Automating Systematic Reviews

John Rathbone

A thesis submitted in total fulfilment of the requirement of the degree of
Doctor of Philosophy (PhD)

May 2017

Centre for Research in Evidence-Based Practice
Faculty of Health Science and Medicine

Professors Paul Glasziou, Tammy Hoffmann & Associate Professor Elaine Beller

This research was supported by an Australian Government Research Training Program Scholarship.

Abstract

Background

Systematic reviews are used as the 'gold standard' to evaluate healthcare, education, and social policies. They are integral to the clinical decision making of healthcare professionals, and funding decisions made by governmental agencies. The rapid growth in primary research has not been matched by a growth in the efficiency of producing systematic reviews and consequently evidence-based decision making is struggling to remain feasible.

Aims

This body of research aimed to develop and evaluate strategies towards the automation of systematic reviews, so that secondary health research can be produced more efficiently and cost effectively. To that end, four research studies were developed: 1. Comparing the performance of biomedical databases to determine the sensitivity and precision for identifying systematic reviews; 2. Developing and evaluating algorithms to detect duplicate records arising from searching biomedical databases; 3. Evaluating the potential benefits from using a semi-automated machine learning predictive algorithm for citation screening; 4. Developing and evaluating strategies to expedite citation screening using title-only keyword searching.

Methods

Different methods were used to answer the research questions. For the first research study (identifying reviews), 7 biomedical databases were searched for systematic reviews of any intervention for hypertension and the performance of each database was assessed and compared for both comprehensiveness and accuracy. For the second research study (deduplication), an iterative approach was needed to develop and evaluate the performance of each algorithm to detect duplicates; the results acquired from each algorithm were used to inform the next iteration until an ideal algorithm was produced that achieved higher duplicate detection than current methods, but without compromising accuracy. For the third research study (predictive screening), 4 datasets from the literature searches of

published systematic reviews were used to evaluate an online machine learning predictive algorithm by replicating the screening decisions of the original reviews; sensitivity analyses were performed to determine if the reduction in screening effort could be further improved by including non-relevant citations that were closely matched to the review inclusion criteria. For the fourth study (expediting screening), 10 datasets from the literature searches of published systematic reviews were used to evaluate title-only screening. Datasets were screened using title-only keywords searching based upon the inclusion criteria of each systematic review. The results were compared against the published reviews for reduction in screening effort and recall of included studies.

Results

In the first study, the biomedical database, EMBASE, retrieved the largest number of relevant citations (69% sensitivity), but also was the least specific (7% specificity), retrieving many irrelevant citations. The Cochrane Library had 60% sensitivity and was the most precise (30%) of all the databases. None of the databases identified all the relevant records, but a combination of EMBASE, the Cochrane Library and Epistemonikos identified 83% of all the relevant systematic reviews.

In the second study, the iteratively developed deduplication algorithm increased duplicate detection by an average of 42% compared with duplicate detection using EndNote™ bibliographic reference management software. Additionally, all unique citations were correctly classified, whereas EndNote™ classified some unique citations wrongly as duplicate records.

In study 3, the evaluation found that the predictive screening tool (Abstrackr) reduced the screening effort in a range from 9% to 57% depending on the complexity of the systematic review. The reliability to retrieve included studies was good, with most relevant citations found, but in 2 datasets one included study was not retrieved by Abstrackr. Sensitivity analyses found that workload savings could be further increased by including closely matched non-relevant citations, and very large datasets (≥15,000 citations) could achieve as much as 80% reduction in screening.

In study 4, the interest was to reduce screening effort using title-only screening. This ranged from 11% to 78% with a median reduction in screening effort of 53%. In

9 systematic reviews the recall of included studies was 100%. In one review, 4 of 5 reviewers did not identify the same included study (median recall: 67%, total included studies *n*=3).

Discussion and implications

Automation tools are increasingly being developed and interest in the subject continues to grow with new automation methods and literature overviews being published. Some of the automation tools have not been fully tested and this is likely to be a barrier to implementation by systematic reviewers. Other tools show promise but have not been developed into consumer level products. As a response to these challenges, working parties have been established to overcome these barriers and establish a set of principles and goals. The findings from this body of research have shown that more efficient working practices are possible through improved duplicate detection and can be made available to the systematic review community without a prolonged research and development period. The clear potential for machine learning algorithms to automate decisions and reduce screening was demonstrated, but has not been realised into a consumer ready product, and therefore is worthy of further research and development. Biomedical databases offer different products which vary in scale and content and researchers should be prepared to search several databases rather than relying on a single database. The title-only screening developed during this research was shown to be effective and demonstrated similar reliability to both predictive screening tools and human screening, and could be used with other automation tools to assist with screening. Progress with automation tools will be accelerated once technical barriers are overcome, and by pursuing proof of concept technologies into consumer ready products and thoroughly evaluating automation tools for reliability.

Keywords

- Abstrackr
- Algorithm
- Automation
- Biomedical database
- Citation screening
- Deduplication
- Expediting Evidence Synthesis
- Machine learning
- PICO based title-only screening
- Rapid Review
- Scoping Search
- Semi-automation
- Systematic Review Assistant-Deduplication Module
- SRA
- Systematic Review

Declaration by Author

This thesis is submitted to Bond University in fulfilment of the requirements of the degree of *Doctor of Philosophy*. I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes three original papers published in peer reviewed journals and one *in press* publication. The core theme of the thesis is developing and investigating methods to increase the efficiency of producing evidence-based medicine research, specifically systematic reviews. The publications and thesis represents my own original work under the supervision of Paul Glasziou, Tammy Hoffmann and Elaine Beller. The inclusion of co-authors is demonstrative of input between multidisciplinary researchers and acknowledges collaboration with team-based research methods.

John Rathbone 31/08/2017

Research outputs

Peer-reviewed publications:

1) Rathbone J, Carter M, Hoffmann T, Glasziou P. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. **Systematic Reviews (2015)** 4**:6.**

2) Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. **Systematic Reviews (2015) 4:80.**

3) Rathbone J, Carter M, Hoffmann T, Glasziou P. A comparison of the performance of seven key bibliographic databases in identifying all relevant systematic reviews of interventions for hypertension. **Systematic Reviews (2016) 5:27.**

4) Rathbone J, Albarqouni L, Bakhit M, Beller E, Byambasuren O, Hoffmann T, Scott AM, Glasziou P. Expediting citation screening using PICo based title-only screening for identifying rapid reviews **Systematic Reviews (2017)** (*In Press*).

Published and Presented Conference Abstracts

5) Rathbone J, Carter M, Hoffmann T, Glasziou P. Solving research waste with better duplicate detection. **October 2015. European Journal of Public Health Conference,** 25 (suppl. 3).

Declaration of authors contribution

Publications co-authored	Statement of contribution
Rathbone, J., M. Carter, T. Hoffmann, P. Glasziou (2015). *Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module.* Syst Rev 4(1)	JR (70%), MC (20%), TH (5%), PG (5%)
Rathbone, J., T. Hoffmann, P. Glasziou (2015). *Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers.* Syst Rev 4(1)	JR (90%), TH (5%), PG (5%)
Rathbone, J., M. Carter, T. Hoffmann, P. Glasziou (2016). A comparison of the performance of seven key bibliographic databases in identifying all relevant systematic reviews of interventions for hypertension. Syst Rev 5(27)	JR (80%), MC (10%), TH (5%), PG (5%)
John Rathbone, Loai Albarqouni, Mina Bakhit, Elaine Beller, Oyungerel Byambasuren, Tammy Hoffmann, Anna Mae Scott, Paul Glasziou *(in Press).* Expediting citation screening using PICo based title-only screening for identifying studies in Rapid Reviews	JR (74%), LA (5%), MB (5%), EB (2%), OB (5%), TH (2%), AMS (5%), PG (2%)

Copyright declaration

Acknowledgements

I would like to express my gratitude to my supervisors Paul Glasziou, Tammy Hoffmann, and Elaine Beller for their continuous support and guidance during my PhD and for providing such an enjoyable academic study environment.

I would also like to thank my colleagues at the Centre for Evidenced-Based Practice (CREBP), Bond University for their support and collegiality, and for making the research environment so informal and engaging. Special thanks to Matt Carter whom I collaborated with during my research. In addition, I wish to acknowledge with thanks the Australia Fellowship grant from the National Health and Medical Research Council, and the academic support from the school of Health Sciences and Medicine, Bond University.

My thanks to the external examiners - Hans Lund, Iain Marshall and Karen Robinson, for giving-up so freely their precious time, and for their thoughtful consideration of the thesis.

And to my wife, Evelyne, for the ongoing support and planting the idea that led to uprooting to the antipodes and embarking upon a PhD programme.

Table of Contents

List of Tables

Chapter 3 Biomedical database coverage

Chapter 4 Duplicate detection within bibliographic records

List of Figures

Chapter 1 Introduction

Chapter 3 Biomedical database coverage

Chapter 5 Semi-automated citation screening

Chapter 6 Screening citations using PICo based title-only screening

List of Abbreviations

aHUS – Atypical Haemolytic Uraemic Syndrome

API - Application Programming Interface

CDSR - Cochrane Database of Systematic Reviews

CENTRAL - Cochrane Central Register of Controlled Trials

CINAHL - Cumulative Index to Nursing and Allied Health Literature

CONSORT - Consolidated Standards of Reporting Trials

CREBP – Centre for Research in Evidence based Practice

DARE - Database of Abstracts of Reviews of Effects

DICOM - Digital Imaging and Communications in Medicine

ECHO - Echocardiography

EMBASE - Excerpta Medica Database

EBM – Evidence Based Medicine

FN – False Negative

FP – False Positive

HTA - Health Technology Appraisal

IBM - International Business Machines Corporation

ICASR - International Collaboration for the Automation of Systematic Reviews

LILACS - Latin American and Caribbean Health Sciences Literature

MEDLINE - Medical Literature Analysis and Retrieval System Online

MeSH - Medical Subject Headings

PICO - Participants, Interventions, Comparators, Outcomes

PICo - Participants, Interventions, Comparators, outcomes

PICOS - Participants, Interventions, Comparators, Outcomes, Study design

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analysis

RSE – Reduction in Screening Effort

SE – Screening Effort

SRA-DM – Systematic Review Assistant – Deduplication algorithm

TN – True Negative

TRIP - Turning Research Into Practice

"*In 18th century England, James Hargreaves [an illiterate weaver] invented the Spinning Jenny, and Richard Arkwright [wig maker & inventor] pioneered the water-propelled spinning frame which led to the mass production [automation] of cotton. This was truly revolutionary. The cotton manufacturers created a whole new class of people - the urban proletariat. The structure of society itself would never be the same*."

A. N. Wilson on Society