE, non-validated Spanish versions where necessary.	Index test: MMSE Index test: Minimal state evaluation
Target condition and reference standard(s):	Dementia diagnosed according to DSM-IV Participants consented to a 3-4 h structured assessment conducted in-home, including a medical examination with a nurse and a neuropsychological battery with a trained psychometrician A panel of 3 expert scientists, including a neurologist, cognitive neuroscientist, and geropsychiatrist determined the participants' initial DSM-IV cognitive status based on the in-home diagnostic evaluation, which assessed several cognitive domains The final cognitive status was made by a consensus panel of experts based on a review of the information collected through the neuropsychological, medical, and neurological assessment measures We assess this as meaning that not all 701 participants were clinically evaluated by a specialist	Target condition: Dementia  Reference standard: DSM-IV Reference standard: DSM IV Reference standard: Consensus evaluation
Note: DSM-IV is the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders published by the APA. The non-abbreviated form is almost never used and would just give false positives.		

223

Table 16.1 – Examples of raw data from three data extractions forms in unstructured format (left) and a structured summary of the data intended for distant supervision by pattern matching (right).

16.1.2 Objectives

In this study we seek to:

1. Construct a dataset for training machine learning models to identify and extract data from full-text articles on diagnostic test accuracy. We focus on the target condition, index test, and reference standard.
2. Train models to identify specific data items in full-text articles on diagnostic test accuracy

One of the main aims of our study is to determine how such a dataset should be constructed to allow for training well performing models. In particular, do we need

Although **typhoid fever** is confirmed by **culture** of **Salmonella enterica serotype Typhi**, rapid and simple diagnostic **serologic tests** would be useful in developing countries. We examined the performance of **Widal test** in a community field site and compared it with **Typhidot** and **Tubex tests** for diagnosis of **typhoid fever**. Blood samples were collected from 6697 patients with fever for  $\leq 3$  days for microscopy, culture, and serologic testing and from randomly selected 172 consenting healthy individuals to assess the baseline Widal anti-Typhi O lipopolysaccharide antibody (anti-TO) and anti-Typhi H flagellar antibody (anti-TH) titers. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the **3 serologic tests** were calculated using **culture-confirmed typhoid fever** cases as "true positives" and paratyphoid fever and malaria cases as "true negatives". Comparing cutoff values for the **Widal test**, an anti-TO titer of 1/80 was optimal with 58% sensitivity, 85% specificity, 69% PPV, and 77% NPV. Sensitivity was increased to 67% when the Widal test was done on the 5th day of illness and thereafter. The sensitivity, specificity, PPV, and NPV of **Typhidot** and **Tubex** were not better than **Widal test**. There is a need for more efficient rapid **diagnostic test for typhoid fever** especially during the acute stage of the disease. Until then, **culture** remains the method of choice.

Legend: **Target condition** **Index Test** **Reference standard**

Figure 16.1 – Examples of data items highlighted in text, with supporting context underlined. Based on the manual annotation by one expert (ML) on a study by Dutta et al. (2006).

directly supervised data, or can we build reliable models with distantly supervised data? If we do need directly supervised data, how much is necessary?

### 16.1.3 *Related Work*

There have been attempts to extract several types of data relevant to systematic reviews, most notably extracting PICO<sup>1</sup> statements from article text (Kim et al., 2011; Kiritchenko et al., 2010; Nye et al., 2018; Wallace et al., 2016). Other data items include background and study design (Kim et al., 2011), as well as automatically performing risk of bias assessments (Marshall et al., 2014). There is also a recent TAC track for data extraction in systematic reviews of environmental agents.<sup>2</sup> Similarly, previous work by Kiritchenko et al. (2010) aimed to extract 21 different kinds of data from articles, including treatment name, sample size, as well as the primary and secondary outcome from article text. Furthermore, the key criterion for

<sup>1</sup> Population, intervention, control group, and outcome.

<sup>2</sup> <https://tac.nist.gov/2018/SRIE/index.html>

extraction in a systematic review is not the actual data, but the context it appears in. For instance, both intervention studies and a diagnostic studies have target conditions, but these refer to different things: the intervention study seek to *treat* the condition while the diagnostic study seeks to *diagnose* it. As a consequence, in an intervention study the inclusion criterion often mentions the disease, while in a diagnostic study inclusion criteria may mention symptoms rather than the actual disease. This means that a data extraction system trained on interventions may not work as well (or at all) for systematic reviews of diagnostic test accuracy, even though it may seem that the same data is extracted in both. Furthermore, unlike the data required in diagnostic reviews, many previously considered data items are mentioned once in articles, often using formulaic expressions (e.g. sex, blinding, randomization).

Conventional methods for automated data extraction split articles into sentences and classify these individually using conventional machine learning methods (e.g. SVM, Naive Bayes) (Jonnalagadda et al., 2015), or label spans in the text and classify these using sequence tagging (e.g. CRF, LSTM) (Nye et al., 2018). Despite the body of previous work on automation, many data items relevant to systematic reviews have been overlooked. A 2015 systematic review of data extraction found 26 articles describing the attempted extraction of 52 different data items, but almost all focused on interventions (Jonnalagadda et al., 2015). No study considered any data item specific to diagnostic studies, except for general data items common to both interventions and diagnostic studies, such as age, sex, blinding, or the generation of random allocation sequences. The likely reason for this is that traditional data extraction systems require bespoke training data for each particular data item to extract, which is generally only available through expensive, manual annotation by experts. A cheaper way to construct datasets for data extraction is to use distant supervision, where the dataset is annotated per article or per review, rather than per sentence or

Target Condition			
	pos	neg	total
Distant train	11,336	63,204	74,540
test	2,884	13,572	16,456
total	14,220	77,776	90,996
Annotated by ML	92	889	981
Annotated by RS	48	983	1,031

Index Test			
	pos	neg	total
Distant train	14,280	63,343	77,623
test	2,675	13,992	16,667
total	16,955	77,335	94,290
Annotated by ML	93	888	981
Annotated by RS	87	944	1,031

Reference Standard			
	pos	neg	total
Distant train	7,006	56,638	63,644
test	1,258	14,602	15,860
total	8,264	71,240	79,504
Annotated by ML	26	955	981
Annotated by RS	26	1,005	1,031

Table 16.2 – The number of sentences in our dataset, broken into distantly annotated training and test sets, as well as a manually annotated subset. Distant annotations for each data type were not available for all studies, and the total number of labelled sentences are therefore different for each data type.

per text span. Supervised methods are then trained on fuzzy annotations derived heuristically for each sentence. For instance, Wallace et al. (2016) used supervised distant supervision to learn to identify PICO statements in full text, and Marshall et al. (2014) used supervised distant learning with SVMs to identify risk of bias assessments.

There is likely a trade-off between quality and data size. All else being equal, direct supervision is generally better than distant supervision (distantly supervised training data adds a source of noise not present for direct supervision). At the same time, it may not be feasible for experts to annotate large amounts of data. Crowd-sourcing is sometimes used as an alternative to a group of known experts, but if a high degree of expertise is necessary to annotate, crowd-sourcing may not give sufficient guarantees about the expertise of the annotators.

## 16.2 MATERIAL

We used data from a previous dataset, the LIMS1-Cochrane dataset (Norman et al., 2018a),<sup>1</sup> to identify references included in previous systematic reviews of diagnostic test accuracy. The LIMS1-Cochrane dataset comprises 1,738 references to DTA studies from 63 DTA systematic reviews. The dataset includes the data extraction forms for each study completed by the systematic review authors.

The dataset itself does not contain abstracts or full-texts, but include identifiers in the form of PubMed IDs and DOIs which can be used to retrieve abstracts or full-texts.

We used the reference identifiers (PMID and/or DOI) taken from the LIMS1-Cochrane dataset to construct a collection of PDF articles. We used EndNote's 'find full text' feature, which retrieves PDF articles from a range of publishers.<sup>2</sup> The PDF articles were then converted into XML format using Grobid (Lopez, 2009).

We randomly split the dataset into dedicated training and evaluation sets, where we used 48 of the systematic reviews as the training set, and we kept the remaining 15 systematic reviews for evaluation. For each of the 15 systematic reviews in the evaluation set, we randomly selected one article to be annotated manually. The remaining articles in the evaluation set were not used for training, since training and testing on the same systematic review is known to overestimate classification performance (Cohen, 2008). The goal of this work is to learn the semantics of the context, rather than the semantics of particular terms, and these contexts should be consistent across reviews.

---

<sup>1</sup> DOI: [10.5281/zenodo.1303259](https://doi.org/10.5281/zenodo.1303259)

<sup>2</sup> <https://endnote.com/>

*Distant annotation*

The data forms from the systematic reviews were intended to be read by and be useful to the human systematic review authors. The contents are therefore usually semi-structured rather than structured, and will include different kinds of data depending on what is relevant to the systematic review (see Table 16-1).

We create a dataset of distant annotations from the LIMS1-Cochrane dataset by manually converting the semi-structured data into structured data items, and by ensuring that these items can be found in the corresponding article using pattern matching (see Table 16-1).

We split each of the XML documents into sentences using the nltk sentence splitter.<sup>1</sup> The sentences are then divided into positive and negative depending on whether the relevant data items occur as a partial match in the sentence. Partial matches were calculated using *tfidf* cosine similarity between the data item and the sentence, where we took the 20 top ranking sentences for each pair of data item and article, with a similarity score of 0.1 or higher. We chose 20 as a target number of sentences since we felt this was a reasonable upper limit on the number of relevant sentences in a single article. We added an absolute threshold of 0.1 to keep the system from annotating obviously non-relevant sentences (scores close to zero) when no matches could be found in the article. For articles that have multiple data items we used the concatenation of all data items. For example, in Table 16-1, the data items for 'Schwartz 1997b' would be: target condition: 'Group A streptococcus; Group A streptococcal infection', index test: 'QuickVue In-Line Strep A; EIA; ELISA Immunoassays', and reference standard 'Microbial culture; Bacterial culture'.

We excluded all articles where the data items were not provided in the data form (because the reviewers did not extract this data), or where data forms were missing from the systematic review. Since we do not know which sentences were relevant or not in these articles we did not use these articles as either positive or negative data. As a consequence the total amount of sentences differ for the target condition, index test and reference standard.

We repeated the matching procedure for the target condition, the index test and the reference standard, resulting in three distinct datasets.

*Expert annotation*

We randomly split the evaluation set into three sets of five systematic reviews. Two experts (ML and RS) on systematic reviews of diagnostic test accuracy manually an-

<sup>1</sup> <https://www.nltk.org/>

Target condition				Index test				Reference standard			
	Auto	ML	RS		Auto	ML	RS		Auto	ML	RS
Auto	1.00	0.07	0.04	Auto	1.00	0.09	0.07	Auto	1.00	0.01	0.03
ML	0.90	1.00	0.38	ML	1.00	1.00	0.61	ML	1.00	1.00	0.86
RS	1.00	0.62	1.00	RS	0.93	0.70	1.00	RS	1.00	0.40	1.00

Table 16.3 – Agreement in terms of recall where columns are considered ground truth, e.g. annotator RS chose 62% of ML’s annotations for the target condition.

notated the 15 articles by highlighting all sentences in the text that 1) mentions the target condition, index test, and reference standard 2) makes it clear that these are the target condition, index test and reference standard, and 3) do not simply mention these same items in an unrelated context. The annotation instructions were written and adjusted twice to remove ambiguity, and the reasons for disagreement were discussed and resolved after two rounds of annotation. As a compromise between getting more data and being able to use the agreement between the experts as baseline for the performance, one expert annotated the first five studies, the second expert annotated the next five studies, and both annotated the last five studies.

16.3 METHOD

We construct three pipelines, one for each of the target condition, index test, and reference standard, and we train and evaluate these separately. We varied our experiments in three dimensions: We tried A) two machine learning algorithms, B) two levels of preprocessing, and C) distantly supervised training data versus directly supervised training data. The directly and distantly supervised models were evaluated on the same data.

A1: BioBERT We here used a pointwise learning-to-rank approach, where we trained a sentence ranking model by using BioBERT, a version of BERT pretrained on PubMed and PMC (Lee et al., 2019), and fine-tuned the model by training it to regress probability scores. This model was thus trained to map sentences to relevance scores. To train and evaluate, we used the default BERT setup for the GLUE datasets,<sup>1</sup> modified to output a relevance score rather than a binary value. We used default parameters.

<sup>1</sup> <https://github.com/google-research/bert>

**A2: LOGISTIC REGRESSION** We here used a pairwise learning-to-rank approach, where we trained a logistic regression model using stochastic gradient descent (sklearn). As features we used 1) lowercased, *tf-idf* weighted word *n*-grams, 2) lowercased, binary word *n*-grams, 3) lowercased, *tf-idf* weighted, stemmed word *n*-grams, 4) lowercased, stemmed, binary word *n*-grams, as well as i) lowercased, *tf-idf* weighted character *n*-grams, and ii) *non*-lowercased, *tf-idf* weighted character *n*-grams. We used word *n*-grams up to length 3, and character *n*-grams up to length 6. The first set of features is intended to capture contextual information ('for the diagnosis of ...'); the second set of features is intended to capture medical technical terms, which are often distinctive at the morpheme level (e.g. 'ischemia', 'anemia'). We deliberately did not use stop-words, since doing so would discard almost all the contextual information. This results in a sparse feature matrix consisting of approximately 1.8 million features for the distantly supervised experiments, and approximately 300,000 features for the directly supervised experiments. We handled class imbalance by setting the weight for the positive class to 80. This was previously determined to be a reasonable weight in experiments on screening automation in diagnostic test accuracy systematic reviews, a problem with similar class imbalance.

**B1: RAW SENTENCES** Here we used the sentences as they appear in the articles.

**B2: SENTENCES WITH UMLS CONCEPTS** In this setup we used the *Unified Medical Language System*, a large ontology of medical concepts maintained by the National Library of Medicine (Bodenreider, 2004; Lindberg et al., 1993). We used MetaMap<sup>1</sup> to locate concept mentions in the sentences, and to replace these with their corresponding UMLS semantic types. For instance the sentence '*Typhoid fever is a febrile and often serious systemic illness caused by Salmonella enterica serotype Typhi*' was transformed into '*DSYN is a FNDG and TMCO serious DSYN caused by BACT enterica BACT*'.

**C1: DIRECTLY SUPERVISED TRAINING** We here trained and evaluated on the articles manually annotated by our two experts (ML and RS), using leave-one-out cross-validation. In other words, to evaluate on each of the ten articles annotated by each annotator we used the remaining 9 articles annotated by the same expert as training data. This was done separately for each expert, and the annotations from the other expert was not used.

---

<sup>1</sup> <https://metamap.nlm.nih.gov/>

Target condition										
	<i>n</i> pos	BioBERT				Logistic Regression				As ranked by the other expert (rs)
		Distant		Supervised		Distant		Supervised		
		Raw	UMLS	Raw	UMLS	Raw	UMLS	Raw	UMLS	
CD007394	1	1.000	0.500	0.143	0.250	1.000	0.500	1.000	0.500	0.500
CD007427	14	0.228	0.267	0.500	0.588	0.423	0.573	0.462	0.509	—
CD008054	10	0.197	0.353	0.060	0.182	0.167	0.118	0.170	0.148	—
CD008782	2	1.000	1.000	0.283	0.567	0.500	0.417	0.500	0.583	0.700
CD008892	29	0.182	0.274	0.384	0.247	0.368	0.439	0.290	0.333	0.338
CD009372	29	0.110	0.117	0.461	0.543	0.328	0.250	0.378	0.276	—
CD010173	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
CD010339	16	0.192	0.179	0.642	0.513	0.537	0.432	0.482	0.495	0.154
CD010653	2	0.053	0.035	0.023	0.015	0.107	0.112	0.062	0.086	—
CD011420	6	0.070	0.074	0.239	0.175	0.189	0.138	0.254	0.157	0.190
mean:		0.336	0.311	0.304	0.342	0.402	0.331	0.400	0.343	0.376

Index test										
	<i>n</i> pos	BioBERT				Logistic Regression				As ranked by the other expert (rs)
		Distant		Supervised		Distant		Supervised		
		Raw	UMLS	Raw	UMLS	Raw	UMLS	Raw	UMLS	
CD007394	2	1.000	1.000	0.643	0.361	0.750	0.500	0.583	0.583	1.000
CD007427	17	0.354	0.225	0.580	0.568	0.551	0.526	0.534	0.484	—
CD008054	10	0.388	0.305	0.449	0.281	0.170	0.161	0.195	0.218	—
CD008782	2	0.833	1.000	0.079	0.523	0.750	0.750	0.750	0.750	0.700
CD008892	34	0.342	0.473	0.458	0.391	0.471	0.484	0.496	0.529	0.524
CD009372	8	0.269	0.351	0.194	0.225	0.261	0.270	0.303	0.390	—
CD010173	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
CD010339	1	0.167	0.050	0.067	0.067	0.071	0.100	0.013	0.017	0.010
CD010653	0	NA	NA	NA	NA	NA	NA	NA	NA	—
CD011420	19	0.251	0.342	0.284	0.218	0.288	0.266	0.280	0.256	0.391
mean:		0.450	0.468	0.344	0.329	0.414	0.382	0.394	0.403	0.525

Reference standard										
	<i>n</i> pos	BioBERT				Logistic Regression				As ranked by the other expert (rs)
		Distant		Supervised		Distant		Supervised		
		Raw	UMLS	Raw	UMLS	Raw	UMLS	Raw	UMLS	
CD007394	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
CD007427	2	0.145	0.032	0.081	0.034	0.052	0.037	0.035	0.041	—
CD008054	6	0.215	0.108	0.239	0.076	0.635	0.619	0.525	0.515	—
CD008782	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
CD008892	13	0.112	0.097	0.152	0.154	0.408	0.351	0.264	0.255	0.201
CD009372	3	0.052	0.095	0.253	0.414	0.681	0.692	0.679	0.729	—
CD010173	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
CD010339	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
CD010653	1	0.020	0.016	0.020	0.059	0.029	0.034	0.067	0.067	—
CD011420	1	0.034	0.100	1.000	0.014	1.000	1.000	0.500	0.500	0.333
mean:		0.097	0.075	0.291	0.125	0.467	0.455	0.345	0.351	0.267

Table 16.4 – Average precision results for the 8 different machine learning models on the data annotated by the first annotator (ML), compared to the performance of an independent human expert (annotator RS). The ‘Raw’ columns denote results for models trained and evaluated on raw sentences. The ‘UMLS’ columns denote results for models trained and evaluated on sentences where the concept mentions have been replaced with their corresponding UMLS semantic types. The ‘*n* pos’ column denotes the number of positive sentences labeled by ML for each article. Cells are marked ‘NA’ if no result could be computed because no sentences were labeled positive. In the baseline results, cells are marked ‘—’ if the article was not annotated by the other expert (RS).

Target condition

	<i>n</i> pos	BioBERT				Logistic Regression				As ranked by the other expert (ML)
		Distant		Supervised		Distant		Supervised		
		Raw	UMLS	Raw	UMLS	Raw	UMLS	Raw	UMLS	
CD007394	2	0.750	0.500	0.667	0.040	0.833	0.500	1.000	0.833	0.667
CD008081	8	0.136	0.198	0.213	0.371	0.504	0.380	0.394	0.388	—
CD008760	5	0.200	0.144	0.283	0.163	0.252	0.300	0.481	0.300	—
CD008782	1	1.000	1.000	0.500	1.000	0.500	0.333	1.000	0.500	0.500
CD008892	15	0.170	0.270	0.088	0.342	0.440	0.505	0.667	0.542	0.564
CD009647	2	0.036	0.021	0.021	0.047	0.020	0.026	0.012	0.023	—
CD010339	2	0.061	0.040	0.066	0.062	0.044	0.029	0.063	0.023	0.019
CD010360	2	0.089	0.080	0.093	0.261	0.181	0.083	0.244	0.064	—
CD010705	7	0.189	0.269	0.127	0.341	0.382	0.359	0.254	0.402	—
CD010420	4	0.036	0.044	0.209	0.097	0.210	0.214	0.302	0.132	0.178
mean:		0.267	0.257	0.227	0.273	0.337	0.273	0.412	0.321	0.386

Index test

	<i>n</i> pos	BioBERT				Logistic Regression				As ranked by the other expert (ML)
		Distant		Supervised		Distant		Supervised		
		Raw	UMLS	Raw	UMLS	Raw	UMLS	Raw	UMLS	
CD007394	2	1.000	1.000	0.417	0.393	0.750	0.500	0.700	0.750	1.000
CD008081	11	0.464	0.229	0.463	0.454	0.431	0.412	0.394	0.447	—
CD008760	9	0.357	0.411	0.512	0.475	0.457	0.470	0.481	0.476	—
CD008782	1	1.000	1.000	1.000	0.500	1.000	1.000	1.000	1.000	0.500
CD008892	27	0.499	0.539	0.717	0.758	0.740	0.666	0.667	0.474	0.692
CD009647	1	0.053	0.015	0.020	0.006	0.006	0.009	0.012	0.040	—
CD010339	6	0.085	0.054	0.040	0.047	0.053	0.041	0.063	0.047	0.058
CD010360	8	0.154	0.119	0.233	0.278	0.222	0.202	0.244	0.242	—
CD010705	14	0.599	0.533	0.292	0.270	0.352	0.327	0.254	0.327	—
CD010420	8	0.234	0.296	0.280	0.251	0.259	0.235	0.302	0.257	0.328
mean:		0.444	0.420	0.397	0.343	0.427	0.386	0.412	0.406	0.516

Reference standard

	<i>n</i> pos	BioBERT				Logistic Regression				As ranked by the other expert (ML)
		Distant		Supervised		Distant		Supervised		
		Raw	UMLS	Raw	UMLS	Raw	UMLS	Raw	UMLS	
CD007394	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
CD008081	3	0.254	0.132	0.134	0.177	0.867	0.698	1.000	1.000	—
CD008760	2	0.101	0.553	0.529	0.013	0.667	0.833	0.667	0.833	—
CD008782	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
CD008892	11	0.110	0.212	0.283	0.108	0.356	0.286	0.334	0.225	0.417
CD009647	0	N/A	N/A	N/A	N/A	0.036	N/A	N/A	N/A	—
CD010339	1	0.012	0.010	0.029	0.009	0.224	0.031	0.071	0.028	N/A
CD010360	1	0.200	0.037	0.111	0.038	0.810	0.023	0.167	0.143	—
CD010705	5	0.150	0.152	0.194	0.086	0.224	0.122	0.172	0.125	—
CD010420	3	0.167	0.347	0.358	0.019	0.810	0.806	0.692	0.694	0.345
mean:		0.142	0.206	0.234	0.064	0.428	0.400	0.443	0.435	0.381

Table 16.5 – Average precision results for the 8 different machine learning models on the data annotated by the second annotator (RS), compared to the performance of an independent human expert (annotator ML). The 'Raw' columns denote results for models trained and evaluated on raw sentences. The 'UMLS' columns denote results for models trained and evaluated on sentences where the concept mentions have been replaced with their corresponding UMLS semantic types. The '*n* pos' column denotes the number of positive sentences labeled by RS for each article. Cells are marked 'N/A' if no result could be computed because no sentences were labeled positive. In the baseline results, cells are marked '—' if the article was not annotated by the other expert (ML).

**C2: DISTANTLY SUPERVISED TRAINING** We here trained on the distant annotations from the 48 systematic reviews in the training set, and evaluated on the 15 manually annotated articles in the evaluation set, where each annotator provided annotation data for 10 articles (with a 5 article overlap). The articles used for evaluation were the same as in C1.

### 16.3.1 *Evaluation*

Since our model output ranked sentences, rather than a binary classification, we evaluated all experiments in terms of average precision.

As a comparison, we also evaluated the average precision using the ranking given by the other annotator. In plain language, we tried to evaluate how useful it would have been for the expert to highlight sentences for each other. The expert annotations were binary (Yes/No), rather than a ranking score, so we calculated the average precision by interpolating ties in the ranking.

### 16.4 RESULTS

Out of the 1,738 references in the LIMS1-Cochrane dataset, 1152 had either a PMID or DOI assigned. EndNote was able to retrieve PDF articles for 666 of these references. A total of 90,996 sentences were distantly labeled for target condition, 94,290 sentences were distantly labeled for index test, and 79,504 sentences were distantly labeled for reference standard. The first annotator (ML) annotated 981 sentences and the second annotator (RS) annotated 1,031 sentences (Table 16.2).

We present the results of our algorithm evaluated on the annotations by ML in Table 16.4, and evaluated on the annotations by RS in Table 16.5.

The ranking performance exhibited large variations. Neither BioBERT or logistic regression were consistently better than the other, neither distant supervision or direct supervision were consistently better than the other, and neither raw sentence nor sentences augmented with UMLS concepts were consistently better than the other. For the target condition, the best performance was achieved by logistic regression on raw sentences using either distant or direct supervision, with a maximum at 0.412 compared to human performance at 0.376 and 0.386 respectively. For the index test, the performance fell within the range 0.344–0.468 compared to human performance at 0.525 and 0.516 respectively. For the reference standard, BioBERT exhibited substantially inferior results on the reference standard compared to logistic regression, while logistic regression performance fell within the range 0.345–0.467, compared to human performance at 0.267 and 0.381 respectively.

The performance also varied between systematic reviews, with consistently close

to perfect performance on a few reviews (CDO07394 and CDO008782), and consistently very low performance on a few (CDO09647 and CDO10339). These also correspond to the articles with the highest and lowest inter-annotator agreement. The consensus of the two experts is that CDO10339 is not a diagnostic test accuracy study.

## 16.5 DISCUSSION

Raw sentences worked consistently better for logistic regression on the target condition (8/8), and worked better than UMLS concepts as a general trend (20/24). While general concepts could theoretically improve performance by helping the models generalize, this may also remove important semantic information from the sentences, keeping the models from ranking accurately. We also note that BioBERT already encodes a language model (similar to word embeddings), and concepts may therefore be unhelpful for the model.

BioBERT performed consistently better than logistic regression on the index test when using distant supervision (4/4), but not when using direct supervision (0/4). Logistic regression performed consistently better than BioBERT on both the target condition and the reference standard (16/16). On the reference standard the difference in performance is substantial, with BioBERT scoring very poorly, and logistic regression performing much better than human performance. The reason for BioBERT's poor performance on the reference standard may be due to the relative sparsity of the annotations for this subtask (see Table 16-2).

Distant supervision was consistently on par with or better than direct supervision. The top performing models also outperformed the human annotators on the target condition and the reference standard, and came comparatively close on the index test (0.468 versus 0.525 and 0.444 versus 0.516).

### 16.5.1 Limitations

We only manually annotated a small sample of the dataset. The small size is further compounded by problems with converting PDF to text, which may also bias the training and evaluation in favor of articles where the conversion works better (mainly articles from big publishers).

The dataset was constructed from articles included in previous systematic reviews of diagnostic test accuracy. These include articles that contain diagnostic results, while not being diagnostic test accuracy studies. Arguably, these should be excluded from training or evaluation, and possibly even from the dataset.

## 16.6 CONCLUSIONS

Our results suggest that distant supervision is sufficient to train models to identify target condition, index test, and reference standard in diagnostic articles. Our results also suggest that such models can perform on par with human annotators. We constructed a dataset of full-text articles of diagnostic test accuracy studies, with distant annotations for target condition, index test and reference standard, that can be used to train machine learning models. We also provide a subset of the data manually annotated by experts for evaluation. Our dataset cannot be publicly distributed due to copyright restrictions, but will be available upon request. We also plan to distribute the code for the distant annotations and data preprocessing, as well as the cleaned data extraction forms.

### 16.6.1 *Future Work*

The dataset is being updated, and we plan to increase the amount of manually annotated data to improve the statistical reliability of the experiments. We also plan to let all experts annotate the same articles to simplify the comparisons.

**T**HROUGH THIS SECTION WE HAVE PRESENTED a dataset documenting the data extraction, data synthesis, and meta-analysis stages of systematic reviews of diagnostic test accuracy, encompassing 63 systematic reviews of 1,738 DTA studies. In total, it encompasses 589 meta-analyses of 5,848 diagnostic test evaluations. This dataset is to our knowledge the first of its kind. We hope it will be of aid for better understanding how the process is undertaken by human reviewers, as well as for modelling the process with automated methods.

As remarked in the introduction of this part, comparing studies to determine which are similar enough to be group in each comparison could be assisted by graphing and tabulation tools. The homogenization could similarly likely be automated (Tsafnat et al., 2014). Only the data extraction is an NLP problem however, and will be treated in this thesis.

We will however build a pipeline for automatically performing meta-analyses, even though this is not an NLP problem. The reason we do this is because this allows us to later perform cumulative meta-analyses to measure the impact screening automation methods have on the results of the systematic review (see chapter 12).

235

#### 17.1 AUTOMATED DATA EXTRACTION FROM DTA STUDIES

To make any data extraction system relevant and useful, it must target items that are necessary or useful in the review process. There are to our knowledge no definite lists of items that must be extracted in a DTA systematic review. A systematic review by Jonnalagadda et al. (2015) determined relevance and usefulness by two published guidelines

The first of these guidelines is the checklist of items available from the Cochrane Handbook of Systematic Reviews of Interventions (Jonnalagadda et al., 2015; Li et al., 2019, in Higgins et al., 2019). No similar checklist is available in the Cochrane Handbook of DTA Reviews (Deeks et al., 2013a). Due to the differences in design and purpose of RCTs and DTA studies however, few data items are relevant to both RCTs and DTA studies. Additionally, due to differences in study aims and design of primary studies, a DTA systematic review will extract different data items than a systematic review of interventions.

The second guidelines is the STARD checklist of items that should be reported by DTA studies (Cohen et al., 2016; Jonnalagadda et al., 2015). However, STARD is intended as an aid to human authors to improve reporting standards, not as a laundry list of items to be extracted in systematic reviews. In practice, several items that should be mentioned in a DTA study are not routinely extracted in systematic reviews, and vice versa. Furthermore, STARD bundles several items commonly extracted into the flow diagrams. While the flow diagram includes all data necessary

		Published Methods	STARD	Cochrane
Flow and timing	Conflict of interest	No	No	Yes
	Key conclusions by the authors	Yes [280]	2 <sup>1</sup>	Yes
	Flow and timing	No	19, 22	No
	Diagnosis at baseline	No	19 <sup>2</sup>	No
	Diagnosis at follow up	No	19 <sup>2</sup>	No
	Loss to follow-up	No	19 <sup>2</sup>	No
	Time interval between index test and reference standard	No	22	No
	Withdrawals	No	19 <sup>2</sup>	For RTCS
	Index test description and parameters	No	10a	No
	Threshold for positive result	No	12a	No
Index tests	Examiners	No	No	No
	Blinding of examiners	For RCTs [201]	13a	For RTCS
	Interobserver variability	No	No	No
	Sequence of tests	No	19 <sup>2</sup>	No
	Target condition	No	No	No
Target condition and ref. standard	Reference standard description and parameters	No	10b	No
	Blinding of examiners	For RCTs [201]	13b	For RTCS
	Positive case definition by reference standard	No	12b	No
	Examiners	No	No	No
	Prevalence of target condition in the sample	No	No	No
Patient characteristics and setting / Patient Sampling	Number of participants	For RCTs [33; 82; 87; 140; 141; 144; 153; 154; 169; 174; 191; 312; 330; 332]	20 <sup>3</sup>	No
	Number of participants available for analysis	No	19	For RTCS
	Country/Place of Study/Location	For RCTs [191; 332]	No	For RTCS
	Sources of recruitment/referral	No	8	No
	Age	For RCTs [169; 191; 332; 333]	20 <sup>3</sup>	For RTCS
	Gender	For RCTs [169; 332; 333]	20 <sup>3</sup>	For RTCS
	Ethnicity	For RCTs [333]	20 <sup>3</sup>	For RTCS
	Education	No	20 <sup>3</sup>	For RTCS
	Clinical presentation	No	21a, 21b	No
	Clinical setting	No	8	For RTCS
	Selection criteria	For RCTs [34; 175; 252]	7	No
	Inclusion criteria	No	7	No
	Exclusion criteria	No	7	No
	Language	Yes <sup>4</sup>	No	No
	Period of study	No	No	For RTCS
	Primary objective	No	4	No
	Study design	No	2	For RTCS

Table 17.1 – Structure of the entries in the data extraction forms in the LIMS1-Cochrane dataset. Only entries found in two or more systematic review have been included. ‘Published methods’ denote whether each data item has been addressed by previous literature. ‘STARD’ denote whether each data item is required by the STARD checklist (Cohen et al., 2016). ‘Cochrane’ denote whether each data item is required by the Cochrane ‘Checklist of items to consider in data collection’ (Li et al., 2019, in Higgins et al., 2019). We do not list QUADAS items.

<sup>1</sup> In abstract

<sup>2</sup> As part of the diagram of flow

<sup>3</sup> STARD item 20 does not specify specific characteristics to extract

<sup>4</sup> Language detection does not require methods tailored for scientific articles

to extract, it does so at a different level of granularity, along with data that may not be necessary for the systematic review (table 17.1). Also, ironically, the target condition does not appear as a separate item in the STARD list, presumably because authors are unlikely to fail to report this.

To give a clearer and more relevant overview of the data items that are routinely extracted in systematic reviews of diagnostic test accuracy, we therefore list all data items that were extracted at least twice in the 63 systematic reviews currently in the LIMS1-Cochrane dataset (Norman et al., 2018a) (table 17.1). We excluded all data items that are only relevant to specific target conditions or index tests, such as ‘history of TB’ or ‘APOE ε4 carrier’. Many data extraction forms included a ‘note’ field which is not relevant for automated extraction and therefore omitted. Most of data extraction forms also included a ‘comparative’ field, which was not filled for any of the primary studies.

Out of the 38 data item identified as common in the LIMS1-Cochrane dataset, only 14 are mentioned by the Cochrane checklist, and only 29 by STARD (table 17.1). Eight of the data items have been addressed by previous methods for RCTs, one (key conclusions) by previous methods for systematic reviews of several types including DTA, and one (article language) by methods not specific to systematic reviews.

Since the majority of previous methods have focused on interventions, there is a lack of relevant methods for DTA systematic reviews. Among previous methods, only one study is directly applicable to extract key conclusions from DTA studies (Song et al., 2013). Article language can also easily be determined using standard methods, and does not require specialized methods for DTA studies.

A number of data items have been addressed by previous literature for data extraction in systematic reviews of interventions. A few of these – age, gender, ethnicity – may be described by similar language in DTA studies and RCTs. Others, such as the number of participants use markedly different language, and it is not clear how well models trained on RCTs will perform on DTA studies.

In the second study in this section, we have attempted to automatically extract the index test, target condition, and reference standard from DTA studies. These items have not been previously considered. These items are also important to extract in systematic reviews, since they are necessary for determining whether the record should be included in the review, as well as for determining what meta-analyses the studies will be included in.

We implemented models using both logistic regression, and deep learning using BioBERT. We have also compared the use of data preprocessing by replacing medical terms with their corresponding UMLS semantic types. No single method was consistently better than the other ones.

The results were on par with the inter-annotator performance among human annotators. In other words, the labeling method was approximately as useful as high-

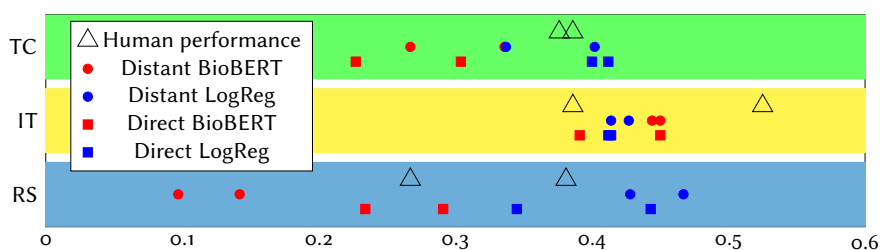


Figure 17.1 – Summary average precision results for automated data extraction, compared to inter-annotator performance. TC denote the target condition, IT denote the index test, and RS denote the reference standard.

lighting done by another expert.

Distant supervision performed roughly as well as direct supervision, presumably due to the much larger amounts of data available through this method.

At the same time, inter-annotator agreement was consistently low, even after several rounds of adjustments to the annotation instructions. Part of the reason appear to be that it is difficult to clearly delineate which sentence unambiguously specify or do not specify study characteristics. Often this is a judgement call with a large degree of subjectivity. Furthermore, the exhaustiveness of the systematic review process leads to systematic reviews often including studies of highly variable reporting quality. For instance, one of the included studies in the manually annotated sample was not a diagnostic test accuracy study according to the annotators.

## 17.2 AUTOMATED META-ANALYSES FOR DTA SYSTEMATIC REVIEWS

The purpose of this subsection of the thesis is to construct a pipeline where tabulated data from DTA studies can be used to perform meta-analyses without human intervention. The primary purpose of this pipeline was to be able to integrate meta-analysis software in the screening automation process, so that the meta-analysis results can be calculated cumulatively, and ultimately to use these cumulative meta-analyses to measure the impact of screening automation. This was one of the main purposes of part III.

Software to perform meta-analyses from tabulated data exist in several packages, such as the SAS NLMixed procedure,<sup>1</sup> the Stata xtmelogit or meqrlogit routines,<sup>2</sup> or the reitsma function from the mada R package (Doebler and Holling, 2015). SAS and Stata are widely used by the systematic review community, but are primarily

<sup>1</sup> <https://support.sas.com/documentation/onlinedoc/stat/141/nlmixed.pdf>

<sup>2</sup> <https://www.stata.com/help.cgi?xtmelogit>

Reported results					
Test Strategy	Studies	Women (cases)	Sensitivity (95% CI)	Specificity (95% CI)	Threshold
Inhibin	3	2098 (184)	19 (4 to 58)	95	5% FPR
Total hCG	3	2098 (184)	19 (4 to 58)	95	5% FPR

Correct results					
Test Strategy	Studies	Women (cases)	Sensitivity (95% CI)	Specificity (95% CI)	Threshold
Inhibin	3	2,098 (184)	19 (4 to 58)	95	5% FPR
Total hCG	2	2,482 (109)	7 (3 to 18)	95	5% FPR

Table 17.2 – Comparison between the replicated summary scores for ‘Total hCG’ in review CD011975 and the reported summary scores. Top: Results reported in the summary of findings for ‘Total hCG’ and ‘Inhibin.’ Bottom: Expected results based on the reported data tables.

239

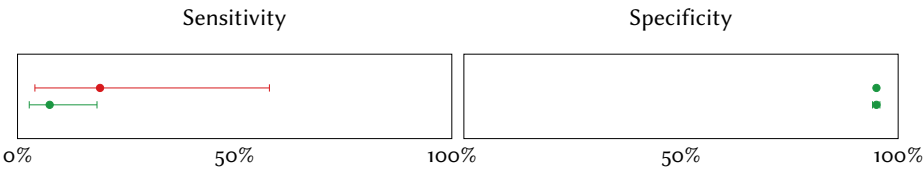


Figure 17.2 – Forest plot of the reported (top) versus the replicated scores (bottom) for ‘Total hCG’ in review CD011975.

intended for use with an interactive GUI. Their support for automated integration with other software remain relatively limited. Furthermore, both are proprietary software, which prohibited their use in this project. In contrast, mada is an open source R package, and is straightforward to call from the command line or interfaced directly from several programming languages, including python, which was used in this project.

Meta-analysis models for DTA systematic reviews are considered too complex to be implemented in RevMan, and there is consequently no pipeline to perform meta-analyses automatically from within RevMan (Macaskill et al., 2010). As a consequence, review authors need to perform meta-analyses in external software. Beyond increasing the workload, this has been hypothesized to lead to mistakes (Tsafnat et al., 2014).

To test this, we used the automated meta-analysis pipeline to recalculate the reported meta-analyses in the 63 systematic reviews. Since none of the original reviews used mada, but SAS or Stata, we cannot expect to achieve exactly the same numerical results. In practice, minor implementation details may influence the results. However, mada is still methodologically sound, and should give results that are methodologically equivalent to those reported.

On average, we observed approximately 2% discrepancies from the reported results,

Reported results			
Test / subgroup	Sensitivity (95% ci)	Specificity (95% ci)	No. of participants (studies)
Diagnosis of schizophrenia...	58.0 (50.3, 65.3)	74.7 (85.2, 82.3)	4070 (16)

Correct results			
Test / subgroup	Sensitivity (95% ci)	Specificity (95% ci)	No. of participants (studies)
Diagnosis of schizophrenia...	58.0 (50.3, 65.3)	74.7 (65.2, 82.3)	4070 (16)

Table 17.3 – Comparison between the replicated summary scores for ‘Diagnosis of schizophrenia from other types of psychosis’ in review CD010653 and the reported summary scores. Top: Results reported in the summary of findings. Bottom: Expected results based on the reported data tables.

240

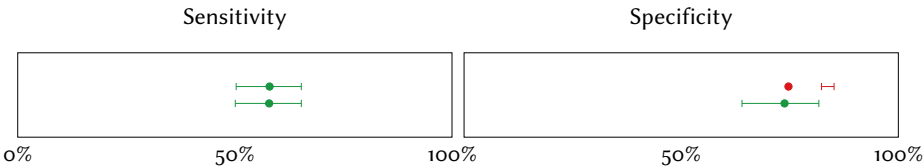


Figure 17.3 – Forest plot of the reported (top) versus the replicated scores (bottom) for ‘Diagnosis of schizophrenia from other types of psychosis’ in review CD010653.

and we have used this as an indication of the lower bound on the accuracy that can be expected by automated screening methods in one of our studies (chapter 12) (Norman et al., 2019c). Second, large discrepancies may be an indication that there may be an error in either of the calculations. In this way we identified two errors in the original summary of findings tables.

One of the errors (CD011975 Total hCG in table 15.2) was apparently due to a copy-paste error, where the authors of the review copied the same meta-analysis results for two different diagnostic tests (table 17.2, figure 17.2). This error could have been spotted in several ways: 1) our replicated score was off by more than 10 point, 2) two rows in the summary of findings were identical, 3) the summary of findings table in the review described a different number of included studies than what was reported in ‘data and analyses’, and 4) the summary of findings table in the review described a different number of participants than what was reported in ‘data and analyses’.

The second error (CD010653 Diagnosis of Schizophrenia from other types of psychosis in table 15.2), also appears to be a typo: ‘74.7 [85.2, 82.3]’ instead of ‘74.7 [65.2, 82.3]’ (table 17.3, figure 17.3). This error could also have been spotted in several ways: 1) our replicated score was again off by more than 10 point, 2) [85.2, 82.3] is not a legal confidence interval, and 3) the mean 74.7 lies outside the interval [82.3, 85.2]. As a further irony, we also overlooked this second error in our paper – which is present unremarked upon in table 15.2 – and only noticed it after the paper was

published (Norman et al., 2018a).

All considered, such errors were rare in the summary of findings tables (2 in 63 reviews), but both could have been spotted automatically using very simple methods. Alternatively, automatically populating summary of findings tables from the data would also have avoided these errors.

### 17.3 CONCLUSIONS

Automated data extraction can perform as well as human experts, even with heuristically annotated training data. At the same time, inter-annotator agreement among experts was low, and the results of the automated extraction was less than perfect. The major reason for the low performance of both humans and machines appears to be that highlighting sentences describing e.g. a target condition is a much more contextual and subjective task than highlighting non-contextual, non-subjective elements such as diseases.

Arguably, it does not make sense to try to automate a labelling task where experts cannot agree on a gold standard. Future research for extracting these items from text therefore may need to forsake sentence highlighting – or at least keep in mind that sentence highlighting will necessarily be hit and miss – in favor of direct extraction of the target items, where experts largely agree.

Major obstacles to automated extraction include variable quality of reporting, and the relatively high number of references from non-mainstream journals and publishers, including gray literature. Better and more consistent reporting in primary studies would likely increase inter-annotator agreement as well as data extraction performance. These obstacles do not just hinder the use of data extraction methods, but severely degrade the performance of existing automated full-text retrieval methods and PDF to text converters like Grobid.

Manual handling of data when performing meta-analyses and limited integration between review managers like RevMan and external software have been hypothesized to cause errors in systematic reviews. We found only 2 instances of such errors. However, both of these could have been avoided or spotted using very simple data consistency checks.



PART V

SUMMARY

**T**HE DEMAND AND PRODUCTION of systematic reviews is increasing rapidly. PubMed indexed 17,254 new systematic reviews in 2018 alone, and this number has increased more than five-fold since 2009. While the demand for systematic reviews is growing, the number of publications that systematic reviews need to sift through is also increasing at a similarly break-neck pace. We today spend more time and money producing new systematic reviews than we ever have.

Authors conducting systematic reviews face issues throughout the systematic review process. It is difficult and time-consuming to search and retrieve, collect data, write manuscripts, and perform statistical analyses.

The thesis included seven articles (chapters 6–8, 11, 12, 15, and 16) published between 2017–2019. In these papers we have attempted to explore methods for performing systematic reviews quicker, cheaper, and more efficiently. At the same time, systematic reviews still require a thorough, objective, and reproducible methodology to avoid bias. While we are attempting to make the process more expedient, we are also striving to uphold the same methodological rigor of the process.

Most previous work have focused on screening automation in the title and abstract screening stage. Comparatively less work have been done automating the other parts of the review process, including the article retrieval, article screening, data extraction, data synthesis, and the analysis stages. Furthermore, the work that does exist have largely focused on systematic reviews of interventions, with little previous work on systematic reviews of diagnostic test accuracy.

Search queries to identify diagnostic studies tend to have low accuracy, and are discouraged for use in systematic reviews. This results in relatively larger numbers of candidate references to screen. Systematic reviews of diagnostic test accuracy may therefore be a prime target for alternative approaches to cope with the rapidly increasing workloads.

In this thesis we have examined how machine learning methods can be used to reduce this workload, how such methods can be made to work, and how they can fit into different systematic review contexts and settings.

## 18.1 SCREENING AUTOMATION

In chapter 6–8 we presented three papers on screening automation methods for systematic reviews, with a particular focus on systematic reviews of diagnostic test accuracy.

Screening automation methods need to cope with several technical constraints, including extreme class imbalance. Diagnostic test accuracy systematic reviews may include only one per thousand studies retrieved from the database search.

Furthermore, among the 50 systematic reviews, the median number of included studies was 14 (range: 0 to 99). Eleven of the reviews included four studies or less. The training data that we can expect from many systematic reviews are therefore far below the numbers necessary to reach data-saturation in machine learning models. It is unclear how to effectively train screening automation models on such limited data, particularly to reach the close to perfect recall commonly required by systematic reviews.

#### 18.1.1 *Differences Between Studies Included by Abstract and Full-Text*

References included in the systematic review and references provisionally included in the systematic review but excluded based on full-text were not sufficiently different in terms of language or word choice that general machine learning algorithms such as logistic regression could distinguish the two based on title and abstract. However, training data where gold standard labels were based on titles, abstracts, and full-text appear to lead to slight performance improvements over training data where gold standard labels were based only on titles and abstracts. The differences appeared to be minor, however.

#### 18.1.2 *Using Training Data from Both Screening Stages*

Training data should preferentially use gold standard labels based on full-text if such labels are available. However, out of the 50 systematic reviews examined, only 19 included at least 20 studies, and in practice the best quality examples of included studies may be too few to use effectively. To construct well performing screening automation methods for diagnostic test accuracy systematic reviews it may therefore be necessary to complement the data with gold standard labels based only on titles and abstracts. We have also observed performance improvements when complementing training data with training data from similar systematic review topics (i.e. by using transfer learning).

To take advantage of gold standard labels from both stages of screening as well as transfer learning from similar topics, we have presented a stacked model, which uses meta-regression to combined decisions from multiple models.

#### 18.1.3 *Screening Approaches*

We have presented three different models, for slightly different systematic review contexts.

We have presented a static model, trained on the inclusion/exclusion decisions of references screened in previous systematic reviews. This model thus requires

training data to be available at the time the screening is started. This typically limits the applicability of this model to systematic review updates, or to train general models, e.g. to identify general diagnostic test accuracy studies.

We have presented an active model, which uses active learning to improve its performance throughout screening. This model does not require training data at the time the screening is started, and can therefore be used also in systematic reviews conducted de novo, where training data is not available. This process can be started from scratch, with no training data at all, but if some quantity of training data is available, or if training data can be constructed artificially, such data can be used as a starting point.

We have presented a stacked model, which combines the static and active models to achieve the best of both. It uses a static (intertopic) model as a base and then uses the more targeted intratopic data collected through the screening process to improve the model further, using active learning.

## 18·2 SCREENING PERFORMANCE

In chapters 6–8 we compared the performance of our models with the current state of the art.

### 18·2·1 *Static Model (Intratopic)*

When we constructed our static model, the results were better than the then state-of-the-art in terms of wss@95 (0.392 on average), but worse in terms of AUC. This suggests that our model works well for finding all relevant studies, whereas the competing approaches are better at finding the first relevant studies, but struggles to find the last ones.

Two more studies have been published evaluating five new models on this dataset, and have since pushed the state-of-the-art further. The two studies report average wss@95 ranging from 0.347 to 0.408. Our model performed better than the reported results on 5, 6, 8, 9, and 10 out of 15 topics. Consequently, the performance of our model still compares favorably with the current state of the art.

### 18·2·2 *Static Model (Intertopic)*

The results of our static model performed better than the state-of-the-art across all topics for transfer learning. We are not aware of subsequent studies examining transfer learning, and this model thus appear to remain the current state-of-the-art. Despite the simplicity of the approach, the static approach combined with transfer learning frequently gave performance comparable to current state-of-the-art active learning models.

18.2.3 *Active Learning*

The best active learning approach consistently outperformed the transfer learning approach, although the differences were modest.

18.2.4 *Stacked Model*

The stacked model combines the high initial performance of the transfer learning approach with the performance improvements that can be gained over time using active learning. In our experiments, the stacked model was the best performing model for systematic reviews conducted de novo, with slight performance improvements over the static intertopic model and standard active learning.

247

## 18.3 SCREENING REDUCTION COMPATIBLE WITH CURRENT PRACTICE

In chapter 11, we presented a prospective study where we documented the use of the static method in the 2019 update of the COMET database. In this study the cut-off was determined retrospectively on previous review updates. We identified a threshold that would have resulted in an acceptable balance between workload reduction and screening exhaustiveness in previous review updates, and applied this criterion in the screening for the 2019 review update.

We judged that missing 2% of the references was an acceptable trade-off for a 75% workload reduction. This particular review is intended to populate a literature database, and recall is therefore a direct and appropriate measure of the impact screening automation had on the review.

References without abstracts could not be ranked with acceptable performance guarantees, and were therefore ineligible for screening automation. There were however only a small number of such references, corresponding to a workload of approximately 2–4 hours per screener.

Since we applied the model prospectively, we could only estimate the loss in recall. Relevant studies for inclusion in the COMET database are however identified from multiple sources, and time will tell whether the estimated number of missed articles matches the number of studies that were actually missed. Screening was done on a small sample (1%) of the excluded references to verify the results, and all of these references were found to be correctly excluded.

The application of screening automation was done to adhere to the established process as closely as possible. We used a screening reduction approach where the inclusion threshold could be determined as part of the protocol. We applied the model before starting the screening, and the records were randomized to avoid rank order bias prior to screening. The screening was then performed as normal

in EndNote. Unlike previous years, the 2019 update only involved two screeners, but the remainder of the process was unchanged apart for the use of screening reduction. No specialized software was required by the screeners.

### 18·3·1 *Better Metrics for Screening Automation*

In chapter 12, we tried to measure ‘information loss’ directly for systematic reviews of diagnostic test accuracy. We tried to address what it means for an abridged method to yield the ‘same’ systematic review as with exhaustive screening.

We tried to satisfy three criteria with the measure:

- ❖ The measure should be possible to calculate cumulatively through the screening process
- ❖ It should be possible to stop screening once we are confident that further screening will not change the conclusions of the review; and
- ❖ It should be possible to determine criteria for stopping as part of the review protocol to avoid bias

If screening automation methods are to be used in systematic reviews, reviewers need to judge what amount and kind of loss are acceptable. A loss in recall or exhaustiveness may yield a review that does not ‘look like’ a systematic review, but may not meaningfully impact the results and conclusions of the review – provided a sufficient selection of studies are identified to address all review questions, and provided the selection of studies is unbiased. To avoid reviewer bias and ad-hoc decisions during the screening process there should be a clear, pre-specified protocol for judging when the screening is complete.

To this end, we have attempted to use meta-analysis accuracy as a performance metric during screening, by performing cumulative meta-analyses through the screening process. This accuracy can be estimated prospectively and thresholds can be decided as part of the protocol. This however requires the screening process to be performed in parallel with the data extraction, synthesis and meta-analysis stages of the systematic review process, and would thus result in an unconventional systematic review process.

The benefit of the measure is that it is conservative and reliable, and interrupting screening once the accuracy falls within prespecified limits is unlikely to lead to wrong results or conclusions in the systematic review. Furthermore, this allows the screening to be interrupted much earlier in the process and reduce the workload by orders of magnitude more than with conventional stopping criteria.

## 18.4 DATA EXTRACTION &amp; SYNTHESIS

In part IV we presented a dataset documenting the data extraction, data synthesis, and meta-analysis stages of systematic reviews of diagnostic test accuracy, encompassing 63 systematic reviews of 1,738 diagnostic test accuracy studies. In total, it encompasses 589 meta-analyses of 5,848 diagnostic test evaluations. This dataset is to our knowledge the first of its kind. We hope it will be of aid for better understanding how the process is performed by human reviewers, as well as for modelling the process with automated methods.

While several parts of the process could conceivably be automated, only the data extraction is a natural language processing problem, and was treated in this thesis.

249

18.4.1 *Automated Data Extraction from diagnostic test accuracy Studies*

Since the majority of previous methods have focused on interventions, there is a lack of relevant methods for diagnostic test accuracy systematic reviews. The only directly applicable previous methods extract key conclusions and study language. In chapter 16 we have attempted to automatically extract the index test, target condition, and reference standard from diagnostic test accuracy studies. These items have not been previously considered by data extraction methods, and are important to extract in systematic reviews, since they often take a similar role as PICO for randomized controlled trials. They are therefore often necessary for determining whether the record should be included in the review, as well as for determining what meta-analyses the studies will be included in.

We implemented models using both logistic regression, and deep learning using BioBERT. We have also compared the use of data preprocessing by replacing medical terms with their corresponding UMLS semantic types. No single method was consistently better than the other ones.

The results were on par with the inter-annotator performance among human annotators. In other words, the labeling method was approximately as useful as highlighting done by another expert.

Distant supervision performed roughly as well as direct supervision, presumably due to the much larger amounts of data available through this method.

Inter-annotator agreement was consistently low, even after several rounds of adjustments to the annotation instructions. Part of the reason appear to be that it is difficult to clearly delineate which sentence unambiguously specify or do not specify study characteristics. Often this is a judgement call with a large degree of subjectivity. Furthermore, the exhaustiveness of the systematic review process leads to systematic reviews often including studies of highly variable reporting quality. For instance, one of the included studies in the manually annotated sample was not a diagnostic test accuracy study according to the annotators.

18.4.2 *Automated Meta-Analyses for diagnostic test accuracy systematic reviews*

In chapter 15, we presented work constructing a pipeline where tabulated data from diagnostic test accuracy studies can be used to perform meta-analyses without human intervention.

Meta-analysis models for diagnostic test accuracy systematic reviews are considered too complex to be implemented in RevMan, and there is consequently no pipeline to perform meta-analyses automatically from within RevMan. As a consequence, review authors need to perform meta-analyses in external software. This increases the workload, and has been hypothesized to lead to mistakes.

To test whether calculating meta-analyses in external software lead to mistakes, we used the automated meta-analysis pipeline to recalculate the reported meta-analyses in the 63 systematic reviews.

On average, we observed approximately 2% discrepancies from the reported results, and we have used this as an indication of the lower bound on the accuracy that can be expected by automated screening methods in one of our studies (chapter 12).

Furthermore, we used large discrepancies to screen for potential errors in the summary of findings tables in the systematic reviews. We identified two errors in 103 eligible meta-analyses.

One of the errors appears to be due to a copy-paste error, where the authors of the review copied the same meta-analysis results for two different diagnostic tests. This error could have been spotted in several ways: 1) our replicated score was off by more than 10 point, 2) two rows in the summary of findings were identical, 3) the summary of findings table in the review described a different number of included studies than what was reported in ‘data and analyses’, and 4) the summary of findings table in the review described a different number of participants than what was reported in ‘data and analyses’.

The second error also appears to be a typo: ‘74.7 [85.2, 82.3]’ instead of ‘74.7 [65.2, 82.3]’. This error could also have been spotted in several ways: 1) our replicated score was again off by more than 10 point, 2) [85.2, 82.3] is not a legal confidence interval, and 3) the mean 74.7 lies outside the interval [82.3, 85.2].

All considered, such errors were rare in the summary of findings tables (2 in 103 eligible meta-analyses in 63 reviews), but both could have been spotted automatically using very simple methods. Alternatively, automatically populating summary of findings tables from the data would also have avoided these errors.

## 18.5 CONCLUSIONS

We have presented a screening automation system that can be used in a variety of systematic review contexts – ranging from review updates to reviews conducted

de novo. The system is general in purpose, and performs well on several types of systematic reviews, including diagnostic test accuracy reviews. The system is furthermore highly customizable, and the underlying preprocessing pipeline and classification or ranking algorithms can be changed to fine-tune the system for specific systematic review topics or contexts.

Systematic review automation method can be used in systematic reviews without fundamentally altering the process. Screening reduction method can be used as an extra search filter, leaving the remainder of the review process identical to the conventional process, including screening in random order, and the use of standard reference managers like EndNote.

The accuracy of the screening process, and the impact it has on the results and conclusions of the review can be measured prospectively through the screening process using cumulative meta-analyses. This requires modifying the systematic review process to perform data extraction and meta-analyses concurrently, but can lead to substantial improvements over traditional stopping criteria for screening automation.

Automated data extraction can perform as well as human experts, even with heuristically annotated training data.

Manual handling of data when performing meta-analyses and limited integration between review managers like RevMan and external software have been hypothesized to cause errors in systematic reviews. We found only 2 instances of such errors. However, both of these could have been avoided or spotted using very simple data consistency checks.

Major obstacles in systematic review automation include variable quality of reporting, and the relatively high number of references from non-mainstream journals and publishers, as well as gray literature. Better and more consistent reporting in primary studies would likely increase inter-annotator agreement as well as data extraction performance with automated methods. These obstacles do not just hinder the use of screening automation and automated data extraction, but severely degrade the performance of existing automated full-text retrieval methods and PDF to text converters like Grobid.

De vraag naar en productie van systematische reviews neemt snel toe. PubMed heeft alleen al in 2018 17.254 nieuwe systematische reviews geïndexeerd, en dit aantal is sinds 2009 meer dan vervijfvoudigd. Terwijl de vraag naar systematische reviews toeneemt, neemt ook het aantal publicaties dat auteurs van systematische reviews moeten lezen toe. We besteden vandaag de dag meer tijd en geld aan het produceren van nieuwe systematische reviews dan ooit tevoren.

Auteurs die systematische reviews uitvoeren, worden gedurende het hele systematische reviewproces geconfronteerd met problemen. Het is moeilijk en tijdrovend om te zoeken en op te halen, gegevens te verzamelen, manuscripten te schrijven en statistische analyses uit te voeren.

Dit proefschrift bevat zeven artikelen (hoofdstukken 6–8, 11, 12, 15, en 16), gepubliceerd tussen 2017 en 2019. In deze artikelen hebben we methoden onderzocht om systematische reviews sneller, goedkoper en efficiënter uit te voeren. Tegelijkertijd vereisen systematische reviews nog steeds een grondige, objectieve en reproduceerbare methodologie om vertekening te voorkomen. Terwijl we probeerden het proces doelmatiger te maken, streefden we er naar om dezelfde methodologische striktheid van het proces te handhaven.

Eerder onderzoek was meestal gericht op de automatisering van de screening op basis van titel en samenvatting van de referenties. Er is relatief minder onderzoek gedaan naar het automatiseren van de andere onderdelen van het reviewproces, waaronder het ophalen van referenties, het screenen van fulltext referenties, het extraheren van gegevens, gegevenssynthese, en de analysefase. Bovendien is het eerdere onderzoek grotendeels gericht op systematische reviews van interventies en slechts weinig op systematische reviews van diagnostische accuratesse.

Zoekstrategieën om diagnostische onderzoeken te identificeren, hebben vaak een lage nauwkeurigheid en hun gebruik in systematische reviews wordt ontmoedigd. Dit resulteert in relatief grote aantallen referenties die op basis van titel en samenvatting in eerste instantie geselecteerd zullen worden voor verdere beoordeling. Systematische reviews van diagnostische accuratesse zijn daarom een belangrijk doelwit voor de ontwikkeling van alternatieve benaderingen om de snel toenevende werklast het hoofd te bieden.

In dit proefschrift hebben we onderzocht hoe geautomatiseerde leermethoden kunnen worden gebruikt om deze werklast te verminderen, hoe dergelijke methoden kunnen werken en hoe ze kunnen worden ingepast in verschillende systematische reviewcontexten en -instellingen.

---

This section was translated by a non-native speaker (CN) with computer assisted translation using a semi-automated deep neural machine translator (<https://www.deepl.com/translator>). The text was subsequently post-edited by a native speaker (ML).

## 19.1 AUTOMATISERING VAN SCREENING

In hoofdstuk 6–8 presenteerden we drie onderzoeken over methoden om het screeningsproces in systematische reviews te automatiseren, met bijzondere aandacht voor systematische reviews van diagnostische accuratesse.

Van alle referenties die zijn gevonden met het de initiële zoektocht in elektronische databases, kan het zo zijn dat slechts één op de duizend in het systematische review wordt opgenomen. Dit betekent dat er een extreme onbalans is tussen de aantallen relevante en niet-relevante referenties, hetgeen één van de technische uitdagingen is waarmee rekening gehouden moet worden wanneer we het screeningsproces van diagnostische reviews automatiseren.

Onder de 50 systematische reviews die wij hebben geanalyseerd, was het mediane aantal opgenomen onderzoeken 14 (bereik: 0 tot 99). Elf van de reviews omvatten vier onderzoeken of minder. De trainingsgegevens die we kunnen verwachten van veel systematische reviews liggen dan ook ver onder de aantallen die nodig zijn om de dataverzadiging in automatische leermodellen te bereiken. Het is onduidelijk hoe we deze modellen effectief kunnen trainen op dergelijke beperkte gegevens, met name om (bijna) alle beschikbare relevante artikelen terug te vinden, zoals gewoonlijk vereist is bij systematische reviews.

253

19.1.1 *Verschillen tussen referenties geselecteerd op basis van de samenvatting en op basis van de volledige tekst*

Artikelen die in de uiteindelijke review zijn opgenomen en referenties die op basis van de titel en samenvatting waren meegenomen, maar later op basis van de volledige tekst waren uitgesloten, verschilden onvoldoende in termen van taal of woordkeuze om door algemene algoritmen voor machinaal leren, zoals logistieke regressie, op basis van titel en samenvatting van elkaar te onderscheiden. Trainingsgegevens waarbij gouden standaardlabels gebaseerd waren op titels, samenvattingen en volledige tekst lijken echter te leiden tot lichte resultaatverbeteringen ten opzichte van opleidingsgegevens waarbij gouden standaardlabels alleen op titels en samenvattingen gebaseerd waren. De verschillen bleken echter gering te zijn.

19.1.2 *Gebruik van trainingsgegevens uit beide screeningsfasen*

Voor trainingsgegevens moet bij voorkeur gebruik worden gemaakt van labels op basis van volledige tekst, die - indien dergelijke labels beschikbaar zijn - kunnen fungeren als gouden standaard. Van de 50 onderzochte systematische reviews omvatten er echter slechts 19 ten minste 20 onderzoeken en in de praktijk zijn de beste

kwaliteitsvoorbeelden van opgenomen onderzoeken wellicht te weinig om effectief te gebruiken. Om goed presterende automatiseringsmethoden voor het screeningsproces van diagnostische systematische reviews te ontwikkelen, kan het daarom nodig zijn om de gegevens aan te vullen met gouden standaardlabels die alleen op titels en samenvattingen zijn gebaseerd. We hebben ook resultaatverbeteringen waargenomen bij het aanvullen van trainingsgegevens met trainingsgegevens uit soortgelijke systematische review-onderwerpen (d.w.z. door gebruik te maken van transductieleren).

Om te profiteren van de gouden standaardlabels uit beide screeningfasen en om het leren van soortgelijke onderwerpen over te dragen, hebben we een stacking-model gepresenteerd, dat gebruik maakt van meta-regressie naar gecombineerde beslissingen van meerdere modellen.

254

### 19.1.3 *Screening benaderingen*

We hebben drie verschillende modellen gepresenteerd, voor verschillende contexten.

We hebben een statisch model gepresenteerd, getraind in de inclusie/exclusiebeslissingen van referenties die in eerdere systematische reviews zijn gescreend. Dit model vereist dus dat de trainingsgegevens beschikbaar zijn op het moment dat de screening wordt gestart. Dit beperkt de toepasbaarheid van dit model tot systematische review updates, of om algemene modellen te trainen, bv. om algemene diagnostische onderzoeken te identificeren.

We hebben een actief model voorgesteld, dat gebruik maakt van actief leren om de resultaten tijdens de screening te verbeteren. Dit model vereist geen trainingsgegevens op het moment dat de screening wordt gestart en kan daarom ook worden gebruikt in systematische reviews die de novo worden uitgevoerd, waar geen trainingsgegevens beschikbaar zijn. Dit proces kan van nul af aan worden gestart, zonder enige trainingsgegevens, maar als er een bepaalde hoeveelheid trainingsgegevens beschikbaar is, of als de trainingsgegevens kunstmatig kunnen worden geconstrueerd, kunnen deze gegevens als uitgangspunt worden gebruikt.

We hebben een stackingmodel gepresenteerd, dat de statische en actieve modellen combineert om het beste van beide te bereiken. Het gebruikt een statisch (interthematisch: over meerdere reviews heen) model als basis en gebruikt vervolgens de meer gerichte intrathematische (binnen het review waaraan gewerkt wordt) gegevens die tijdens het screeningproces zijn verzameld om het model verder te verbeteren door actief te leren.

## 19.2 RESULTATEN

In de hoofdstukken 6–8 hebben we de resultaten van onze modellen vergeleken met de huidige state-of-the-art.

19.2.1 *Statistisch model (intrathematisch)*

Ons statische model presteerde beter dan de toenmalige state-of-the-art op het gebied van wss@95 (gemiddeld 0,392), maar slechter op het gebied van AUC. Dit suggereert dat ons model goed werkt voor het vinden van alle relevante onderzoeken, terwijl de concurrerende benaderingen beter zijn in het vinden van de eerste relevante onderzoeken, maar moeite hebben om de laatste te vinden.

Er zijn nog twee onderzoeken gepubliceerd waarin vijf nieuwe modellen op deze dataset worden geëvalueerd, en die sindsdien de state-of-the-art verder hebben ontwikkeld. De twee onderzoeken rapporteren gemiddelde wss@95 variërend van 0,347 tot 0,408. Ons model presteerde beter dan de gerapporteerde resultaten op 5, 6, 8, 9 en 10 van de 15 onderwerpen. Bijgevolg zijn de resultaten van ons model nog steeds gunstig in vergelijking met de huidige state-of-the-art.

19.2.2 *Statistisch model (interthematisch)*

Ons statische model presteerde beter dan het state-of-the-art model over alle onderwerpen voor transductieleren. We zijn niet op de hoogte van latere onderzoeken die het transductieleren onderzoeken, en dit model lijkt dus het huidige state-of-the-art model te blijven.

Ondanks de eenvoud van de aanpak, leverde de statische benadering in combinatie met transductieleren vaak resultaten op die vergelijkbaar zijn met de huidige state-of-the-art actieve leermodellen.

19.2.3 *Actief leren*

De beste actieve leerbenadering presteerde consequent beter dan de transductie-leerbenadering, hoewel de verschillen bescheiden waren.

19.2.4 *Stackingmodel*

Het stackingmodel combineert de hoge initiële resultaten van de transductie-leerbenadering met de resultaatverbeteringen die in de loop van de tijd kunnen worden bereikt door actief leren. In onze experimenten was het stackingmodel het best presterende model voor systematische reviews die de novo werden uitgevoerd, met

kleine resultaatverbeteringen ten opzichte van het statische interthematische model en het standaard actieve leren.

### 19.3 SCREENINGREDUCTIE COMPATIBEL MET DE HUIDIGE PRAKTIJK

In hoofdstuk 11 hebben we een verkennend onderzoek gepresenteerd waarin we het gebruik van de statische methode hebben gedocumenteerd in de update van 2019 van de COMET-database. In dit onderzoek hebben we de cut-off met terugwerkende kracht bepaald op basis van eerdere review-updates. We hebben een drempel vastgesteld die zou hebben geleid tot een aanvaardbaar evenwicht tussen vermindering van de werklast en de volledigheid van de screening in eerdere review-updates, en we hebben dit criterium toegepast in de screening voor de review-update van 2019.

Wij waren van mening dat het ontbreken van 2% van de potentieel relevante referenties een aanvaardbare afweging was voor een vermindering van de werklast met 75%. Dit specifieke review is bedoeld om een literatuurdatabank te vullen, en het fangst is daarom een directe en passende metriek voor de impact die de screeningautomatisering had op het review.

Als een artikel in de databases alleen is opgenomen met titel en niet met samenvatting, dan kan deze niet worden gerangschikt met aanvaardbare resultaatgaranties. Daarom kwamen referenties zonder samenvatting niet in aanmerking voor automatisering van de screening. Er was echter slechts een klein aantal van dergelijke referenties, wat overeenkomt met een werklast van ongeveer 2-4 uur per screener als deze handmatig gescreend moeten worden.

Omdat we het model prospectief toepasten, konden we het verlies in fangst alleen maar schatten. Relevante onderzoeken voor opname in de COMET-database worden echter uit meerdere bronnen geïdentificeerd en de tijd zal uitwijzen of het geschatte aantal gemiste referenties overeenkomt met het aantal onderzoeken dat daadwerkelijk gemist werd. Een kleine steekproef (1%) van de uitgesloten referenties werd gescreend om de resultaten te verifiëren, en al deze referenties bleken correct te zijn uitgesloten.

We hebben geprobeerd ons zoveel mogelijk aan het vastgestelde proces te houden. We gebruikten een screening reductie aanpak waarbij de inclusiedrempel kon worden bepaald als onderdeel van het protocol. We pasten het model toe voordat we met de screening begonnen, en de gegevens werden gerandomiseerd om te voorkomen dat de rangorde vooringenomenheid voor de screening zou optreden. De handmatige screening werd vervolgens uitgevoerd zoals gewoonlijk in EndNote. In tegenstelling tot voorgaande jaren waren er bij de update van 2019 slechts twee screeners betrokken, maar de rest van het proces was ongewijzigd voor het gebruik van screeningreductie. De screeners hadden geen gespecialiseerde software nodig.

19.3.1 *Betere metrieke voor screeningautomatisering*

In hoofdstuk 12 hebben we geprobeerd om het ‘informatieverlies’ direct te meten voor systematische reviews van de diagnostische accuratesse. We hebben geprobeerd om te onderzoeken wat het betekent voor een verkorte methode om ‘dezelfde’ systematische review te krijgen als bij een uitgebreide screening.

We probeerden met de metriek aan drie criteria te voldoen:

- ❖ De metriek moet via het screeningproces cumulatief kunnen worden berekend
- ❖ Het zou mogelijk moeten zijn om met de screening te stoppen zodra we er vertrouwen in hebben dat verdere screening de conclusies van de evaluatie niet zal veranderen; en
- ❖ Het moet mogelijk zijn om in het kader van het reviewprotocol criteria voor stopzetting vast te stellen om vertekening te voorkomen

Als methoden voor het automatiseren van het screeningsproces worden gebruikt bij systematische reviews, moeten beoordelaars inschatten hoeveel en wat voor soort verlies aanvaardbaar is. Een verlies in fangst of volledigheid kan de indruk wekken dat niet aan de criteria van een systematische review is voldaan, maar hoeft geen significante invloed te hebben op de resultaten en conclusies van het review – op voorwaarde dat een voldoende selectie van onderzoeken wordt geïdentificeerd om alle evaluatievragen aan te pakken, en op voorwaarde dat de selectie van onderzoeken onbevooroordeeld is. Om te voorkomen dat de beoordelaar vooringenomenheid en ad-hocbeslissingen tijdens het screeningproces meeneemt, moet er een duidelijk, vooraf gespecificeerd protocol zijn om te beoordelen wanneer de screening is voltooid.

Hiertoe hebben we getracht de nauwkeurigheid van de meta-analyse te gebruiken als resultaatmetriek tijdens de screening, door cumulatieve meta-analyses uit te voeren via het screeningsproces. Deze nauwkeurigheid kan prospectief worden ingeschat en drempels kunnen worden bepaald als onderdeel van het protocol. Dit vereist echter dat het screeningsproces parallel met de dataextractie, synthese en meta-analyse van het systematische reviewproces wordt uitgevoerd, en zou dus resulteren in een onconventioneel systematisch reviewproces.

Het voordeel van de metriek is dat deze conservatief en betrouwbaar is, en het onderbreken van de screening zodra de nauwkeurigheid binnen de vooraf vastgestelde grenzen valt, zal waarschijnlijk niet leiden tot verkeerde resultaten of conclusies in de systematische review. Bovendien kan het onderzoek daardoor veel eerder in het proces worden onderbroken en kan de werkbelasting veel eerder worden verminderd dan bij conventionele stopcriteria.

## 19·4 GEGEVENSEXTRACTIE &amp; SYNTHESE

In deel IV presenteren wij een dataset die de gegevensextractie, de gegevenssynthese, en de meta-analysestadiën van de systematische reviews van diagnostische accuratesse van de test documenteert, die 63 systematische reviews van 1.738 diagnostische accuratesseonderzoeken omvat. In totaal, omvat het 589 meta-analyses van 5.848 diagnostische testevaluaties. Wij hopen het van hulp voor beter begrip zal zijn voor hoe het proces door menselijke reviewers, evenals voor het modelleren van het proces met geautomatiseerde methodes wordt uitgevoerd.

Hoewel verschillende onderdelen van het proces denkbaar geautomatiseerd kunnen worden, is alleen de gegevensextractie een natuurlijk taalverwerkingsprobleem, dat in deze dissertatie aan de orde is gekomen.

19·4·1 *Geautomatiseerde gegevensextractie van diagnostische onderzoeken*

Aangezien de meeste eerdere methoden zich hebben gericht op interventies, is er een gebrek aan relevante methoden voor systematische reviews van de diagnostische accuratesse. De enige direct toepasbare eerdere methoden halen de belangrijkste conclusies en taal.

In hoofdstuk 16 hebben we geprobeerd om automatisch de index test, aandoening en referentiestandaard te extraheren uit diagnostische accuratesseonderzoeken. Deze items zijn niet eerder overwogen door gegevensextractiemethoden, en zijn belangrijk om te extraheren in systematische reviews, omdat ze de kern van een accuratessestudie vormen. Ze zijn daarom vaak nodig om te bepalen of het onderzoek/artikel/referentie? moet worden opgenomen in de evaluatie, maar ook om te bepalen in welke meta-analyses de onderzoeken zullen worden opgenomen.

We hebben modellen geïmplementeerd die gebruik maken van zowel logistische regressie als *deep learning* met BioBERT. Ook hebben we het gebruik van data preprocessing vergeleken door medische termen te vervangen door de bijbehorende semantische UMLS-typen. Geen enkele methode was steeds beter dan de andere methoden.

De verschillen in annotaties tussen de modellen en menselijke annotators waren vergelijkbaar met de verschillen tussen de menselijke annotators onderling. Met andere woorden, de geautomatiseerde methode was ongeveer net zo nuttig als het labelen? van de resultaten door een tweede expert.

Toezicht op afstand gaf ruwweg dezelfde resultaten als direct toezicht, vermoedelijk als gevolg van de veel grotere hoeveelheden gegevens die beschikbaar zijn via deze methode.

De overeenstemming tussen de menselijke annotators was constant laag, zelfs na verschillende rondes van aanpassingen aan de annotatie-instructies. Een deel van

de reden lijkt te zijn dat het moeilijk is om duidelijk af te bakenen welke zin een-  
duidig de kenmerken (index test, aandoening, referentiestandaard) specificeert of  
niet specificeert. Vaak is dit een oordeelsvorming met een grote mate van subjec-  
tiviteit. Bovendien leidt de volledigheid van het systematische reviewproces tot  
systematische reviews die vaak ook onderzoeken met een zeer wisselende kwali-  
teit van de verslaglegging omvatten. Eén van de referenties die in de handmatig  
geannoteerde steekproef zijn opgenomen, was volgens de annotators bijvoorbeeld  
geen diagnostisch accuratessestudie.

#### 19.4.2 *Geautomatiseerde meta-analyses voor systematische reviews van diagnostische accu- ratesse*

259

In hoofdstuk 15 ontwikkelden we een pijplijn waarin de gegevens uit diagnostische  
accuratesseonderzoeken in tabelvorm kunnen worden gebruikt voor het uitvoeren  
van meta-analyses, zonder menselijke tussenkomst.

Meta-analysemodellen voor diagnostische accuratesse worden te complex geacht  
om in RevMan te implementeren en er is dan ook geen pijplijn om automatisch  
meta-analyses uit te voeren vanuit RevMan. Als gevolg hiervan moeten review  
auteurs meta-analyses uitvoeren in externe software. Dit verhoogt de werklast en  
kan tot fouten leiden.

Om te testen of het berekenen van meta-analyses in externe software tot fouten  
leidt, hebben we onze geautomatiseerde meta-analyse pijplijn gebruikt om de ge-  
rapporteerde meta-analyses in de 63 systematische reviews opnieuw te berekenen.  
Gemiddeld hebben we ongeveer 2% afwijkingen waargenomen ten opzichte van  
de gerapporteerde resultaten, en dit hebben we gebruikt als indicatie van de on-  
dergrens van de nauwkeurigheid die mag worden verwacht van geautomatiseerde  
screeningmethoden in een van onze onderzoeken (hoofdstuk 12).

Verder hebben we mogelijke fouten in de summary-of-findingstabellen in de sys-  
tematische reviews opgespoord. Over 103 in aanmerking komende meta-analyses  
hebben we in totaal twee fouten geïdentificeerd.

Een van de fouten lijkt te wijten te zijn aan een copy-paste fout, waarbij de auteurs  
van de review dezelfde meta-analyseresultaten hebben gekopieerd voor twee ver-  
schillende diagnostische tests. Deze fout had op verschillende manieren kunnen  
worden gesignaleerd: 1) onze gerepliceerde score verschilde met meer dan 10 pro-  
centpunten, 2) twee rijen in de summary-of-findings waren identiek, 3) in de tabel  
met de summary-of-findings in het review beschreven een ander aantal opgeno-  
men onderzoeken dan wat werd gerapporteerd in 'data and analyses', en 4) in de  
summary-of-findingstabel in de evaluatie werd een ander aantal deelnemers be-  
schreven dan werd gerapporteerd in de 'data and analyses'.

De tweede fout blijkt ook een copy-paste fout te zijn: '74,7 [85,2; 82,3]' in plaats van

‘74,7 [65,2; 82,3]’. Deze fout kan ook op verschillende manieren zijn gesignaleerd: 1) onze gerepliceerde samenvattende schatter verschilde opnieuw met meer dan 10 procentpunten, 2) [85,2; 82,3] is geen geldig betrouwbaarheidsinterval: de lagere grens is hoger dan de hogere grens, en 3) het gemiddelde 74,7 ligt buiten het interval [82,3; 85,2].

Al met al komen deze fouten zelden voor in de summary-of-findingstabellen (2 op 103 in aanmerking komende meta-analyses in 63 reviews), maar beide fouten hadden automatisch kunnen worden gesignaleerd met behulp van zeer eenvoudige methoden. Een andere mogelijkheid zou zijn geweest om deze fouten te voorkomen door automatisch summary-of-findingstabellen uit de gegevens in te vullen.

## 19.5 CONCLUSIES

We hebben een geautomatiseerd screeningsstelsel gepresenteerd dat kan worden gebruikt in een verscheidenheid van systematische reviewcontexten – variërend van review-updates tot reviews die de novo worden uitgevoerd. Het stelsel is algemeen in doel, en presteert goed op verscheidene types van systematische reviews, met inbegrip van systematische reviews van diagnostische accuratesse. Het stelsel is bovendien zeer aanpasbaar, en de onderliggende voorbewerking pijplijn en classificatie- of rangschikkingsalgoritmen kunnen worden gewijzigd om het stelsel te verfijnen voor specifieke systematische review-onderwerpen of contexten. Methoden die het proces van systematische reviews automatiseren kunnen worden gebruikt zonder het proces fundamenteel te veranderen. De screeningsreductiemethoden kunnen gebruikt worden als een extra zoekfilter, waardoor de rest van het reviewproces identiek blijft aan het conventionele proces, met inbegrip van screening in willekeurige volgorde, en het gebruik van standaard referentiemanagers zoals EndNote.

De nauwkeurigheid van het screeningsproces en de impact ervan op de resultaten en conclusies van de review kan prospectief gemeten worden via het screeningsproces met behulp van cumulatieve meta-analyses. Dit vereist een aanpassing van het systematische reviewproces om gegevensextractie en meta-analyses gelijktijdig uit te voeren, maar kan leiden tot aanzienlijke verbeteringen ten opzichte van de traditionele stopcriteria voor screeningautomatisering.

Geautomatiseerde gegevensextractie kan net zo goed presteren als menselijke deskundigen, zelfs met heuristisch geannoteerde trainingsgegevens.

De handmatige verwerking van gegevens bij het uitvoeren van meta-analyses en de beperkte integratie tussen review managers zoals RevMan en externe software zijn verondersteld fouten te veroorzaken in systematische reviews. We vonden slechts 2 gevallen van dergelijke fouten. Beide fouten hadden echter voorkomen of ontdekt kunnen worden met behulp van zeer eenvoudige controles van de gege-

vensconsistentie.

Belangrijke obstakels in de systematische reviewautomatisering zijn onder meer de wisselende kwaliteit van de rapportage en het relatief hoge aantal referenties uit niet-*mainstream* tijdschriften en uitgevers, alsook grijze literatuur. Betere en consistentere rapportage in primaire onderzoeken zou waarschijnlijk leiden tot een betere overeenkomsten tussen de beoordelaars en tot betere resultaten op het gebied van gegevensextractie met geautomatiseerde methoden. Deze obstakels belemmeren niet alleen het gebruik van screeningautomatisering en geautomatiseerde gegevensextractie, maar tasten ook de resultaten van bestaande geautomatiseerde full-text retrieval-methoden en tekstconverters zoals Grobid sterk aan.

La demande et la production de revues systématiques augmentent rapidement. 17 254 nouvelles revues systématiques ont été indexées dans PubMed en 2018 seulement, et ce nombre a plus que quintuplé depuis 2009. Bien que la demande en revues systématiques augmente, le nombre de publications que les revues systématiques doivent prendre en compte augmente également à un rythme effréné. Nous consacrons aujourd'hui plus de temps et d'argent que jamais à la production de nouvelles revues systématiques.

Les auteurs de revues systématiques font face à des problèmes tout au long du processus production des revues. Il est difficile et fastidieux d'identifier les articles pertinents, d'en récupérer le texte intégral, de recueillir les données, d'effectuer les analyses statistiques afférentes, et de rédiger le manuscrit final.

Cette thèse a donné lieu à sept articles (chapitres 6–8, 11, 12, 15, et 16) publiés entre 2017 et 2019. Dans ces documents, nous avons tenté d'explorer des méthodes permettant d'effectuer les revues systématiques plus rapidement, à moindre coût et plus efficacement. En même temps, les revues systématiques exigent toujours une méthodologie consciencieuse, objective et reproductible pour éviter tout biais. Tout en essayant d'accélérer le processus, nous nous efforçons également de maintenir la même rigueur méthodologique du processus. La plupart des travaux antérieurs ont porté sur l'automatisation de la présélection à l'étape de la présélection des titres et des résumés.

Comparativement, peu de travaux ont porté sur l'automatisation des autres parties du processus de production des revues systématiques, y compris l'extraction des articles, la sélection des articles, l'extraction des données, la synthèse des données et les étapes de l'analyse. De plus, les travaux existant se sont focalisés sur les revues systématiques consacrées aux interventions, et peu de travaux antérieurs ont porté sur les revues systématiques consacrées à l'exactitude des tests diagnostiques.

Les requêtes soumises à des moteurs de recherche pour identifier les études diagnostiques ont tendance à être peu précises et leur utilisation est découragée dans le cadre de revues systématiques. Il en résulte un nombre relativement plus élevé de références d'articles candidats à la présélection. Les revues systématiques consacrées à l'exactitude des tests diagnostiques peuvent donc être une cible de choix pour des approches alternatives permettant de faire face à l'augmentation rapide de la charge de travail.

Dans cette thèse, nous avons examiné comment les méthodes d'apprentissage automatique peuvent être utilisées pour réduire cette charge de travail, comment ces

---

This section was translated by a non-native speaker (CN) with computer assisted translation using a semi-automated deep neural machine translator (<https://www.deepl.com/translator>). The text was subsequently post-edited by a native speaker (AN).

méthodes peuvent fonctionner et comment elles peuvent s'intégrer dans différents contextes et paramètres des revues systématiques.

## 20.1 AUTOMATISATION DE LA SÉLECTION D'ARTICLES

Dans les chapitres 6–8, nous avons présenté trois articles sur les méthodes d'automatisation de la sélection d'articles pour les revues systématiques, avec un accent particulier sur les revues systématiques consacrées à l'exactitude des tests diagnostiques.

Les méthodes d'automatisation de la sélection doivent faire face à plusieurs contraintes techniques, dont un déséquilibre de classe extrême. Les revues systématiques consacrées à l'exactitude des tests diagnostiques peuvent inclure un seul article sur mille initialement issus de la recherche par requête dans les bases de données.

De plus, parmi les 50 revues systématiques de notre étude, le nombre médian d'articles inclus était de 14 (intervalle : 0 à 99). Onze des revues systématiques comprenaient quatre articles inclus ou moins. Les données d'entraînement que l'on peut tirer de nombreuses revues systématiques sont donc bien en deçà des volumes nécessaires pour atteindre la saturation des données dans les modèles d'apprentissage. On ne sait pas très bien comment entraîner efficacement les modèles d'automatisation de la sélection sur des données aussi limitées, en particulier pour atteindre le rappel presque parfait généralement requis par les revues systématiques.

263

### 20.1.1 *Différences entre les articles inclus d'après le résumé et le texte intégral*

Les articles inclus dans une revue systématique et les articles provisoirement inclus sur la base du résumé mais exclus sur la base du texte intégral ne sont pas suffisamment différents en termes de contenu linguistique ou de choix de mots pour que des algorithmes génériques d'apprentissage tels que la régression logistique puissent distinguer les deux catégories en fonction du titre et du résumé. Toutefois, les données d'entraînement pour lesquelles le gold-standard était fondées sur les titres, les résumés et le texte intégral semblent offrir de meilleures performances par rapport aux données pour lesquelles le gold-standard était fondées uniquement sur les titres et les résumés. Les différences semblaient cependant mineures.

### 20.1.2 *Utilisation des données d'entraînement des deux étapes de sélection*

Les données d'entraînement devraient reposer de préférence sur un gold standard issu du texte intégral, si cette information est disponibles. Toutefois, sur les 50 revues systématiques utilisées dans notre étude, seulement 19 comprenaient au moins 20 articles inclus. Dans la pratique, les exemples d'articles inclus de la meilleure

qualité peuvent s'avérer trop rares pour être utilisés efficacement. Pour élaborer des méthodes d'automatisation de la sélection performantes pour les revues systématiques consacrées à l'exactitude des tests diagnostiques, il peut donc être nécessaire de compléter les données par un gold standard fondé uniquement sur les titres et les résumés des articles. Nous avons également observé des améliorations de performance en complétant les données d'entraînement sur une thématique par des données d'entraînement issues de revues systématiques consacrées à une thématique similaire (c.-à-d. en utilisant l'apprentissage par transfert).

Pour tirer parti du gold standard issus des deux étapes de sélection préliminaire ainsi que de l'apprentissage par transfert sur des sujets similaires, nous avons présenté un modèle de stacking, qui utilise la méta-régression à des décisions combinées à partir de modèles multiples.

### 20.1.3 *Approches de sélection d'articles*

Nous avons présenté trois modèles différents, pour des contextes de revues systématiques légèrement différents.

Nous avons présenté un modèle statique, fondé sur les décisions d'inclusion/exclusion d'articles sélectionnés lors d'éditions précédentes d'une même revue systématique. Ce modèle nécessite donc que les données d'entraînement soient disponibles au moment où le processus de sélection est amorcé. Cela limite généralement l'applicabilité de ce modèle à la mise à jour des revues systématiques ou à l'entraînement de modèles génériques, par exemple pour identifier des articles sur l'exactitude des tests diagnostiques.

Nous avons présenté un modèle dynamique, qui utilise l'apprentissage actif pour améliorer ses performances tout au long du tri. Ce modèle ne nécessite pas de données d'entraînement disponibles au début de la sélection et peut donc être utilisé également dans des revues systématiques portant sur de nouvelles thématiques pour lesquelles aucune donnée d'entraînement ad-hoc ne sont disponibles. Ce processus peut être commencé sans aucune donnée d'entraînement. Cependant, si des données d'entraînement sont disponibles ou peuvent être construites artificiellement, ces données peuvent être utilisées comme point de départ.

Nous avons présenté un modèle de stacking, qui combine les modèles statique et dynamique pour tirer parti de leur complémentarité. Le modèle statique (inter-thématique) est utilisé comme base et les données intra-thématiques plus ciblées recueillies dans le cadre du processus de sélection sont ensuite utilisées pour améliorer le modèle grâce à l'apprentissage actif.

## 20.2 PERFORMANCE DE LA SÉLECTION AUTOMATIQUE

Dans les chapitres 6 à 8, nous avons comparé la performance de nos modèles avec l'état de l'art actuel.

### 20.2.1 *Modèle statique (intra-thématique)*

Lorsque nous avons construit notre modèle statique, il offrait de meilleures performances que l'état de l'art en termes de « Worked Saved Under Sampling » au niveau de rappel de 0,95 : wss@95 (0,392 en moyenne), mais se positionnait en dessous de l'état de l'art en termes d'AUC. Cela suggère que notre modèle fonctionne bien pour identifier toutes les études pertinentes, alors que les approches concurrentes sont meilleures pour trouver les premières études pertinentes, mais peinent à identifier les dernières.

Deux autres études ont été publiées pour évaluer cinq nouveaux modèles de cet ensemble de données et ont depuis amélioré l'état de l'art. Ces deux nouvelles études font état d'un wss@95 compris entre 0,347 et 0,408 en moyenne. Notre modèle a donné de meilleurs résultats que les résultats rapportés sur 5, 6, 8, 9 et 10 des 15 sujets. Ainsi, la performance de notre modèle se compare toujours favorablement à l'état actuel de l'art.

### 20.2.2 *Modèle statique (inter-thématique)*

Les résultats de notre modèle statique ont été meilleurs que l'état de l'art dans tous les domaines de l'apprentissage par transfert. Nous n'avons pas connaissance d'autres études portant sur l'apprentissage par transfert, et ce modèle semble donc demeurer à la pointe de la technologie.

Malgré sa simplicité, l'approche statique combinée à l'apprentissage par transfert a souvent donné des performances comparables aux modèles d'apprentissage actif les plus performants.

### 20.2.3 *Apprentissage actif*

La meilleure approche d'apprentissage actif a constamment surpassé l'approche d'apprentissage par transfert, bien que les différences soient modestes.

### 20.2.4 *Modèle de Stacking*

Le modèle de stacking combine les performances élevées de l'approche d'apprentissage par transfert et les améliorations de performance qui peuvent être obtenues

au fil du temps grâce à l'apprentissage actif. Dans nos expériences, le modèle de stacking était le modèle le plus performant pour les revues systématiques effectuées de novo, avec de légères améliorations de performance par rapport au modèle inter-thématique statique et l'apprentissage actif standard.

### 20·3 RÉDUCTION DE LA CHARGE DE TRAVAIL ASSOCIÉE À LA SÉLECTION D'ARTICLES COMPATIBLE AVEC LA PRATIQUE

266

Le chapitre chapter :database présente une étude prospective dans laquelle nous avons documenté l'utilisation de la méthode statique dans la mise à jour 2019 de la base de données COMET, via une revue systématique sur les *Core Outcome Sets*. Dans cette étude, le seuil de sélection du modèle a été déterminé rétrospectivement d'après les mises à jour antérieures de la revue. Nous avons choisi un seuil qui aurait permis d'atteindre un équilibre acceptable entre la réduction de la charge de travail et l'exhaustivité de la sélection dans les mises à jour précédentes et nous avons appliqué ce critère dans l'examen des articles pour la mise à jour de 2019. Nous avons estimé que la perte de 2 % des articles constituait un compromis acceptable pour une réduction de 75 % de la charge de travail. Cette revue vise à alimenter une base de données documentaire, et le rappel est donc une mesure directe et appropriée de l'impact de l'automatisation de la sélection des articles inclus dans la revue – et donc dans la base de données.

Les articles sans résumés ne pouvaient être classés avec une garantie de performance acceptable et n'étaient donc pas candidats à la sélection automatique. Ils ont été examinés entièrement manuellement. Il n'y avait cependant qu'un petit nombre de ces articles, ce qui correspond à une charge de travail d'environ 2 à 4 heures par examinateur.

Comme nous avons appliqué le modèle de façon prospective, nous n'avons pu qu'estimer la perte de rappel. Les études pertinentes à inclure dans la base de données COMET sont toutefois identifiées à partir de sources multiples, et le temps nous dira si le nombre estimé d'articles manqués correspond au nombre d'études qui ont été effectivement manquées. Un petit échantillon (1 %) des articles exclus a été examiné manuellement pour vérifier les prédictions du modèle, et tous ces articles ont été correctement exclus.

L'application de l'automatisation de la sélection a été faite de manière à respecter le plus fidèlement possible le processus établi. Nous avons utilisé une approche de réduction de la sélection où le seuil d'inclusion pouvait être déterminé dans le cadre du protocole. Nous avons appliqué le modèle avant de commencer la sélection, et l'ordre des articles a été randomisé pour éviter tout biais d'inclusion. La sélection a ensuite été effectuée dans EndNote conformément à la pratique habituelle. Contrairement aux années précédentes, la mise à jour de 2019 a été faite

par seulement deux examinateurs. Le reste du processus est demeuré inchangé, à l'exception de l'utilisation de la sélection automatique. Aucun logiciel spécialisé n'était requis pour les examinateurs.

### 20.3.1 *Meilleures mesures pour l'automatisation de la sélection*

Le chapitre 12 présente une étude dans laquelle nous avons essayé de mesurer directement la « perte d'information » pour les revues systématiques portant sur l'exactitude des tests diagnostiques. Nous avons essayé d'examiner ce que cela signifie pour une méthode de sélection automatique d'articles produire la « même » revue systématique qu'un examen exhaustif de l'ensemble des articles.

Nous avons essayé de satisfaire trois critères avec cette mesure :

- ❖ La mesure devrait pouvoir faire l'objet d'un calcul cumulatif dans le cadre du processus de tri
- ❖ Il devrait être possible d'arrêter la sélection dès que nous serons convaincus que la poursuite ne modifiera pas les conclusions de la revue ; et
- ❖ Il devrait être possible de définir un critère d'arrêt dans le cadre du protocole de revue afin d'éviter tout biais.

Pour qu'il soit possible d'utiliser des méthodes d'automatisation de sélection d'articles dans des revues systématiques, les examinateurs doivent juger quel volume et quel type de perte sont acceptables.

Une perte de rappel ou d'exhaustivité peut donner lieu à une revue qui ne « ressemble » pas à une revue systématique, mais peut ne pas avoir d'incidence significative sur les résultats et les conclusions de la revue, pourvu que la sélection d'études soit suffisante pour permettre de répondre à toutes les questions de la revue et que le choix des études soit impartial. Afin d'éviter tout parti pris de la part des examinateurs et des décisions ad hoc pendant le processus de sélection, il devrait y avoir un protocole clair et préétabli pour juger quand la revue préalable est terminée.

À cette fin, nous avons tenté d'utiliser l'exactitude des méta-analyses comme mesure de performance de la présélection, en effectuant des méta-analyses cumulatives au cours du processus de sélection. Cette mesure peut être estimée prospectivement et des seuils peuvent être fixés dans le cadre du protocole. Cela demande toutefois que le processus de sélection soit effectué parallèlement aux étapes d'extraction, de synthèse et de méta-analyse des données, ce qui donnerait lieu à un processus de revue systématique non conventionnel.

L'avantage de la mesure est qu'elle est fiable, et l'interruption de la sélection dès que l'exactitude se situe dans les limites prescrites est peu susceptible d'entraîner

des résultats ou des conclusions erronés dans la revue systématique. De plus, cela permet d'interrompre la sélection beaucoup plus tôt dans le processus et de réduire la charge de travail de plusieurs ordres de grandeur de plus qu'avec les critères d'arrêt conventionnels.

#### 20.4 EXTRACTION & SYNTHÈSE DE DONNÉES

268

Dans la partie IV, nous présentons un corpus documentant les étapes d'extraction de données, de synthèse de données et de méta-analyse pour les revues systématiques portant sur l'exactitude des tests diagnostiques, comprenant 63 revues systématiques sur 1 738 études portant sur l'exactitude des tests diagnostiques. Au total, le corpus comprend 589 méta-analyses de 5 848 évaluations de tests diagnostiques. À notre connaissance, il s'agit du premier corpus partageant ce genre de données avec la communauté scientifique. Nous espérons qu'il aidera à mieux comprendre les étapes mises en œuvre dans l'élaboration des revues systématiques par leur auteurs, ainsi qu'à modéliser le processus à l'aide de méthodes automatisées. Bien que plusieurs parties du processus puissent être automatisées, l'extraction des données est un problème particulièrement intéressant du point du traitement du langage naturel, que nous avons traité dans cette thèse.

##### 20.4.1 *Extraction de données des études portant sur l'exactitude des tests diagnostiques*

La majorité des travaux antérieurs en extraction d'information pour les revues systématiques portaient sur les interventions. Ainsi, aucune méthode disponible n'est directement applicable pour les revues systématiques portant sur l'exactitude des tests diagnostiques. Les seuls éléments présents dans les études de test diagnostiques ayant fait l'objet d'extraction automatique sont les conclusions clés et le langage d'étude.

Dans le chapitre 16, nous avons tenté d'extraire automatiquement le test de référence, la maladie et le gold standard des études portant sur l'exactitude des tests diagnostiques. L'extraction automatique de ces éléments n'a pas été étudiée auparavant. Il est important d'être en mesure de les extraire car ils jouent souvent un rôle similaire aux éléments « PICO » (Patient, Intervention, Comparator, Outcome) dans les essais contrôlés randomisés. Ils sont donc souvent nécessaires pour déterminer si l'article doit être inclus dans la revue systématique, ainsi que pour déterminer dans quelles méta-analyses les études doivent être incluses.

Nous avons mis en œuvre deux modèles utilisant la régression logistique, et l'apprentissage profond utilisant BioBERT. Nous avons également comparé l'utilisation du prétraitement des données en remplaçant les termes médicaux par leurs types sémantiques Dans l'UMLS. Aucune méthode ne s'est montré meilleure que

les autres dans toutes les configurations. Les résultats obtenus étaient néanmoins comparables à l'accord inter-annotateur humain. En d'autres termes, la méthode d'annotation automatique des test de référence, maladie et gold standard est à peu près aussi utile que les annotations faites par un expert humain.

L'apprentissage distant offre des performances équivalentes à celles de l'apprentissage direct, probablement en raison de la quantité beaucoup plus importante de données disponibles grâce à cette méthode.

L'accord inter-annotateur est globalement faible, même après plusieurs cycles d'ajustement du guide d'annotation. Cela s'explique en partie par le fait qu'il est difficile de délimiter clairement les phrases qui précisent sans ambiguïté ou ne précisent pas les caractéristiques de l'étude. Il s'agit souvent d'une question de jugement avec un grand degré de subjectivité. En outre, l'exhaustivité du processus de revue systématique conduit à des revues systématiques incluant souvent des études dont la qualité des rapports est très variable. Par exemple, l'une des études incluses dans l'échantillon annoté manuellement n'était pas une étude portant sur l'exactitude d'un test diagnostique, selon les experts humains.

269

#### 20.4.2 *Méta-analyses automatisées pour les revues systématiques portant sur l'exactitude des tests diagnostiques.*

Au chapitre 15, nous avons présenté les travaux de construction d'une chaîne de traitement dans laquelle les données tabulées des études portant sur l'exactitude des tests diagnostiques peuvent être utilisées pour effectuer des méta-analyses sans intervention humaine.

Les modèles de méta-analyse pour les revues systématiques portant sur l'exactitude des tests diagnostiques sont considérés comme trop complexes pour être implémentés dans l'outil RevMan (utilisé par le centre Cochrane pour la mise en forme de revues systématiques), et il n'y a donc pas de chaîne de traitement pour effectuer automatiquement des méta-analyses à partir de RevMan. Par conséquent, les auteurs de la revue doivent effectuer des méta-analyses dans des logiciels externes. Cela augmente la charge de travail et on a émis l'hypothèse qu'il pourrait en résulter des erreurs.

Pour vérifier si le calcul des méta-analyses dans un logiciel externe entraîne des erreurs, nous avons utilisé la chaîne de traitement automatisée de méta-analyses pour recalculer les méta-analyses rapportées dans les 63 revues systématiques.

En moyenne, nous avons observé des écarts d'environ 2 % par rapport aux résultats originaux, ce qui nous a permis d'établir la limite inférieure de la précision que l'on peut attendre des méthodes de sélection automatique dans une de nos études (chapitre 12).

De plus, nous avons utilisé d'importantes divergences pour déceler les erreurs po-

tentielles dans les tableaux du « summary of findings » des revues systématiques. Nous avons relevé deux erreurs dans 103 méta-analyses admissibles.

L'une des erreurs semble être due à une erreur de copier-coller, où les auteurs de la revue ont copié les mêmes résultats de méta-analyse pour deux tests diagnostiques différents. Cette erreur aurait pu être décelée de plusieurs façons : 1) notre score répliqué était décalé de plus de 10 points, 2) deux lignes du « summary of findings » étaient identiques, 3) le tableau du « summary of findings » de la revue décrit un nombre différent d'études incluses que ce qui a été rapporté dans « data and analyses », et 4) le tableau du « summary of findings » de la revue décrit un nombre de participants différent de celui qui a été rapporté dans « data and analyses ».

La deuxième erreur semble également être une faute de frappe : « 74,7 [85,2 ; 82,3] » au lieu de « 74,7 [65,2 ; 82,3] ». Cette erreur aurait également pu être décelée de plusieurs façons : 1) notre score répliqué était encore une fois décalé de plus de plus de 10 points, 2) [85,2 ; 82,3] n'est pas un intervalle de confiance légal, et 3) la moyenne 74,7 se situe en dehors de l'intervalle [82,3 ; 85,2].

Toutes proportions gardées, de telles erreurs étaient rares dans les tableaux de synthèse des résultats (2 sur 103 méta-analyses admissibles dans 63 revues), mais les deux auraient pu être repérées automatiquement au moyen de méthodes très simples. Par ailleurs, le fait de remplir automatiquement les tableaux de synthèse des résultats à partir des données aurait également permis d'éviter ces erreurs.

## 20·5 CONCLUSIONS

Nous avons présenté un système d'automatisation de la sélection d'articles qui peut être utilisé dans divers contextes de revue systématique, allant de la mise à jour des revues à la mise en œuvre de revues effectuées de novo. Le système est générique et donne de bons résultats pour plusieurs types de revues systématiques, y compris les revues portant sur l'exactitude des tests diagnostiques. De plus, le système est hautement configurable et peut être adossé à des algorithmes de prétraitement et de classification afin d'adapter le système à des sujets ou contextes de revues systématiques précis. La méthode de tri automatique d'articles peut être utilisée dans les revues systématiques sans en modifier fondamentalement le processus. Le filtrage automatique d'articles peut être utilisée comme un filtre de recherche supplémentaire, laissant le reste du processus d'examen identique au processus conventionnel, y compris la présélection dans un ordre aléatoire, et l'utilisation de questionnaires de référence standard comme EndNote.

L'exactitude du processus de sélection et l'impact qu'il a sur les résultats et les conclusions de la revue peuvent être mesurés de façon prospective au moyen de méta-analyses cumulatives dans le cadre du processus de sélection. Pour ce faire, il faut modifier le processus de revue systématique afin d'effectuer l'extraction des

données et les méta-analyses simultanément, mais cela peut mener à des améliorations substantielles par rapport aux critères d'arrêt traditionnels pour l'automatisation de la sélection. L'extraction automatisée des données peut fonctionner à niveau comparable à celui d'experts humains, même avec des données d'entraînement annotées automatiquement à l'aide d'heuristiques.

Des hypothèses ont indiqué que le traitement manuel des données lors des méta-analyses et l'intégration limitée entre les gestionnaires de revue comme RevMan et les logiciels externes pourraient être à l'origine d'erreurs dans les revues systématiques. Nous n'avons trouvé que deux cas d'erreurs de ce genre. Cependant, ces deux erreurs auraient pu être évitées ou repérées à l'aide de contrôles de cohérence des données très simples.

Les principaux obstacles à l'automatisation de la revue systématique sont la qualité variable des rapports et le nombre relativement élevé d'articles provenant de revues et d'éditeurs non traditionnels, ainsi que la littérature grise. Des rapports de meilleure qualité et plus cohérents dans les études primaires permettraient probablement d'accroître l'accord inter-annotateurs ainsi que les performances de l'extraction des données grâce à des méthodes automatisées. Ces obstacles ont un impact négatif sur l'utilisation de l'automatisation pour la sélection d'articles et l'extraction de données, ainsi que sur la performance des méthodes existantes de récupération automatisée du texte intégral à l'aide de convertisseurs PDF en texte comme Grobid.



PART VI

APPENDICES





## PUBLICATION LIST

---

*Several of these publications are full papers published in conferences. Note that conference proceedings are the main venue for publication in computer science, are published in full in the conference proceedings, and are peer reviewed based on the full text.*

### A·1 PEER REVIEWED PUBLICATIONS

275

#### A·1·1 *Letters to Editor*

(**Norman et al., 2017d**): Norman, C., Van Nguyen, T., and Névéol, A. (2017d). Contribution of natural language processing in predicting rehospitalization risk. *Medical care*, 55(8):781

#### A·1·2 *Journal Publications*

(**Norman et al., 2019f**): Norman, C. R., Gargon, E., Leeﬂang, M. M. G., Névéol, A., and Williamson, P. R. (2019f). Evaluation of an automatic article selection method for timelier updates of the COMET core outcome set database. *Database*. Oxford University Press

(**Norman et al., 2019c**): Norman, C., Leeﬂang, M., Porcher, R., and Névéol, A. (2019c). Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. *BMC Systematic Reviews*. Springer Nature

#### A·1·3 *Shared Task Papers*

(**Norman et al., 2017b**): Norman, C., Leeﬂang, M., and Névéol, A. (2017b). LIMS1@CLEF eHealth 2017 task 2: Logistic regression for automatic article ranking. *Working Notes of CLEF*

(**Norman et al., 2018b**): Norman, C., Leeﬂang, M., and Névéol, A. (2018b). LIMS1@CLEF eHealth 2018 task 2: Technology assisted reviews by stacking active and static learning. *Working Notes of CLEF*, pages 10–14

#### A·1·4 *Conference Papers*

([Norman et al., 2018c](#)): Norman, C., Leeﬂang, M., Zweigenbaum, P., and Névéol, A. (2018c). Automating document discovery in the systematic review process: How to use chaff to extract wheat. In Calzolari, N. et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA)

([Norman et al., 2017c](#)): Norman, C., Leeﬂang, M., Zweigenbaum, P., and Névéol, A. (2017c). Tri automatique de la littérature pour les revues systématiques. *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 234–41

([Norman et al., 2019e](#)): Norman, C., Spijker, R., Kanoulas, E., Leeﬂang, M., and Névéol, A. (2019e). A distantly supervised dataset for automated data extraction from diagnostic studies. *ACL BioNLP*

([Norman et al., 2018a](#)): Norman, C., Leeﬂang, M., and Névéol, A. (2018a). Data extraction and synthesis in systematic reviews of diagnostic test accuracy: A corpus for automating and evaluating the process. In *AMIA Annual Symposium Proceedings*, volume 2018, page 817. American Medical Informatics Association

#### A·2 OTHER PUBLICATIONS

##### A·2·1 *Peer Reviewed Conference Abstracts*

([Norman et al., 2019d](#)): Norman, C., Leeﬂang, M., Porcher, R., and Névéol, A. (2019d). Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. In *AMIA Annual Symposium Proceedings*

##### A·2·2 *Other Conference Abstracts*

([Norman et al., 2019b](#)): Norman, C., Leeﬂang, M., Porcher, R., and Névéol, A. (2019b). Does screening automation negatively impact meta-analyses in systematic reviews of diagnostic test accuracy? In *Cochrane Colloquium*

(**Norman et al., 2019a**): Norman, C., Leeflang, M., and Névéol, A. (2019a). Automated checking for human errors in meta-analyses of diagnostic test accuracy. In *Cochrane Colloquium*

A·2·3 *System Demonstrations*

(**Norman et al., 2017a**): Norman, C., Grouin, C., Lavergne, T., Zweigenbaum, P., and Névéol, A. (2017a). Traitement de la langue biomédicale au LIMSI. *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, page 33



## BIBLIOGRAPHY

---

- Alharbi, A., Briggs, W., and Stevenson, M. (2018). Retrieving and ranking studies for systematic reviews: The university of sheffield's approach to CLEF eHealth 2018 task 2. In *CEUR Workshop Proceedings*, volume 2125. CEUR Workshop Proceedings. (Cited on page 55, 76)
- Alharbi, A. and Stevenson, M. (2017). Ranking abstracts to identify relevant evidence for systematic reviews: The university of sheffield's approach to CLEF eHealth 2017 task 2. In *CLEF (Working Notes)*. (Cited on page 55, 73)
- Alharbi, A. and Stevenson, M. (2019). A dataset of systematic review updates. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1257–1260. ACM. (Cited on page 35)
- Allen, I. E. and Olkin, I. (1999). Estimating time to conduct a meta-analysis from number of citations retrieved. *JAMA*, 282(7):634–635. (Cited on page 1, 23, 25, 41)
- Allen, V. B., Gurusamy, K. S., Takwoingi, Y., Kalia, A., and Davidson, B. R. (2013). Diagnostic accuracy of laparoscopy following computed tomography (ct) scanning for assessing the resectability with curative intent in pancreatic and peri-ampullary cancer. *Cochrane Database of Systematic Reviews*, 11. (Cited on page 36, 39, 198)
- Allers, K., Hoffmann, F., Mathes, T., and Pieper, D. (2018). Systematic reviews with published protocols compared to those without: more effort, older search. *Journal of clinical epidemiology*, 95:102–110. (Cited on page 28)
- Van Altena, A. and Olabarriaga, S. D. (2017). Predicting publication inclusion for diagnostic accuracy test reviews using random forests and topic modelling. In *CLEF (Working Notes)*. (Cited on page 55, 71)
- Van Altena, A., Spijker, R., and Olabarriaga, S. (2019). Usage of automation tools in systematic reviews. *Research synthesis methods*, 10(1):72–82. (Cited on page 1, 28, 82, 84)
- Altman, D. G. (1994). The scandal of poor medical research. (Cited on page 17)
- Anagnostou, A., Lagopoulos, A., Tsoumakas, G., and Vlahavas, I. P. (2017). Combining inter-review learning-to-rank and intra-review incremental training for title and abstract screening in systematic reviews. In *CLEF (Working Notes)*. (Cited on page 55, 73, 74)

- Aphinyanaphongs, Y., Tsamardinos, I., Statnikov, A., Hardin, D., and Aliferis, C. F. (2005). Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association*, 12(2):207–216. (Cited on page 168)
- Aromataris, E. and Munn, Z., editors (2017). *Joanna Briggs Institute Reviewer's Manual, 4th edition*. The Joanna Briggs Institute. Available from <https://reviewersmanual.joannabriggs.org/>. (Cited on page 28)
- Bachmann, L. M., Puhan, M. A., Ter Riet, G., and Bossuyt, P. M. (2006). Sample sizes of studies on diagnostic accuracy: literature survey. *Bmj*, 332(7550):1127–1129. (Cited on page 19)
- Bahaadinbeigy, K., Yogesan, K., and Wootton, R. (2010). MEDLINE versus EMBASE and CINAHL for telemedicine searches. *Telemedicine and e-Health*, 16(8):916–919. (Cited on page 30)
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452. (Cited on page 207, 208)
- Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., and Macleod, M. R. (2019). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic reviews*, 8(1):23. (Cited on page 51, 76, 169, 222)
- Banta, H., Behney, C., Andrulis, D., et al. (1978). *Assessing the efficacy and safety of medical technologies*. United States. Congress. Office of Technology Assessment. (Cited on page 14)
- Bastian, H., Glasziou, P., and Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326. (Cited on page 14, 20, 21, 22)
- Bekhuis, T. and Demner-Fushman, D. (2010). Towards automating the initial screening phase of a systematic review. *Studies in Health Technology and Informatics*, 160(PART 1):146–150. (Cited on page 50, 52, 53, 55, 61, 91, 143)
- Bekhuis, T. and Demner-Fushman, D. (2012). Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artificial intelligence in medicine*, 55(3):197–207. (Cited on page 50, 65)
- Bekhuis, T., Tseytlin, E., and Mitchell, K. J. (2015). A prototype for a hybrid system to support systematic review teams: A case study of organ transplantation.

- In 2015 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 940–947. IEEE. (Cited on page 50, 68)
- Bekhuis, T., Tseytlin, E., Mitchell, K. J., and Demner-Fushman, D. (2014). Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PLoS ONE*, 9(1):1–10. (Cited on page 50, 68, 134)
- Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., Ouzzani, M., Thayer, K., Thomas, J., Turner, T., et al. (2018). Making progress with the automation of systematic reviews: principles of the international collaboration for the automation of systematic reviews (icasr). *Systematic reviews*, 7(1):77. (Cited on page 23, 49, 150)
- Beller, E. M., Chen, J. K.-H., Wang, U. L.-H., and Glasziou, P. P. (2013). Are systematic reviews up-to-date at the time of publication? *Systematic reviews*, 2(1):36. (Cited on page 23)
- Beltagy, I., Cohan, A., and Lo, K. (2019). SciBERT: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*. (Cited on page 222)
- Betrán, A. P., Say, L., Gülmezoglu, A. M., Allen, T., and Hampson, L. (2005). Effectiveness of different databases in identifying studies for systematic reviews: experience from the who systematic review of maternal morbidity and mortality. *BMC medical research methodology*, 5(1):6. (Cited on page 30)
- Beynon, R., Leeflang, M. M., McDonald, S., Eisinga, A., Mitchell, R. L., Whiting, P., and Glanville, J. M. (2013). Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. *Cochrane Database of Systematic Reviews*, 9. (Cited on page 2, 31, 32, 80, 164)
- Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl\_1):D267–D270. (Cited on page 229)
- Booth, A. (2010). How much searching is enough? comprehensive versus optimal retrieval for technology assessments. *International journal of technology assessment in health care*, 26(4):431–435. (Cited on page 166)
- Booth, A. (2016a). Evident guidance for reviewing the evidence: a compendium of methodological literature and websites. *University of Sheffield, Sheffield*. (Cited on page 13)
- Booth, A. (2016b). Over 85% of included studies in systematic reviews are on MEDLINE. *Journal of clinical epidemiology*, 79:165–166. (Cited on page 30, 168)

- Borah, R., Brown, A. W., Capers, P. L., and Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545. (Cited on page 22, 25, 41)
- Boudin, F., Nie, J.-Y., Bartlett, J. C., Grad, R., Pluye, P., and Dawes, M. (2010). Combining classifiers for robust pico element detection. *BMC medical informatics and decision making*, 10(1):29. (Cited on page 197, 236)
- De Bruijn, B., Carini, S., Kiritchenko, S., Martin, J., and Sim, I. (2008). Automated information extraction of key trial design elements from clinical trial publications. In *AMIA Annual Symposium Proceedings*, volume 2008, page 141. American Medical Informatics Association. (Cited on page 197, 236)
- Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81. (Cited on page 46, 151)
- Burns, P. B., Rohrich, R. J., and Chung, K. C. (2011). The levels of evidence and their role in evidence-based medicine. *Plastic and reconstructive surgery*, 128(1):305. (Cited on page 15, 16)
- Campbell Collaboration (2019). Campbell systematic reviews: Policies and guidelines. *Campbell Systematic Reviews: Policy and guidelines*. DOI: [10.4073/cpg.2016.1](https://doi.org/10.4073/cpg.2016.1). (Cited on page 28)
- Canadian Medical Association (1979). The periodic health examination. canadian task force on the periodic health examination. (Cited on page 15)
- Castaldi, P. J., Cho, M. H., Cohn, M., Langerman, F., Moran, S., Tarragona, N., Moukhachen, H., Venugopal, R., Hasimja, D., Kao, E., et al. (2009). The copd genetic association compendium: a comprehensive online database of copd genetic associations. *Human molecular genetics*, 19(3):526–534. (Cited on page 55)
- Center for History and New Media, George Mason University. (2019). Zotero. Version 5.0.73 (2019-07-05), available from <https://www.zotero.org/>. (Cited on page 27)
- Chalmers, I. (2003). Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations. *The Annals of the American Academy of Political and Social Science*, 589(1):22–40. (Cited on page 12, 17, 30, 41, 42)

- Chalmers, I. (2007). The lethal consequences of failing to make full use of all relevant evidence about the effects of medical treatments: the importance of systematic reviews. *Treating individuals: from randomised trials to personalized medicine*. London: *The Lancet*, pages 37–58. (Cited on page 16)
- Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A. M., Howells, D. W., Ioannidis, J. P., and Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet*, 383(9912):156–165. (Cited on page 38)
- Chalmers, I., Glasziou, P., and Godlee, F. (2013). All trials must be registered and the results published. (Cited on page 18, 32)
- Chandler, J. and Hopewell, S. (2013). Cochrane methods-twenty years experience in developing systematic review methods. *Systematic reviews*, 2(1):76. (Cited on page 19, 26)
- Chapman, W. W., Nadkarni, P. M., Hirschman, L., D’Avolio, L. W., Savova, G. K., and Uzuner, O. (2011). Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions. (Cited on page 54, 85, 169)
- Chassin, M. R., Kosecoff, J., Solomon, D. H., and Brook, R. H. (1987). How coronary angiography is used: clinical determinants of appropriateness. *JAMA*, 258(18):2543–2547. (Cited on page 13)
- Chen, J., Chen, S., Song, Y., Liu, H., Wang, Y., Hu, Q., He, L., and Yang, Y. (2017). Ecnu at 2017 eHealth task 2: Technologically assisted reviews in empirical medicine. In *CLEF (Working Notes)*. (Cited on page 55, 72)
- Cheng, S., Augustin, C., Bethel, A., Gill, D., Anzaroot, S., Brun, J., DeWilde, B., Minnich, R., Garside, R., Masuda, Y., et al. (2018). Using machine learning to advance synthesis and use of conservation and environmental evidence. *Conservation biology: the journal of the Society for Conservation Biology*, 32(4):762. (Cited on page 74)
- Choi, S., Ryu, B., Yoo, S., and Choi, J. (2012). Combining relevancy and methodological quality into a single ranking for evidence-based medicine. *Information Sciences*, 214:76–90. (Cited on page 55, 64)
- Chung, G. Y.-C. (2009). Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. *Journal of biomedical informatics*, 42(5):790–800. (Cited on page 197)

- Chung, M., Balk, E. M., Ip, S., Raman, G., Yu, W. W., Trikalinos, T. A., Lichtenstein, A. H., Yetley, E. A., and Lau, J. (2009). Reporting of systematic reviews of micronutrients and health: a critical appraisal. *The American journal of clinical nutrition*, 89(4):1099–1113. (Cited on page 55)
- Clarivate Analytics (2019). EndNote. Version x9.2 (2019-06-11), available from <https://endnote.com/>. (Cited on page 27, 33)
- Clarke, M., Alderson, P., and Chalmers, I. (2002). Discussion sections in reports of controlled trials published in general medical journals. *JAMA*, 287(21):2799–2801. (Cited on page 17, 25)
- Cochrane (2014). RevMan. Version 5.3 (2014-06-13), available from <https://community.cochrane.org/help/tools-and-software/revman-5>. (Cited on page 26)
- Cochrane, A. L. et al. (1972). *Effectiveness and efficiency: random reflections on health services*, volume 900574178. Nuffield Provincial Hospitals Trust London. (Cited on page 13, 15, 16)
- Cochrane Community (2019). Search filters. Cochrane Community list of validated search filters for database searches in systematic reviews. Accessed October 2019. Available from <https://community.cochrane.org/search-filters>. (Cited on page 31)
- Cohen, A. M. (2006). An effective general purpose approach for automated biomedical document classification. *AMIA Annual Symposium proceedings*, pages 161–165. (Cited on page 94, 136)
- Cohen, A. M. (2008). Optimizing feature representation for automated systematic review work prioritization. *AMIA Annual Symposium proceedings*, pages 121–5. (Cited on page 55, 57, 93, 94, 95, 97, 104, 109, 113, 136, 145, 226)
- Cohen, A. M. (2011). Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association : JAMIA*, 18(1):author reply 104. (Cited on page 57)
- Cohen, A. M., Ambert, K., and McDonagh, M. (2009). Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. *Journal of the American Medical Informatics Association*, 16(5):690–704. (Cited on page 54, 55, 58, 59, 94, 95, 99, 136, 142, 145)

- Cohen, A. M., Ambert, K., and McDonagh, M. (2010). A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. *AMIA Annual Symposium Proceedings*, 2010:121–125. (Cited on page 54, 55, 56, 94, 136)
- Cohen, A. M., Ambert, K., and McDonagh, M. (2012). Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC medical informatics and decision making*, 12(1):33. (Cited on page 175)
- Cohen, A. M., Bhupatiraju, R. T., and Hersh, W. R. (2004). Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In *TREC*. (Cited on page 154)
- Cohen, A. M., Hersh, W. R., Peterson, K., and Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219. (Cited on page 35, 53, 54, 55, 56, 62, 71, 77, 91, 93, 94, 95, 97, 98, 99, 104, 136, 137, 164, 168, 175, 208)
- Cohen, A. M. and Smalheiser, N. R. (2018). Uic/OHSU CLEF 2018 task 2 diagnostic test accuracy ranking using publication type cluster similarity measures. In *CEUR Workshop Proceedings*, volume 2125. (Cited on page 55, 75)
- Cohen, J. F., Korevaar, D. A., Altman, D. G., Bruns, D. E., Gatsonis, C. A., Hooft, L., Irwig, L., Levine, D., Reitsma, J. B., De Vet, H. C., et al. (2016). Stard 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ open*, 6(11):e012799. (Cited on page 235, 236)
- Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T. J., Hargraves, O., Goss, F., Ide, N., Névél, A., Grouin, C., and Hunter, L. E. (2018). Three dimensions of reproducibility in natural language processing. In *LREC. International Conference on Language Resources and Evaluation*, volume 2018, page 156. N1H Public Access. (Cited on page 207)
- Collberg, C. and Proebsting, T. A. (2016). Repeatability in computer systems research. *Communications of the ACM*, 59(3):62–69. (Cited on page 207)
- Collins, M. (2019). Pubmed updates february 2019. NLM technical bulletin. Available from [https://www.nlm.nih.gov/pubs/techbull/jf19/jf19\\_february\\_pubmed\\_updates.html](https://www.nlm.nih.gov/pubs/techbull/jf19/jf19_february_pubmed_updates.html). (Cited on page 20)
- Cook, D. J., Mulrow, C. D., and Haynes, R. B. (1997). Systematic reviews: synthesis of best evidence for clinical decisions. *Annals of internal medicine*, 126(5):376–380. (Cited on page 1, 12, 18, 25, 41)

- Cormack, G. and Grossman, M. (2014). Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. *SIGIR*, pages 153–162. (Cited on page 67, 68, 91)
- Cormack, G. V. and Grossman, M. R. (2015). Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv preprint arXiv:1504.06868*. (Cited on page 67, 124, 125, 172)
- Cormack, G. V. and Grossman, M. R. (2016). Engineering quality and reliability in technology-assisted review. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 75–84. ACM. (Cited on page 136, 138, 171, 182, 184)
- Cormack, G. V. and Grossman, M. R. (2017). Technology-assisted review in empirical medicine: Waterloo participation in CLEF eHealth 2017. In *CLEF (Working Notes)*. (Cited on page 55, 67, 84, 124, 125, 169)
- Cormack, G. V. and Grossman, M. R. (2018). Technology-assisted review in empirical medicine: Waterloo participation in CLEF eHealth 2018. In *CLEF (Working Notes)*. (Cited on page 55, 67)
- Davidoff, F., Case, K., and Fried, P. W. (1995a). Evidence-based medicine: why all the fuss? *Annals of Internal Medicine*, 122(9):727–727. (Cited on page 1, 12, 15)
- Davidoff, F., Haynes, B., Sackett, D., and Smith, R. (1995b). Evidence based medicine. (Cited on page 15)
- Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM. (Cited on page 117)
- Davis, K., Gorst, S. L., Harman, N., Smith, V., Gargon, E., Altman, D. G., Blazeby, J. M., Clarke, M., Tunis, S., and Williamson, P. R. (2018). Choosing important health outcomes for comparative effectiveness research: An updated systematic review and involvement of low and middle income countries. *PLoS ONE*, 13(2):e0190695. (Cited on page 150, 151, 152)
- Davis-Desmond, P. and Mollá, D. (2012). Detection of evidence in clinical research papers. In *Proceedings of the Fifth Australasian Workshop on Health Informatics and Knowledge Management-Volume 129*, pages 13–20. Australian Computer Society, Inc. (Cited on page 197)

- Dawes, M., Pluye, P., Shea, L., Grad, R., Greenberg, A., and Nie, J.-Y. (2007). The identification of clinically important elements within medical journal abstracts: Patient\_population\_problem, exposure\_intervention, comparison, outcome, duration and results (pecodr). *Journal of Innovation in Health Informatics*, 15(1):9–16. (Cited on page 197, 236)
- De Vet, H., Eisinga, A., Riphagen, I., Aertgeerts, B., Pewsner, D., and Mitchell, R. (2008). Chapter 7: searching for studies. *Cochrane handbook for systematic reviews of diagnostic test accuracy version 0.4 [updated September 2008]*. The Cochrane Collaboration. (Cited on page 164, 166, 222)
- Deeks, J. J., Bossuyt, P. M., and Gatsonis, C. A., editors (2013a). *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. The Cochrane Collaboration. Available from <http://srdta.cochrane.org/>, accessed Aug. 2019. (Cited on page 2, 19, 28, 32, 33, 37, 197, 201, 202, 212, 214, 235, 287, 298, 304, 310)
- Deeks, J. J., Higgins, J. P., and Altman, D. G. (2019). Cochrane handbook for systematic reviews of interventions, chapter 10: Analysing data and undertaking meta-analyses. draft version (29 january 2019). In Higgins et al. (2019). Available from <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 16, 27, 201, 202)
- Deeks, J. J., Wisniewski, S., and Gatsonis, C. A. (2013b). Cochrane handbook for systematic reviews of diagnostic test accuracy, chapter 4: Guide to the contents of a cochrane diagnostic test accuracy protocol. In Deeks et al. (2013a). Available from <http://srdta.cochrane.org/>, accessed Aug. 2019. (Cited on page 28)
- Demner-Fushman, D. and Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103. (Cited on page 197, 236)
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. (Cited on page 222)
- Di Nunzio, G. M., Beghini, F., Vezzani, F., and Henrot, G. (2017). An interactive two-dimensional approach to query aspects rewriting in systematic reviews. ims unipd at CLEF eHealth task 2. In *CLEF (Working Notes)*. (Cited on page 55, 72)
- Di Nunzio, G. M., Ciuffreda, G., and Vezzani, F. (2018). Interactive sampling for systematic reviews. ims unipd at CLEF 2018 eHealth task 2. In *CLEF (Working Notes)*. (Cited on page 55, 72)

- Dickersin, K., Min, Y.-I., and Meinert, C. L. (1992). Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA*, 267(3):374–378. (Cited on page 30)
- Dobrokhoto, P. B., Goutte, C., Veuthey, A.-L., and Gaussier, E. (2005). Assisting medical annotation in swiss-prot using statistical classifiers. *International journal of medical informatics*, 74(2-4):317–324. (Cited on page 168)
- Doebler, P. and Holling, H. (2015). Meta-analysis of diagnostic accuracy with mada. Retrieved from <https://cran.rproject.org/web/packages/mada/vignettes/mada.pdf>. (Cited on page 171, 177, 202, 212, 238)
- Donoso-Guzmán, I. and Parra, D. (2018). An interactive relevance feedback interface for evidence-based health care. In *23rd International Conference on Intelligent User Interfaces*, pages 103–114. ACM. (Cited on page 74)
- Dowell, K. G., McAndrews-Hill, M. S., Hill, D. P., Drabkin, H. J., and Blake, J. A. (2009). Integrating text mining into the MGI biocuration workflow. *Database*, 2009. (Cited on page 149)
- Dubinsky, M. and Ferguson, J. H. (1990). Analysis of the national institutes of health medicare coverage assessment. *International journal of technology assessment in health care*, 6(3):480–488. (Cited on page 14)
- Dunn, A. G., Coiera, E., and Bourgeois, F. T. (2018). Unreported links between trial registrations and published articles were identified using document similarity measures in a cross-sectional analysis of clinicaltrials. gov. *Journal of clinical epidemiology*, 95:94–101. (Cited on page 51)
- Dutta, S., Sur, D., Manna, B., Sen, B., Deb, A. K., Deen, J. L., Wain, J., Von Seidlein, L., Ochiai, L., Clemens, J. D., et al. (2006). Evaluation of new-generation serologic tests for the diagnosis of typhoid fever: data from a community-based surveillance in calcutta, india. *Diagnostic microbiology and infectious disease*, 56(4):359–365. (Cited on page 224)
- Easterbrook, P. J., Gopalan, R., Berlin, J., and Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, 337(8746):867–872. (Cited on page 30)
- Eddy, D. (1980). ACS report on the cancer-related health checkup. *CA: a cancer journal for clinicians*, 30(4):193–240. (Cited on page 15)
- Eddy, D. M. (1990). Practice policies: where do they come from? *JAMA*, 263(9):1265–1275. (Cited on page 15)

- Eddy, D. M. (2005). Evidence-based medicine: a unified approach. *Health affairs*, 24(1):9–17. (Cited on page 12, 14, 15)
- Eddy, D. M. (2011). The origins of evidence-based medicine: A personal perspective. *AMA Journal of Ethics*, 13(1):55–60. (Cited on page 12, 13, 15)
- Edwards, P., Clarke, M., DiGuseppi, C., Pratap, S., Roberts, I., and Wentz, R. (2002). Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statistics in medicine*, 21(11):1635–1640. (Cited on page 33)
- Egger, M., Juni, P., Bartlett, C., Hoenstein, F., Sterne, J., et al. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? empirical study. *Health Technol Assess*, 7(1):1–76. (Cited on page 30, 167)
- Egger, M. and Smith, G. D. (1998). Bias in location and selection of studies. *Bmj*, 316(7124):61–66. (Cited on page 38, 167)
- Egger, M., Zellweger-Zähner, T., Schneider, M., Junker, C., Lengeler, C., and Antes, G. (1997). Language bias in randomised controlled trials published in english and german. *The Lancet*, 350(9074):326–329. (Cited on page 30)
- Eisinga, A., Siegfried, N., and Clarke, M. (2007). The sensitivity and precision of search terms in phases i, ii and iii of the cochrane highly sensitive search strategy for identifying reports of randomized trials in medline in a specific area of health care—hiv/aids prevention and treatment interventions. *Health Information & Libraries Journal*, 24(2):103–109. (Cited on page 32)
- Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., and Gruen, R. L. (2014). Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, 11(2):e1001603. (Cited on page 164)
- Elsevier (2019). Elsevier terms and conditions. Elsevier Terms and Conditions. Accessed October 2019. Available from <https://www.elsevier.com/legal/elsevier-website-terms-and-conditions>. (Cited on page 34)
- The European Medicines Agency (2019). Eu clinical trials register. Accessed October 2019. Available from <https://www.clinicaltrialsregister.eu/>. (Cited on page 32)
- Evidence Partners (2019). Distiller SR. Web based service, available at <https://www.evidencepartners.com/products/distillersr-systematic-review-software/>. (Cited on page 26)

- F1000 (2019). F1000 Workspace. Web based service, available at <https://f1000workspace.com/>. (Cited on page 27)
- Feinstein, A. R. and Horwitz, R. I. (1997). Problems in the “evidence” of “evidence-based medicine”. *The American journal of medicine*, 103(6):529–535. (Cited on page 15)
- Felizardo, K., Maldonado, J., Minghim, R., MacDonell, S., and Mendes, E. (2012a). Appendix d: An extension of the systematic literature review process with visual text mining: a case study on software engineering. In Felizardo (2012). (Cited on page 51)
- Felizardo, K. R. (2012). *Evidence-based software engineering: systematic literature review process based on visual text mining*. PhD thesis, Universidade de São Paulo. (Cited on page 290)
- Felizardo, K. R., Andery, G. F., Paulovich, F. V., Minghim, R., and Maldonado, J. C. (2012b). A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Information and Software Technology*, 54(10):1079–1091. (Cited on page 51)
- Felizardo, K. R., Salleh, N., Martins, R. M., Mendes, E., MacDonell, S. G., and Maldonado, J. C. (2011). Using visual text mining to support the study selection activity in systematic literature reviews. In *2011 International Symposium on Empirical Software Engineering and Measurement*, pages 77–86. IEEE. (Cited on page 51)
- Felizardo, K. R., Souza, S. R., and Maldonado, J. C. (2013). The use of visual text mining to support the study selection activity in systematic literature reviews: a replication study. In *2013 3rd International Workshop on Replication in Empirical Software Engineering Research*, pages 91–100. IEEE. (Cited on page 51)
- Fiszman, M., Bray, B., Shin, D., Kilicoglu, H., Bennett, G., Bodenreider, O., and Rindfleisch, T. (2010). Combining Relevance Assignment with Quality of the Evidence to Support Guideline Development. *Stud Health Technol Inform*, 160(1):709–713. (Cited on page 61)
- Frunza, O., Inkpen, D., and Matwin, S. (2010). Building systematic reviews using automatic text classification techniques. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 303–311. Association for Computational Linguistics. (Cited on page 50, 55, 60)
- Frunza, O., Inkpen, D., Matwin, S., Klement, W., and O’Blenis, P. (2011). Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*, 51(1):17–25. (Cited on page 50, 55, 63, 134)

- Fuhr, N. (1992). Probabilistic models in information retrieval. *The computer journal*, 35(3):243–255. (Cited on page 45, 49, 81, 151)
- García Adeva, J. J., Pikatza Atxa, J. M., Ubeda Carrillo, M., and Ansuategi Zengotitabengoa, E. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, 41(4 PART 1):1498–1508. (Cited on page 50, 69, 91)
- Gargon, E., Gorst, S. L., Harman, N. L., Smith, V., Matvienko-Sikar, K., and Williamson, P. R. (2018). Choosing important health outcomes for comparative effectiveness research: 4th annual update to a systematic review of core outcome sets for research. *PLoS ONE*, 13(12):e0209869. (Cited on page 150, 151, 152)
- Gargon, E., Gorst, S. L., and Williamson, P. R. (2019). The use of automated article ranking in the fifth annual update to a systematic review of core outcome sets for research. *PLoS ONE*. (Cited on page 7, 145, 148)
- Gargon, E., Gurung, B., Medley, N., Altman, D. G., Blazeby, J. M., Clarke, M., and Williamson, P. R. (2014). Choosing important health outcomes for comparative effectiveness research: a systematic review. *PLoS ONE*, 9(6):e99111. (Cited on page 150, 151, 152, 166)
- Gargon, E., Williamson, P. R., and Clarke, M. (2015). Collating the knowledge base for core outcome set development: developing and appraising the search strategy for a systematic review. *BMC medical research methodology*, 15(1):26. (Cited on page 150)
- Gehanno, J.-F., Rollin, L., and Darmoni, S. (2013). Is the coverage of google scholar enough to be used alone for systematic reviews. *BMC medical informatics and decision making*, 13(1):7. (Cited on page 30)
- Gelband, H. (1983). *The impact of randomized clinical trials on health policy and medical practice: background paper*. United States. Congress. Office of Technology Assessment. (Cited on page 14)
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10):3–8. (Cited on page 41)
- Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., and Zuccon, G. (2017). CLEF 2017 eHealth evaluation lab overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 291–303. Springer. (Cited on page 104)

- Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, 8(341):341ps12–341ps12. (Cited on page 207)
- Gorst, S. L., Gargon, E., Clarke, M., Blazeby, J. M., Altman, D. G., and Williamson, P. R. (2016a). Choosing important health outcomes for comparative effectiveness research: an updated review and user survey. *PLoS One*, 11(1):e0146444. (Cited on page 150, 151, 152)
- Gorst, S. L., Gargon, E., Clarke, M., Smith, V., and Williamson, P. R. (2016b). Choosing important health outcomes for comparative effectiveness research: an updated review and identification of gaps. *PLoS One*, 11(12):e0168403. (Cited on page 150, 151, 152)
- Gough, D., Oliver, S., and Thomas, J. (2017). *An introduction to systematic reviews*. Sage. (Cited on page 12, 13, 17)
- Graves, P. M., Choi, L., Gelband, H., and Garner, P. (2018). Primaquine or other 8-aminoquinolines for reducing plasmodium falciparum transmission. *Cochrane Database of Systematic Reviews*, (2). (Cited on page 199)
- Greenhalgh, T. and Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *Bmj*, 331(7524):1064–1065. (Cited on page 32, 37, 40)
- Halladay, C. W., Trikalinos, T. A., Schmid, I. T., Schmid, C. H., and Dahabreh, I. J. (2015). Using data sources beyond pubmed has a modest impact on the results of systematic reviews of therapeutic interventions. *Journal of clinical epidemiology*, 68(9):1076–1084. (Cited on page 30, 167, 168, 171, 188)
- Hansen, M. J., Rasmussen, N. Ø., and Chung, G. (2008). A method of extracting the number of trial participants from abstracts describing randomized controlled trials. *Journal of Telemedicine and Telecare*, 14(7):354–358. (Cited on page 197, 236)
- Hara, K. and Matsumoto, Y. (2007). Extracting clinical trial design information from MEDLINE abstracts. *New Generation Computing*, 25(3):263–275. (Cited on page 197, 236)
- Hariton, E. and Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716. (Cited on page 15)

- Hartling, L., Featherstone, R., Nuspl, M., Shave, K., Dryden, D. M., and Vandermeer, B. (2017). Grey literature in systematic reviews: a cross-sectional study of the contribution of non-english reports, unpublished studies and dissertations to the results of meta-analyses in child-relevant reviews. *BMC medical research methodology*, 17(1):64. (Cited on page 167)
- Hassanzadeh, H., Groza, T., and Hunter, J. (2014). Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of biomedical informatics*, 49:159–170. (Cited on page 197, 236)
- He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284. (Cited on page 133)
- Higgins, J. P. and Deeks, J. J. (2011). Selecting studies and collecting data. *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series*. (Cited on page 26, 33)
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Tianjing, L., Page, M. J., and Welch, V. A., editors (2019). *Cochrane Handbook for Systematic Reviews of Interventions, version 6 DRAFT*. Cochrane. Available from <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 16, 17, 21, 23, 25, 26, 27, 28, 29, 30, 31, 32, 36, 37, 38, 40, 41, 42, 145, 197, 201, 202, 235, 236, 287, 296, 297, 300, 303, 308)
- Hill, A. B. (1965). The reasons for writing. *BMJ*, 2(8701.15). (Cited on page 17)
- Horsley, T., Dingwall, O., and Sampson, M. (2011). Checking reference lists to find additional studies for systematic reviews. *Cochrane Database of Systematic Reviews*, 8. (Cited on page 37)
- Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., Holmgren, S., Pelch, K. E., Walker, V., Rooney, A. A., et al. (2016). Swift-review: a text-mining workbench for systematic review. *Systematic reviews*, 5(1):87. (Cited on page 55, 70, 151, 168)
- Hsu, W., Speier, W., and Taira, R. K. (2012). Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature. In *AMIA Annual Symposium Proceedings*, volume 2012, page 350. American Medical Informatics Association. (Cited on page 197)
- Huang, C.-C. and Lu, Z. (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, 17(1):132–144. (Cited on page 54, 85, 169)

- Huang, K.-C., Chiang, I.-J., Xiao, F., Liao, C.-C., Liu, C. C.-H., and Wong, J.-M. (2013). Pico element detection in medical text without metadata: Are first sentences enough? *Journal of biomedical informatics*, 46(5):940–946. (Cited on page 197, 236)
- Huang, K.-C., Liu, C. C.-H., Yang, S.-S., Xiao, F., Wong, J.-M., Liao, C.-C., and Chiang, I.-J. (2011). Classification of pico elements by text features systematically extracted from pubmed abstracts. In *2011 IEEE International Conference on Granular Computing*, pages 279–283. IEEE. (Cited on page 197, 236)
- Hull, D., Pettifer, S. R., and Kell, D. B. (2008). Defrosting the digital library: bibliographic tools for the next generation web. *PLoS computational biology*, 4(10):e1000204. (Cited on page 30, 31, 32)
- Ioannidis, J. P. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, 94(3):485–514. (Cited on page 20, 21, 25, 38, 85)
- Jackson, J. L. and Kuriyama, A. (2018). From the editors’ desk: bias in systematic reviews—let the reader beware. (Cited on page 29)
- Jadad, A. R., Cook, D. J., Jones, A., Klassen, T. P., Tugwell, P., Moher, M., and Moher, D. (1998). Methodology and reports of systematic reviews and meta-analyses: a comparison of cochrane reviews with articles published in paper-based journals. *JAMA*, 280(3):278–280. (Cited on page 23, 26, 188, 209)
- Jauhiainen, T., Lindén, K., and Jauhiainen, H. (2017). Evaluation of language identification methods using 285 languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 183–191. (Cited on page 198)
- Ji, X., Ritter, A., and Yen, P.-Y. (2017). Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. *Journal of biomedical informatics*, 69:33–42. (Cited on page 55, 71, 137)
- Jonnalagadda, S. and Petitti, D. (2013). A new iterative method to reduce workload in systematic review process. *International journal of computational biology and drug design*, 6(1-2):5–17. (Cited on page 55, 66)
- Jonnalagadda, S. R., Goyal, P., and Huffman, M. D. (2015). Automating data extraction in systematic reviews: a systematic review. *Systematic reviews*, 4(1):78. (Cited on page 197, 225, 235)

- Jonnalagadda, S. R. and Petitti, D. (2014). A new iterative method to reduce workload in the systematic review process. *Int J Comput Biol Drug Des*, 6(o):5–17. (Cited on page 91, 94, 136)
- Kalphov, V., Georgiadis, G., and Azzopardi, L. (2017). Sis at clef 2017 ehealth task. In *CEUR Workshop Proceedings*, volume 1866, pages 1–5. (Cited on page 55, 73)
- Kanoulas, E., Li, D., Azzopardi, L., and Spijker, R. (2017a). CLEF 2017 technologically assisted reviews in empirical medicine overview. In *CEUR Workshop Proceedings*, volume 1866, pages 1–29. (Cited on page 35, 53, 54, 85, 99, 104, 122, 142, 151, 153, 164, 167, 168, 169, 170, 171)
- Kanoulas, E., Li, D., Azzopardi, L., and Spijker, R. (2017b). Overview of the CLEF technologically assisted reviews in empirical medicine. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017*, CEUR Workshop Proceedings. CEUR-WS.org. (Cited on page 7, 55, 80, 196, 207)
- Kanoulas, E., Li, D., Azzopardi, L., and Spijker, R. (2018). Overview of the CLEF technologically assisted reviews in empirical medicine 2018. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*, CEUR Workshop Proceedings. (Cited on page 7, 32, 35, 54, 55, 80, 85, 122, 153, 164, 167, 168, 169, 170, 171, 196, 208)
- Kellermeyer, L., Harnke, B., and Knight, S. (2018). Covidence and Rayyan. *Journal of the Medical Library Association: JMLA*, 106(4):580. (Cited on page 26)
- Kelly, C. and Yang, H. (2013). A system for extracting study design parameters from nutritional genomics abstracts. *Journal of integrative bioinformatics*, 10(2):82–93. (Cited on page 197, 236)
- Khabsa, M., Elmagarmid, A., Ilyas, I., Hammady, H., and Ouzzani, M. (2016). Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, 102(3):465–482. (Cited on page 55, 57, 70, 71, 77, 91, 93, 94, 95, 97, 104, 108, 109, 113, 136, 137, 151, 168)
- Khangura, S., Konnyu, K., Cushman, R., Grimshaw, J., and Moher, D. (2012). Evidence summaries: the evolution of a rapid review approach. *Systematic reviews*, 1(1):10. (Cited on page 23)
- Kim, S. and Choi, J. (2012). Improving the performance of text categorization models used for the selection of high quality articles. *Healthcare informatics research*, 18(1):18–28. (Cited on page 66)

- Kim, S. and Choi, J. (2014). An svm-based high-quality article classifier for systematic reviews. *Journal of biomedical informatics*, 47:153–159. (Cited on page 66)
- Kim, S. N., Martinez, D., Cavedon, L., and Yencken, L. (2011). Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, page S5. BioMed Central. (Cited on page 197, 208, 224, 236)
- Kiritchenko, S., De Bruijn, B., Carini, S., Martin, J., and Sim, I. (2010). Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10(1):56. (Cited on page 197, 208, 224, 236)
- Kouznetsov, A. and Japkowicz, N. (2010). Using classifier performance visualization to improve collective ranking techniques for biomedical abstracts classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6085 LNAI:299–303. (Cited on page 55, 59)
- Kouznetsov, A., Matwin, S., Inkpen, D., Razavi, A. H., Frunza, O., Sehatkar, M., Seaward, L., and O’Blenis, P. (2009). Classifying biomedical abstracts using committees of classifiers and collective ranking techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5549 LNAI:224–228. (Cited on page 55, 59)
- Krallinger, M., Vazquez, M., Leitner, F., Salgado, D., Chatr-Aryamontri, A., Winter, A., Perfetto, L., Briganti, L., Licata, L., Iannuccelli, M., et al. (2011). The protein-protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC bioinformatics*, 12(8):S3. (Cited on page 149)
- Kung, J., Chiappelli, F., Cajulis, O. O., Avezova, R., Kossan, G., Chew, L., and Maida, C. A. (2010). From systematic reviews to clinical recommendations for evidence-based health care: validation of revised assessment of multiple systematic reviews (r-amstar) for grading of clinical relevance. *The open dentistry journal*, 4:84. (Cited on page 167)
- Lasserson, T. J., Thomas, J., and Higgins, J. P. T. (2019). Cochrane handbook for systematic reviews of interventions, chapter 1: Starting a review. draft version (29 january 2019). In Higgins et al. (2019). Available from <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 21, 23, 25, 26, 27, 28, 29, 145)

- Lau, J. (2019). Systematic review automation thematic series. (Cited on page 1, 22, 80)
- Lee, G. E. (2017). A study of convolutional neural networks for clinical document classification in systematic reviews: sysreview at CLEF eHealth 2017. In *CLEF (Working Notes)*. (Cited on page 55, 72)
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*. (Cited on page 222, 228)
- Leeflang, M. M., Deeks, J. J., Gatsonis, C., and Bossuyt, P. M. (2008). Systematic reviews of diagnostic test accuracy. *Annals of internal medicine*, 149(12):889–897. (Cited on page 2, 16, 19, 32, 38, 40, 165, 201, 206)
- Leeflang, M. M., Scholten, R. J., Rutjes, A. W., Reitsma, J. B., and Bossuyt, P. M. (2006). Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *Journal of clinical epidemiology*, 59(3):234–240. (Cited on page 2, 32, 80, 164)
- Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M.-I., Noel-Storr, A., Rader, T., Shokraneh, F., Thomas, J., and Wieland, L. S. (2019). Cochrane handbook for systematic reviews of interventions, chapter 4: Searching for and selecting studies. draft version (29 january 2019). In Higgins et al. (2019). Available from <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 23, 27, 30, 31, 32, 36, 37)
- Leizorovicz, A., Haugh, M., Chapuis, F., Samama, M., and Boissel, J. (1992). Low molecular weight heparin in prevention of perioperative thrombosis. *Bmj*, 305(6859):913–920. (Cited on page 38)
- Lemeshow, A. R., Blum, R. E., Berlin, J. A., Stoto, M. A., and Colditz, G. A. (2005). Searching one or two databases was insufficient for meta-analysis of observational studies. *Journal of clinical epidemiology*, 58(9):867–873. (Cited on page 30)
- Lerner, I., Créquit, P., Ravaud, P., and Atal, I. (2019). Automatic screening using word embeddings achieved high sensitivity and workload reduction for updating living network meta-analyses. *Journal of clinical epidemiology*, 108:86–94. (Cited on page 76, 169, 222)
- Li, T., Higgins, J. P., and Deeks, J. J. (2019). Cochrane handbook for systematic reviews of interventions, chapter 5: Collecting data. draft version (29 january 2019). In Higgins et al. (2019). Available from

- <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 27, 37, 197, 235, 236)
- Lin, S., Ng, J.-P., Pradhan, S., Shah, J., Pietrobon, R., and Kan, M.-Y. (2010). Extracting formulaic and free text clinical research articles metadata using conditional random fields. In *Proceedings of the NAACL HLT 2010 second Louhi workshop on text and data mining of health documents*, pages 90–95. Association for Computational Linguistics. (Cited on page 197, 236)
- Lind, J. (1753). *A treatise on 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*. A. Millar. (Cited on page 37)
- Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). The unified medical language system. *Yearbook of Medical Informatics*, 2(01):41–51. (Cited on page 229)
- Liu, J., Timsina, P., and El-Gayar, O. (2018). A comparative analysis of semi-supervised learning: the case of article selection for medical systematic reviews. *Information Systems Frontiers*, 20(2):195–207. (Cited on page 69)
- Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer. (Cited on page 226)
- Lorenzetti, D. L. and Ghali, W. A. (2013). Reference management software for systematic reviews and meta-analyses: an exploration of usage and usability. *BMC medical research methodology*, 13(1):141. (Cited on page 27, 30)
- Lorenzetti, D. L., Topfer, L.-A., Dennett, L., and Clement, F. (2014). Value of databases other than MEDLINE for rapid health technology assessments. *International journal of technology assessment in health care*, 30(2):173–178. (Cited on page 30, 32)
- Ma, Y. (2007). Text classification on imbalanced data: Application to systematic reviews automation. Master’s thesis, University of Ottawa (Canada). (Cited on page 57)
- Macaskill, P., Gatsonis, C. A., Deeks, J. J., Harbord, R., and Takwoingi, Y. (2010). Cochrane handbook for systematic reviews of diagnostic test accuracy, chapter 10 analysing and presenting results. In Deeks et al. (2013a). Available from <http://srdta.cochrane.org/>, accessed Aug. 2019. (Cited on page 19, 165, 171, 201, 202, 212, 214, 239)

- Malheiros, V., Hohn, E., Pinho, R., Mendonca, M., and Maldonado, J. C. (2007). A visual text mining approach for systematic reviews. In *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, pages 245–254. IEEE. (Cited on page 51)
- Marshall, I. J., Kuiper, J., and Wallace, B. C. (2014). Automating risk of bias assessment for clinical trials. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '14*, pages 88–95. (Cited on page 197, 224, 226, 236)
- Marshall, I. J., Marshall, R., Wallace, B. C., Brassey, J., and Thomas, J. (2019). Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. *Journal of clinical epidemiology*, 109:30–41. (Cited on page 30, 167, 168, 171, 188)
- Martin, P., Surian, D., Bashir, R., Bourgeois, F. T., and Dunn, A. G. (2019). Trial2rev: Combining machine learning and crowd-sourcing to create a shared space for updating systematic reviews. *JAMIA Open*, 2(1):15–22. (Cited on page 50)
- Martín-Martín, A., Costas, R., van Leeuwen, T., and López-Cózar, E. D. (2018). Evidence of open access of scientific publications in google scholar: A large-scale analysis. *Journal of Informetrics*, 12(3):819–841. (Cited on page 31, 34)
- Martinez, D., Karimi, S., Cavedon, L., and Baldwin, T. (2008). Facilitating biomedical systematic reviews using ranked text retrieval and classification. In *Australasian Document Computing Symposium (ADCS)*, pages 53–60. (Cited on page 55, 57)
- Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., and O’Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association : JAMIA*, 17(4):446–53. (Cited on page 55, 62, 91, 93)
- Matwin, S., Kouznetsov, A., Inkpen, D., and O’Blenis, P. (2011). Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association : JAMIA*, 1(18):author reply 105. (Cited on page 62, 95)
- Matwin, S. and Sazonova, V. (2012). Direct comparison between support vector machine and multinomial naive Bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, 19(5):917–917. (Cited on page 57, 62, 77, 91, 94, 136, 137)

- McKenzie, J. E., Brennan, S. E., Ryan, R. E., Thomson, H. J., and Johnston, R. V. (2019a). Cochrane handbook for systematic reviews of interventions, chapter 9: Summarizing study characteristics and preparing for synthesis. draft version (29 january 2019). In Higgins et al. (2019). Available from <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 27, 38, 40, 201)
- McKenzie, J. E., Brennan, S. E., Ryan, R. E., Thomson, H. J., Johnston, R. V., and Thomas, J. (2019b). Cochrane handbook for systematic reviews of interventions, chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis. draft version (29 january 2019). In Higgins et al. (2019). Available from <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 27, 29, 40, 201)
- Minas, A., Lagopoulos, A., and Tsoumakas, G. (2018). Aristotle university's approach to the technologically assisted reviews in empirical medicine task of the 2018 CLEF eHealth lab. In *CLEF (Working Notes)*. (Cited on page 55, 73)
- Miwa, M., Thomas, J., O'Mara-Eves, A., and Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, 51:242–253. (Cited on page 55, 68)
- Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M., and Altman, D. G. (2007). Epidemiology and reporting characteristics of systematic reviews. *PLoS medicine*, 4(3):e78. (Cited on page 23, 26)
- Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307. (Cited on page 177, 179)
- Monni, G., Iuculano, A., and Zoppi, M. A. (2014). Screening and invasive testing in twins. *Journal of clinical medicine*, 3(3):865–882. (Cited on page 18)
- Névéol, A. and Zweigenbaum, P. (2017). Making sense of big textual data for health care: Findings from the section on clinical natural language processing. *Yearbook of medical informatics*, 26(01):228–233. (Cited on page 94)
- Névéol, A., Zweigenbaum, P., et al. (2016). Clinical natural language processing in 2015: leveraging the variety of texts of clinical interest. *Yearbook of medical informatics*, 25(01):234–239. (Cited on page 94)
- Norman, C., Grouin, C., Lavergne, T., Zweigenbaum, P., and Névéol, A. (2017a). Traitement de la langue biomédicale au LIMS1. *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, page 33. (Cited on page 79, 277)

- Norman, C., Leeflang, M., and Névél, A. (2017b). LIMS@CLEF eHealth 2017 task 2: Logistic regression for automatic article ranking. *Working Notes of CLEF*. (Cited on page v, 43, 79, 99, 102, 124, 125, 132, 133, 275)
- Norman, C., Leeflang, M., and Névél, A. (2018a). Data extraction and synthesis in systematic reviews of diagnostic test accuracy: A corpus for automating and evaluating the process. In *AMIA Annual Symposium Proceedings*, volume 2018, page 817. American Medical Informatics Association. (Cited on page ix, 34, 170, 177, 184, 195, 204, 226, 237, 241, 276)
- Norman, C., Leeflang, M., and Névél, A. (2018b). LIMS@CLEF eHealth 2018 task 2: Technology assisted reviews by stacking active and static learning. *Working Notes of CLEF*, pages 10–14. (Cited on page vi, 32, 43, 79, 120, 133, 138, 139, 171, 275)
- Norman, C., Leeflang, M., and Névél, A. (2019a). Automated checking for human errors in meta-analyses of diagnostic test accuracy. In *Cochrane Colloquium*. (Cited on page 48, 195, 277)
- Norman, C., Leeflang, M., Porcher, R., and Névél, A. (2019b). Does screening automation negatively impact meta-analyses in systematic reviews of diagnostic test accuracy? In *Cochrane Colloquium*. (Cited on page 141, 276)
- Norman, C., Leeflang, M., Porcher, R., and Névél, A. (2019c). Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. *BMC Systematic Reviews*. Springer Nature. (Cited on page viii, 141, 162, 240, 275)
- Norman, C., Leeflang, M., Porcher, R., and Névél, A. (2019d). Measuring the impact of screening automation on meta-analyses of diagnostic test accuracy. In *AMIA Annual Symposium Proceedings*. (Cited on page 141, 195, 276)
- Norman, C., Leeflang, M., Zweigenbaum, P., and Névél, A. (2017c). Tri automatique de la littérature pour les revues systématiques. *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 234–41. (Cited on page 43, 79, 276)
- Norman, C., Leeflang, M., Zweigenbaum, P., and Névél, A. (2018c). Automating document discovery in the systematic review process: How to use chaff to extract wheat. In Calzolari, N. et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA). (Cited on page v, 35, 43, 79, 88, 133, 153, 276)

- Norman, C., Spijker, R., Kanoulas, E., Leeflang, M., and Névél, A. (2019e). A distantly supervised dataset for automated data extraction from diagnostic studies. *ACL BioNLP*. (Cited on page x, 29, 34, 195, 197, 220, 276)
- Norman, C., Van Nguyen, T., and Névél, A. (2017d). Contribution of natural language processing in predicting rehospitalization risk. *Medical care*, 55(8):781. (Cited on page 275)
- Norman, C. R., Gargon, E., Leeflang, M. M. G., Névél, A., and Williamson, P. R. (2019f). Evaluation of an automatic article selection method for timelier updates of the COMET core outcome set database. *Database*. Oxford University Press. (Cited on page vii, 34, 43, 141, 148, 275)
- Nurmohamed, M., Buller, H. R., Dekker, E., Hommes, D., Rosendaal, F., Briet, E., and Vandenbroucke, J. (1992). Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. *The Lancet*, 340(8812):152–156. (Cited on page 38)
- Nussbaumer-Streit, B., Klerings, I., Wagner, G., Heise, T. L., Dobrescu, A. I., Armijo-Olivo, S., Stratil, J. M., Persad, E., Lhachimi, S. K., Van Noord, M. G., et al. (2018). Abbreviated literature searches were viable alternatives to comprehensive searches: a meta-epidemiological study. *Journal of clinical epidemiology*, 102:1–11. (Cited on page 30, 167, 168)
- Nye, B., Li, J. J., Patel, R., Yang, Y., Marshall, I. J., Nenkova, A., and Wallace, B. C. (2018). A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access. (Cited on page 224, 225)
- Olofsson, H., Brolund, A., Hellberg, C., Silverstein, R., Stenström, K., Österberg, M., and Dagerhamn, J. (2017). Can abstract screening workload be reduced using text mining? user experiences of the tool Rayyan. *Research synthesis methods*, 8(3):275–280. (Cited on page 70, 71)
- Olorisade, B. K., Brereton, P., and Andras, P. (2019). The use of bibliography enriched features for automatic citation screening. *Journal of biomedical informatics*, 94:103202. (Cited on page 55, 76, 77, 137)
- Olorisade, B. K., de Quincey, E., Brereton, P., and Andras, P. (2016). A critical analysis of studies that address the use of text mining for citation screening in systematic reviews. In *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, page 14. ACM. (Cited on page 55, 85, 151, 169)

- Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, 5(1):210. (Cited on page 26)
- The Oxford Centre for Evidence-Based Medicine (2016). OCEBM levels of evidence 2. Accessed October 2019. Available from <https://www.cebm.net/index.aspx?o=5653>. (Cited on page 14, 16)
- O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):5. (Cited on page 1, 35, 43, 47, 49, 50, 51, 53, 80, 81, 85, 90, 91, 103, 121, 133, 134, 142, 145, 151, 164, 165, 168, 175, 206, 221)
- Page, M. J., Cumpston, M., Chandler, J., and Lasserson, T. J. (2019a). Cochrane handbook for systematic reviews of interventions, chapter iii: Reporting the review. draft version (august 2019). In Higgins et al. (2019). Available from <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 26, 42)
- Page, M. J., Higgins, J. P., and Sterne, J. A. (2019b). Cochrane handbook for systematic reviews of interventions, chapter 13: Assessing risk of bias due to missing results in a synthesis. draft version (29 january 2019). In Higgins et al. (2019). Available from <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 40, 41)
- Parkhill, A. F., Clavisi, O., Pattuwage, L., Chau, M., Turner, T., Bragge, P., and Gruen, R. (2011). Searches for evidence mapping: effective, shorter, cheaper. *Journal of the Medical Library Association: JMLA*, 99(2):157. (Cited on page 30)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830. (Cited on page 44, 95, 107, 157)
- Petersen, H., Poon, J., Poon, S. K., and Loy, C. (2014). Increased workload for systematic review literature searches of diagnostic tests compared with treatments: Challenges and opportunities. *JMIR medical informatics*, 2(1):e11. (Cited on page 80, 103, 121, 133, 164, 206)
- Pham, B., Bagheri, E., Rios, P., Pourmasoumi, A., Robson, R. C., Hwee, J., Isaranuwachai, W., Darvesh, N., Page, M. J., Tricco, A. C., et al. (2018). Improving the conduct of systematic reviews: a process mining perspective. *Journal of clinical epidemiology*, 103:101–111. (Cited on page 1, 23, 25, 37, 197)

- ProQuest (2019). RefWorks. Web based service, available at <http://www.refworks-cos.com/refworks>. (Cited on page 27)
- Przybyła, P., Brockmeier, A. J., Kontonatsios, G., LePogam, M.-A., McNaught, J., von Elm, E., Nolan, K., and Ananiadou, S. (2018). Prioritising references for systematic reviews with robotanalyst: A user study. *Research synthesis methods*, 9(3):470–488. (Cited on page 75, 151, 168, 169, 222)
- Rathbone, J., Hoffmann, T., and Glasziou, P. (2015). Faster title and abstract screening? evaluating abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic reviews*, 4(1):80. (Cited on page 62, 63)
- Razavi, A. H., Matwin, S., Inkpen, D., and Kouznetsov, A. (2009). Parameterized contrast in second order soft co-occurrences: a novel text representation technique in text mining and knowledge extraction. In *2009 Ieee International Conference on Data Mining Workshops*, pages 471–476. IEEE. (Cited on page 55, 58)
- Reitsma, J., Rutjes, A., Whiting, P., Vlassov, V., Leeflang, M. M., and Deeks, J. J. (2008). Cochrane handbook for systematic reviews of diagnostic test accuracy, chapter 9: Assessing methodological quality. In Deeks et al. (2013a). Available from <http://srdta.cochrane.org/>, accessed Aug. 2019. (Cited on page 37, 197)
- Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M., and Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*, 58(10):982–990. (Cited on page 19, 165, 177, 201)
- Restificar, A. and Ananiadou, S. (2012). Inferring appropriate eligibility criteria in clinical trial protocols without labeled data. In *Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics*, pages 21–28. ACM. (Cited on page 197, 236)
- Rindflesch, T. C. and Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of biomedical informatics*, 36(6):462–477. (Cited on page 61)
- Roberts, D., Brown, J., Medley, N., and Dalziel, S. R. (2017). Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. *Cochrane database of systematic reviews*, 3. (Cited on page 12)
- Robinson, D. A. (2012). *Finding patient-oriented evidence in PubMed abstracts*. PhD thesis, University of Georgia Athens. (Cited on page 197)

- Rombey, T., Allers, K., Mathes, T., Hoffmann, F., and Pieper, D. (2019). A descriptive analysis of the characteristics and the peer review process of systematic review protocols published in an open peer review journal from 2012 to 2017. *BMC medical research methodology*, 19(1):57. (Cited on page 28, 31)
- Roth, D., Pace, N. L., Lee, A., Hovhannisyan, K., Warenits, A.-M., Arrich, J., and Herkner, H. (2018). Airway physical examination tests for detection of difficult airway management in apparently normal adult patients. *Cochrane Database of Systematic Reviews*, 5. (Cited on page 27)
- Royle, P. and Milne, R. (2003). Literature searching for randomized controlled trials used in cochrane reviews: rapid versus exhaustive searches. *International journal of technology assessment in health care*, 19(4):591–603. (Cited on page 30)
- Royle, P. and Waugh, N. (2005). A simplified search strategy for identifying randomised controlled trials for systematic reviews of health care interventions: a comparison with more exhaustive strategies. *BMC Medical Research Methodology*, 5(1):23. (Cited on page 30)
- Royle, P. L., Bain, L., and Waugh, N. R. (2005). Sources of evidence for systematic reviews of interventions in diabetes. *Diabetic medicine*, 22(10):1386–1393. (Cited on page 30)
- Rutter, C. M. and Gatsonis, C. A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in medicine*, 20(19):2865–2884. (Cited on page 19, 165, 201)
- Sackett, D., Ellis, J., Mulligan, I., and Rowe, J. (1995). Inpatient general medicine is evidence based. *The Lancet*, 346(8972):407–410. (Cited on page 14, 15)
- Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. (Cited on page 13, 15)
- Sager, J. C., Dungworth, D., and McDonald, P. F. (1980). *English special languages: principles and practice in science and technology*. John Benjamins Pub Co. (Cited on page 222)
- Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432. (Cited on page 117)

- Sampson, M., Barrowman, N. J., Moher, D., Klassen, T. P., Platt, R., John, P. D. S., Viola, R., Raina, P., et al. (2003). Should meta-analysts search embase in addition to medline? *Journal of clinical epidemiology*, 56(10):943–955. (Cited on page 30, 167, 168)
- Satopää, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171. IEEE. (Cited on page 175)
- Scells, H., Zuccon, G., Deacon, A., and Koopman, B. (2017). Qut ielab at CLEF 2017 technology assisted reviews track: Initial experiments with learning to rank. In *CLEF (Working Notes)*. (Cited on page 55, 72)
- Schünemann, H. J. and Moja, L. (2015). Reviews: rapid! rapid! rapid!... and systematic. (Cited on page 1, 18, 22, 25)
- Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., Henry, D. A., and Boers, M. (2009). Amstar is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of clinical epidemiology*, 62(10):1013–1020. (Cited on page 167)
- Shekelle, P. G., Dalal, S. R., and Shetty, K. D. (2012). A Pilot Study Using Machine Learning and Domain Knowledge To Facilitate Comparative Effectiveness Review Updating. *AHRQ*. (Cited on page 64, 91)
- Shekelle, P. G., Shetty, K., Newberry, S., Maglione, M., and Motala, A. (2017). Machine learning versus standard techniques for updating searches for systematic reviews: a diagnostic accuracy study. *Annals of internal medicine*, 167(3):213–215. (Cited on page 71)
- Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O’Mara-Eves, A., Kelly, M. P., and Thomas, J. (2014). Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 5(1):31–49. (Cited on page 67)
- Shojania, K. G., Sampson, M., Ansari, M. T., Ji, J., Doucette, S., and Moher, D. (2007). How quickly do systematic reviews go out of date? a survival analysis. *Annals of internal medicine*, 147(4):224–233. (Cited on page 1, 23)
- Si, Y., Wang, J., Xu, H., and Roberts, K. (2019). Enhancing clinical concept extraction with contextual embedding. *arXiv preprint arXiv:1902.08691*. (Cited on page 222)
- Simpson, R. and Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *The British Medical Journal*, pages 1243–1246. (Cited on page 41)

- Singh, G., Marshall, I., Thomas, J., and Wallace, B. (2017). Identifying diagnostic test accuracy publications using a deep model. In *CEUR Workshop Proceedings*, volume 1866. CEUR Workshop Proceedings. (Cited on page 55, 73)
- Siontis, K. C., Hernandez-Boussard, T., and Ioannidis, J. P. (2013). Overlapping meta-analyses on the same topic: survey of published studies. *Bmj*, 347:f4501. (Cited on page 21)
- Slobogean, G. P., Verma, A., Giustini, D., Slobogean, B. L., and Mulpuri, K. (2009). MEDLINE, EMBASE, and cochrane index most primary studies but not abstracts included in orthopedic meta-analyses. *Journal of clinical epidemiology*, 62(12):1261–1267. (Cited on page 30)
- Song, M. H., Lee, Y. H., and Kang, U. G. (2013). Comparison of machine learning algorithms for classification of the sentences in three clinical practice guidelines. *Healthcare informatics research*, 19(1):16–24. (Cited on page 197, 199, 236, 237)
- Stein, B., Hoppe, D., and Gollub, T. (2012). The impact of spelling errors on patent search. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 570–579. Association for Computational Linguistics. (Cited on page 91)
- Stern, J. M. and Simes, R. J. (1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *Bmj*, 315(7109):640–645. (Cited on page 30)
- Sterne, J. A., Egger, M., and Smith, G. D. (2001). Investigating and dealing with publication and other biases in meta-analysis. *Bmj*, 323(7304):101–105. (Cited on page 30, 32)
- Stevinson, C. and Lawlor, D. (2004). Searching multiple databases for systematic reviews: added value or diminishing returns? *Complementary therapies in medicine*, 12(4):228–232. (Cited on page 30)
- Stewart, L., Moher, D., and Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense. (Cited on page 28)
- Subirana, M., Solá, I., Garcia, J. M., Gich, I., and Urrútia, G. (2005). A nursing qualitative systematic review required MEDLINE and CINAHL for study identification. *Journal of Clinical Epidemiology*, 58(1):20–25. (Cited on page 30)
- Summerscales, R., Argamon, S., Hupert, J., and Schwartz, A. (2009). Identifying treatments, groups, and outcomes in medical abstracts. In *The Sixth Midwest Computational Linguistics Colloquium (MCLC 2009)*. (Cited on page 197)

- Summerscales, R. L., Argamon, S., Bai, S., Hupert, J., and Schwartz, A. (2011). Automatic summarization of results from clinical trials. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377. IEEE. (Cited on page 197)
- Sun, Y. B., Yang, Y., Zhang, H., Zhang, W., and Wang, Q. (2012). Towards evidence-based ontology for supporting systematic literature review. *Evaluation and Assessment in Software Engineering*, 2012(1):171–175. (Cited on page 65, 91)
- Suominen, H., Kelly, L., Goeuriot, L., Névél, A., Ramadier, L., Robert, A., Kanoulas, E., Spijker, R., Azzopardi, L., Li, D., et al. (2018). Overview of the CLEF ehealth evaluation lab 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 286–301. Springer. (Cited on page 122, 151)
- Surian, D., Dunn, A. G., Orenstein, L., Bashir, R., Coiera, E., and Bourgeois, F. T. (2018). A shared latent space matrix factorisation method for recommending new trial evidence for systematic review updates. *Journal of biomedical informatics*, 79:32–40. (Cited on page 50)
- Terasawa, T., Dvorak, T., Ip, S., Raman, G., Lau, J., and Trikalinos, T. A. (2009). Systematic review: charged-particle radiation therapy for cancer. *Annals of internal medicine*, 151(8):556–565. (Cited on page 55)
- Thomas, J. (2013). Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation? *OA Evidence-Based Medicine*, 1(2):1–6. (Cited on page 1, 142, 145, 164, 165)
- Thomas, J. and Brunton, J. (2007). Eppi-reviewer: software for research synthesis. (Cited on page 151, 168)
- Thomas, J., Kneale, D., McKenzie, J. E., Brennan, S. E., and Bhaumik, S. (2019). Cochrane handbook for systematic reviews of interventions, chapter 2: Determining the scope of the review and the questions it will address. draft version (29 january 2019). In Higgins et al. (2019). Available from <https://training.cochrane.org/handbook/version-6>, accessed Aug. 2019. (Cited on page 17, 27, 29, 201)
- Thomson Reuters (1999). ProCite. Version 5 (1999-10-26), discontinued. (Cited on page 27)
- Thomson Reuters (2008). Reference Manager. Version 12 (2008-02-09), discontinued. (Cited on page 27)

- Thorlund, K., Imberger, G., Johnston, B. C., Walsh, M., Awad, T., Thabane, L., Gluud, C., Devereaux, P., and Wetterslev, J. (2012). Evolution of heterogeneity (i2) estimates and their 95% confidence intervals in large meta-analyses. *PLoS ONE*, 7(7):e39471. (Cited on page 174)
- Timsina, P., El-Gayar, O. F., and Liu, J. (2015). Leveraging advanced analytics techniques for medical systematic review update. In *2015 48th Hawaii International Conference on System Sciences*, pages 976–985. IEEE. (Cited on page 69)
- Timsina, P., Liu, J., and El-Gayar, O. (2016a). Advanced analytics for the automation of medical systematic reviews. *Information Systems Frontiers*, 18(2):237–252. (Cited on page 70)
- Timsina, P., Liu, J., El-Gayar, O., and Shang, Y. (2016b). Using semi-supervised learning for the creation of medical systematic review: An exploratory analysis. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1195–1203. IEEE. (Cited on page 69)
- Tomassetti, F., Rizzo, G., Vetro, A., Ardito, L., Torchiano, M., and Morisio, M. (2011). Linked data approach for selection process automation in systematic reviews. *15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011)*, pages 31–35. (Cited on page 63, 64)
- Tran, V.-T., Porcher, R., Tran, V.-C., and Ravaud, P. (2017). Predicting data saturation in qualitative surveys with mathematical models from ecological research. *Journal of clinical epidemiology*, 82:71–78. (Cited on page 175, 176)
- Tricco, A. C., Antony, J., Zarin, W., Striffler, L., Ghassemi, M., Ivory, J., Perrier, L., Hutton, B., Moher, D., and Straus, S. E. (2015). A scoping review of rapid review methods. *BMC medicine*, 13(1):224. (Cited on page 167)
- Tricco, A. C., Brehaut, J., Chen, M. H., and Moher, D. (2008). Following 411 cochrane protocols to completion: a retrospective cohort study. *PLoS One*, 3(11):e3684. (Cited on page 23)
- Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., and Coiera, E. (2014). Systematic review automation technologies. *Systematic reviews*, 3(1):74. (Cited on page 32, 35, 36, 40, 49, 90, 150, 202, 206, 235, 239)
- Tsafnat, G., Glasziou, P., Karystianis, G., and Coiera, E. (2018). Automated screening of research studies for systematic reviews using study characteristics. *Systematic reviews*, 7(1):64. (Cited on page 23, 75)

- Tsertsvadze, A., Chen, Y.-F., Moher, D., Sutcliffe, P., and McCarthy, N. (2015). How to conduct systematic reviews more expeditiously? *Systematic reviews*, 4(1):160. (Cited on page 23)
- Tsoumakas, G. (2018). Learning-to-rank and relevance feedback for literature appraisal in empirical medicine. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, volume 11018, page 52. Springer. (Cited on page 73)
- The US National Library of Medicine (2019). Clinicaltrials.gov. Accessed October 2019. Available from <https://clinicaltrials.gov/>. (Cited on page 32)
- The US Preventive Services Task Force (1989). *Guide to clinical preventive services: report of the US Preventive Services Task Force*. DIANE publishing. (Cited on page 15)
- Verbeke, M., Van Asch, V., Morante, R., Frasconi, P., Daelemans, W., and De Raedt, L. (2012). A statistical relational learning approach to identifying evidence based medicine categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 579–589. Association for Computational Linguistics. (Cited on page 197, 236)
- Veritas Health Innovation (2019). Covidence. Web based service available at <https://www.covidence.org/>. (Cited on page 26)
- De Vet, H., Eisinga, A., Riphagen, I., Aertgeerts, B., Pewsner, D., and Mitchell, R. (2008). Cochrane handbook for systematic reviews of diagnostic test accuracy, chapter 7: Searching for studies. In Deeks et al. (2013a). Available from <http://srdta.cochrane.org/>, accessed Aug. 2019. (Cited on page 2, 32, 33)
- Wallace, B. C., Kuiper, J., Sharma, A., Zhu, M. B., and Marshall, I. J. (2016). Extracting pico sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research*, 17(132):1–25. (Cited on page 208, 224, 226)
- Wallace, B. C., Small, K., Brodley, C. E., Lau, J., Schmid, C. H., Bertram, L., Lill, C. M., Cohen, J. T., and Trikalinos, T. A. (2012a). Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in Medicine*, 14(7):663–669. (Cited on page 66)
- Wallace, B. C., Small, K., Brodley, C. E., Lau, J., and Trikalinos, T. a. (2010a). Modeling Annotation Time to Reduce Workload in Comparative Effectiveness Reviews

- Categories and Subject Descriptors Active Learning to Mitigate Workload. *Proceedings of the 1st ACM International Health Informatics Symposium*. ACM,, pages 28–35. (Cited on page 51, 55, 62)
- Wallace, B. C., Small, K., Brodley, C. E., Lau, J., and Trikalinos, T. a. (2012b). Deploying an interactive machine learning system in an evidence-based practice center. *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12*, page 819. (Cited on page 62, 151, 168)
- Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. (2010b). Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 173–182. ACM. (Cited on page 53, 55, 60, 143)
- Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. (2011). Who should label what? instance allocation in multiple expert active learning. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 176–187. SIAM. (Cited on page 55, 64)
- Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., and Schmid, C. H. (2010c). Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):55. (Cited on page 34, 53, 55, 60, 143)
- Wennberg, J. and Gittelsohn, A. (1973). Small area variations in health care delivery: a population-based health information system can guide planning and regulatory decision-making. *Science*, 182(4117):1102–1108. (Cited on page 13)
- Wetterslev, J., Jakobsen, J. C., and Gluud, C. (2017). Trial sequential analysis in systematic reviews with meta-analysis. *BMC medical research methodology*, 17(1):39. (Cited on page 166, 184)
- Wijedoru, L., Mallett, S., and Parry, C. M. (2017). Rapid diagnostic tests for typhoid and paratyphoid (enteric) fever. *The Cochrane Library*. (Cited on page 209, 211)
- Williamson, J., Goldschmidt, P., and Jillson, I. (1979). Medical practice information demonstration project: final report. *Contract*, pages 282–77. (Cited on page 14)
- Williamson, P. R., Altman, D. G., Bagley, H., Barnes, K. L., Blazeby, J. M., Brookes, S. T., Clarke, M., Gargon, E., Gorst, S., Harman, N., et al. (2017). The comet handbook: version 1.0. *Trials*, 18(3):280. (Cited on page 150)
- The World Health Organization (2019). WHO International Clinical Trials Registry Platform (ICTRP). Accessed October 2019. Available from <https://www.who.int/ictpr/en/>. (Cited on page 32)

- Wright, R. W., Brand, R. A., Dunn, W., and Spindler, K. P. (2007). How to write a systematic review. *Clinical orthopaedics and related research*, 455:23–29. (Cited on page 89)
- Wu, H., Wang, T., Chen, J., Chen, S., Hu, Q., and He, L. (2018). Ecnua at 2018 ehealth task 2: Technologically assisted reviews in empirical medicine. *Methods*, 4(5):7. (Cited on page 55, 75)
- Xu, R., Garten, Y., Supekar, K. S., Das, A. K., Altman, R. B., Garber, A. M., et al. (2007). Extracting subject demographic information from abstracts of randomized clinical trial reports. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 550. IOS Press. (Cited on page 197, 236)
- Yu, W., Clyne, M., Dolan, S. M., Yesupriya, A., Wulf, A., Liu, T., Khoury, M. J., and Gwinn, M. (2008). GAPscreener: An automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*, 9:205. (Cited on page 58)
- Zhao, J., Bysani, P., and Kan, M.-Y. (2012). Exploiting classification correlations for the extraction of evidence-based practice information. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1070. American Medical Informatics Association. (Cited on page 197, 236)
- Zhu, H., Ni, Y., Cai, P., Qiu, Z., and Cao, F. (2012). Automatic extracting of patient-related attributes: disease, age, gender and race. *Studies in health technology and informatics*, 180:589–593. (Cited on page 197, 236)

**Titre :** Méthodes d'automatisation des revues systématiques

**Mots clés :** Traitement automatique des langues, apprentissage automatique, revues systématiques

**Résumé :** Les récentes avancées en matière d'intelligence artificielle ont vu une adoption limitée dans la communauté des auteurs de revues systématiques, et une grande partie du processus d'élaboration des revues systématiques est toujours manuelle, longue et coûteuse. Les auteurs de revues systématiques rencontrent des défis tout au long du processus d'élaboration d'une revue. Il est long et difficile de chercher, d'extraire, de collecter des données, de rédiger des manuscrits et d'effectuer des analyses statistiques. L'automatisation de la sélection d'articles a été proposé comme un moyen de réduire la charge de travail, mais son adoption a été limitée en raison de différents facteurs, notamment l'investissement important de prise en main, le manque d'accompagnement et les décalages par rapport au flux de travail. Il est nécessaire de mieux harmoniser les méthodes actuelles avec les besoins de la communauté des revues systématiques.

Les études sur l'exactitude des tests diagnostiques sont rarement indexées de façon à pouvoir être facilement retrouvées dans les bases de données

bibliographiques. La variabilité terminologique et l'indexation lacunaire ou incohérente de ces études sont autant de facteurs augmentant le niveau de difficulté de réalisation des revues systématiques qui s'y intéressent. Les requêtes de recherche méthodologique visant à repérer les études diagnostiques ont donc tendance à être peu précises, et leur utilisation dans les études méthodiques est déconseillée. Par conséquent, il est particulièrement nécessaire d'avoir recours à d'autres méthodes pour réduire la charge de travail dans les études méthodiques sur l'exactitude des tests diagnostiques.

Dans la présente thèse, nous avons examiné l'hypothèse selon laquelle les méthodes d'automatisation peuvent offrir un moyen efficace de rendre le processus d'élaboration des revues systématique plus rapide et moins coûteux, à condition de pouvoir cerner et surmonter les obstacles à leur adoption. Les travaux réalisés montrent que les méthodes automatisées offrent un potentiel de diminution des coûts tout en améliorant la transparence et la reproductibilité du processus.

**Title :** Systematic Review Automation Methods

**Keywords :** Natural language processing, machine learning, systematic reviews

**Abstract :** Recent advances in artificial intelligence have seen limited adoption in systematic reviews, and much of the systematic review process remains manual, time-consuming, and expensive. Authors conducting systematic reviews face issues throughout the systematic review process. It is difficult and time-consuming to search and retrieve, collect data, write manuscripts, and perform statistical analyses. Screening automation has been suggested as a way to reduce the workload, but uptake has been limited due to a number of issues, including licensing, steep learning curves, lack of support, and mismatches to workflow. There is a need to better align current methods to the need of the systematic review community.

Diagnostic test accuracy studies are seldom indexed in an easily retrievable way, and suffer from variable terminology and missing or inconsistently applied database labels. Methodological search queries to identify diagnostic studies therefore tend to have low accuracy, and are discouraged for use in systematic reviews. Consequently, there is a particular need for alternative methods to reduce the workload in systematic reviews of diagnostic test accuracy.

In this thesis we have explored the hypothesis that automation methods can offer an efficient way to make the systematic review process quicker and less expensive, provided we can identify and overcome barriers to their adoption. Automated methods have the opportunity to make the process cheaper as well as more transparent, accountable, and reproducible.