



The Adoption and Effectiveness of Automation in Health Evidence Synthesis

Anneliese Downey Arno

University College London

Submitted for the degree of Doctor of Philosophy (PhD)

September 2021

Declarations

I, Anneliese Downey Arno, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

25 September 2021

Acknowledgements

First I'd like to extend my deepest thanks to my three advisors: James Thomas, Julian Elliott, and Byron Wallace. You have each provided me with invaluable wisdom, both academically and personally, and I feel truly privileged to have had each of you as mentors in my PhD journey.

I'd also like to thank Ian Shemilt and Tari Turner for their collaborations during this research. I am a better researcher because of your guidance: Ian, in helping me to re-learn all of my economic analysis skills and then some, and Tari, in helping a quantitative mind see the beauty in qualitative analysis.

Many at Covidence deserve my sincere gratitude, but Andy O'Neil and Bo Jeanes in particular. Their generous help was essential to the trial presented in this thesis.

Thank you to my family for providing me with so many opportunities and so much support.

Finally, to my amazing partner Rich: you have somehow managed to make pandemic lockdowns combined with the stress of research feel easy. This PhD would not have happened without you.

Abstract

Background:

Health systems worldwide are often informed by evidence-based guidelines which in turn rely heavily on systematic reviews. Systematic reviews are currently hindered by the increasing volume of new research and by its variable quality. Automation has potential to alleviate this problem but is not widely used in health evidence synthesis. This thesis sought to address the following: why is automation adopted (or not), and what effects does it have when it is put into use?

Methods:

Roger's Diffusion of Innovations theory, as a well-established and widely used framework, informed the study design and analysis. Adoption barriers and facilitators were explored through a thematic analysis of guideline developers' opinions towards automation, and by mapping the adoption journey of a machine learning (ML) tool among Cochrane Information Specialists (CISs). A randomised trial of ML assistance in Risk of Bias (RoB) assessments and a cost-effectiveness analysis of a semi-automated workflow in the maintenance of a living evidence map each evaluated the effects of automation in practice.

Results:

Adoption decisions are most strongly informed by the professional cultural expectations of health evidence synthesis. The stringent expectations of systematic reviewers and their users must be met before any other characteristic of an automation technology is considered by potential adopters. Ease-of-use increases in importance as a tool becomes more diffused across a population.

Results of the randomised trial showed that ML-assisted RoB assessments were non-inferior to assessments completed entirely by human researcher effort. The cost-effectiveness analysis showed that a semi-automated workflow identified more relevant studies than the manual workflow and was less costly.

Conclusions:

Automation can have substantial benefits when integrated into health evidence workflows. Wider adoption of automation tools will be facilitated by ensuring they are aligned with professional values of the field and limited in technical complexity.

Impact statement

The results of this research will support the implementation of automation in evidence synthesis practice with strong empirical evidence. These findings justify and will promote a significant shift in methods of health evidence synthesis. The integration of such a powerful approach into the evidence synthesis pipeline will translate to higher quality in health guidelines, as they are able to connect the most current knowledge to day-to-day practice more quickly than ever before. As this knowledge translation grows more efficient, outcomes will improve for individuals accessing health services worldwide.

Health guidelines play a crucial role in the everyday lives of many millions of people around the world. Efficient identification and summarisation of current knowledge – evidence synthesis – is vital to the support of these health guidelines. Evidence is currently produced at an ever-increasing rate, outpacing the ability of evidence syntheses to keep up. Applying automation to the evidence synthesis process presents a potential solution to this issue, but care must be taken to ensure that the high quality of health guidelines is maintained. Automation is also seldom used, and the reasons for this are unclear. The research presented in this thesis addresses both topics, and therefore addresses several barriers which had previously been hindering the broader adoption of automation.

Cultural factors are the greatest contributor to the trust placed in automation processes and must be reliably demonstrated before implementation of a semi-automated process is considered. Once individuals are open to integrating an automation tool, the first group to do so highly value the ability to control the parameters of the tool. This value drops over time as additional groups adopt a tool, while an emphasis on a tool's ease of use grows in importance. This finding provides clarity to key stakeholders regarding the prerequisite for a shift in research methodology as it relates to the use of automation.

This research also examined two implementations of automation in evidence synthesis. In both cases, it was found that automation either maintained equal quality to a manual method, or it improved quality. Concerns that automation would require a trade off in quality in favour of time or resource savings turned out to be

unnecessary in these instances, thus removing a critical barrier to future implementation.

To summarise, this research has removed several important barriers to the adoption and implementation of automation in health evidence synthesis. By doing so, it supports a shift in the methodological practices of evidence synthesis professionals. In addition, developers of automation tools should take note of these findings in order to better validate their product development, as well as to better target the needs of their intended audience. Finally, key organisations which influence the field of health evidence synthesis are now equipped with empirical evidence to inform their ongoing discussions relating to the potential use of automation.

Table of Contents

<i>Declarations</i>	2
<i>Acknowledgements</i>	3
<i>Abstract</i>	4
<i>Impact statement</i>	5
<i>List of Tables</i>	10
<i>List of Figures</i>	10
<i>List of Abbreviations</i>	11
Chapter 1. Introduction	12
<i>Historical context of Evidence-based Medicine</i>	13
<i>Conceptual definitions</i>	15
<i>My pathway to EBM and to this PhD</i>	18
<i>Automation and systematic reviews</i>	20
<i>Thesis themes and structure</i>	21
<i>Conclusions of this thesis</i>	24
<i>Chapter references</i>	25
Chapter 2. Literature review	27
<i>Chapter overview</i>	27
<i>Systematic reviews are resource-intensive</i>	28
<i>Increasing data growth</i>	30
<i>Innovations in systematic reviews</i>	32
<i>Adoption of automation in systematic reviews</i>	40
<i>Conclusions</i>	41
<i>Chapter references</i>	43
Chapter 3. Methodological frameworks	49
<i>Chapter overview</i>	49
<i>The conceptual frameworks used in this thesis</i>	50
<i>Levels of automation</i>	53
<i>Trust in automation</i>	57
<i>Diffusion of Innovations</i>	62
<i>Research questions</i>	66
<i>Summary</i>	71
<i>Chapter references</i>	72
Chapter 4. Acceptability	74
<i>Chapter overview</i>	74

<i>Introduction</i>	75
<i>Methods</i>	77
<i>Results</i>	81
<i>Discussion</i>	93
<i>Conclusion</i>	105
<i>Chapter references</i>	106
Chapter 5. The User Journey	108
<i>Chapter overview</i>	108
<i>Introduction</i>	109
<i>Methods</i>	112
<i>Results</i>	117
<i>Discussion</i>	124
<i>Conclusion</i>	130
<i>Chapter references</i>	131
Chapter 6. Validity	132
<i>Chapter overview</i>	132
<i>Introduction</i>	133
<i>Methods</i>	137
<i>Results</i>	147
<i>Discussion</i>	151
<i>Conclusion</i>	156
<i>Chapter references</i>	157
Chapter 7. Economic evaluation	159
<i>Chapter overview</i>	159
<i>Introduction</i>	160
<i>Methods</i>	163
<i>Results</i>	171
<i>Discussion</i>	176
<i>Conclusion</i>	183
<i>Chapter references</i>	185
Chapter 8. Discussion	186
<i>An evidence-based road map of automation in health evidence synthesis</i>	188
<i>Implications for research and practice</i>	197
<i>Conclusions</i>	200
<i>Chapter references</i>	201

Appendix A. COREQ checklist	202
Appendix B. Interview instrument.....	204
Appendix C. CROSS checklist.....	205
Appendix D. Survey instrument.....	208
Appendix E. CONSORT checklists.....	219
Appendix F. CHEERS checklist.....	224
Appendix G. Data availability statements	227

List of Tables

Table 2.1. Automation approaches to screening reduction, from O'Mara-Eves (2015).....	37
Table 3.1. Levels of automation, from Sheridan and Verplank (1978).....	54
Table 4.1. Participant characteristics.....	82
Table 5.1. Survey participant characteristics	118
Table 6.1. Potential accuracy outcomes of an individual Risk of Bias assessment	145
Table 6.2. Accuracy data for Risk of Bias.....	150
Table 7.1. Characteristics of study arms.....	164
Table 7.2. Effectiveness results by study arm.....	172
Table 7.3. Cost data by study arm.....	173
Table 7.4. Cost results by study arm.....	174

List of Figures

Figure 1.1. The principles of guidelines	16
Figure 1.2. Stages in a systematic review.....	17
Figure 2.1. RCTs indexed in PubMed, estimated by manual indexing (yellow) versus automation (blue), from Marshall et al (2020)	30
Figure 3.1. Diffusion of Innovations adoption curve.....	63
Figure 4.1. Visual overview of guideline developers' opinions via the Diffusion of Innovations characteristics.....	84
Figure 5.1. Observed and expected distributions among Cochrane Information Specialists	118
Figure 5.2. Average rating of RCT classifier versus adopter category.....	120
Figure 6.1. An example RobotReviewer assessment.....	134
Figure 6.2. Time in seconds to complete Risk of Bias using machine learning versus standard (fully manual), from Soboczenski et al (2019).....	136
Figure 6.3. Blank Cochrane Risk of Bias template in Covidence.....	141
Figure 6.4. Risk of Bias form pre-populated with RobotReviewer assistance.....	143
Figure 6.5. Trial study design.....	143
Figure 6.6. Regions of inferiority and non-inferiority.....	146
Figure 6.7. Trial flow diagram	148
Figure 6.8. Effect of RobotReviewer assistance on RoB accuracy, overall and by domain.....	149
Figure 6.9. Effect of RobotReviewer assistance on time to complete RoB	150
Figure 6.10. Person-time results grouped by review.....	151
Figure 7.1. Living COVID-19 evidence map published by the EPPI-Centre.....	161
Figure 7.2. Results of cost-effectiveness analysis.....	172
Figure 7.3. Results of cost-effectiveness sensitivity analysis for time-on-task	175
Figure 7.4. Results of cost-effectiveness sensitivity analysis for precision, lower limit.....	176
Figure 7.5. Results of cost-effectiveness sensitivity analysis for precision, upper limit.....	176

List of Abbreviations

AHRQ	Agency for Healthcare Research and Quality (United States)
AQE	Automated query expansion
CDSR	Cochrane Database of Systematic Reviews
CHEERS	Consolidated Health Economics Evaluation Reporting Standards
CIS	Cochrane Information Specialist
Cochrane IRMG	Cochrane Information Retrieval Methods Group
CRS	Cochrane Registry of Studies
DHSC	Department of Health and Social Care (England)
EBM	Evidence-based medicine
G-I-N	Guidelines International Network
ICASR	International Collaboration for the Automation of Systematic Reviews
ICER	Incremental cost-effectiveness ratio
IQWiG	Institute for Quality and Efficiency in Health Care (Germany)
JBI	Joanna Briggs Institute
LOA	Levels of automation
MAG	Microsoft Academic Graph
ML	Machine learning
MUHREC	Monash University Human Research Ethics Committee
NHMRC	National Health and Medical Research Council (Australia)
NICE	National Institute for Health and Care Excellence
NIH	National Institute of Health (United States)
NLP	Natural language processing
PICO	Participants, interventions, comparisons, and outcomes
RCT	Randomised controlled trial
RoB	Risk of Bias
SME	Subject matter expert
VDM	Visual data mining
WHO	World Health Organization

Chapter 1. Introduction

Behind the scenes of every evidence-based healthcare decision that is made there are years, if not decades, of research that informed the training and the policies guiding that decision. Health evidence syntheses, such as health technology assessments, rapid reviews, recommendations, guidelines, and others, are the bridges that connect that research to practice. The systematic review, one of the most widely known and used forms of evidence synthesis, aims not only to summarise all of the existing evidence on a topic, but also to do so in a reproducible and transparent way [1]. Systematic reviews form the foundations of clinical guidelines, of drug licensing and regulation, and of health policy; the efficiency of systematic review production and their ultimate quality are therefore of enormous societal significance.

Due to inefficiencies in research conduct and reporting, however, health evidence synthesis is both time- and resource-intensive, leading to suboptimal translation of knowledge into practice [2]. Decisions taken using out-of-date evidence syntheses, or taken without access to evidence syntheses at all, risk missing out on the latest benefits of scientific pursuit, needlessly harming population health in the process.

Automation has been proposed as a solution (or partial solution) to this problem, as new natural language technologies may have the potential to aid or to complete many of the tasks currently performed by people. Automation tools exist that support much of the systematic review process, from search strategy development, assessing studies for eligibility, risk of bias assessment, data extraction, and even synthesis [3, 4]. With such a wealth of available automation tools, I originally intended for my PhD research to focus on the validation of a select number of them to inform best practice in the integration of automation to systematic reviews. It became clear, however, that despite the apparent widespread availability of tools, the overall state of the health evidence field was not inclined towards the uptake of automation: systematic reviewers generally did not, and do not, use automation.

Such a realisation necessitated asking why this is the case, and partially shifted the direction of this work. The mere provision of evidence that automation

tools ‘work’ is clearly not sufficient to push those in decision-making positions to choose to adopt them, therefore a deeper understanding of that decision-making process is needed. My original focus and interest did remain, however, resulting in the two broad aims which determined the direction of the research presented here.

The two aims of this thesis, both undertaken within the context of health evidence synthesis, are: 1) to understand why automation is or is not adopted; and 2) to discover what can happen in practice when it is adopted. These two aims are interrelated: I aim to describe and analyse decision-making in this context, and to provide evidence to inform those decisions and the people or systems making them. In other words, I will first create a map which describes these decisions and their influences, and then I will create new knowledge to add to this map, which will improve future navigation of the use of automation across health evidence syntheses.

Before the detailed exploration of health evidence methodologies and automation presented in the remainder of this thesis, it is useful to position this discussion within the historical precedents that have led to the current moment. In addition to this historical context and orientation, I will provide my personal history which led me to this research, and which should be considered in framing my unique relationship to the research and its results. The rest of this introduction will outline the structure and main conclusions of the remaining chapters and conclude with what I expect to demonstrate by the conclusion of this thesis.

Historical context of Evidence-based Medicine

Before Evidence-based Medicine: 1753 through the 1960s

The first record of a randomised controlled trial (RCT), generally considered the bedrock of most health-focused evidence synthesis, is widely described as having been conducted by James Lind aboard a Scottish naval vessel in 1753. Seeking to determine which of six different popular treatments was the most effective against scurvy, Dr Lind split sailors suffering from the disease into groups, attempted to control for stage of the disease and living conditions, and randomly assigned one of

the treatments to each group. The cohort randomly allocated to receive fresh fruits improved, proving it to be the most effective treatment.

This concept of a ‘fair comparison’ [5] has since been used as a foundational principle in the scientific method and was most significantly developed throughout the twentieth century. For the purposes of the medical field, the corollary concept of randomisation has become increasingly entrenched in research design.

Randomisation’s generation of groups comparable in invisible and visible characteristics is crucial to accurately determining causation. When groups differ from one another only by random chance, any changes observed are more likely to be attributable to the intervention of interest. The Institute for Quality and Efficiency in Health Care (IQWiG, Germany) identifies a 1948 tuberculosis study as the first modern drug trial to use randomisation. Importantly, this trial included descriptions of the randomisation procedure and baseline characteristics of the participants when it was published. Put into the context of research today, this might be viewed as the beginning of standardised study or trial quality metrics. By the 1960s, randomisation was practically a requirement in medical trial reporting worldwide.

The Rise of Evidence-based Medicine: 1971 through the early 1990s

While trials utilising scientific comparisons had become commonplace by the 1970s, the translation of basic science to medical practice was determined entirely by practitioners’ opinions. Archie Cochrane, the future namesake of the Cochrane Collaboration established in 1993, famously criticised this situation in 1971 in an evaluation of the British National Health Service [6]. In this seminal text, he advocated for scientific evidence to provide the basis of medical care rather than opinion. Several years later, Cochrane made the further suggestion that the lack of “a critical summary, by speciality and subspeciality, adapted periodically, of all randomised controlled trials” was medicine’s “greatest criticism” [7]. Systematic reviews, then termed ‘meta-analyses’, had already been produced in the area of psychology [8, 9], but researchers in obstetrics heeded Cochrane’s call in the late 1980s and published systematic reviews of obstetric practice. Cochrane later praised these and encouraged that their approach should be “copied by other medical specialties” [10].

Occurring in parallel to the academic discussions spearheaded by Archie Cochrane, the 1980s saw the beginning of a shift away from experience and individual opinions holding the most influence over patient care towards a more systematic and evidence-based approach. Many discussions of the evidence-based medicine movement begin with Dr David Sackett and Gordon Guyatt. They initially promoted the term ‘scientific medicine’, but were quickly met with derision from clinicians unimpressed by the implication that current practice was unscientific [11]. The term ‘evidence-based medicine’ (EBM) was introduced instead and has now become the standard term.

EBM Today

Today, evidence-based medicine is widely practiced and formally recognised via educational institutions and a number of international organisations. Many top universities have leading academic centres focused exclusively on evidence-based medicine, training, and production. Important organisations not based in universities include the aforementioned Cochrane Collaboration (now known simply as Cochrane), the Campbell Collaboration, the Joanna Briggs Institute (JBI), and Guidelines International Network (G-I-N). Further, many government bodies have dedicated resources to the development of health guidelines and to the systematic reviews that support these; such organisations include the World Health Organization (WHO), National Institute for Health and Care Excellence (NICE, United Kingdom), the National Medical Health and Research Council (NHMRC, Australia), and the Agency for Healthcare Research and Quality (AHRQ, United States). The links between evidence synthesis and health guidelines will be examined in more detail in the next section.

Conceptual definitions

Having considered both the broader global history that precedes my research, I will now define the concepts underlying EBM. Sackett and Guyatt defined EBM as “the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients” [12]. It has been more succinctly described elsewhere as “the principle of integrating the powerful methods of science into the practice of medicine” [13].

In addition to the connection between scientific evidence and care in practice, Chung and Ram (2009) write that five concepts are central to EBM: “gathering evidence, integrating evidence with experience to arrive at a clinical decision, implementing this decision at the bedside, assessing performance, and staying current with research” [13]. This might be thought of as a simplified cycle in which knowledge is gathered, knowledge is implemented, the implementation is assessed, and knowledge is gathered again to re-start the cycle. This thesis will focus on two of the five elements described: gathering evidence and staying current with research. ‘Health evidence synthesis’ is used throughout this thesis to jointly denote these two concepts.

The research presented in this thesis relates to two major components of synthesising health evidence: health or clinical guidelines, and systematic reviews. The former are defined in the literature as “systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances” [14], and they aim to improve quality of care and patient outcomes. More commonly, guidelines (and evidence-based medicine in general) are described as the combination of clinical experience, patient preferences, and the best research evidence.

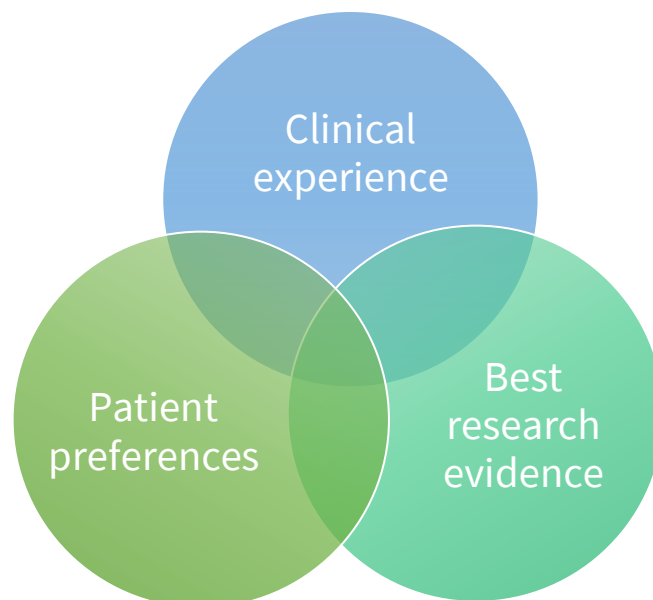


Figure 1.1. The principles of guidelines

Systematic reviews are the evidentiary foundation for guidelines, forming the ‘best research evidence’ arm as described above. Cochrane defines a systematic review as “attempting to identify, appraise, and synthesise all the empirical evidence that meets pre-specified eligibility criteria to answer a specific research question” [15]. It is key that a systematic review has a well-defined and reproducible search strategy, and this characteristic is an important distinction between a systematic review and other types of literature review. By following a protocol established prospectively, systematic reviews aim to eliminate or minimise bias to approach near-complete knowledge on a focused question; literature reviews have no such methodological obligation and instead aim for broad knowledge on a topic. In theory, a meta-analysis, the statistical synthesis of data from comparable sources [16], may or may not be conducted using the studies resulting from a systematic review. In practice, however, it is now expected that meta-analyses also synthesise the results from studies identified through a systematic, and not systematically biased, search strategy. Stages of a systematic review may slightly vary, but generally follow a standard set of steps, which will be referenced throughout this thesis (Figure 1.2).



Figure 1.2. Stages in a systematic review

These steps will be preceded by the formulation of an initial question and of a review protocol which defines the eligibility criteria and methods for the review. In the initial stages, a search strategy must be formulated. A review team, often in consultation with an information specialist or a search specialist, must decide which databases to use, which terms to search and which to exclude, when to conduct these searches, and how to manage and de-duplicate the results. Screening stages see the review team applying their pre-defined eligibility criteria. It is worth noting that it is not uncommon for teams to revise their original search strategy at this stage after conducting a sub-set of initial screening, which has implications for the role automation may play with this task.

During the extraction stage, reviewers will collect information from eligible studies, typically in a structured format. This is particularly true of health

intervention focused systematic reviews (e.g., the efficacy of X treatment for Y condition in Z population) which often restrict inclusion of studies to RCTs and assess quantitative outcomes. Reviewers then assess the quality of included studies. In this way systematic reviews not only collate the results of relevant studies, but also describe the quality of these studies. Various appraisal tools are available [17] and will be discussed in more detail in the relevant chapters.

In the analysis stage, results from included studies are collectively assessed. If conducting a meta-analysis, the review authors will pool the results of included studies to derive a final estimate of the magnitude of the intervention effect (if any). Finally, reviewers publish their findings using their preferred method of dissemination, typically a peer-reviewed journal article.

‘Automation’ in this thesis refers to any tool, process, or system which operates independently of human labour, or with limited human supervision. In relation to systematic reviews, an ‘automation tool’ is “any computer tool that can fully or semi-automate a systematic review task” [18]. Further sub-categories of automation, such as machine learning and natural language processing (NLP), will be discussed in more detail in the following chapter.

My pathway to EBM and to this PhD

A proper historical orientation would be lacking without some personal historical context as well. Though academics ostensibly seek objective knowledge, I am closely intertwined with the nature of my own research. By simply asking the questions I ask, I have positioned myself within my own PhD research, and therefore this should be explicitly described.

On a descriptive level, I consider myself a younger woman emerging from a fairly privileged (in my view) background. As above, while seemingly this should not factor into findings derived from a scientific methodology, on a practical level it cannot be unwound from the pursuit of this PhD. I grew up in a house with a computer and internet access. I am comfortable with smartphones. Automation is not, to me, hypothetical, but rather it is present in my daily life and existence. The framing of my research questions and my interpretation of the meaning of the results might be different if constructed by someone from a different set of life experiences.

There is also the presence and influence of my background in highly quantitative fields: biology and economics. These might seem unrelated, but to me they have always been similarly interested in system-wide behaviour given a discrete set of resources. Biology considers questions such as how the human body allocates units of energy, whereas economics considers questions such as how money supply affects allocation of wages. While both fields are immeasurably more complex than these descriptions, my fascination with them both centres on a sense of efficiency. This fixation on efficiency connects my backgrounds in biology and economics to my current research into EBM and automation. I believe a more efficient system to be a better one. Once again, the entire framing of this thesis is intrinsically entwined with that normative assumption.

After my undergraduate degrees in the fields mentioned in the previous paragraph, and a few years spent working in biological research, I pursued a Master of Science (MSc.) in Global Health. It was here that I was first introduced to guidelines, systematic reviews, and evidence-based medicine. Like many, I was surprised to learn that many interventions initiated by governments or non-profit organisations were not evidence-based at all. My passion for health evidence grew out of the hope that better use of data would not only mean better health outcomes, but also more equitable ones. This value-driven aspiration continues to drive my career and my research today, and so in addition to my belief outlined above that a more efficient system is a better one, I also believe that a more equitable system is a better one (at least when it comes to health outcomes). The fact that guidelines and the research that feeds them are so resource-intensive is an undesirable state, in my perspective, because it constricts the availability of the highest quality evidence, often to those with the highest means. My pursuit of automation is at least in part to remedy this, and to make evidence-based practices more readily available to the entire global community, regardless of their resources and privileges, or lack thereof.

The final piece of my personal pathway to this PhD was my connection with Covidence, an online systematic review tool. I began working for Covidence in June 2015, and this work served as my introduction to many individuals in the systematic review and Cochrane community, including my supervisors, each of whom leads one or more systematic review automation tools. Though Covidence contains no automation itself, I view it as laying the groundwork for future automation potential.

Covidence plays a significant role in this work, particularly in Chapter 6, while tools managed by my other supervisors are featured heavily as well (RobotReviewer, also in Chapter 6; EPPI-Reviewer in Chapter 7; the Cochrane RCT classifier in Chapter 5). Given my position and my connection to Covidence, to Cochrane, and to my supervisors and their respective tools, I was not operating in the world of health evidence synthesis as an outsider. Rather, I may have been known to those who either provided data or collaborated for my research as someone associated with individuals in this field who are leading advocates of automation tools.

The convergence of these influences led me to pursue this PhD. I see automation as a potential opportunity to further the pursuit of my core beliefs in efficiency and equity in health; the ultimate goal, however, remains the promotion of an evidence-based approach to health. It must be determined whether automation furthers this goal with efficient and accurate evidence synthesis, or if it undermines the accuracy or trustworthiness of EBM. My aim with this PhD is to further the evidence base towards the appropriate use of automation, and to determine empirically whether it improves efficiency while also meeting the standards of EBM.

Automation and systematic reviews

As discussed above, evidence-based medicine and systematic reviews are integral to the administration of health systems worldwide. In recent years, however, they have increasingly struggled to stay up to date, both because of increasingly rigorous methodological standards, and because of the explosion in the rate of primary research publication. This situation will be examined in more detail in Chapter 2.

Given these pressures, researchers have looked to automation as a potential solution, or at least mitigating assistant, to keeping health evidence both current and high-quality. Each of the systematic review steps outlined in the previous section (Figure 1.2) may be automated to a certain extent and will be explored in detail in the literature review in the next chapter. The stages described intentionally excluded the protocol stages on the basis of a previous argument in the literature regarding what should and should not be automated [19]. Tsafnat argued that every review is part creative process and part technical process, that automation should focus on the

technical processes, and that “the review protocol is developed much like a recipe that can then be executed by machine.” Automation cannot, and should not, try to write these recipes, and my research incorporates this value assumption. Similarly, automation’s influence on guidelines (Figure 1.1) is currently limited to the best research evidence, as it cannot itself provide human opinions on patient preferences nor clinical experience.

Significant work has been completed in the development of tools for the automation of health evidence synthesis, but automation is not yet widely used [20-22]. The reasons for this are unclear. This situation requires research to determine why automation is not yet widely used, and it is of significant impact to the current and future state of health evidence. Without a methodological paradigm shift, the current rate of evidence production outpaces the capacity of resources available for its analysis and integration into guidelines [2]. That is, without automation, systematic reviews are likely to fall out of date, causing guidelines to decrease in quality, ultimately resulting in a negative impact on health outcomes.

Thesis themes and structure

The current model of evidence synthesis is failing to keep up with the rate of new research, leading to the overall aims of this thesis: to explore the adoption and the effectiveness of automation technologies in health evidence synthesis. Why do individuals, teams, or organisations choose to adopt automation? What happens if they do adopt automation? By exploring these questions, my PhD research aims to strengthen the evidence base used to inform decision-making by identifying and addressing key barriers, facilitators, benefits, and harms of applying automation to health evidence synthesis, and to document and analyse that decision-making in detail. It is further expected that the results of this research may be used to inform technology development and research and development priorities.

The next chapter’s discussion of the state of the field of health evidence synthesis and of systematic reviews will establish an urgent need for more efficient systems. The literature described will also show there are numerous potential solutions available, as well as additional tools which are in ongoing development, but that these are not being widely adopted and put into practice.

With the current state of knowledge established and discussed, I will then draw further on existing literature describing the analytical frameworks which underpin my research. These are used in the design of my projects and in the analysis of their results, and most importantly in translating these results into recommendations for practice and creating meaningful connections among my findings. These frameworks also serve to place my PhD within the wider scope of research methodologies and to locate it within existing knowledge.

Two projects were conducted to explore adoption of automation, and two to explore the effectiveness of automation. Each of these four individual projects will be presented in its own chapter and are outlined in the following sections.

Adoption of automation

Guideline developers are key gatekeepers in the translation of evidence into practice, including evidence produced using automation. Given this important role and the limited adoption of automation, I sought to understand their attitudes towards automation tools in order to better understand potential barriers and facilitators to adoption, and to make recommendations based on these findings.

This qualitative exploration will be detailed in Chapter 4, underpinned by the thematic frameworks presented in Chapter 3. My results will show that guideline developers hold the values of their professional field above all other considerations. In short, guideline developers believe automation must show itself to be as uncompromising in systematic review quality as a human researcher, and any benefits it may or may not offer come second to this criterion.

Like guideline developers, information specialists are key players in the evidence production pipeline. In Cochrane Information Specialists, I had a unique opportunity to analyse and understand their adoption of the Cochrane RCT classifier. As one of the few widely available and relatively widely adopted automation tools advocated within official Cochrane channels, this study is a unique case study providing insights into the adoption of automation in a high-profile systematic review organisation.

This project will be presented in Chapter 5, and my results will show first that my selected thematic framework yields new insights into who trusts in

automation and why. I will show that as health evidence automation diffuses across a population, user experience (the positive or negative subjective quality of overall interactions of a user with a system) increases in importance, while user technical control decreases in importance.

Effectiveness of automation

The second half of my research chapters shift focus towards the effectiveness of automation in the context of health evidence synthesis.

Much of the published research on systematic review automation has focused on screening, a relatively easier task than the appraisal and extraction stages of a systematic review. Even among the studies examining the use of automation in quality assessment, these have focused on efficacy rather than effectiveness. Understanding the effects of automation in ‘real-world’ practice, rather than its efficacy in controlled validation experiments during development, may be helpful to the uptake and implementation of automation. To begin to address this gap, I conducted a randomised trial of the effectiveness of integrating automation into quality appraisal in real-world, ongoing systematic reviews.

This trial will be presented in Chapter 6, and the results will show that assistance from an automation tool did not negatively impact the accuracy of quality assessments in health-focused systematic reviews.

Finally, the COVID-19 pandemic offered an opportunity to examine the cost-effectiveness of automation. With the global emergency, more researchers than ever turned to automation to expedite their syntheses. I worked with researchers maintaining a living map of COVID-19 evidence to investigate the cost-effectiveness of integrating several automation tools into their weekly workflow.

These results are presented in Chapter 7 and will show that switching from an entirely manual workflow to one assisted by automation resulted in both lower costs and higher effectiveness, as measured by study recall (i.e., search sensitivity). Two unique contributions arise from these results: first, cost-effectiveness of automation in health evidence synthesis is rarely presented in the previous literature on this topic. Second, my data showed that there was not only no sacrifice in the systematic review quality, but there was actually an improvement to it.

Conclusions of this thesis

Priorities shift as adoption of automation in health evidence synthesis becomes more widespread, while consistently underpinned by a requirement for sound methodologies. Further, automation shows benefits even under real-world conditions. Combining the results of my research with several analytical frameworks to structure new connections in knowledge, an evidence-based roadmap for the adoption and implementation of automation in health evidence synthesis will be created by the conclusion of this thesis. By doing so, this PhD will newly equip the field of health evidence synthesis to move forward with automation as an indispensable part of its toolkit

The following chapter will provide a thorough grounding in the existing literature on the automation of health evidence synthesis. We will see that the data deluge situation urgently calls for a shift in methodologies, and that automation has potential to address these.

Chapter references

1. Khan, K.S., et al. *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews*: NHS Centre for Reviews and Dissemination; 2001.
2. Elliott, J.H., et al. Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of Clinical Epidemiology* 2017; 91:23-30.
3. Tsafnat, G., et al. Systematic review automation technologies. *Systematic Reviews* 2014; 3(1):74.
4. O'Mara-Eves, A., et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 2015; 4(1):5.
5. InformedHealth.org. The history of evidence-based medicine. 8 September 2016 2016. <https://www.ncbi.nlm.nih.gov/books/NBK390299/> (accessed 1 April 2021).
6. Cochrane, A.L. Effectiveness and efficiency: random reflections on health services. 1971.
7. Cochrane, A., *A critical review, with particular reference to the medical profession. Medicines for the Year 2000*. 1979, London: Office of Health Economics.
8. Smith, M., G. Glass, and T. Miller. Meta-analysis of psychotherapy. *American Psychologist* 1980; 41:165-180.
9. Smith, M.L. and G.V. Glass. Meta-analysis of psychotherapy outcome studies. *American Psychologist* 1977; 32(9):752.
10. Cochrane, A. Foreword in I Chalmers, M Enkin, MJNC Keirse, eds. *Effective Care in Pregnancy and Childbirth* 1989.
11. Sur, R.L. and P. Dahm. History of evidence-based medicine. *Indian Journal of Urology* 2011; 27(4):487.
12. Sackett, D.L. Evidence-based medicine. *Seminars in Perinatology*; 1997: Elsevier; 1997. p. 3-5.
13. Chung, K.C. and A.N. Ram. Evidence-based medicine: the fourth revolution in American medicine? *Plastic and reconstructive surgery* 2009; 123(1):389-398.
14. Woolf, S.H., et al. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. *BMJ: British Medical Journal* 1999; 318(7182):527-530.
15. Library, C. About Cochrane Reviews. 2021. <https://www.cochranelibrary.com/about/about-cochrane-reviews> (accessed 1 April 2021).
16. John, M. *A dictionary of epidemiology*: Oxford university press; 2001.
17. Critical Appraisal Tools. 2020. <https://www.unisa.edu.au/research/Health-Research/Research/Allied-Health-Evidence/Resources/CAT/>.

18. Bannach-Brown, A. Automated Technologies: Systematic Review & Meta-Analysis. *Statistical Society of Australia: Queensland Branch Seminar*; 2019; Online; 2019.
19. Tsafnat, G., et al. The automation of systematic reviews. *BMJ: British Medical Journal* 2013; 346.
20. Scott, A.M., et al. Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. *Journal of Clinical Epidemiology* 2021.
21. van Altena, A.J., R. Spijker, and S.D. Olabarriaga. Usage of automation tools in systematic reviews. *Research Synthesis Methods* 2019; 10(1):72-82.
22. Thomas, J. Diffusion of innovation in systematic review methodology: why is study selection not yet assisted by automation. *OA Evidence-Based Medicine* 2013; 1(2):1-6.

Chapter 2. Literature review

Why is this research needed?

Chapter overview

This literature review will build upon the definitions and historical context laid out in the introduction and provide the background knowledge critical to understanding and contextualising this thesis. Academic literature will be presented showing that systematic reviews are resource-intensive and becoming unsustainable due to the exponential growth in data production. Existing knowledge about innovations in evidence synthesis will be discussed, and gaps in knowledge will be highlighted. The chapter will conclude with available evidence on current adoption of automation in systematic reviews. Overall the literature will show that health evidence synthesis would benefit from automation, but that adoption is slow.

Systematic reviews are resource-intensive

The introduction of this thesis established the important role systematic reviews play in translating knowledge into practice in health. Guidelines are often publicly funded and they impact the general public's access to and experience of healthcare; systematic reviews that support guidelines therefore need to be robust and transparent to future examination. In addition to feeding into health guidelines, the focus systematic reviews have brought to primary research methods has contributed to an increase in primary research quality [1]. The high quality of systematic reviews is therefore of significant importance [2, 3]. Quality, in this case, should be used to refer not only to the content of the review, but also of the timeliness of its content – that is, how up-to-date and currently applicable is the information contained?

The resources required to complete a systematic review of high quality are substantial [4]. Each of the main steps of a review (Figure 1.2) requires individuals with high levels of training and expertise. With respect to the time required, a 2017 study published in the *British Medical Journal* found the mean time to complete and publish a systematic review was 67.3 weeks [5]. The study also concluded that funded reviews took more than 50% more time to complete. In addition, the median time from the latest search of a systematic review until its date of publication was 8 months [6], further risking that the information contained in the review is not entirely up-to-date.

Perhaps most relevant to the research presented in this dissertation is that the mean yield rate (the proportion of studies included out of the total screened) was 2.94%. Using an estimate of 1 minute per study screened [7] and a 40-hour work week, every 10,000 studies retrieved during the search stage represents over four weeks of researcher time spent on screening studies that do not result in any helpful data for the review. This time-use estimate is even greater when it is considered that many tasks are done in duplicate in a review, with further time required to resolve discrepancies in researchers' screening results (approximately 5 minutes per conflicting decision, according to one study [7]). While streamlining to a single reviewer workflow might appear tempting, a 2019 systematic review of the literature found studies were consistently missed when only one reviewer screened studies [8].

Of equal or greater importance, the authors found that eligible studies missed when using a single reviewer often would have changed the conclusions of the resulting meta-analysis had they been correctly identified and included. More recent publications continue to confirm this conclusion; Gartlehner et al (2020) showed that screening by a single reviewer missed 13% of eligible studies, while screening by two reviewers only missed 3% [9]. Perhaps as a result of these issues, the literature shows that the median time from a primary study's publication to its inclusion in a systematic review ranges from 2.5 to 6.5 years [10].

Updating a review requires further resources and consequently difficult judgements in the prioritisation of research resources [11, 12]. Following completion of a review, it will need to be updated at regular intervals so the information continues to be relevant and accurate, but only a minority of reviews are updated within the recommended time frames [13]. Shojania et al conducted a survival analysis which examined the average time to changes in evidence sufficient to substantially change the conclusions of an existing systematic review; they found that within 2 years, 23% of reviews required updating based on new evidence [14]. This likely contributed to Cochrane changing its former guidance of a two-year updating schedule in favour of one focused on prioritisation of systematic reviews most in need of updating [15].

Health guideline developers face challenges owing to the difficulties of keeping systematic reviews up to date. Prior research has shown that in some fields, some systematic reviews were already out of date by the time that they were published [6, 16]. Consequently, guidelines run a risk of working from out-of-date information, and thus risking inaccurate recommendations for patient care.

Considering the typical time to include primary research in a systematic review, to complete and to publish a systematic review, the lag time between searches and review publication, and the difficulty of prioritising and completing review updates, the evidence begins to draw a picture of a slow-moving health evidence pipeline.

Increasing data growth

Alongside the challenge already presented by the time-intensive evidence-to-practice pipeline, current rates of data growth often exceed the ability of researchers to keep up with screening tasks [17]. The growth rate of articles published each year has been increasing for decades, and does not show any signs of stopping [18]. As of 2010, eleven systematic reviews of trials were published per day, along with 75 trials [19]. Consider this in comparison to the previous decade: in 1995, 429 meta-analyses were published in PubMed, compared to 4739 in 2011 [20]. In the nearly three decades since the founding of the Cochrane Collaboration in 1993, the rate of evidence publication has increased more than eleven times over. Figure 2.1, reproduced from Marshall et al (2020) shows the rate of randomised trial publication on PubMed; the trend is clear, whether examined by manual effort or by automated estimates [21].

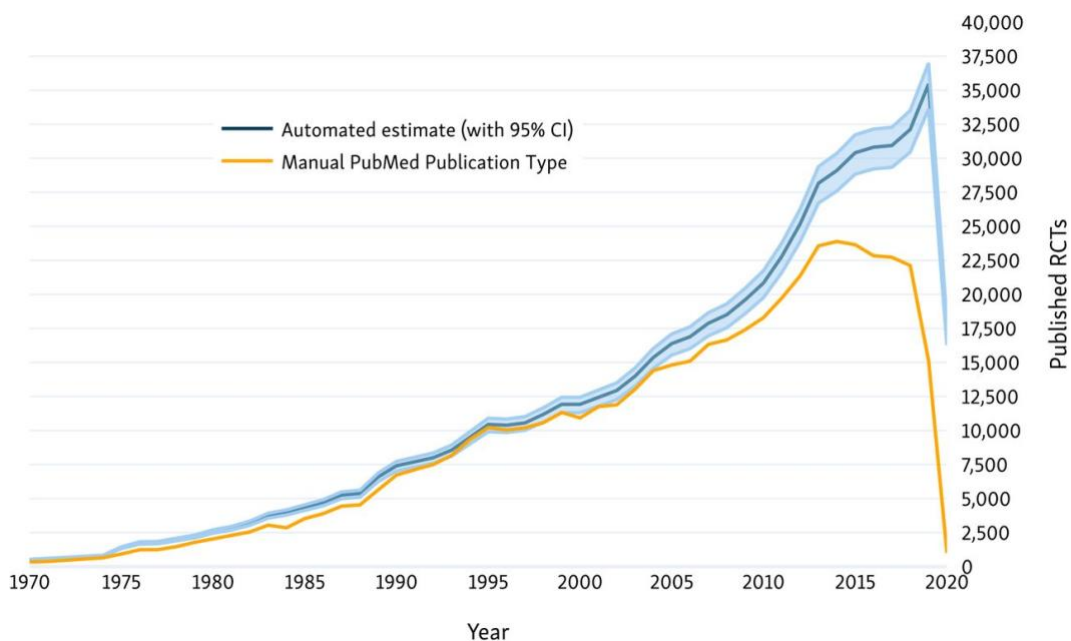


Figure 2.1. RCTs indexed in PubMed, estimated by manual indexing (yellow) versus automation (blue), from Marshall et al (2020)¹

¹ Note: This paper was published in May 2020, and this accounts for the seeming drop-off for 2020 RCTs.

Systematic reviews and evidence synthesis are not limited to trials, of course, and including other types of evidence also contributes to data growth. Growth in case reports and non-systematic reviews has been even faster, and these too potentially contribute to an increased workload in health evidence synthesis [19].

Superficially, a wealth of information might seem to be inherently positive. Surely with so much information, decisions can always be well-informed. However, high availability does not necessarily indicate high quality of data, and indeed can mean the opposite (discussed further in the following section). Bountiful data is also negative in that it is labour-intensive to process it all. It is easy to wonder if such a circumstance as that we see today was ever foreseen in the early days of EBM, but it has been the subject of academic discussion at least since the 1990s. In a 1994 editorial, Doug Altman called for “less research, [but] better research” [22]. From the perspective of systematic reviewers and health guideline developers, this ‘era of data deluge’ – information becoming available at a faster rate than it can be put into use and/or usable formats – is unsustainable and worsening without a major paradigm shift in health evidence synthesis methodology. The evidence-to-practice pipeline needs to adapt.

The dangers of messy methods

The era of data deluge was clearly evident before 2020, but the COVID-19 pandemic threw a sharp spotlight onto the problems that can occur when evidence is needed quickly. Though the scientific community sprang into action at the onset of the pandemic and quickly produced evidence syntheses, the quality and consequent utility of such synthesis was variable to a problematic degree. Abbot et al (2021) conducted a systematic review of reviews to “explore the relationship between review quality and researcher, policy, and media interest” [23]. This study provides three noteworthy and relevant findings: (1) reviews were often of low quality, (2) review quality had no association with review impact, as measured by altmetrics and citations, and (3) a significant number of reviews had conflicting findings. Stamm et al (2021) similarly found most COVID-19 evidence syntheses “fell short of basic methodological standards” [24]. In short, in the case of COVID-19, rapidly conducted reviews were not conducted well, and their low quality was no hindrance on the reach of their results. Policy makers and the public trust in systematic reviews

overall and put their evidence into action even when that evidence is discordant and of low quality.

Innovations in systematic reviews

Conventional wisdom dictates that you can only select two of the three attributes of fast, good, and cheap. Continued innovations in machine learning and natural language processing, however, may eventually challenge that assumption, but significant caution is warranted before trusting these tools. In the context of health evidence syntheses, by reducing the resources required to conduct and maintain systematic reviews and meta-analyses, automation presents one possible means of addressing the data deluge challenge, and one way to discourage the creation and impact of low-quality evidence. Alternatively, automation may actually have the opposite effect and enable mass production of low-quality evidence. Research is therefore needed to ensure proper consideration has been taken in the validation of any automation tools.

The following sections will describe various innovations in automation which aim to address some of the specific challenges in systematic reviews established in the preceding section. All of these innovations are potential targets for further research. Where available, primary literature investigating efficacy and effectiveness is described, and attention drawn to gaps in the literature which might be useful research targets.

Each stage of a systematic review, as defined in Chapter 1 (Figure 1.2), presents a potential target for automation. That is, the automation strategies examined in this thesis operate on the existing framework of how to complete a systematic review. Discussion on how and whether to change the broader format of evidence synthesis exist elsewhere [25, 26].

Most often, the aim of using automation in systematic reviews is to reduce the time required from specialists on lower level tasks, such as screening, so they may redirect effort to more complex tasks [27]. As automation becomes more sophisticated, however, it may also be used for more complex tasks, allowing human researcher effort to be rendered unnecessary or to be redirected to increasingly

nuanced tasks. Some call for the field to move towards a future vision in which an entire systematic review can be conducted by a computer programme, with the only human input being the protocol [28].

Living Systematic Reviews

Living Systematic Reviews (LSRs) present one such scenario in which the entire systematic review process could be expedited via automation, or even entirely automated. LSRs aim to create a review which is kept constantly up-to-date, incorporating the latest available evidence in real time as it becomes available [29]. Automating or partially automating every stage of a typical systematic review would certainly make this endeavour more achievable, with human effort shifted towards final judgements informed by the preceding automated steps. Though not yet automated, LSRs represent an important shift in the framing of systematic reviews: from a static and discrete data synthesis to a dynamic and real-time summary of current knowledge. Activities thus far have largely focused on the monitoring of evidence availability. Several LSRs have already been published [30-32], and main challenges identified have included how to prioritise key topics and maintain consistent methodology [33].

Search

The search stages of a systematic review – searching for references and managing them prior to beginning screening – present a simultaneously highly specialised and time-consuming task. Recall and precision are typically two metrics used to assess the strength of a search. Recall, or sensitivity, refers to the proportion of studies identified out of the total number of relevant studies, while precision, or specificity, is the proportion of identified studies which are relevant to the search question. Perfect recall means every eligible study has been found, while perfect precision would mean the search retrieved zero ineligible studies. It is practically impossible to achieve perfect recall and precision, so trade-offs are in order.

Systematic reviewers often operate on the assumption that current approaches achieve close to perfect recall and tend to be reluctant to sacrifice this ostensibly ‘perfect’ model by accepting lower precision, despite the investment of time required to screen large reference sets [34-36]. Automation could, in theory, achieve the same

or improved recall with the same time resource allocation as a fully manual search. As importantly, and technological development depending, this maintained or improved recall could simultaneously improve precision, saving reviewer time without the methodologically costly trade-offs.

Because trials are sometimes found only on one database, multiple databases should be searched to maximise recall under the current standard workflow [37-39]. Each database, however, tends to have its own specialist terms and operators, requiring extensive training for researchers to make the best use of the databases [40-42]. Automated translation between databases is therefore another target of automation development [43].

Microsoft Academic Graph (MAG) is one example of an effort to shift away from the limitations of a multi-database standard workflow by making use of automation. MAG uses a graph structure to map citation information, as would be found on a database, with information about its connections to other citations. This relationship information about the relationships between documents is frequently updated by software which automatically extracts structured identifying information from publications. The availability of a single unified data set of not only publications, but also the meta-data associated with them, unlocks myriad possibilities of further automation in terms of searching. MAG also has been found to be a less costly and more effective alternative to human-curated content [18].² More broadly, MAG represents a shift away from databases which rely on publishers indexing their data, and instead relies on wide automated searches of information on the internet. Such a model has the potential to be more comprehensive than manual searches, avoiding selection biases such as location and language in search

² Shortly before submission of this thesis, it was announced that MAG would close at the end of 2021; SemanticScholar and OpenAlex are taking over the existing data feed [44].

strategies, but creates new challenges in creating clean data from highly variable web page sources.

Tsafnat et al (2014) identified two, non-mutually exclusive approaches of automation systems during the development of a search strategy: first in improving the precision and recall of queries, and second in prompting the user to improve their query themselves [43]. Search precision and recall might be improved through Automatic Query Expansion (AQE): algorithms that modify the user's query before it is processed [45]. AQE can be integrated into the database search interface itself or can be third-party programs. Strategies of AQE include automated inclusion of synonyms, clarification of their word choices, and spelling corrections [43]. User prompts can be observed in several search engines, in which keyword suggestions might appear as a user enters their query.

Deduplication of search results is another critical part of the search stage and could potentially be supported by automation programs. Many reference management programs already include partially automated deduplication algorithms, and evidence shows they perform fairly accurately [46, 47].

Screening

There are various potential approaches to applying automation to screening. As a consequence of the preference for high recall/sensitivity noted previously, in addition to fragmented bibliographic data and absence of quality metadata, search results tend to have low precision/specificity. Throughout the screening stages, the overwhelming majority of records retrieved in the previous step are excluded under typical review conditions [48]; significant reviewer time is thus spent on ultimately irrelevant work. The screening tasks of systematic reviews have been the subject of the majority of published studies concerned with the automation of health evidence synthesis [49]. Despite this, disparate methods and conclusions mean it remains unclear which approach is best.

Machine learning (ML) generally refers to the approach of providing an algorithm with a set of inputs ('training data') and desired outputs, from which it derives a new general rule ('learning'). This can then be applied to a new set of inputs to predict new outputs. In the context of screening in a systematic review, an

example use case might be a ML model created by a review author to align with their prespecified review protocol; it would then be able to screen records on its own based on the training data (a sample of included and excluded records) created prior to beginning the review. Active learning [50], a specific type of ML, can continually update its training inputs and consequent predicted outputs rather than following static program instructions, as is typical for conventional statistical model. In the context of screening during systematic reviews, active learning usually refers to the continual use of previous screening decisions to train the ML prediction algorithm [51]. The idea is to reviewer decisions in real-time and continually use these as training data, ideally building a better ability to predict screening decisions over time until a point when it can screen autonomously.³

A systematic review of current automation approaches [49] identified four methods in the published literature of semi-automating screening stages of systematic reviews: (1) reducing the number of references to manually screen; (2) using text mining (e.g., ML) as a second screener; (3) increasing screening speed; and (4) prioritisation of references. Prioritisation and reduction in the number of references to screen were the two most common approaches examined in the literature; see Table **2.1**, reproduced from the O'Mara-Eves (2015) [49]. One additional application could be to use automation as a tiebreaking vote on previous conflicting votes, though this was not identified in any existing literature.

³ Note that this relies on the assumption that reviewer decisions are correct – ethicists and commentators have raised valid concerns throughout the literature about the risk of ML further entrenching flawed methodologies and/or existing human biases. Health evidence synthesis should by no means be considered exempt from this risk.

Table 2.1. Automation approaches to screening reduction, from O'Mara-Eves (2015)

Workload reduction approach	Number of studies
Reducing number needed to screen	30
Text mining as a second screener	6
Increasing the speed of screening	7
Improving workflow through screening prioritisation	12

In a prioritisation approach, the full list of references retrieved during the search is sorted such that those at the top are the most likely to be relevant. This may or may not include an active learning approach which observes reviewer decisions and periodically updates the prioritisation of the reference list. While technically this approach does not itself reduce overall workload, it can potentially expedite reviews and reduce the overall time required to complete them by better focusing researcher effort and permitting full text screening to commence before title and abstract screening is complete in the knowledge that most relevant records will be found early in the process [49]. However, there is also risk that prioritisation could prejudice human decisions.

Prioritisation classifiers can also be used by applying a threshold to the list and excluding references within a 'negative prediction zone' [35, 52-54]; when used in this way, they can indeed reduce overall workload, though the overall reduction is dependent on the threshold selected, which is itself dependent on previous human decisions. The Cochrane RCT classifier is one such tool that works in this manner by discarding studies below a particular prediction threshold [55]. Chapter 5 reports on a qualitative study relating to the RCT classifier; the tool will be described in greater detail in that chapter. Similarly, Chapter 7, which presents an economic evaluation of automation, makes use of prioritised screening and will discuss its specific application in that chapter.

Several tools which use the active learning approach are already available. Abstrackr is one such tool that records human screening decisions and uses these data to build automated screening models [56]. EPPI-Reviewer Web (ER-Web) also includes in-built machine learning tools as well as active learning tools which can be user-constructed [57].

DistillerSR, a commercially available systematic review software, has also recently introduced DistillerAI which uses user decisions to predict and apply (with

user permission) future decisions [58]. This tool is able to act independently or to act as a second reviewer, and the latter approach was independently evaluated by Gartlehner et al (2019) [59]. This study first trained DistillerAI with the decisions of dual screeners on 300 randomly selected citations out of a set of 2472 total. After training the automation tool, the remaining studies used DistillerAI as its second screener; this approach resulted in reduced sensitivity, and the authors concluded that DistillerAI's accuracy is currently insufficient to replace human screening.

Evidence has shown that use of ML for screening can reduce reviewer workload substantially [60]. Some published reports have shown no adverse effect on recall at all [61], while others have shown recall lowered to 70% [62]; this variation undoubtedly affects researcher confidence in automation as a whole. Each of these cases targeted specific sets of citations, and it will be important for the field moving forward to evaluate ML performance against a broad range of reviews and topics, as well as to develop open evaluation datasets to encourage internally valid comparisons.

In addition to these approaches, natural language processing (NLP), which aims to teach computers to intelligently analyse language and word associations, can be used in various ways to expedite this process. Currently, this includes decision support tools which use NLP to highlight content to inform an individual reviewer's decision [63-65]. Visual data mining (VDM) may also be applied to create a visual representation of connections between already classified documents, and then to provide the screeners with prompts to reduce their screening time. O'Mara-Eves et al identified five evaluations of VDM [66-70], the results of which suggested that humans can indeed screen more quickly with VDM assistance without substantially changing their screening accuracy. A weakness of current NLP technologies, however, is the lack of algorithms for non-English languages [43], potentially further reinforcing the North American publication bias identified elsewhere [49, 71].

Extraction and Appraisal

In common with the range of potential approaches in screening, automation may be applied in various ways to evidence extraction and evaluation of data: as a second reviewer, as a tiebreaker, or as a prompter for a human decision-maker, to

name a few. Given the complexity of extraction and subsequent meta-analysis, however, these stages of a systematic review are relatively difficult to automate and are limited by current technology. One challenge to the advancement of automation in data extraction is a lack of training data [72]. Current machine learning systems need manual input in order to train the system to perform future tasks, and completing this work is expensive and time-consuming. As described in Chapter 1, the “recipe” of a systematic review should not be automated [28], and moreover currently cannot be automated; while a computer can perform complex calculations when directed to do so, it cannot decide on the most appropriate method of analysis for a given dataset, nor make an entirely autonomous decision on what data to collect. These limitations do not mean that extraction automation offers nothing of use for these stages, however.

Automation may be used to reduce the amount of text to be reviewed [60], or it may be used to collate data from the study in a structured manner which can then be used in the meta-analysis [73]. Automation could also be used to convert distinct data types so they may be compared with one another [43]. A systematic review of methods to automate data extraction was published in 2015 [74]. While 26 published reports were identified in this review, it was found that the scope of extraction was limited in the elements extracted, and it did not provide information on the performance of the extractions in terms of accuracy.

Methods previously assessed for partially automated data extraction include a template [73] or a statistical model [75]; these can be used to extract information on number of patients in each arm and number of events in each arm. There are also tools available and in development to digitise graphs [76-78]. Unfortunately, these do not currently support survival curves, a common data type in clinical trials and thus in systematic review data extractions. One tool currently available for the reduction of the amount of text to be reviewed is ExaCT [60]. This tool classifies PICO characteristics, randomisation, and select intervention and outcome elements.

A further attempt to automate the quality appraisal stage of a review was reported in 2015. In this study, supervised machine learning was used to train two models for three of the seven different domains of the Cochrane Risk of Bias (RoB) template. One model dealt with prediction of sentences relevance to the assessment,

while the other model attempted to assign a RoB score for each study. They found that a third of studies could be assessed by one reviewer only by applying machine learning, saving significant reviewer time [79].

A significant tool to note among the efforts to automate synthesis, including Risk of Bias assessments and PICO (participants, interventions, comparators, and outcomes) extraction, is RobotReviewer [80]. RobotReviewer is an open-access platform designed to partially automate elements of data extraction using ML and NLP. RobotReviewer produces RoB reports on uploaded trial PDFs, along with identifying relevant supporting annotations. In addition, RobotReviewer automatically assesses study sample size and PICO (participants, interventions, comparisons, and outcomes) characteristics of uploaded study reports. A more extended discussion of RobotReviewer will be presented in Chapter 6, which reports on a randomised trial of the tool.

Adoption of automation in systematic reviews

Though there are numerous automation tools available for systematic reviews, as described in the previous sections, uptake of these tools has been limited in scope and slow in pace [81, 82]. Marshall and Wallace (2019) write that potential reasons for this include the lack of interoperability of automation systems, poor performance, opaque methodologies, or unintuitive user interfaces, but could also come down to a lack of clarity to practitioners regarding when to use and when not to use such tools [72].

The most thorough research to date on why automation is not widely used is van Altena et al (2019) [82]. They found that less than a third of reviewers surveyed had used automation tools, and also that reviewers often started using a tool but then stopped. Although participants in this study indicated that automation reduced their workload, the researchers did not find any relationship between size of workload and use of automation; in other words, even those who would benefit the most from uptake of automation in their workflows did not appear to be using it. The study also identified poor usability and a steep learning curve as barriers to automation adoption. Other barriers identified were lack of support, a lack of time to properly validate the tool, insufficient literature validating the performance and benefits of a

given tool, and a lack of transparency on how a tool works. Organisational and peer endorsement were identified as facilitators; respondents tended to learn about automation tools from others in their environment.

These results have been supported by more recent literature. Scott et al (2021) also found that while approximately half of the respondents to their survey reported using a tool during their review, the tools were also frequently abandoned [83]. Among those who reported using a tool, a strong majority (80%) thought they saved time, while 54% felt they had increased accuracy. In the opinion of O'Connor et al (2019), a general perception among the evidence synthesis community of the non-inferiority or superiority of automation methods is key to their adoption [84]. In other words, while the perception of time savings appears to be encouraging of adoption, the accuracy and quality perception is currently less certain, according to existing academic opinions. Two caveats need to be noted in relation to both of the above-cited articles. First, the 2021 survey examined 'automation tools', but would have more accurately been described as 'technology tools'; many of the tools examined do not replace any human decisions with ones autonomously nor semi-autonomously taken by a computer. Second, the 2019 discussion from O'Connor et al is presented as commentary; while this is helpful in informing my research of the current thinking in the field, it is not empirically based, and its utility in informing adoption strategies is therefore unclear.

Conclusions

In surveying the available literature, a chain of insights becomes clear: Automation tools exist for most stages of the systematic review process, and these tools are more advanced for the early search and screening stages of reviews than they are for the synthesis stages. A wide range of screening automation techniques are available but are limited in their availability (e.g., DistillerAI and EPPI-Reviewer) to typical users.

Finally, automation is not being widely used for systematic reviews, and robust evidence to explain the potential reasons why is limited at best. While at least one systematic review using automation throughout its process has been documented [85], the remaining evidence on the use of automation tools is narrow in its scope,

usually isolated to validation studies of specific tools in case study settings, thus limiting its generalisability. Though these validation studies are necessary and invaluable, they lack insight into the real-world outcomes of adopting automation.

This PhD will therefore seek to fill these gaps in knowledge. I aim to enhance academic understanding of why the adoption of automation has been lagging despite widely available tools, and to further understand the outcomes of adoption of automation in current practice of health evidence synthesis.

Chapter references

1. Clarke, M. and I. Chalmers. Reflections on the history of systematic reviews. *BMJ Evidence-Based Medicine* 2018; 23(4):121-122.
2. Gough, D. and D. Elbourne. Systematic research synthesis to inform policy, practice and democratic debate. *Social Policy and Society* 2002; 1(3):225-236.
3. Gough, D., S. Oliver, and J. Thomas. *An Introduction to Systematic Reviews*: Sage; 2017.
4. Nussbaumer-Streit, B., et al. Resource use during systematic review production varies widely: a scoping review. *Journal of Clinical Epidemiology* 2021.
5. Borah, R., et al. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017; 7(2).
6. Beller, E.M., et al. Are systematic reviews up-to-date at the time of publication? *Systematic Reviews* 2013; 2(1):36.
7. Shemilt, I., et al. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews* 2016; 5(1):1-13.
8. Waffenschmidt, S., et al. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Medical Research Methodology* 2019; 19(1):132.
9. Gartlehner, G., et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *Journal of Clinical Epidemiology* 2020; 121:20-28.
10. Elliott, J.H., et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Medicine* 2014; 11(2):e1001603.
11. Takwoingi, Y., et al. A multicomponent decision tool for prioritising the updating of systematic reviews. *BMJ: British Medical Journal* 2013; 347.
12. Garner, P., et al. When and how to update systematic reviews: consensus and checklist. *BMJ: British Medical Journal* 2016; 354:i3507.
13. Jadad, A.R., et al. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. *JAMA* 1998; 280(3):278-280.
14. Shojania, K.G., et al. How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine* 2007; 147(4):224-233.
15. MacLehose, H. Updating Classification System. 25 June 2019 2016.
16. Johnston, E. How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis. *Journal of Emergency Medicine* 2008; 34(2):231.

17. Thomas, J. How the "Pipeline" project can speed up the identification of studies (Plenary II: Challenges and different approaches to improve quality, timeliness, and usability). *Cochrane Colloquium*; 2016; Seoul, South Korea; 2016.
18. Wang, K., et al. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 2020; 1(1):396-413.
19. Bastian, H., P. Glasziou, and I. Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Medicine* 2010; 7(9):e1000326.
20. Ioannidis, J.P., et al. The geometric increase in meta-analyses from China in the genomic era. *PloS One* 2013; 8(6):e65602.
21. Marshall, I.J., et al. Trialstreamer: a living, automatically updated database of clinical trial reports. *medRxiv* 2020.
22. Altman, D.G. The scandal of poor medical research. *BMJ: British Medical Journal* 1994; 308.
23. Abbott, R., et al. Characteristics, quality and volume of the first 5 months of the COVID-19 evidence synthesis infodemic: a meta-research study. *BMJ Evidence-Based Medicine* 2021.
24. Stamm, T.A., et al. The methodological quality is insufficient in clinical practice guidelines in the context of COVID-19: systematic review. *Journal of Clinical Epidemiology* 2021; 135:125-135.
25. Thomas, J., A. Noel-Storr, and M. Steve. Evidence surveillance to keep up to date with new research. *Systematic Searching: Practical ideas for improving results* 2019:189.
26. Nakagawa, S., et al. A new ecosystem for evidence synthesis. *Nature Ecology & Evolution* 2020; 4(4):498-501.
27. Adams, C.E., S. Polzmacher, and A. Wolff. Systematic reviews: work that needs to be done and not to be done. *Journal of Evidence-Based Medicine* 2013; 6(4):232-235.
28. Tsafnat, G., et al. The automation of systematic reviews. *BMJ: British Medical Journal* 2013; 346.
29. Living systematic reviews. 2018. <https://community.cochrane.org/review-production/production-resources/living-systematic-reviews>.
30. Akl, E.A., et al. Parenteral anticoagulation in ambulatory patients with cancer. *Cochrane Database of Systematic Reviews* 2017; (9).
31. Kahale, L.A., et al. Oral anticoagulation in people with cancer who have no therapeutic or prophylactic indication for anticoagulation. *Cochrane Database of Systematic Reviews* 2017; (12).
32. Hodder, R.K., et al. Interventions for increasing fruit and vegetable consumption in children aged five years and under. *Cochrane Database of Systematic Reviews* 2018; (5).

33. Akl, E.A., et al. Living systematic reviews: 4. Living guideline recommendations. *Journal of Clinical Epidemiology* 2017; 91:47-53.
34. Hammerstrøm, K., A. Wade, and A. Jørgensen. *Searching for studies: A guide to information retrieval for Campbell Systematic Reviews*. Keystone, Colorado: Campbell Collaboration; 2010.
35. Choi, S., et al. Combining relevancy and methodological quality into a single ranking for evidence-based medicine. *Information Sciences* 2012; 214:76-90.
36. Eysenbach, J.T., Thomas L. Diepgen, Gunther. Evaluation of the usefulness of Internet searches to identify unpublished clinical trials for systematic reviews. *Medical Informatics and the Internet in Medicine* 2001; 26(3):203-218.
37. Wilkins, T., R.A. Gillies, and K. Davies. EMBASE versus MEDLINE for family medicine searches: can MEDLINE searches find the forest or a tree? *Canadian Family Physician* 2005; 51(6):848-849.
38. Wolf, F., et al. Comparison of Medline and Embase retrieval of RCTs of the effects of educational interventions on asthma-related outcomes. *Abstracts of the 3rd Cochrane Colloquium*; 1995; Oslo, Norway; 1995.
39. Woods, D. and K. Trewheellar. Medline and Embase complement each other in literature searches. *BMJ: British Medical Journal* 1998; 316(7138):1166.
40. Hull, D., S.R. Pettifer, and D.B. Kell. Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Computational Biology* 2008; 4(10):e1000204.
41. McKibbin, A. Systematic reviews and librarians. *Library Trends* 2006; 55(1):202-215.
42. Harris, M.R. The librarian's roles in the systematic review process: a case study. *Journal of the Medical Library Association* 2005; 93(1):81.
43. Tsafnat, G., et al. Systematic review automation technologies. *Systematic Reviews* 2014; 3(1):74.
44. Next Steps for Microsoft Academic – Expanding into New Horizons. 4 June 2021 2021. <https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/>.
45. Carpineto, C. and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)* 2012; 44(1):1.
46. Qi, X.-S., et al. Duplicates in systematic reviews: A critical, but often neglected issue. *World Journal of Meta-Analysis* 2013; 1(3):97-101.
47. Qi, X., et al. Find duplicates among the PubMed, EMBASE, and Cochrane Library Databases in systematic review. *PloS One* 2013; 8(8):e71838.
48. Lefebvre, C., E. Manheimer, and J. Glanville. Searching for studies. *Cochrane Handbook for Systematic Reviews of Interventions*. 2008: 95-150.

49. O'Mara-Eves, A., et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 2015; 4(1):5.
50. Settles, B. Curious machines: Active learning with structured instances. University of Wisconsin--Madison; 2008.
51. Wallace, B.C., et al. Active learning for biomedical citation screening. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*; 2010; 2010. p. 173-182.
52. Shemilt, I., et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods* 2014; 5(1):31-49.
53. Fiszman, M., et al. Combining relevance assignment with quality of the evidence to support guideline development. *Studies in Health Technology and Informatics* 2010; 160(0 1):709.
54. Kouznetsov, A. and N. Japkowicz. Using classifier performance visualization to improve collective ranking techniques for biomedical abstracts classification. *Canadian Conference on Artificial Intelligence*; 2010: Springer; 2010. p. 299-303.
55. Thomas, J., et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *Journal of Clinical Epidemiology*.
56. Wallace, B.C., et al. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*; 2012; 2012. p. 819-824.
57. Thomas, J. and C. Stansfield. Automation technologies for undertaking HTAs and systematic reviews. *EAHIL*; 2018; Cardiff, Wales, UK; 2018.
58. O'Blenis, P., *No Second Screener? There's A Robot For That*. 2018, Evidence Partners.
59. Gartlehner, G., et al. Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. *Systematic Reviews* 2019; 8(1):277.
60. Kiritchenko, S., et al. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Medical Informatics and Decision Making* 2010; 10(1):56.
61. Wallace, B.C., et al. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics* 2010; 11(1):55.
62. Cohen, A.M., K. Ambert, and M. McDonagh. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Medical Informatics and Decision Making* 2012; 12(1):33.
63. Thomas, J., J. McNaught, and S. Ananiadou. Applications of text mining within systematic reviews. *Research Synthesis Methods* 2011; 2(1):1-14.
64. Covidence. 2021. www.covidence.org.

65. Chung, G.Y. and E. Coiera. A study of structured clinical abstracts and the semantic classification of sentences. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*; 2007: Association for Computational Linguistics; 2007. p. 121-128.
66. Felizardo, K., et al. Evidence-based software engineering: systematic literature review process based on visual text mining. Brazil: USP São Carlos; 2012.
67. Felizardo, K.R., et al. A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Information and Software Technology* 2012; 54(10):1079-1091.
68. Felizardo, K.R., et al. Using visual text mining to support the study selection activity in systematic literature reviews. *Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on*; 2011: IEEE; 2011. p. 77-86.
69. Felizardo, K.R., S.R. Souza, and J.C. Maldonado. The use of visual text mining to support the study selection activity in systematic literature reviews: a replication study. *Replication in empirical software engineering research (RESER), 2013 3rd International Workshop On*; 2013: IEEE; 2013. p. 91-100.
70. Malheiros, V., et al. A visual text mining approach for systematic reviews. *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*; 2007: IEEE; 2007. p. 245-254.
71. Gomersall, A. and C. Cooper. Database selection bias and its affect on systematic reviews: a United Kingdom perspective. *Workshop at the Joint Cochrane and Campbell Colloquium*; 2010; 2010. p. 18-22.
72. Marshall, I.J. and B.C. Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews* 2019; 8(1):163.
73. Summerscales, R.L., et al. Automatic summarization of results from clinical trials. *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*; 2011: IEEE; 2011. p. 372-377.
74. Jonnalagadda, S.R., P. Goyal, and M.D. Huffman. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews* 2015; 4(1):78.
75. Hsu, W., W. Speier, and R.K. Taira. Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature. *AMIA Annual Symposium Proceedings*; 2012: American Medical Informatics Association; 2012. p. 350.
76. Huwaldt, J. and S. Steinhorst, *Plot digitizer (Version 2.6.6)*. 2014.
77. Egner, J., et al., *Engauge Digitizer: convert graphs or map files into numbers*. 2013.
78. Rohatgi, A., *WebPlotDigitizer*. 2011.
79. Millard, L.A., P.A. Flach, and J.P. Higgins. Machine learning to assist risk-of-bias assessments in systematic reviews. *International Journal of Epidemiology* 2015; 45(1):266-277.

80. RobotReviewer. 2018. www.robotreviewer.net.
81. Thomas, J. Diffusion of innovation in systematic review methodology: why is study selection not yet assisted by automation. *OA Evidence-Based Medicine* 2013; 1(2):1-6.
82. van Altena, A.J., R. Spijker, and S.D. Olabarriaga. Usage of automation tools in systematic reviews. *Research Synthesis Methods* 2019; 10(1):72-82.
83. Scott, A.M., et al. Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. *Journal of Clinical Epidemiology* 2021.
84. O'Connor, A.M., et al. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Systematic Reviews* 2019; 8(1):1-8.
85. Clark, J., et al. A full systematic review was completed in 2 weeks using automation tools: a case study. *Journal of Clinical Epidemiology* 2020; 121:81-90.

Chapter 3. Methodological frameworks

Chapter overview

Building on the foundational knowledge presented in the previous chapter, this chapter will first provide the rationale for using specific existing theoretical frameworks to locate my research in the academic literature. Three frameworks are then described: one to classify and describe automation as it is used throughout this thesis, and two to inform the design and analysis of each of the individual projects presented in this dissertation. To describe and classify different levels of automation, a hierarchy of human-automation interactions is presented; examinations of specific examples of automation will be classified according to this hierarchy in future chapters. Second, the three factors which influence trust in automation will be presented. This three-layered trust model was generated from a previous systematic review focused on empirical research on human-computer interactions. Finally, Rogers' Diffusion of Innovations framework will be detailed and related to the research to be presented in the remainder of the dissertation. The chapter will conclude with my research questions and discussion on how the chosen theoretical frameworks will be used to analyse my results.

The conceptual frameworks used in this thesis

Before presenting the research and results chapters, the next sections will establish three frameworks. First, a framework is presented to define automation and describe the tools I will examine. Existing literature on systematic review automation is often framed as semi-automation (as compared to full automation); the distinction between semi- and full automation is necessary, but the current standard descriptions are insufficiently detailed. My selected descriptive framework addresses this lack of detail and will help to define the types of human-automation interactions investigated. Next, I will present the two central frameworks which were used to inform my studies' designs and analyses. These will act together to describe the underlying mechanisms of human trust in technology, and my findings on user behaviour, beliefs, trust, and preferences will be mapped within existing theory on adopter persona and innovation characteristics.

As stated in my introductory chapter, the first of my research themes focusses on the adoption of automation in health evidence synthesis. In other words, I wanted to explore the decision-making process with respect to the use of automation in this field. Decision-making can be examined through a variety of lenses. One might focus on an individual's decision-making process and the factors influencing that single person. One could also focus on decision-making at an organisational level; in this context, we might look at why Cochrane recommends a specific piece of software, for example, but not others. In examining decision-making around automation for health evidence synthesis, both individual decisions and organisational decisions are present, located in the professional environment, small team dynamics among a guideline or systematic review team, and other contextual factors, including the expectations and demands of funders and users of systematic reviews.

Broadening our view, all of these influences operate in the wider cultural context of health evidence synthesis. There are rigorous methodological standards and expectations around accountability: an individual or an organisation does not decide merely for themselves; they are accountable to the culture within which they operate. The dynamics of the wider system may both shape individual adoption decisions, but also be shaped over time by those decisions. Moreover, health

evidence synthesis is not a field which benefits from a long history on which to draw insights to answer these normative questions, having really only taken off since 1993. Though we can observe the evolution of methodological standards and expectations to some degree over these 30 years, they are under constant development and re-evaluation. This places my own research, as well as the subjects of it, in a situation in which they are both subject to the cultural context, but also highly influential in continually shaping it.

No one level of these perspectives is sufficient when considering my research themes. Exclusively focusing on an individual would miss out on the organisational decision-making which is influencing that individual. Exclusively focusing on the organisational level would leave out the building blocks of individuals first making decisions that then combine to make up the organisation of interest; each level is interactive with another, and further is subject to but also influential in the construction of a wider cultural context. We therefore need a way to consider multiple levels of decision-making and simultaneously consider the interactions between those levels. To accomplish this in a structured way, it can be useful to use existing theoretical frameworks which locate an individual's decisions within the multiple levels of influence and contextual factors described above.

As well as providing a multi-faceted and interconnected analytical structure for my thesis, the use of theoretical frameworks assists in applying my findings to advance the field of evidence synthesis. In observing and explaining adoption decisions as they relate to automation and health evidence synthesis, this thesis will make recommendations for practice in how to influence decision-makers, as well as providing evidence to be consumed by those same decision-makers. Seeking to influence decision-making incorporates the assumption that such an impact is possible, and that insights from one context can be used to inform and predict behaviours in another context. For this to be possible, the knowledge gathered in my PhD must be structured, categorised, and constructed in such a way that it can be transferred to other settings. This can be facilitated by the application of theoretical frameworks. Clear categorisation and identification of interactions between elements described in these frameworks first allows my results to be used to form hypotheses about causal pathways; because of the structured relationship of the ideas, I can

reasonably propose which elements lead to which results within my own findings, and these hypotheses can provide the basis for further empirical research. Describing such causation explicitly facilitates the transfer of this knowledge elsewhere and supports prediction of outcomes in other settings. By prospectively planning for these explanatory mechanisms and pathways in the findings of each of these projects, the potential impact of the results is strengthened.

Further, use of these frameworks allowed for a unified structure through a methodologically diverse PhD. My descriptive and qualitative chapters (Chapters 4 and 5) aimed to gain a thorough understanding of the existing priorities of the decision-makers I am studying, while the experimental and quantitative chapters (Chapter 6 and 7) aimed to evaluate and understand the effects of specific adoption decisions. Having this breadth of study designs is a strength of this thesis, and the theoretical frameworks enabled me to tie the empirical studies together in a coherent way. They allowed me first to analyse and understand the context within which participants framed adoption decisions, and then to perform empirical studies which can inform adoption decisions with robust evidence. The greatest strength in the end will be in the combination of these analytical frameworks moving forward to form an evidence-based guide with a strong theoretical grounding in existing knowledge. With this structured guide to user decisions, trust, and behaviour, in addition to the evidence to further inform each of these, software developers and researchers will be empowered to prioritise the development and the advocacy of automation tools in the most effective ways.

These analytical models provide further benefits. First, by using these frameworks in combination, it is hoped that it will make this thesis easier for the reader and the wider audience to understand. I believe that clear reporting of science is not only a value-driven best practice, but also maximises the likelihood of greater impact of my research. Second, I can assess how well these frameworks fit the needs of this area of research. This is a relatively novel area of research, and therefore methodological refinement is required. This thesis can either encourage future research to use a similar approach, or it can discount this approach as less than useful and one to be avoided in future experimentation. Either outcome would be a benefit for the literature on this topic. Finally, these frameworks locate my thesis within an

existing field of scholarship. In this way, my research is related to existing knowledge enabling future researchers to relate to and to build upon my work as part of a wider body of literature. With my aim to analyse the decision-making process around the adoption of automation, it was helpful to build upon existing knowledge rather than to set out with no anchoring guide for any discoveries I made. In building my map of decision-making processes, and then contributing new insights for its navigation, these theoretical frameworks kept me accountable to broader knowledge rather than simply following my own assumptions.

To summarise, I have chosen to establish my PhD research within these theoretical frameworks in order to better describe and communicate my findings, to locate them within a wider field of academic research, to test the effectiveness of my approach, and most importantly to maximise the ability of my findings to translate across multiple contexts.

The next section will first consider a descriptive framework for categorising the degree of automation present in the tools I will inspect.

Levels of automation

The academic literature offers a highly relevant and helpful foundation for building a framework for defining the levels of automation (LOA) in this research. Specifically, a 2016 literature review presented an overview of the evolution of taxonomies describing the levels of automation, starting from the 1950s [1]. Twelve approaches to LOA are described in detail in this review. While the authors sought to understand the strengths and weaknesses of each, they eventually concluded that there “does not exist such a thing as a ‘best taxonomy’”. Their discussion, however, provides an excellent starting point to determine which is the best taxonomy for the purposes of this PhD.

As previously mentioned, the frameworks (or taxonomies) presented in this chapter were selected to inform my design and analysis, but also to maximise impact of results by aiming to propose explanatory variables for my results, and to test the applicability of these frameworks and of their combination. Therefore, I selected the framework which seemed to be both the most exhaustive categorisation of levels of


automation, and that which seemed to have the broadest uptake in the current academic literature. With these goals in mind, this research will use the levels of automation taxonomy proposed by Sheridan and Verplank [2, 3].

Description of the framework

Ten levels of potential automation are described in the taxonomy, from an entirely human-run system at the ‘low’ end of the scale to an entirely autonomous computer at the ‘high’ end (Table 3.1).

Table 3.1. Levels of automation, from Sheridan and Verplank (1978)

Level	Description
10	Full computer control
9	Computer informs human only if it decides to
8	Computer informs human only if asked
7	Computer executes automatically, then informs the human
6	Computer allows the human a restricted time to veto an automatic decision
5	The computer executes a suggestion if the human approves
4	The computer suggests one alternative
3	The computer narrows the selection down to a few
2	The computer offers a complete set of decision/action alternatives
1	Full human control



In the case of screening automation technologies currently described in literature (and summarised in Chapter 2), many of these fit into level 3: narrowing the selection down to a few. For instance, automation may filter out citations below a certain threshold of anticipated relevance. The Abstrackr system potentially provides an example of level 7 automation, as part of the screening process may take place without explicit human input (if the operator opts into this mode), as well as fitting in to level 3 as it narrows down the work to be completed by human effort. The Cochrane RCT classifier is essentially a level 7 tool: it automatically decides on the identification of a citation as a likely RCT or as a non-RCT. However, the subsequent action is taken by the human researchers: they might double-check the results of the classifier, then falling into a level 4 system, or they may allow for the classifier to completely make the decision for them, more closely aligning with a level 7.

Shifting focus to the extraction phases of a review, RobotReviewer would fit into level 4 or level 5, depending on the nature of the integration to the evaluation process. For example, a person completing a Risk of Bias assessment might be

presented with suggested text annotations as they are working; it does not truncate the individual's workflow, but instead provides an option which the human may then decide to use, or to ignore. Much like the first scenario described above in relation to the Cochrane RCT classifier, this would be a level 4 use-case. For a level 5 use-case, the RobotReviewer system may be integrated in such a way that the Risk of Bias assessment is more or less completed before the human accesses the assessment form, and their input is limited to approving or rejecting the suggested judgement and annotations. The difference is subtle, but subtle prompts can have profound effects on human behaviour, which in turn might have profound effects on the evidence synthesis output and on the amount of human effort required.

From the previous two paragraphs, a clear theme is already emerging from the choice of this framework: with all the currently available technologies, the level of automation category is highly dependent on how the technology is applied. This realisation already has important implications for future research in this area. First, when applying this taxonomy, it will be helpful to specify which level of automation is being investigated, most importantly in cases where a single technology might fit into several levels depending on use. Second, when analysing results, it may be helpful to examine results through multiple lenses; that is, I suggest analysing results through the potentially appropriate levels of automation for a given tool. It is very possible that implications for the adoption and use of a single tool will vary according to the level of automation selected with the nature of its integration.

Contribution to this thesis

The most notable contribution of this framework in relation to this PhD is to clarify that automation exists as a spectrum rather than as an all-or-nothing process. A later author who borrowed heavily from this framework for their own proposed taxonomy rightly pointed out that the evidence shows that “automation does not simply supplant human activity but rather changes it” [4]. Given the range of available automation design options, Parasuraman (2000) also posited that “system designers [must] consider some hard choices regarding what to automate and to what extent, given that there is little that cannot be automated.” While Parasuraman tended to operate in the more mechanical sphere of automation (e.g., manufacturing), a

similar position might be taken in health evidence automation: what to automate, and to what extent, will be a hard but crucial choice in this process.

Placed into the context of health evidence synthesis research, the utility of applying these classifications is to clarify that automation need not mean wholesale erasure of human effort in a systematic review or in a guideline. This clarification is important to describing the results of this research, but also in disseminating its results more broadly. To explore the use and the adoption of automation in health evidence synthesis is not to examine the difference between a level 1, entirely manual, process versus a level 10, entirely automated process. Parts of a systematic review may be automated while others are entirely manual, or parts may be partially automated in combination with human effort. The combinations of potential applications and levels are extensive.

Throughout this thesis, this framework will be applied to all experimentation to classify the level of automation being explored. Doing so facilitates orientation of further discussion of the technology and builds a practical basis for any future experimentation. As I explore the consequences of the choice to adopt automation in a particular workflow, results may show that these consequences differ along the LOA axis. Results may also show that adoption decisions similarly hinge on the level of automation proposed.

It is worth noting that many of the previous authors who have proposed LOA taxonomies have included stages of automation functions in their discussions. For example, Parasuraman included a four-stage model of potential automation functions in a combined human-computer interactive workflow: sensory processing, perception / working memory, decision making, and response selection. However, given that this research largely focuses on systematic reviews, which themselves have discrete and well-defined stages, these stages are more applicable to the context of this research and will be used instead. Each of these may have its own optimal automation level and therefore requires individual research to build evidence to best determine the current ideal.

To summarise the use of my chosen levels of automation framework: I have selected the taxonomy proposed by Sheridan and Verplank because of its relatively

more exhaustive description of levels of automation, and on its time-tested and wide-reaching influence in the automation literature. Because of the context in which this research is conducted, I have opted not to use the models of information processing included with many LOA taxonomies and instead continue to use the defined stages of systematic reviews. Use of this framework will allow for clearer discussion of my results, and additionally highlights the spectrum on which automation exists.

I will now examine two analytical frameworks concerning trust and the diffusion of innovations, which inform my analysis and conclusions.

Trust in automation

Connecting back to the broad themes which underpin this thesis – why do individuals choose to adopt automation, and what happens once they do? – this research will also use a framework to define trust in automation. When deciding on automation and how to interact with it, trust is a foundational concept [5]. Therefore, it is useful to apply a framework of human-automation trust, and I will use the framework detailed by Hoff and Bashir [5].

While ‘trust’ may seem a rather nebulous concept to define, let alone to measure and to predict, the literature on trust is extensive. For the purposes of my research, I will use the definition of trust which the authors of my selected trust framework also used, namely from Lee and See [6]: “the attitude that an agent will help achieve an individual’s goals in a situation characterised by uncertainty and vulnerability.” In this context, the ‘agent’ would likely refer to one or more automation technologies or software programs. ‘Uncertainty’ and ‘vulnerability’ are likely to play large roles when considering the effects of automation in health evidence synthesis. While ‘vulnerability’ may initially conjure up emotions of personal exposure, in terms of health evidence synthesis it may imply more system-level effects. By giving control – even incrementally as outlined in the levels of automation framework – to automation systems, the people who had formerly taken charge of setting health standards may feel vulnerable and uncertain.

The three-layered model proposed by Hoff and Bashir was selected for this research because it is situated in the same context as this PhD. That is, their

framework directly concerns trust in the context of human interaction with automation technologies. To build their model, Hoff and Bashir conducted a systematic review on empirical research on the factors that influence trust in automation. The authors found three variables which interact to influence human-automation trust: the human operator, the environment, and the automated system. These three variables in turn build upon and reflect the three layers previously described by Marsh and Dibben and reiterated by Hoff and Bashir [7]: dispositional trust, situational trust, and learned trust.

Dispositional trust

“Dispositional trust represents an individual’s overall tendency to trust automation, independent of context of a specific system.”

Dispositional trust refers to the inherent characteristics of an individual which influence their ability to trust a system; dispositional trust corresponds to the automation-specific trust source of the human operator. Hoff and Bashir identified four primary inputs to dispositional trust: culture, age, gender, and personality. They notably describe these as an “enduring tendency”; such traits do not change over the long-term. They are notably distinct from other characteristics which may change in the short-term – mood, self-confidence, or social influences.

Culture was identified as “a particularly important variable,” and it is anticipated that this assessment is likely to remain true throughout the analysis of this research. First, it is inherently important to note variations in trust across human cultures (e.g., geographic regions, national identities, religions, etc.). Substantial evidence supports this assertion in the realm of interpersonal trust, with some selected evidence relating to automation as well [8, 9]. Second, like all work environments, the evidence-based medicine community has its own values, and it should be anticipated that these will influence individuals’ trust levels within that community. These may also share significant overlap with personality traits, as personal interests naturally play a significant role in choice of professional direction.

Dispositional trust may not vary over time, but there is evidence that behaviour may vary over time in alignment with individuals' dispositional trust. Unsurprisingly, several empirical studies found that those with greater dispositional trust in automation tended to more readily accept the initial results of automated programs such as navigational equipment. Merritt and Ilgen [10] conducted an additional study on this observation that is worth detailing. They found that should an automation aid fail or perform poorly, individuals expressed lower trust in the aid just as one might expect. However, individuals with higher baseline dispositional trust experienced a more significant decline in trust in the aid than did those with lower baseline dispositional trust. Hoff and Bashir concluded that “these results suggest that individuals with high levels of dispositional trust in automation are more inclined to trust reliable systems, but their trust may decline more substantially following system errors.”

Situational trust

Situational trust may be thought of as trust that varies depending on the context in which the human-automation interaction takes place. Unlike dispositional trust, these are likely to vary in the short term. Hoff and Bashir linked this layer of their trust model to the environment in which an interaction takes place, and they divided situational trust into two general categories: internal variability and external variability.

Internal variability

Internal variability includes self-confidence, subject matter expertise, mood, and attentional capacity. Subject matter expertise seems the likely candidate to play a relatively more significant role in this research, given that the population targeted in my PhD research is composed of extremely highly trained and skilled individuals. Moreover, many are likely to have invested significant time in their area of expertise, potentially influencing their self-confidence.

Empirical evidence can inform some predictions as to how internal variability may influence the results of studies looking at the integration of automation. Looking first at self-confidence in general, evidence suggests that when individuals have about the same level of self-confidence as they have trust in an automation aid, they prefer manual control; that is, people tend to favour themselves over a computer

when all else is equal [11]. Self-confidence also plays a role more specifically in the form of computer self-efficacy, or an individual's perception of their own ability to use a computer. The current evidence here shows that higher levels of perceived computer self-efficacy are positively correlated with trust in automation [12].

Next examining subject matter expertise (SME), the current evidence shows that those with higher expertise are less likely to trust automation than individuals with less experience in their field [13, 14]. In the context of this thesis, it is important to note that the role of subject matter expertise would relate to established systematic review methods, and not relate to automation tools for systematic reviews or other health evidence. This instead would fall under initial learned trust, discussed in subsequent sections.

External variability

External variability includes the type of system in use, its complexity, the difficulty of the task it has been assigned, workload, perceived risks and benefits, the organisational setting, and the framing of the task. The potential impacts of each of these in the context of health evidence synthesis are significant.

Though a particular automation technology may excel at one task but not at another, perception bias clearly can influence the trust that humans decide to place in automation. A previous experiment found that when a poorly performing automation aid was juxtaposed with a well-performing aid, people tended to put more trust in the poorly performing aid than they previously had [15]. If such an observation were repeated in the context of health evidence synthesis, this could be hugely problematic for evidence quality. For example, if a screening automation system works particularly well, but a data extraction automation system does not, the reviewer might perceive the latter as more effective than it actually is, place undue trust in its results, and disseminate evidence of diminished quality.

Organisational settings and perceived risks and benefits should also be carefully considered in the context of health evidence synthesis. The researchers and academics who provide systematic reviews and guidelines are a highly interactive community, and the expectations therein factor significantly into methodological choices. Moreover, the intended impact of health guidelines is quite personal, and therefore the potential impact of a significant methodological change (such as

automating this research) may be felt quite personally by the individuals in this field. These expectations, and the potential risks introduced by the use of automation, are foundational to the context of health evidence synthesis and of this PhD.

Learned trust

“Humans are creatures of experience.”

Learned trust is the cumulative effect of a user’s past experiences and/or current interactions on their trust in a system. Like situational trust, learned trust is further sub-divided into two categories: initial and dynamic. The former relates to past experiences that an individual may have had prior to the interaction of interest, while the latter relates to the variation throughout the specific interaction with an automation tool. Both of these types of learned trust stem from the characteristics of the tool itself, rather than characteristics of the end user or of the environment.

Initial learned trust

Like people, automation tools may be at the mercy of their pre-established reputation. When a system is portrayed as ‘expert’, people tend to trust it more. However, this may come with a higher risk: like the quick degradation of trust found by Merritt and Ilgen [10], an initially high reputation may suffer greater consequences in operator trust from initial mistakes.

Naturally, the most important factor in initial learned trust is how much experience a person may have had previously with the specific automation tool being considered. Someone with a previously negative experience is less likely to make themselves vulnerable to a further negative experience, while someone with a previously positive experience with automation is more likely to put trust in the tool. Here is where the previously mentioned distinction between subject matter expertise in a field and significant previous experience with a particular automation tool is important to highlight once more. Experience with the automation itself contributes to initial learned trust, while significant experience with the field or subject matter potentially being automated contributes to the internal validity of situational trust.

Dynamic learned trust

Of all of the types of trust presented, dynamic learned trust is the one most impacted by the design features of an automation system [5]. This layer of trust refers to the knowledge gained by a user over the course of a single interaction with a system. Performance can, of course, be highly variable overall as well as within the course of a single interaction.

How information is presented regarding system performance will affect how users perceive the automation system. This may be in design features, aesthetics, or an interface. However, the effect size of system aesthetics on user trust depends on the service being automated. For example, e-commerce websites seem to be more affected by their aesthetics than in-vehicle automation [16]. Similarly, these same websites benefitted from higher ease of use [17]. Hoff and Bashir write that the evidence regarding ease of use's effect on trust in automation, however, is somewhat lacking [5].

Overall, in considering dynamic learned trust in the context of health evidence synthesis, it is likely to be important to consider how an automation system communicates its results and its performance to its end users.

Diffusion of Innovations

As the overarching theme of this thesis is the adoption of automation in health evidence synthesis, the most relevant framework to use in approaching the analysis is the Diffusion of Innovations theory [18]. Diffusion of Innovations theory, as a highly regarded and time-tested and supported theory, was the first and most obvious lens to select of the three frameworks. The utility of its use in this thesis was two-fold: first, to inform the study design, particularly in relation to the qualitative studies; second, collecting data and assessing it within this framework allowed for reflection on how well or how poorly this framework fits this area of research in general. Diffusion of Innovations is particularly important for the first of my broad research questions: why might individuals adopt automation?

This theory, first published by Rogers in 1962, was developed to detail how and why an innovation spreads in a given population or setting over time. In the case

of this thesis, it will be applied to consider how and why the innovation of interest – automation – spreads through a particular setting – health evidence synthesis.

It will be useful to first define certain terms in relation to this theoretical framework for discussion later in this thesis. First, Rogers' theory defines an innovation as a new idea, behaviour, or product. Adoption is used to refer to an individual changing a practice or a behaviour, wherein they now do something differently than they did before. This change happens over time, that is to say it is diffused throughout a population.

Two additional taxonomies, both detailed in Diffusion of Innovations theory must be laid out in more detail: adopter categories, and innovation characteristics.

Adopter categories

As established above, the diffusion of an innovation happens over time. If the proportion of a population that has adopted an innovation is plotted against time, this produces an *adoption curve* (Figure 3.1). This curve can then be split into five sections which represent the five groups or categories of adopters: innovators, early adopters, early majority, late majority, and laggards. Each persona has particular tendencies, and these are essential to consider if one is promoting or discouraging an

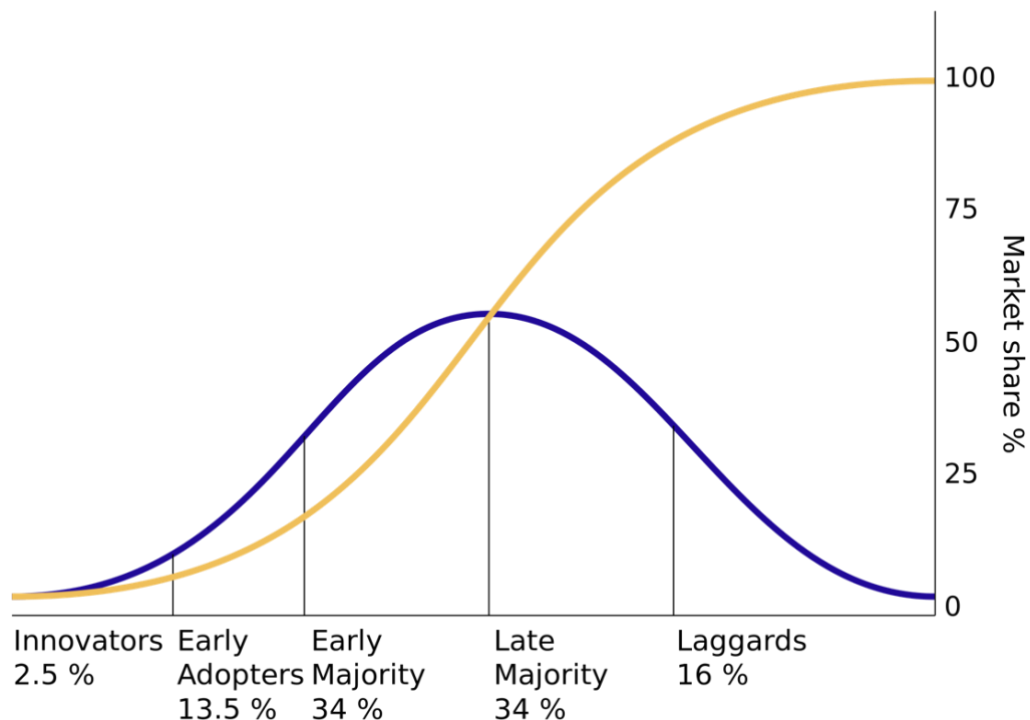


Figure 3.1. Diffusion of Innovations adoption curve

innovation adoption. Human beings tend to not fit into any descriptive box without some leaning into another, but these categories are nonetheless useful descriptors in order to interpret the behaviour observed in a population.

Innovators

The first group to adopt an innovation are, unsurprisingly, termed *innovators*. The first 2.5% to adopt an innovation, they try to be on the forefront of new ideas and technology, and moreover enjoy doing so.

Early Adopters

Early adopters tend to be opinion leaders within a population, but not as risk-taking as innovators. The next 13.5% to adopt an innovation, they enter an adoption decision already having acknowledged the need for a change, and therefore seek out solutions and opportunities to try new things. No information is generally required to convince them of a need to change, but rather information relating to the change process itself such as a how-to manual or documentation is helpful.

Early Majority

Early majority individuals are the final category to fall on the earlier side of the adoption curve. With a predicted 34% of the population, they tend to seek evidence in support of a change before considering it.

Late Majority

The first category on the second half of the adoption curve, late majority adopters are generally quite sceptical of a proposed change. Holding an equal proportion of the population as early majority with 34%, they also tend to need evidence to be convinced that existing methods require an innovation and, like the early majority, seek out evidence of an innovation's effectiveness before considering it. Contrasting with early majority's need for evidence, however, they are more likely to seek out success stories from their peers, rather than generic case studies or trials.

Laggards

As the name implies, laggards are the last adopters, making up 16% of the normal adoption curve. They are described as traditional by nature and highly sceptical of change. Laggards are predicted to take input from peer pressure in considering whether to adopt or not to adopt an innovation.

Innovation characteristics

In addition to adopter personas, Diffusion of Innovations provides a framework of innovation characteristics which influence the adoption and diffusion of the innovation. Like adopter personas, there are five categories identified as most affecting the speed and extent of the widespread adoption of a technology. These five attributes are *relative advantage*, *compatibility*, *complexity*, *trialability*, and *observability*. Note that none of these five alone guarantees widespread adoption, but rather they act in concert with each other, and with the above personas, to influence diffusion. In the context of this thesis, it is likely that some attributes will play a larger role in particular areas of systematic reviews, or perhaps even in conjunction with the LOA framework, some innovation attributes may shift in prominence from one level to the next.

Relative advantage

Relative advantage refers to how much better an innovation is, or is perceived to be, than the system it is replacing. The greater the improvement, the more likely that an innovation will spread and at greater speed.

Compatibility

Compatibility refers to an innovation being in line with existing values and practices, and the needs of potential future users. In the context of health evidence synthesis, compatibility will play a role in two distinct ways.

First, an innovation may need to demonstrate compatibility in a technical sense; a well-established workflow may be disrupted by an incompatible novel method (whether automation or otherwise), and individuals may wish to avoid this disruption. Conversely, a highly compatible novel method would act as a facilitator for this adoption process.

Second, an innovation may need to demonstrate compatibility in terms of values and practices of health evidence synthesis as a professional culture. The discussion on the dispositional layer of trust in automation identified culture as a highly influential aspect of dispositional trust. This influence is likely to extend to the professional culture of systematic reviewers and guideline developers. For an intended population for the potential diffusion of automation, automation's perceived

alignment with cultural and professional values and needs will play significantly into the diffusion of automation systems.

Complexity

Complexity refers to how easily an innovation is to comprehend and to put into use. Real or perceived complexity will inherently depend on the audience; some users may readily understand a new technology, while others may require more guidance and time. Some reviewers may not require that they understand the deeper workings of automation, while others may rely heavily on their ability to understand how something works in order to trust it and consequently put an automation system into practice. These possibilities will be explored in my research.

Trialability

Trialability refers to the ability of users or potential future users to experiment with an innovation prior to adopting it. Currently this is a significant barrier, as most of the automation technologies available require significant technical expertise prior to use and are not well-integrated with commonly used systematic review tools.

Observability

Observability is the degree to which potential users may examine the results of an innovation. In the context of automation in health evidence synthesis, the current primary route of observing results of automation would be via journal articles or conference proceedings. Alternatively, individuals may learn from the experience of their peers (particularly important for the late majority adopter persona), but given the currently low level of automation adoption, this may also be a barrier to adoption in the current climate.

Research questions

These three frameworks are now established as the analytical lenses that will be used in this thesis. The broad level themes have also been specified: to explore why individuals may or may not adopt automation in health evidence synthesis, and to explore what happens if and when they do. Chapter 1 provided an introduction to each of my research chapters, with their overall designs and most impactful conclusions stated. Now that we are able to draw on these theoretical frameworks,

my research questions will be described in more detail and connected to the theoretical frameworks described previously in this chapter.

As mentioned in the introduction, each of the two broad themes of this thesis is aligned more closely to two of the four empirical studies conducted as part of my PhD. Chapter 4 explores the acceptability of automation to guideline developers, and Chapter 5 sought information on the adopter personas found in the community of Cochrane Information Specialists, and their respective adoption (or non-adoption) journeys. These two chapters relate more closely to the theme of why and how individuals, communities, or organisations choose to adopt or not to adopt automation tools. Broadly speaking, these two chapters provide data more qualitative in nature to contribute to the evidence base which will test and strengthen the chosen analytical frameworks. Chapter 6 examines the effects of combining human and ML effort in the Risk of Bias stage of systematic reviews, and Chapter 7 looks at the economic implications of adopting a particular automation tool in a living evidence map. These two chapters are more quantitative than the previous two, and they relate more closely to the broad theme of the effectiveness of automation when it is adopted into health evidence synthesis workflows.

Chapter 4: Acceptability – guideline developer views towards automation

To begin the exploration of the ‘why’ component of this thesis – why individuals or organisations do or do not adopt automation in their health evidence synthesis methods – I conducted a qualitative study gathering information from guideline developers on their views regarding automation. Given that guideline developers are key gatekeepers in the evidence-to-practice pipeline (determining the use or non-use of health interventions) they are a relevant population to consult.

This project used a structured interview for data collection, followed by a combined deductive and inductive thematic analysis of the interview transcripts. Given the focus of the work on understanding the wider context around the adoption (or non-adoption) of automation, I used the Diffusion of Innovations framework to structure this inquiry, and as the deductive themes for my analysis. Research questions for this project were:

RQ1.1) How do the opinions of guideline developers towards automation of health evidence synthesis fit into the Diffusion of Innovations framework?

RQ1.2) Within the Diffusion of Innovations themes, what important concepts were identified by participants?

As this study was undertaken at the broad level, discussing guideline developers' attitudes towards automation in a generalised rather than a specific sense, the project does not fit in to the LOA framework. Finally, the three-layered trust model from Hoff and Bashir was used to inform discussion of the analysis and results, but not in the design of the project itself.

The results of the acceptability component of this PhD directly tested the utility of the Diffusions of Innovations framework, contributed to the knowledge of dispositional and situational trust, identified opportunities to contribute to learned trust, and provided insights into how to do so most effectively. This study has since been peer-reviewed and published in *Systematic Reviews* [19].

Chapter 5: The User Journey – mapping adopter personas among Cochrane Information Specialists

Chapter 5 also falls into the first of my broad themes, exploring why individuals may or may not adopt automation in their health evidence synthesis workflows. Furthering knowledge of how well the Diffusion of Innovations framework fits into this context, Chapter 5 reports on a mixed methods project which sought to map adopter personas among Cochrane Information Specialists (CISs) and to explore their experience in adopting the Cochrane RCT classifier. Like guideline developers, CISs are a key stakeholder group, particularly in methodological standard setting for the search stage of systematic reviews, and therefore an important group to study.

Research questions for this project were:

RQ2.1) How applicable are the Diffusion of Innovations adopter personas to this context?

RQ2.2) How do users interact with the RCT classifier?

RQ2.3) To what extent do users trust the RCT classifier, and what factors inform this trust (or lack thereof)?

The first of the three research questions for this chapter was addressed using a survey. The adopter personas from Diffusion of Innovations were used to inform the design of this survey and to analyse its results, thus answering RQ2.1. The survey results also provided information to address RQ2.2, and interviews were conducted to gather additional detail. This second research question drew from the LOA framework to categorise the RCT classifier usage as described by the research participants. I aimed to identify differences, if any, in the way in which some personas use the tool. Finally, both the survey and interviews provided data for RQ2.3 to present any insights regarding the trust mechanisms of the various personas.

The results of the user journey component of this PhD tested the applicability of the adopter personas from the Diffusion of Innovations framework, used the LOA framework to map RCT classifier user behaviour, and used the trust framework in combination with the mapped personas to explore and explain this behaviour.

Chapter 6: Validity – a clustered non-inferiority randomised trial examining the effect of combined human effort and automation on Risk of Bias assessments

Chapter 6 begins to explore what happens if individuals or teams choose to adopt automation systems into their systematic reviews. Rather than testing the applicability of the selected frameworks, it leans instead towards using them as an analytical lens. That is, rather than testing how well the Diffusion of Innovations characteristics or personas apply to the context of health evidence synthesis, this chapter instead begins to contribute evidence which may influence diffusion, and this evidence will be analysed using the Diffusion of Innovations framework. Specifically, Chapter 6 reports on a randomised controlled trial of combining human

effort with automation in the Risk of Bias stage of a systematic review. Research questions for this trial were:

RQ3.1) Is the accuracy of RobotReviewer-assisted RoB assessments non-inferior to human-only RoB assessments?

RQ3.2) Is the person-time required for RobotReviewer-assisted RoB assessments less than the person-time required for human-only RoB assessments?

By testing the validity of an automation-augmented approach, this trial provides data for the Diffusions of Innovations framework regarding the relative advantage of partially automated systematic review methods. Moreover, by disseminating these results, the trial also contributes to the observability of automation in adopted into a real-world systematic review, and potentially to the initial learned trust of the health evidence synthesis community at large. This study has been submitted to *Annals of Internal Medicine* for peer review.

Chapter 7: Economic evaluation – the cost-effectiveness of a semi-automated workflow to maintain a living evidence map

Chapter 7 is similar to the preceding chapter in terms of its framework use and contributions, and in relating to the second unifying theme of this PhD: what happens when teams adopt automation into their health evidence synthesis workflows. Reported in this chapter are the results of a cost-effectiveness analysis which examined the effects of adopting a partially automated workflow into a living evidence map. This map sought to find and classify all evidence relating to COVID-19. This project's research question was:

RQ4.1) What is the cost-effectiveness in terms of costs, recall, and precision of a semi-automated study identification method for a living evidence map of COVID-19 evidence?

By analysing the cost-effectiveness of this case study adopting automation, this project also contributed to the evidence base for relative advantage within the

Diffusions of Innovations framework. This relative advantage was measured both in the effect of cost, but also in the effect on effectiveness as measured by the recall and precision of the map. That is, this analysis sought to understand whether an evidence map using a partially automated strategy would identify fewer, the same, or more eligible included studies as compared to an entirely manual method. Like the previous chapter, dissemination of the results of this study also contributed to community-wide observability of automation in a health evidence context. Finally, because multiple automation tools were tested, this project represents a more extensive coverage of the LOA framework than the previous chapters. The tools assessed fit in to levels 3, 7, and 8; these will be described in more detail in Chapter 7. This study has produced two publications: one in the *Journal of the European Association for Health Information and Libraries* [20], and one in *Wellcome Open Research* [21].

Summary

This thesis has two overarching themes. First, to explore the adoption of automation in health evidence synthesis, its barriers and its facilitators, or, framed as a question: why do decision-makers choose or not choose to adopt automation in this context? Second, to test the effectiveness of automation in health evidence synthesis, or, framed as a question: what happens when automation is used in this context?

Three frameworks have informed my study design and analysis in answering these questions: Sheridan and Verplank's levels of automation, Hoff and Bashir's three-layered model of trust, and Rogers' Diffusion of Innovation theory. They served two primary purposes: first, they were used to inform the design and analysis of the research projects undertaken during this PhD. Second, I sought to determine how well they fit the context of health evidence synthesis. Building from this second point, there were two possible outcomes: that these frameworks fit flawlessly into this context, or that the results of this research can be used to inform improvements to the frameworks to better suit research into automation in health evidence synthesis. By jointly using these frameworks to provide structure for the reporting of the results of my research, I hope to maximise the clarity and impact of these results.

Chapter references

1. Vagia, M., A.A. Transeth, and S.A. Fjerdingen. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics* 2016; 53:190-202.
2. Sheridan, T.B. and W.L. Verplank, *Human and computer control of undersea teleoperators*. 1978, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
3. Sheridan, T.B. *Telerobotics, automation, and human supervisory control*: MIT press; 1992.
4. Parasuraman, R., T.B. Sheridan, and C.D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 2000; 30(3):286-297.
5. Hoff, K.A. and M. Bashir. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 2014; 57(3):407-434.
6. Lee, J.D. and K.A. See. Trust in automation: Designing for appropriate reliance. *Human Factors* 2004; 46(1):50-80.
7. Marsh, S. and M.R. Dibben. The role of trust in information science and technology. *Annual Review of Information Science and Technology (ARIST)* 2003; 37:465-98.
8. Huerta, E., T. Glandon, and Y. Petrides. Framing, decision-aid systems, and culture: Exploring influences on fraud investigations. *International Journal of Accounting Information Systems* 2012; 13(4):316-333.
9. Li, D., P.P. Rau, and Y. Li. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics* 2010; 2(2):175-186.
10. Merritt, S.M. and D.R. Ilgen. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors* 2008; 50(2):194-210.
11. De Vries, P., C. Midden, and D. Bouwhuis. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies* 2003; 58(6):719-735.
12. Madhavan, P. and R.R. Phillips. Effects of computer self-efficacy and system reliability on user interaction with decision support systems. *Computers in Human Behavior* 2010; 26(2):199-204.
13. Fan, X., et al. The influence of agent reliability on trust in human-agent collaboration. *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction*; 2008; 2008. p. 1-8.
14. Sanchez, J., et al. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical issues in ergonomics science* 2014; 15(2):134-160.

15. Ross, J.M., et al. The effect of automation reliability on user automation trust and reliance in a search-and-rescue scenario. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; 2008: Sage Publications Sage CA: Los Angeles, CA; 2008. p. 1340-1344.
16. Weinstock, A., T. Oron-Gilad, and Y. Parmet. The effect of system aesthetics on trust, cooperation, satisfaction and annoyance in an imperfect automated system. *Work* 2012; 41(Supplement 1):258-265.
17. Zhou, T. The effect of initial trust on user adoption of mobile payment. *Information Development* 2011; 27(4):290-300.
18. Rogers, E.M. *Diffusion of Innovations*. 5th ed: Simon and Schuster; 2003.
19. Arno, A.D., et al. The views of health guideline developers on the use of automation in health evidence synthesis. *Systematic Reviews* 2021; 10(1):16.
20. Shemilt, I., et al. Using automation to produce a 'living map' of the COVID-19 research literature. *JEAHIL* 2021; 17(2):11-15.
21. Shemilt, I., et al. Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research [version 1; peer review: awaiting peer review]. *Wellcome Open Research* 2021; 6(210).

Chapter 4. Acceptability

Guideline developer views towards automation

Chapter overview

This chapter reports on the first component of this PhD: research into the acceptability of automation. It will first briefly summarise relevant context and background described in previous chapters, namely the state of automation and the literature describing its adoption in health evidence synthesis, and the Diffusion of Innovations theory which provides one of the three analytical frameworks for this thesis. The methods of a qualitative study undertaken consulting guideline developers regarding their opinions towards automation will then be described, followed by the results of this analysis. These results will then be framed in the existing literature, key insights for the field will be described, potential weaknesses of the research discussed, and suggestions for future directions made. The chapter will conclude with a summary of this study's most significant findings.

Introduction

Evidence-based guidelines are overwhelmed by data production rate

To encourage the best health outcomes at the population level, as well as to maximise consistency in high-quality patient care, various organisations publish clinical guidelines. Such organisations include the National Institute for Health and Care Excellence (NICE) in the United Kingdom, the National Health and Medical Research Council (NHMRC) in Australia, and the World Health Organization (WHO) globally. Each of these organisations has certain standards of quality and teams dedicated to the production of research and guidelines. Formerly, guidelines were expert-decision-based, but in recent years have trended instead towards evidence-based practice and evidence-based medicine (EBM) [1, 2], supported by a belief that interventions must be supported by ‘unbiased’ clinical research [3]. Systematic reviews are a crucial component of the evidence incorporated into these guidelines [4].

As outlined in the literature review in Chapter 2, running parallel to the shift from expert-decision-based towards evidence-based guidelines, the rate of publication of evidence has increased to the point that researchers are hard-pressed to produce and to keep systematic reviews up to date [5]. With nearly 4000 health research articles published daily, systematic reviews cannot keep up with the deluge of data [6]. Evidence is then at risk of being lost and wasted due to a lack of sufficient resources to process such large-scale data, leading to out-of-date healthcare and guidelines, and consequently risking worsened population health outcomes [7].

Literature addressing adoption of automation is lacking

The use of automation to assist, expedite, or even replace human effort in the production of systematic reviews is one proposed solution for the data deluge challenge [8-11]. This proposition is not without controversy, however, and there is some concern that the potential benefits will come at the cost of evidence quality

[12]. Uptake of automation has been notably slow [10, 13], despite the broad availability of various tools. This leads to the question: what are the barriers and facilitators to the slow uptake of automation into systematic reviews and health guidelines?

Primary research into barriers and facilitators to uptake of automation in health evidence synthesis contexts is limited [13, 14]. Considering the slow adoption rate, filling the gap in the literature addressing barriers and facilitators should be a high research priority. In particular, perceptions of key stakeholders in evidence production towards the uptake of automation will be useful, as they are foundational to the translation of knowledge to practice. These perceptions could aid not only in identifying barriers and facilitators to the adoption of automation, but in designing approaches to most effectively address them. As outlined above, guidelines are a key component in the translation of knowledge to practice, therefore evidence from guideline developers detailing their opinions towards the use of automation could be helpful in elucidating the reasons for slow automation adoption, and in identifying strategies to encourage wider uptake.

Diffusion of Innovations

As discussed in the preceding chapter, Rogers' Diffusion of Innovations [15] is a highly applicable framework for analysis to understand the adoption of automation in health evidence production. This theory describes how and why an innovation spreads, the characteristics of the innovation that play a role in this process, and the typical categories of adopters. This project most heavily drew from the innovation characteristics portion of the theory, as will be further detailed in the following section on methods. Recall from Chapter 3 the five characteristics of an innovation as described by Rogers: *relative advantage*, *compatibility*, *complexity*, *trialability*, and *observability*.

These five elements collectively influence potential adopters' decisions toward adoption of an innovation. In distinct contexts, and with distinct populations, some characteristics may play a greater role than others. Understanding the comparative role of these characteristics in the context of systematic reviews' and health guidelines' potential use of automation should prove useful in describing the

current state of adoption, as well as considering future research concentrations and organisational norm setting.

Methods

In addition to my three PhD advisors, Dr Tari Turner provided methodological guidance for this study.

Research questions

The goal of this project was to gather data from guideline developers regarding their attitudes and perceptions of automation, and specifically automation applied to health evidence production. This aim supported the broader goal of this thesis to explore why individuals or teams do or do not adopt automation in their workflows. As stated in the previous chapter, research questions were:

RQ1.1) How do the opinions of guideline developers towards automation of health evidence synthesis fit into the Diffusion of Innovations framework?

RQ1.2) Within the Diffusion of Innovations themes, what important concepts were identified by participants?

Semi-structured interviews were identified as the most suitable method of data collection. Applying a similar interview instrument for each participant allowed for the easier comparison of data collected from each individual, while also allowing space for follow-up questions which facilitated richer, more explanatory information around each structured question. These follow-up questions were generated spontaneously for each participant in response to the answers they provided for the structured questions. This study is reported in line with the Consolidated Criteria for Reporting Qualitative research (COREQ) checklist (included as Appendix A) [16].

Participants

Potential participants were eligible for study participation if they were previous or current developers of health policy or clinical practice guidelines, or if they had first-hand experience with current practices of guideline development.

Participants whose experience was limited to systematic reviews or other research not including guideline development were excluded. The study aimed for between 15 and 20 participants, and for roughly equal gender distribution of participants.

Recruitment

A convenience sample was used to recruit participants. Potential participants were invited via email from an existing contact list provided by Dr Turner. This list included organisational representation from the Guidelines International Network (G-I-N), NICE, NHMRC, and online evidence-based healthcare discussion groups. Individuals known to the research team were directly emailed invitations to participate in an interview and invited to forward the invitation to other potentially interested contacts.

Consent and data collection

Potential participants were invited to participate in a semi-structured interview conducted via phone or via Skype. Interviews were recorded with permission from the participant. I was not personally acquainted with any of the participants prior to the interviews; each individual was provided with information on my background and PhD research and invited to ask any clarifying questions prior to participation.

Interview questions were based on a pre-formulated interview instrument, with follow-up questions to clarify or to elaborate on responses provided. I initially drafted the interview instrument, followed by feedback and validation from the rest of the study team. The final interview instrument is provided in Appendix B.

Participants were provided with an explanatory statement regarding this specific study prior to the interview and were not provided with the questions in advance of the interview.

Data were collected in relation to:

- Current methods used for collecting or using evidence in the production of guidelines and policy

- Interviewees' knowledge of, and attitudes toward, the use of automation (including text and data mining, and machine learning) in the production of systematic review evidence
- How these technologies might affect the translation of evidence into guidelines and policy, including barriers and facilitators

Following the interview, I transcribed the interview and provided the transcript to the participant for verification. Data were stored on password-protected and encrypted storage devices and managed in accordance with University College London (UCL) research policies.

Data analysis

Following verbatim transcription of the interviews and participant validation, the transcripts were entered into the QSR NVivo 12 data management program [17].

A thematic analysis approach, as outlined by Braun and Clarke [18], was selected as the most appropriate method of data analysis for the study's research questions. The thematic analysis combined deductive and inductive methods. The chosen method of combining a deductive and an inductive approach allowed for a framework analysis to be conducted in addition to the reflexive and iterative insights driven by the resulting grounded data [19].

Rogers' Diffusion of Innovations characteristics are relevant to this work as the leading framework of why new technologies or practices do, or do not, become the new standard of practice, and were therefore applied as the deductive, or pre-existing, framework. This framework has been applied to a wide variety of fields, and empirical data has consistently supported the themes it lays out. Applying it in this context both tested the applicability of the framework, as well as providing structure to the first phase of analysis.

Deductive analysis is best applied when driven by specific research questions [18], and was therefore appropriate for RQ1.1. Once it was clear how the collected data answered this first question and how it fitted into the chosen framework, the grounded inductive approach provided further insights in identifying and explaining the reasons behind these findings.

I performed the analysis in five stages, described in more detail below.

Stage 1: Assignment within predefined frameworks

First, Rogers' Diffusion of Innovation framework was used as the top-level deductive codes, using a line-by-line verbatim assignment of transcripts to one or more of the five themes (*relative advantage, compatibility, complexity, trialability, and observability*). Thematic coverage, referenced throughout the results, indicates the proportion out of total coded material which was coded to a particular theme.

This initial stage also allowed for thorough familiarisation with the data, as suggested as the first phase of analysis by Braun and Clarke.

Stage 2: Open coding within Diffusion of Innovations framework

Once each transcript was coded according to the top-level frameworks (i.e., Diffusion of Innovations), a codebook – a document containing all data belonging to a code or theme – was generated for each of the five themes. These codebooks were then examined with an open coding method. Codebooks were examined in detail to identify grounded open codes; that is, in this stage, all concepts expressed by participants were identified individually.

Stage 3: Generation of themes

The codebook of each Diffusion of Innovation theme was reviewed across all transcripts together to identify shared patterns among the grounded open codes. Each individual code was grouped with others with similar meaning and content, forming preliminary explanatory themes. For example, codes which individually expressed an opinion relating to a trade-off between time and quality were grouped into a preliminary theme.

Following formation of these themes, a further review process was undertaken to reconsider how the themes fit together. In this process, initial groups of themes are re-examined to both ensure that the groups have been formed appropriately, and to identify any connections with other themes or concepts. In addition, outlying codes were identified as those that either had not been grouped with codes from other transcripts, or those that had relatively few grounded codes grouped together.

Stage 4: Generation of matrices

In this stage, a matrix was generated comparing each of the top-level framework themes against the grounded data-driven themes. This approach not only allowed me to describe the relative significance of each overall theme – thus addressing research question 1.1 – but also to examine this significance through different lenses.

This stage resulted in formation of sub-themes for each of the five deductive themes. Identification of shared themes that appeared across multiple participants' transcripts allowed for explanation of the results observed in the top-level themes. That is, a sub-theme may provide insight as to what ideas are important within the context of *relative advantage* or in the context of *trialability*, or insight as to why participants weighted one theme over another.

Stage 5: Identifying patterns and outliers

These matrices were finally used to describe the data in relation to the first research question (i.e., how do guideline developers' opinions on automation relate to the Diffusion of Innovations framework?), and to expand upon these data in relation to the research question 1.2 (i.e., what important concepts were identified by participants?).

Results

Participants

Twenty individuals responded to the email invitations. Eighteen interviews were conducted and varied in length from approximately 30 minutes to 80 minutes. The remaining two respondents were deemed ineligible due to lack of first-hand guideline development experience. Table 4.1 presents the characteristics of the final sample. In addition to NICE and NHMRC, organisations represented included the Agency for Healthcare Research and Quality (AHRQ, United States), the Joanna Briggs Institute (JBI, Australia), the World Health Organization (WHO), and private consultancies. No participants withdrew from the study, and no repeat interviews were required.

Table 4.1. Participant characteristics

Characteristic	n
Years of experience in evidence synthesis	
Less than 5 years	2
5-10 years	9
10-20 years	5
20+ years	2
Gender	
Female	13
Male	5
Age range	
30s	3
40s	8
50s	3
60s	4
Location	
Australia	11
United Kingdom	5
European Union	1
United States	1
Primary affiliation	
Academic	10
Government	7
Private sector	1
Disciplines represented:	
Aged care, Allied health, Cardiovascular, Diabetes, Health promotion, Infectious diseases, Information science, Nutrition, Occupational health, Primary care, Psychology, Research translation, Speech pathology, Stroke, Women's health	

The following sections provide an overview of the results, followed by details of the sub-themes identified within the Diffusion of Innovations framework, and conclude with the contextual factors identified during the analysis.

Overview

Interview transcripts demonstrated high consistency in distribution of themes discussed. Following initial coding (Stage 1 as described in the Methods section), *compatibility* had approximately 45% coverage across all transcripts. *Relative advantage* and *observability* had roughly equal coverage with about 25% each, while *trialability* and *complexity* demonstrated fairly low coverage with approximately 5% each.

Sub-themes identified under *compatibility* were an *ability to double-check* and *transparency as accountability*. When discussing *relative advantage*,

participants focused on the *freeing up of human resources*, and to a lesser extent on time and cost saving. The sub-themes identified within *observability* were *a need for evidence* and *a personal need for double-checking*. *Complexity* and *trialability* were not emphasised in the data provided by participants. Upon reflexive examination of how the data informed the deductive framework, several contextual factors were identified. Participants overwhelmingly cited a lack of hands-on experience with automation tools as a moderating factor prior to providing their input. While this might initially be read as data informing *trialability*, it is important to note that participants did not focus on this as its own theme, but rather used it as a disclaimer.

Figure 4.1 on the following page provides a visual overview of these results.

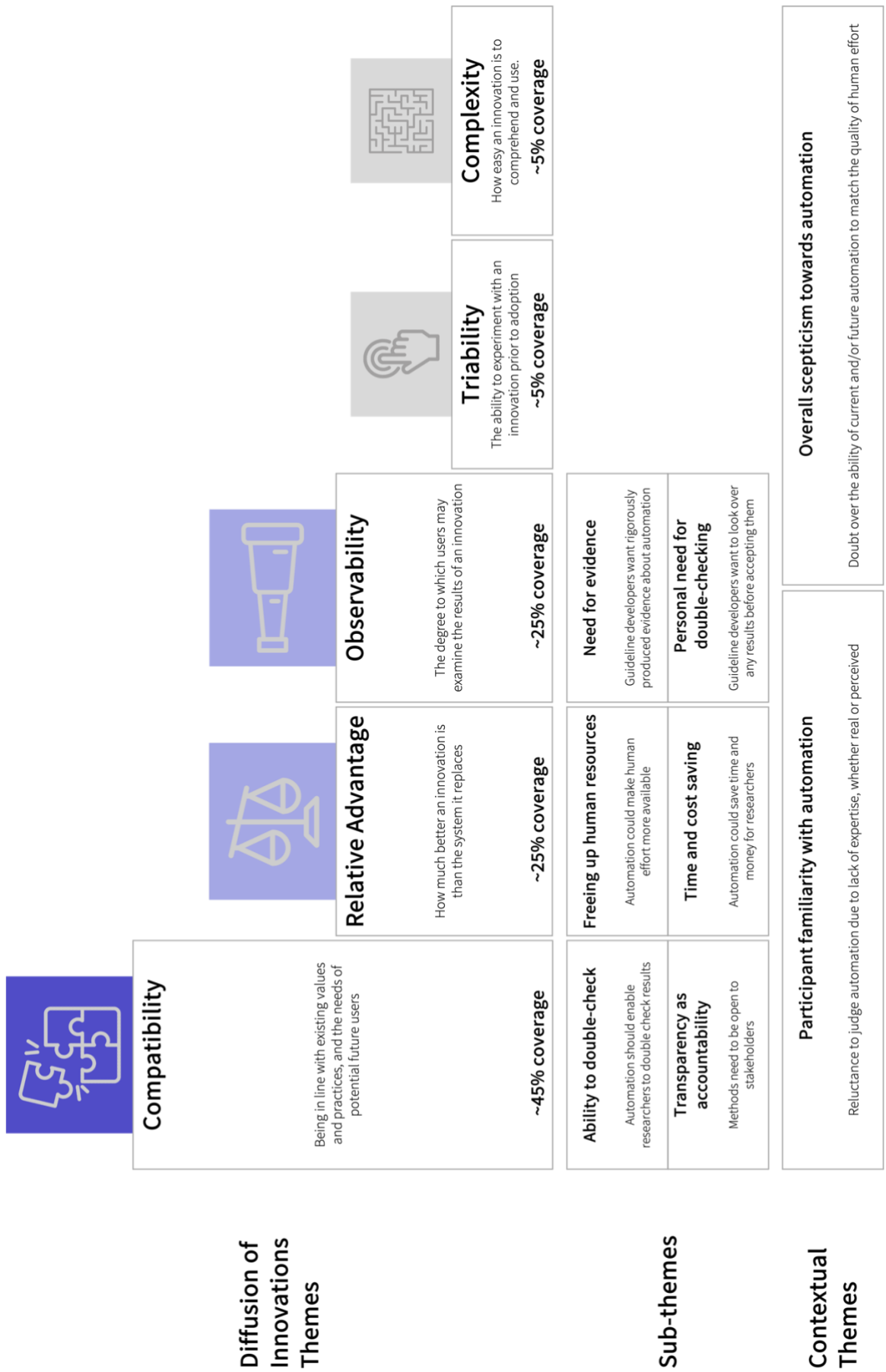


Figure 4.1. Visual overview of guideline developers' opinions via the Diffusion of Innovations characteristics⁴

Compatibility

All participants discussed their values as guideline developers at length. Overall coverage of the *compatibility* theme – how values relate to use and adoption of new tools – was consistently far greater than any of the other four themes in the deductive framework being applied in this analysis. Some examples of values were a rigorous approach to evaluation and synthesis of evidence, careful construction of questions, and a need for human and organisational input. While some of the values important to participants varied, some were consistent and discussed in further detail below.

“How you synthesise it, how you pull it together is kind of key”

Participant 3

“I think it would be a shame if humans weren’t involved in

[synthesis].” Participant 9

In discussing the values involved in their work, existing practices were often described and used to contextualise potential uses of machine learning (ML). Participants expressed a desire to map ML onto existing evidence synthesis practices. For example, it was highlighted that current practices emphasise task completion by multiple researchers (e.g., dual screening, dual extraction), and that for evidence produced by machine learning to be accepted, any results it produces should also be subject to this same double-checking process.

Two sub-themes were identified within the data sorted into the *compatibility* theme which further detailed participants’ desire to match new practices with the values which underpin current practices: *ability to double-check*, and *transparency as accountability*. In discussing the former, participants described how existing practices require that multiple checkpoints are applied to the process of collecting and synthesising data and applied this same requirement to any novel methods. In discussing the latter, guideline developers highlighted that any method, including automation, must be transparent so those ultimately using the results have access to information about the methods and decision processes.

Ability to double-check

Most participants indicated the importance of the *ability to double-check* the work of automation by a human researcher. These discussions often cited as a rationale that current practices usually involve a human double-checking the work of another human and posited that newer workflows should therefore maintain this pattern with a human double-checking the work of a computer.

Even among participants who were optimistic and broadly encouraging of the adoption of automation in health evidence synthesis, their comments were typically qualified by the caveat that results must be double-checked, as they are in current manual methods. The application of this principle to double-checking was extremely clear throughout the collected data; it was less clear which other workflow norms are considered important to maintain. Some indicated that reproducibility was the underlying reason for the double-checking status quo. It is possible that this result might change over time should rigorous research alter overall perceptions of the reproducibility of automated screening and extraction; this is further discussed as a contextual factor in subsequent sections.

Finally, attention should be drawn to a common feature in both of the quotes below: both participants indicated that “someone” should be checking the results, but not necessarily the individual themselves. This point differentiates this sub-theme from the later discussed sub-theme of *personal need for double-checking*.

“I can see it could be done. But surely it would need to be checked by someone anyway. Because even if it’s done by a human with vast experience, it’s always important to have a second person to check it.” Participant 5

“At the minute the standard is for two operators. So you’d want it to have been checked by a second method, if not person. So that would be my only thing – the reproducibility.” Participant 7

Transparency as accountability

Several participants wanted to ensure that the methods used in synthesising evidence were freely accessible and transparent to examination. In particular, many

emphasised that they are accountable to stakeholders who need to be sure they have not missed any information.

Trustworthiness of evidence in general is integral to the professional culture of guideline development, and this was readily apparent throughout the collected data. In the views of participants, trustworthiness and methods to verify it therefore extend to new tools that use automation in the form of transparency and validation.

“We have to make sure that if you’re getting information from, from whatever source, that source should be valid, that source should be credible. And, um, if you have to come up with a tool, or like a short checklist, or background check of some sort, then that’s probably a way to validate the source.” Participant 18

“A group of experts can apply judgement to that body of evidence and needs to know they can trust the evidence that you’d found.” Participant 12

“The key part of working with a face-to-face committee ... is you have they have to have total confidence in what the technical team has done” Participant 16

Relative advantage

Overall, *relative advantage* and *observability* were both prominent themes among participants’ discussions, though not to the same extent as *compatibility*. While some participants did mention that the ability to lessen the time required to develop a health guideline would be desirable, it was not given as much importance in discussion as other points. No participants indicated an openness to a trade-off between accuracy and time. When prompted to discuss ML directly (in contrast to general views of evidence synthesis and guideline development approaches), participants tended to more frequently discuss ideas relating to the *relative advantage* of automation. Participants were interested in freeing time and money, but contingent upon the automation perfectly matching perceived human quality.

Two sub-themes were identified in the data as significant in relation to the potential advantage of switching to automated or partially automated evidence

synthesis practices: direct *time and cost savings* and *freeing up human resources* to be redirected to other tasks.

Freeing up human resources

The primary advantage specified in data relating to *relative advantage* was the potential to make human resources more freely available for rededication to other tasks within the health evidence workflow. While instinctively it might be expected for an individual concerned about the financial burdens of research to aim to minimise person-time required, participants indicated they would instead seek to redirect person-time expenditure given the opportunity. For example, one respondent talked about the “drudgery” of tasks like screening being taken over by computers, with other tasks receiving additional attention as a result.

It may be helpful to view this statement through the lens of the later discussed contextual theme *overall scepticism towards ML*. Many participants were sceptical that a machine could offer the judgement calls that a human could. As they clearly view this as a unique and irreplaceable contribution, unable to be replicated by machine, it makes sense that they would prefer to spend more human resources on these judgements (i.e., “research-related tasks”).

“In research time is always limited and you know there’s never enough grant money to help employ staff. So, then kind of take that load off by having a machine do it, it would be cost-effective, and spare the researchers’ time to do other research-related tasks.”

Participant 17

Time and cost saving

While respondents more typically discussed *relative advantage* in terms of allowing person-time to be better spent on other tasks, some respondents also identified that automation might potentially save time and/or save money.

“No matter how quickly a guideline’s done, everybody always wants it faster and to be of high quality. So anything that can improve on that would be welcome, I think.” Participant 11

It should be noted in the above example that the participant has implied that stakeholders care about the cost and speed, while the data indicate that guideline developers themselves value quality above all other attributes.

Observability

A need for *observability* of potential automation was made clear throughout the interviews. More specifically, users communicated that they would like to see evidence prior to implementing new practices (*need for evidence*), as well as a continued ability to directly observe the behaviour of the technologies (*personal need for double-checking*).

Need for evidence

The need for rigorously produced, disseminated, and easily accessed evidence was clear in the responses. Participants tended to use language indicative of deep unease in the absence of validation, such as “concern”, “risk”, or “distress”. Several participants expressed an openness to automation being integrated into evidence synthesis, on the condition that accuracy has been demonstrated.

“I think at the moment it has a potentially high level of risk of being incorrect. But I don’t really know enough about it. I’d need to be convinced about it I think to consider it.” Participant 9

“If the whole process were done by some machine or machine learning application, I think it would need to be properly trialled.” Participant 5

“As long as there was clear data to support that ... machine learning is a reliable method, but you know, better than or equal to humans doing it.” Participant 17

One notable outlier indicated they were already convinced of automation’s abilities within the specific context of screening. This unusual case raises the possibility that this study’s participants would provide different data if repeated at a later time, pending further evidence production and dissemination.

“I do think it’s been well demonstrated for the screening aspects, for the hit rates of what gets included and what doesn’t, and how correct it is.” Participant 11

Personal need for double-checking

Expanding from the *need for evidence* prior to implementation, participants also often wanted an established and ongoing method of observing the inner workings of the ML processes. This was frequently described as a desire to “check” what the machine had done to ensure it was correct. While similar to the previous theme under *compatibility* of an *ability to double-check*, this is a *personal* desire to look into the methodology of the automation (e.g., “I need to be able to look under the hood myself”), rather than a continuation of the guiding principles of the field (e.g., “it is important to us that someone can check under the hood”). In other words, *compatibility: ability to double-check* states that guideline developers believe the ability to check methods should be available as a matter of principle, while *observability: personal need for double-checking* states that guideline developers want to do such checking themselves.

This need to be able to continually check how the machine learning has processed information could be interpreted as a desire to maintain control over the evidence synthesis process. As previously discussed, guideline developers must convince other stakeholders of their recommendations’ integrity, so personal quality control fits in with the cultural expectations of guideline development.

“The thing that’s sort of a little bit distressing from a novice point of view with machine learning is not feeling like I have a way to check it... I’d need some way to be confident [I’d need] a way to check the algorithms” Participant 3

Complexity and Trialability

Participants did not significantly highlight a need to trial technologies prior to their implementation themselves, though they expressed a need for others to do so as previously described (see *need for evidence* above). A small number of participants briefly mentioned that the *complexity* of any tool, in particular the initial on-boarding cost in person-time, would need to be balanced against the *relative advantage*. In

these instances, participants made clear that a significant *relative advantage* would need to be present in order to justify the change to a new methodology. They also expressed an overall preference for familiarity over the novel.

“Whenever you try and really change things, I think there’s a degree of scepticism anyway...I think that might just be the nature of human beings.” Participant 9

“If they have to learn the process, and if it’s hard, then that sort of discourages them.” Participant 18

“So unless the technology offers a value add that’s substantial enough to overcome the learning curve...however much time it takes to do that has to not be more time than you’re gonna save.” Participant 3

Contextually significant themes

In completing my framework analysis using the Diffusion of Innovations framework, several contextual factors arose which could be notable modifiers on the above-described outcomes.

Participant familiarity with automation

Participants nearly always offered disclaimers prior to commenting further, indicating that they felt they did not have sufficient experience with automation technologies to be able to comment at their desired level of expertise.

Following on from these disclaimers, participants tended to offer what knowledge they did have. These data were of significant interest as they demonstrated a current lack of robust knowledge within the target population of the capabilities of what automation can accomplish.

“I’ve done a very little bit with machine learning.” Participant 3

“It’s just my concern would be that I’ve not had any experience with it.” Participant 7

“I haven’t had much to do with machine learning. Like I’ve kind of heard about it” Participant 17

“I think that’s something I have no personal experience with”

Participant 11

“To be honest I actually haven’t had much experience with it”

Participant 8

“Yeah, I don’t know, I don’t really understand that process.”

Participant 5

Overall scepticism towards ML

Overall scepticism or mistrust towards automation, both towards current technologies and anticipated future ones, was clear in the contributions from participants. They particularly expressed doubt over the ability of a machine to mimic human judgement calls they felt are currently essential to well-formulated health guidelines.

“I guess I’d be dubious about the accuracy of that ... systematic reviewing is very much about value judgement... It would be very difficult to train a machine to make the sort of value decisions that we have to make” Participant 10

“I’m still a bit nervous about some of the interpretation of that ... it just might be a distrust about it, I think?” Participant 13

It is clear that guideline developers feel judgement and interpretation are important elements of their work. The participant quoted above drew attention to the nuanced judgements and interpretation required in health evidence, then juxtaposed this point with a mistrust of applying automation. The proximity in discussion of these two points can be reasonably used to conclude the participant believes automation incapable of producing such nuance.

One participant offered a slightly different perspective from previous points about scepticism towards ML’s ability to replace human judgements. While they indicated some confidence that the analysis itself could be completed by automation, it would rely on human input for the choice of analysis in order for it to be considered accurate. They have also linked back to the previously discussed sub-theme of a *need for evidence*, underlining the interconnectedness of the contextual

factors with the overall Diffusion of Innovations framework results. This contrasts with the two previously mentioned participants' contributions: while they had a blanket mistrust of automation's abilities, the quote below signals a conditional openness, while still maintaining a degree of disbelief.

“Obviously the analysis can be done automatically, but choice of analysis I think would be very suspect... I'd be a little leery of that now until I knew more about the accuracy of the techniques.”

Participant 11

Further illustrative quotes are provided below:

“I don't think it could fully replace a human ... I think there can be subtleties between how things can interact... I think there's always going to be some sort of human element.” Participant 9

“I don't know if we're there yet. Maybe we'll get to the point where we can do that, but to do that, like, quality rating, or to do, um – a level of evidence, or strength of evidence... I mean there's still a lot of value judgements in that. And I don't know how much machine learning could help with that at this point.” Participant 3

“How can a computer apply judgement? ...There's judgement required when it comes to things like quality or – they are not things I expect to be evidence that could be accurate.”

Participant 12

Discussion

This section will describe the key findings of this study, examine some of the potential underlying reasons for those results in the context of existing literature, reflect on this study's limitations, and conclude with recommendations for future research.

Summary

The most compelling conclusion of this study is that guideline developers have deeply held values in relation to their work (*compatibility*), and that these values are foundational when considering transitioning towards novel tools. The culture of guideline developers is highly focused on the expectations and attitudes of their peers and key stakeholders, as demonstrated by the identified sub-themes and the skew towards *compatibility* in the results. Rigorous evaluation prior to adoption (*observability: need for evidence*) and an ongoing ability to examine and re-examine the work of the machine (*compatibility: transparency as accountability* and *personal need for double-checking*) are also important themes throughout participants' contributions. These results must be considered within the context of a lack of expertise within the guideline development community regarding the current and potential abilities of machine learning (*participant familiarity with automation*) and possible reluctance to trust it even in future scenarios (*overall scepticism towards ML*).

While *compatibility* was clearly identified by participants as the most important contributor to their responses, *relative advantage* and *observability* also provide noteworthy contributions which should not be overlooked, according to these results. *Time and cost saving* are important features to develop, but they are not at the core of the decision to adopt or not to adopt a tool, as evidenced by the lack of more significant discussion. Participants spoke in more detail about the *freeing up of human resources*, and generally saw automation tools as allowing for more person-time to be dedicated to deeper and more nuanced analyses. That is, they did not see automation as replacing human effort, but instead as re-directing it.

A high proportion of the participating guideline developers indicated that while they are open to the more widespread use of automation in health evidence production, it was contingent upon the availability of high-quality evaluations supporting the reliability of the tools (*observability: need for evidence*). In addition, participants indicated that throughout the integration of an automated process – before, during, and after – they would want a method of double-checking the methods of the machine themselves (*observability: personal need for double-checking*). Both the *need for evidence* and the *personal need for double-checking*

methods could be interpreted as a need on the part of guideline developers, and most likely on the part of systematic reviewers, to continue to feel in control over all steps of evidence production.

Cultural standards of practice greatly influence decision-making

Perceived cultural standards (i.e., dispositional trust) around the quality of evidence production were the strongest influencer of participants' opinions towards automation. The alignment of the theme of *compatibility* with the framework of dispositional trust is especially important because dispositional trust tends to remain consistent over time; it should not be expected to change this cultural expectation in the short- or even medium-term. Guideline developers demonstrated deeply held core beliefs surrounding the methods of their work, and these will be central to consider in the potential adoption of automation to health evidence synthesis. This fits in with what is typically observed in the field of evidence-based medicine: researchers (particularly in the public sector) greatly emphasise methodology and perceived quality. While the results of this study provide more robust evidence demonstrating this cultural tendency, in isolation they do not provide direct explanation of the underlying reasons for this culture. Examined together with other sources, it is possible to infer potential explanations; these are further discussed later in this section.

In communicating their values around how to create high-quality evidence, a sense of unease was apparent in relation to maintaining a sense of end-to-end control of their work. Participants indicated a need to directly manage evidence synthesis, and they are reluctant to relinquish this control to a computer. This reluctance is largely related to the expectations of regulators, end users, and the demands of the EBM environment, according to the evidence gathered in this study.

This study provides empirical evidence in support of previous hypotheses in the literature. A 2013 paper commenting on the reasons for slow uptake of automation posed the broad question: “why is [automation] not yet widely used?” [20]. At the time, the authors concluded that “further technical and empirical work is needed ... [to] develop solutions which have a demonstrative relative advantage, and

which are clearly compatible with the needs of systematic reviewers and their users.” That is, *relative advantage* and *compatibility* were the most key themes in their opinions, playing a co-equal role in the adoption of automation. Considering the data presented in this study in relation to Thomas et al’s question, the prior conclusions should be adjusted slightly.

The most significant reason appears to be that automation has not fully demonstrated to key stakeholders of EBM that it is compatible with their guiding values, principles, or standards. While *relative advantage* was important, it was secondary to the far more prominent discussion of *compatibility*. Further, the identified sub-themes of *compatibility* focused more on automation’s fitting in with the values behind current practices than on fitting in with existing infrastructure, rather than the current practices themselves.

The preceding points not only represent a shift from the hypotheses presented in the literature, but also from the focus of previous discussions at relevant conferences. The International Collaboration for the Automation of Systematic Reviews (ICASR), formed in 2015, is a global network endeavouring to successfully automate all parts of systematic review production. In the notes of the third ICASR meeting in 2017, the group concluded that the “most pressing needs at present are to develop approaches for validating” automation and integration with existing system architecture [21]. Stated another way, ICASR believed *observability* to be critical to uptake, as well as *compatibility* specifically in reference to fitting into existing practice.

Again, as in the previous case discussed, this research has provided some evidence to support this assertion but also supports reprioritisation of which attributes are best suited to promotion of automation adoption. While evidence gathered in this study reinforced that *compatibility* plays a significant role, it also demonstrates that alignment with values is more highly prioritised than alignment with current practice and system architecture. In addition, the “most pressing need” may not be validation (*observability*), but instead is the demonstration and communication of methodological standards and cultural coordination (*compatibility*).

Potential sources of cultural norms of guideline development

In considering why EBM has such an emphasis on perceived quality, a logical first place to look are the mission statements of prominent guideline organisations. Given that the sample of participants was largely from either Australia or the United Kingdom, NICE and NHMRC were examined.

Mission statements of NICE [22] and NHMRC [23] emphasise quality and transparency, but do not necessarily give obvious insight to contextualise systematic reviewers' and guideline developers' focus on expectations of the regulatory environment. The concept of reproducibility is found in many guiding documents for both evidence synthesis and guideline development (e.g., NICE, Cochrane Handbook [24]); reproducibility may inch closer to identifying the core issue driving the preoccupation with accountability of research results.

Some of the participants' quotes regarding what constitutes high-quality evidence are directly mirrored in NHMRC's "Guidelines for guidelines" [25], making it a helpful resource in contextualising these results. In this publication, it is emphasised that researchers must take care that absolutely nothing is missed and recommends replication of each stage by a second expert as the best method of ensuring this goal. For instance, a search strategy should be reviewed by a second information specialist. Stated another way, it appears that sensitivity (i.e., missing nothing) is valued far more than specificity (i.e., minimising superfluous records) in assessing the quality of literature and database searching.

An intriguing point to note in the data collected during this study is that these two things – double-checking / replicating work, and certainty of an exhaustive search (i.e., near 100% sensitivity) – are a single concept in the NHMRC documentation but are two separate ideas when discussed by guideline developers. In the NHMRC guidance, double-checking is explicitly assigned the purpose of ensuring an exhaustive search; given this, it might be expected that if there were another way of ensuring an exhaustive search, double-checking would then be considered unnecessary. However, in the interviews here, participants indicated that double-checking has become an end in itself as a principle of good practice in and of itself, and in addition to ensuring high-quality searches.

Proponents of automation might have a choice ahead of them here: whether to align with the documented guidance, or whether to align with how this guidance plays out in practice.

Researcher effort will be redirected rather than replaced

Guideline developers anticipate that automation will be most useful in redirecting person-time rather than replacing it. This observed anticipation of automation allowing for refocusing of effort is what should be expected if the results of this study are situated in the historical evidence and context. From the late nineteenth century onwards, there have been repetitive waves of automation of production and consequent population-level job panic [26]. With each wave, however, human effort has not been erased, but rather redirected. In some cases, job opportunities have actually expanded rather than contracted as a by-product of widespread automation. Therefore, in addition to enabling the valuable skills of EBM researchers to be better spent, it is very possible the field will see an expansion of opportunities.

Participants highlighted that a critical (and in their view, irreplaceably human) part of their professional contribution is the nuanced judgements applied to collected evidence, often derived from lived experience. In their view, any person-time freed due to automation would most likely be redirected towards this judgement and consensus process. That is, automation could contribute to an improvement in guideline quality by providing additional resources (namely, person-time) to more difficult aspects of guideline development, and not simply by cutting costs, workload, and human resource demand.

In this case, contributions from participants in this study relating to reluctance to relinquish human judgement align with notes from the previously mentioned ICASR meeting [21]. They stated:

“For example, external stakeholders might believe the current vision is automated reviews devoid of valuable human control and input, that is, a general autonomous artificial intelligence system. That view, however, was neither represented nor sanctioned at the

meeting. Therefore, improving the terminology associated with systematic review automation to reflect the goal more accurately is likely valuable.”

This study provides evidence in support of this proposition: participants were wary of automation in part due to their perception that it might remove crucial human judgement in the process of guideline development. Notably, however, that is unlikely to be the goal of automation in the foreseeable future. Potentially even more important is the point that encouraging complete and total replacement was “neither represented nor sanctioned.” Given that participants in this study echoed this sentiment, it raises the question of why guideline developers hold this view, and also how to best communicate a more accurate representation of the goals of advocates for automation in systematic reviews and guidelines.

Overall, the results of this study support previous conversations surrounding the use automation in the context of EBM, namely that guideline developers inaccurately perceive automation as aiming to entirely replace human effort and would prefer instead to use it to redirect researcher time to more complex tasks. Given historical precedent, it should be both anticipated and encouraged that automation will redirect human contributions in evidence production. This reallocation of human effort will increase efficiency by reducing time spent on some tasks, and it will improve quality by dedicating more human resources to complex tasks.

As in the previous section, perhaps proponents of automation have a choice to steer the general conversation to clarify that expert opinion will not be supplanted, but instead made more accessible by freeing up person-time and other resources. An enabling environment for the promotion and adoption of automation in a manner that redirects rather than replaces research effort could be an effective strategy in building consensus among guideline developers, as key stakeholders of the evidence synthesis process, in accepting, implementing, and promoting automation practices.

Comparison with a previous similar study

The results of this study can be further understood by contrasting the results with a study conducted in 2016 regarding implementation of new technology tools

for dentists [27]. As this study also used the Diffusion of Innovations framework in a health-focused setting, it is an appropriate comparator. *Relative advantage* is the most weighted of the Diffusion of Innovations themes when dentists are considering adoption of new tools, according to the findings of Matthews et al (2016). This differs distinctly from the findings of this study that guideline developers are more influenced by their perceptions of a tool's *compatibility* with the values system underlying their work.

Given the similar designs of each of these studies, these contrasting results provide an interesting opportunity to consider why dentists and guideline developers have offered such differing insights. What is the difference between these two groups?

To begin to understand this, the similarities and differences between the two groups should be examined. Professional associations of both groups typically include better health outcomes and accessibility in their mission statements, so it can be reasonably inferred that the ultimate goals of each are similar in relation to health outcomes. Both dentists and guideline developers appear to feel accountable to themselves in assessing whether they have met their own standards of work and methodology. Dentists and guideline developers occupy distinct positions, however, in the causal pathway leading to this goal, and have different stakeholders to whom they are most immediately accountable. This difference appears to influence how their opinions towards technology adoption fit into the Diffusion of Innovations framework. The dentists in this study indicated repeatedly that they are accountable to their patients; a sub-theme of *relative advantage* described in the report is *will it benefit my patient?* The authors wrote: “[study participants] felt strongly that any new technology should benefit the patient and enable tailoring of treatment to the individual’s needs.” Guideline developers tended instead to express accountability to methodological requirements (see: *compatibility: transparency as accountability*). Enforcement of these requirements was described within the working environment of the guideline developers (i.e., their peers), and at other times in relation to institutional sponsors, if applicable to a specific guideline. The evidence that they integrate into guidelines must be perceived as rigorous, bordering on infallibly trustworthy.

Overall, this stance seems to have led guideline developers to favour *compatibility* rather than *Relative advantage*. These two studies juxtaposed could lead to the hypothesis that the group considered as the most important stakeholders will strongly influence which of the five items outlined in the framework will emerge as the most prominent in adopting an innovation. Further research is warranted to determine if this holds true across multiple contexts, or if the difference in the conclusions of these two studies is situationally unique.

Contribution of analytical frameworks

Hoff and Bashir's trust model [28] is useful to explore potential reasons behind the emphasis in these results on *compatibility*. Two components of trust seem to appear here: external variability in situational trust, and culture as an influence on dispositional trust. Organisational setting is identified as a contributor to external variability in situational trust, and it is therefore reasonable to expect that this is playing a role in bringing *compatibility* to the forefront as an influence on guideline developers' opinions. However, the more significant potential contributor from the trust framework to these data is dispositional trust. These results provide further support for Hoff and Bashir's assertion that culture is a "particularly important variable." Using dispositional trust as a lens through which to view these results brings out a new and significant conclusion: the prominence of *compatibility* in influencing guideline developers' opinions on automation is unlikely to change over time. Future applications of the trust in automation framework in the context of health evidence synthesis should be strengthened by this conclusion.

Of the three conceptual frameworks applied to this thesis, this chapter most heavily relied upon Rogers' Diffusion of Innovations. Guideline developers most prominently expressed interest in the ideological compatibility of automation-enabled workflows with their existing professional values, anticipated the greatest relative advantage in freeing up person-time resources, and generally felt uncertain about their own understanding of automation technologies.

In addition, the levels of automation taxonomy [29, 30] can be applied to the conclusion that guideline developers anticipate the redirection of researcher effort, rather than the replacement of it. It was noted in Chapter 3 that a contribution of the

levels of automation taxonomy was the assertion that automation need not mean wholesale erasure of human effort. This assertion has been shown to be highly applicable to this context by the results of this study. Further, given the ideas put forth by study participants, it might be concluded that guideline developers tend to prefer mid-level automations (i.e., levels 3 through 7). These specific levels appear to align most closely with the guideline developers' need to observe any automated results.

Study limitations

A potential limitation of a study of this kind is the construction of the sample and whether a different sample might lead to different conclusions. The limitations of convenience sampling and the resulting potential impacts on generalisability must be noted.

Convenience sampling was used to efficiently target potential participants from the specified group of interest. This resulted in a sample of 18 participants from a relatively small number of organisations. Though efforts were made to ensure some varied representation across gender, origin, and career stage, identifying similar respondents remains a possibility when using this method [31]. A more diverse sample might have been desirable; however, consistency in data contributed from participants suggests that additional participants might not have changed the results and conclusions.

It might be expected that convenience sampling used in this manner (i.e., using personal direct contacts and networks) would result in respondents with similar views to the investigators. Moreover, as the identity of a researcher is intrinsically linked with any qualitative research they conduct, it also might be expected that my association with known automation researchers would influence the responses of participants. Neither of these risks was readily apparent in the data, however. The generally low awareness of the capabilities of ML, and of the aims of integration of ML into EBM, indicate that participants had minimal prior knowledge of my research associations. They did not appear to hold similar opinions to my own and were able to express those opinions openly in our interviews.

One clear skew in the resulting sample was the inclusion of only 28% male participants. Despite this, distribution of data points within the Diffusion of Innovations framework did not appear to vary according to gender. It is possible that this balance is representative of the current balance in the field of guideline development; for example, according to data available from NICE, they employ 68.63% women and 31.37% men [32].

An additional vulnerability of the sampling technique was the potential over-representation of Australian professionals. Five participants, however, had direct experience in low-resource settings and/or originated from countries other than their current base, broadening the potential perspectives for this analysis. Nevertheless, any use of these results should be tempered by awareness of the strong Australia-, UK-, and US representation in the data collected.

Though the data do not demonstrate many of these negative impacts of convenience sampling, it nevertheless must be highlighted that these risks exist in this study; results should therefore be considered with this context.

Current state on the adoption curve

Finally, as previously outlined in the chapter introduction and in Chapter 3, Diffusion of Innovations theory describes the typical categories of innovators (personas) and provides an approximation of the expected proportions of each category (Figure 3.1 from Chapter 3). As time progresses, successive groups will adopt a given innovation, until a critical mass of the market share is reached.

The finding that many of the participants perceived themselves to be inexperienced with automation in the context of evidence synthesis raises the question of where the field currently resides within this adoption curve. The evidence of this study suggests that the field is in very early stages of adoption (i.e., innovators) with only a small minority taking on use of this new technology.

Once a later stage has been reached, subsequent studies may well return different results. For example, the case from Participant 11 in *observability: need for evidence*, while certainly an outlier in the context of this study, may fall under an innovator or early adopter persona while other participants may fall under late majority or laggards. Additional analysis and/or data collection using persona

categories as the deductive framework could build upon the results of this study. However, until a later stage of diffusion is reached, it may be difficult to find sufficient contributors within each category.

Suggestions for future research

While guideline developers are a crucial group within the field of evidence-based medicine, they are far from the only one. As mentioned in the chapter introduction, patients/consumers, caregivers, healthcare professionals, policymakers, and researchers are also key stakeholders in evidence synthesis. Therefore, it would be logical to repeat this study with different population groups. Systematic reviewers could be considered a high priority for consultation, as these individuals will be using automation software directly, as opposed to guideline developers who act as gatekeepers of the output (i.e., health evidence) of such software. They may also represent a different point on the adoption curve.

Further to the above, patient stakeholders are an integral group to consult. Patients are often involved on guideline panels, and there have been recent pushes to include more consumers and patients in health guidance [33]. Health guidelines should ultimately aim to benefit patients and the community, and organisational mission statements often (and rightly) include statements about patient transparency and empowerment. Finally, policy makers should also be examined, as they were identified by some of the participants in this study as fellow stakeholders in the process of creating guidelines, and whose values influence the practices of evidence synthesis.

In comparing this study to Matthews (2016), it appears that the results of this study were influenced by the most proximal stakeholder group to whom the participants felt accountable. In proposing further work using systematic reviewers as the participant pool, it would also be intriguing to determine who systematic reviewers consider their stakeholders when considering the adoption of automation. That is, to whom are systematic reviewers accountable, and how does this affect their responses in the framework of Diffusion of Innovations? It is possible that with additional data, additional insights into the effect of stakeholders on views of

technology adoption can be drawn, and results could be further generalised to predict Diffusion of Innovations priorities in different contexts.

The above-outlined research should be prioritised and should proceed in parallel to the forms of validation highlighted by participants as crucial to their decision making. Select examples of automation have long been available for evidence synthesis, and several prominent organisations are encouraging automation uptake. Despite this reality, it is clear that they are not being integrated into workflows at large or even medium scale. The data from this study shows justified hesitation from a key stakeholder group, and additional data relating to other user stakeholders will be helpful in identifying barriers and facilitators for these groups.

Conclusion

Analysed via the lens of the Diffusion of Innovations framework, the results of this study strongly conclude that *compatibility* with professional cultural values is the most significant consideration for guideline developers in the potential adoption of automation. Participating guideline developers identified increased availability of person-time as a primary *relative advantage*, and desired rigorous validation (*observability*) to occur both prior to adoption and on an ongoing basis. A lack of knowledge of ML among participants is a contributing contextual factor to the slow uptake of automation, though it was unclear whether this was a real or perceived lack of knowledge. Participants also showed a generalised anxiety around relinquishing human control to a computer. The data demonstrate a common but inaccurate perception that nuanced human judgement is to be removed from evidence synthesis in favour of automation technologies. Future studies may return different results if and when the evidence synthesis field reaches a later stage in the Diffusion of Innovations adoption curve.

The creation and dissemination of empirical evidence that systematically demonstrates automation's alignment with the values and standards of guideline development and EBM should therefore be prioritised. In addition, disseminated evidence and communications around automation tools may benefit from focusing on the combination of human and ML effort, rather than the replacement of human insight.

Chapter references

1. Sackett, D.L. Evidence-based medicine. *Seminars in Perinatology*; 1997: Elsevier; 1997. p. 3-5.
2. Sur, R.L. and P. Dahm. History of evidence-based medicine. *Indian Journal of Urology* 2011; 27(4):487.
3. Herbert, R.D., et al. Evidence-based practice – imperfect but necessary. *Physiotherapy Theory and Practice* 2001; 17(3):201-211.
4. Gough, D. and D. Elbourne. Systematic research synthesis to inform policy, practice and democratic debate. *Social Policy and Society* 2002; 1(3):225-236.
5. Bastian, H., P. Glasziou, and I. Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Medicine* 2010; 7(9):e1000326.
6. Shojania, K.G., et al. How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine* 2007; 147(4):224-233.
7. Chalmers, I., et al. How to increase value and reduce waste when research priorities are set. *The Lancet* 2014; 383(9912):156-165.
8. Elliott, J.H., et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Medicine* 2014; 11(2):e1001603.
9. Marshall, I.J. and B.C. Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews* 2019; 8(1):163.
10. Thomas, J., et al. Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology* 2017; 91:31-37.
11. Tsafnat, G., et al. The automation of systematic reviews. *BMJ: British Medical Journal* 2013; 346.
12. O'Connor, A.M., et al. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Systematic Reviews* 2019; 8(1):1-8.
13. van Altena, A.J., R. Spijker, and S.D. Olabarriaga. Usage of automation tools in systematic reviews. *Research Synthesis Methods* 2019; 10(1):72-82.
14. Cleo, G., et al. Usability and acceptability of four systematic review automation software packages: a mixed method design. *Systematic Reviews* 2019; 8(1):145.
15. Rogers, E.M. *Diffusion of Innovations*. 5th ed: Simon and Schuster; 2003.
16. Tong, A., P. Sainsbury, and J. Craig. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International journal for quality in health care* 2007; 19(6):349-357.
17. *NVivo qualitative data analysis software*. 2018, QSR International Pty Ltd.

18. Braun, V., V. Clarke, and G. Terry. Thematic analysis. *Qualitative Research in Clinical and Health Psychology* 2014; 24:95-114.
19. Gale, N.K., et al. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology* 2013; 13(1):117.
20. Thomas, J. Diffusion of innovation in systematic review methodology: why is study selection not yet assisted by automation. *OA Evidence-Based Medicine* 2013; 1(2):1-6.
21. O'Connor, A.M., et al. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Systematic Reviews* 2019; 8(1):57.
22. What we do. 2020. <https://www.nice.org.uk/about/what-we-do> (accessed 7 January 2020).
23. About us. 2020. <https://www.nhmrc.gov.au/about-us> (accessed 7 January 2020).
24. Higgins, J.P., et al. *Cochrane Handbook for Systematic Reviews of Interventions*: John Wiley & Sons; 2019.
25. Guidelines for guidelines. 2020. <https://www.nhmrc.gov.au/guidelinesforguidelines> (accessed 7 January 2020).
26. David, H. Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives* 2015; 29(3):3-30.
27. Matthews, D., et al. Factors influencing adoption of new technologies into dental practice: a qualitative study. *JDR Clinical & Translational Research* 2016; 1(1):77-85.
28. Hoff, K.A. and M. Bashir. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 2014; 57(3):407-434.
29. Sheridan, T.B. *Telerobotics, automation, and human supervisory control*: MIT press; 1992.
30. Sheridan, T.B. and W.L. Verplank, *Human and computer control of undersea teleoperators*. 1978, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
31. Acharya, A.S., et al. Sampling: Why and how of it. *Indian Journal of Medical Specialties* 2013; 4(2):330-333.
32. NICE. Gender pay gap report. 2020. <https://www.nice.org.uk/about/who-we-are/corporate-publications/gender-pay-gap-report> (accessed 25 February 2020).
33. Rashid, A., et al. Patient and public involvement in the development of healthcare guidance: an overview of current methods and future challenges. *The Patient-Patient-Centered Outcomes Research* 2017; 10(3):277-282.

Chapter 5. The User Journey

Mapping adopter personas among Cochrane Information Specialists

Chapter overview

This chapter will report the results of a study examining the adoption of the Cochrane RCT classifier among Cochrane Information Specialists. The study was undertaken in two parts: a survey followed by individual interviews. The applicability of the Diffusion of Innovations adopter personas to this context was examined primarily by the survey, while the interviews were used to investigate the adoption decisions and trust levels of participants from each adopter category. The chapter will conclude with insights drawn from these interviews, recommendations for practice, and recommendations for future research.

Introduction

In the previous chapter, this thesis began its exploration of the first of its two broad themes: to explore why individuals may or may not adopt automation in health evidence synthesis. The opinions of guideline developers were examined via the Diffusion of Innovations framework in relation to innovation characteristics, and the results demonstrated that compatibility, especially in terms of professional culture norms and practices, was the greatest contributor to guideline developers' willingness to accept evidence which uses automation. Participants also indicated a perceived lack of self-capability around automation tools, though it was unclear whether this perception was accurate or projected. Among the conclusions of the study was that other key stakeholder groups should be consulted, not only to gather additional data which could be used to inform recommendations for practice, but to further test the applicability of the selected framework to the context of health evidence synthesis.

Another key group to consult in the adoption of automation for health evidence synthesis are information specialists, including Cochrane Information Specialists (CISs). Just as key organisations play a significant role in the field of guideline writing and development, certain key organisations are foundational to systematic reviewing. Cochrane is one such organisation and is an important thought leader in the field of evidence-based medicine. Their methodological standards form the basis for many other bodies' practices, as well as the training and education of many systematic reviewers [1]. Other organisations follow Cochrane's methodological lead, therefore, Cochrane's positions towards automation tools are important to consider and likely to be informative of broader practices. Where guideline developers enter the evidence pipeline in the later stages and provide a crucial link in translating knowledge to practice, information specialists are key to the early stages of systematic review production. High levels of expertise and specialisation are required to formulate a search strategy, to execute it on multiple databases, and to manage the resulting information in the most efficient and practical manner.

The search and screening stages of systematic reviews have been a significant target of systematic review automation and include many different

approaches, several of which were described briefly in Chapter 2. Methods to automate search and screening include but are not limited to active learning prioritisation (which will be examined further in Chapter 7's economic evaluation), automated query expansion, and improved deduplication. As detailed in the literature review in Chapter 2, a 2015 systematic review identified 55 studies which targeted workload reduction in screening, 30 of which focused on reduction of the number of records needed to be manually screened [2]. The conclusions of Chapter 4 noted the possibility that the adoption curve for automation in overall evidence synthesis is in early stages, and it further suggested future research might examine the same results once a later stage is reached. Given the more advanced state of automation research relevant to search and screening stages, at least in comparison to the rest of the evidence pipeline, it is possible these stages – and by extension information specialists – may be further progressed on the adoption curve, and therefore are a useful group in which to undertake this research.

The RCT classifier is one of the first machine learning tools widely available and advocated within the Cochrane community. It uses a ML routine to generate a score indicating the likelihood a study record describes an RCT. Records with a score below a pre-defined threshold score are discarded, reducing the number of records needing to be screened. Several characteristics of this tool make it a unique case study for assessing automation adoption. First and foremost, the RCT classifier is built on a large, high-quality dataset, and these datasets and source code are available for user examination [3, 4]. Further strengthening this point, the rigorous standards of the classifier were determined in collaboration with the Cochrane Information Retrieval Methods Group (IRMG), prospectively encouraging organisation-level endorsement of the tool. Cochrane IRMG set quite stringent standards during development; previous work on the classifier had set recall (i.e., sensitivity) at 0.95, whereas in response to input from the IRMG, recall was raised to 0.99. The IRMG also requested transparency of classifier scores, along with validation work using previously published Cochrane reviews. The RCT classifier is built into the tools already used by many (though not all) CISs, namely CRS Web, the online program used by CISs to interact with the Cochrane Registry of Studies (CRS) and to manage their study registers. While many automation tools are

available, few are sanctioned in the way that Cochrane has promoted usage of the RCT classifier.

As might be expected in these favourable circumstances, the RCT classifier is relatively widely adopted among CISs. If the Diffusion of Innovations personas are applicable to this population and this context, then we are at a relatively later stage on the adoption curve. This consequently presents a valuable and timely situation in which we may be able to identify multiple adopter personas and discuss their adoption process with them or to seek information from non-adopters about their rationales. Either option could provide useful data which could then be combined with the trust framework to provide structured insights about how different personas approach the process of adoption of and trust in automation. These insights could then be used to make recommendations for practice in other contexts, including those earlier on the adoption curve. However, it is also possible that the population of CISs is skewed towards one end of the adoption curve. In this case, it would be less useful – though still informative – to examine their user journeys, as it would not be possible to reliably compare and contrast the journey of one adopter persona over another.

Therefore, to analyse CIS insights to adopter persona user journeys, it must first be established whether the adopter personas are normally distributed among this population. Broadly, two outcomes are possible from such an endeavour. First, that the persona categories present in this population are significantly skewed, and it is therefore not useful to consider adoption of the RCT classifier from the perspective of multiple personas. With this outcome, it might also be expected that the adoption curve would not follow the pattern predicted in the Rogers framework. If the population is skewed towards innovators, the total market share would rise quickly and level off, whereas a population skewed towards laggards would take longer to achieve a majority market share of adoption. The second potential outcome is that the CIS populations maps as expected against the persona categories, and that advancement along the adoption curve can be informed by all the respective personas according to the current stage of diffusion.

Methods

Research questions

The goal of this project was to describe and to understand the user journey of each Diffusion of Innovations persona in adopting automation in systematic reviews. To achieve this, Cochrane Information Specialists' use of the RCT classifier was identified as a relevant and informative case study. To achieve the project goal, it was first necessary to test the applicability of the Diffusion of Innovations personas to this context; I needed to know the observed distribution of adopter personas among the population of interest (i.e., CISs) to then subsequently gather information about their behaviour and attitudes towards the tool of interest (i.e., the RCT classifier) and whether they behaved in a manner consistent with their persona. Specific research questions for this project were:

RQ2.1) How applicable are the Diffusion of Innovations adopter personas to this context?

RQ2.2) How do users interact with the RCT classifier?

RQ2.3) To what extent do users trust the RCT classifier, and what factors inform this trust (or lack thereof)?

A multi-phase mixed methods approach was used in this study. An initial survey was used to gather information about adopter persona characteristics and general engagement and behaviour with the RCT classifier. These results were used to identify potential participants for interviews from each of the persona categories and to inform the approach of the subsequent unstructured interviews.

To answer RQ2.1, descriptive data of the population of interest (CISs) was required. Given the presence of a discrete, clearly defined research question relating to a similarly well-defined target population, a survey was selected as an appropriate research method for the first phase of this project. This approach made use of the advantages of surveys: the collection of empirical, descriptive information which is likely to be generalisable [5]. Surveys also have disadvantages, namely the lack of explanatory data, and the unstructured interviews in the second phase of this project were used to address this weakness, as well as to answer RQ2.2 and RQ2.3. A

further advantage to this mixed methods approach was that in addition to answering RQ2.1, survey data could be used to formulate hypotheses in relation to RQ2.2 and RQ2.3 which could then be tested via the unstructured interviews in the second phase of the project.

This study was explicit in classifying no correct nor incorrect use, but strictly to describe the current state of use and user experience in interacting with the tool. More detail on each study phase is provided in the following sections. This study is reported in line with the Checklist for Reporting of Survey Studies (CROSS) checklist (included as Appendix C) [6].

Phase 1: Survey

Design

The survey facilitated the collection of standardised data points used to assign respondents into adopter persona categories. By inviting the entire target population to respond, it was hoped to achieve maximum population coverage for the most accurate – and generalisable – data collection possible. Guidance was taken from Kelley et al [5] to optimise survey quality, which also provided guidance on the weaknesses of surveys which should be addressed by the subsequent interviews. The survey was hosted on Google Forms; by using Google’s feature requiring sign in, participants were limited to one response.

Survey questions covered two general topic areas. first, respondents’ interaction, if any, with CRS Web and/or the RCT classifier. Second, the Diffusions of Innovations persona framework was used to draft questions aimed at revealing the dispositional characteristics of the respondents. That is, the questions were designed to reveal to which, if any, adopter persona designation the respondent belonged. This designation acts independently of the behaviour covered by the first topic area of the survey; adopter categories were positioned as personal tendencies that are not dependent on context or on the tool being examined, and certainly not tied to usage of the RCT classifier nor to CRS Web. The distinction between these two topic areas should be highlighted and reiterated: user interactions with technology tools were not used to inform their persona category during data analysis; adopter persona categorisation was informed solely by the dispositional characteristic questions in the

second portion of the survey. Questions relating to dispositional characteristics were marked as required to avoid missing data points.

Using such an approach maximised the potential implications derived from this research; insights into user trust in automation mapped against adopter persona should be able to translate across similar populations, even if the automation tool being examined varies. It also enabled me to form hypotheses about user behaviour towards automation according to adopter persona category which could then be partially tested against the interaction questions and further tested using the qualitative data collected in phase two.

Participants and recruitment

Following initial development, the survey instrument was piloted with the Cochrane Information Specialists Executive, a group of CISs which functions as advisory body for the wider group, and which liaises with the Cochrane Central Executive. Its feedback was incorporated into the final version of the survey.

The survey was then circulated using the CIS listserv, an email list which reaches all current individuals working as a CIS for all Cochrane Review Groups. The survey instrument is included in Appendix D.

Participants were provided with an explanatory statement at the beginning of the survey. This statement identified the research team, and specified that they may withdraw at any time, including withdrawal post-participation. Contact information was made available to participants, and I was readily available to participants throughout the study for any enquiries.

The survey could be completed anonymously according to individual preference. Any questions which contained potentially identifiable information included an anonymised option (“Prefer not to say”).

In order to proceed with the survey, individuals had to indicate their understanding of the explanatory statement and consent to proceed.

Data collection and analysis

The survey was active for six weeks, from mid-May 2020 through to the end of June 2020. After the survey was closed to new responses, results were downloaded and coded.

Responses from the second portion of the survey, relating to persona characteristics, were used to assign adopter categories; responses relating to CRS Web and/or RCT classifier behaviour were not used to inform coding of persona. Individuals were coded into the adopter category for which their self-reported behaviour and characteristics most closely aligned. To validate coding results, two rounds of analysis were performed. First, survey responses were coded in the numerical order in which they were completed: first respondent coded first, second respondent coded second, and so on. The second round of coding was performed in a randomised order in order to blind outcome assessment. The results of these two rounds were then assessed for agreement; where there was an initial disagreement, a third round of coding was used to determine the final identification for the respondent.

Once each respondent was assigned into the appropriate adopter persona, data relating to behaviour were assessed to identify broad patterns. The unstructured interviews from Phase 2 were informed by these commonalities among responses, whether among the whole participant pool or solely within a persona category. Data from open response questions on the survey were also used to address RQ2.2.

Phase 2: Interviews

Design

To explore RQ2.3, as well as to provide further data and explanatory reasoning for the results of RQ2.2 from the survey phase, an unstructured interview was used. This interview was designed as a discussion of how the participant uses the RCT classifier, and to gather qualitative data around how they had arrived at their current stage of trust (or lack thereof) in the classifier.

Individual survey responses were also used to frame the discussion with each of the invited interview participants. As each participant was distinct, interview questions varied among each. A common theme for all participants, however, was to

talk through their experience from their first use of the RCT classifier through the present day. The purpose of going over this history was to identify their initial trust level in the classifier, if they had evolved in their trust since then, and what had caused this shift. For example, a user of the RCT classifier might explain that they had a personalised set of tests that they like to run on any new reference technology or software, while another might instead completely trust in the RCT classifier output from the beginning.

Participants and recruitment

Only participants who had positively indicated their availability for a follow up interview during the survey phase were contacted in relation to interview phase participation. One respondent from each of the adopter persona categories was invited via email for a brief interview on their preferred online platform. The audio of each interview was recorded, and key segments were transcribed for analysis.

Participants were free to withdraw at any time, including after completing their participation.

Data collection and analysis

A combined framework analysis was conducted using both Diffusion of Innovations adopter personas in conjunction with the Hoff and Bashir three-layered trust model.

Analysis was conducted using the guidance for framework analysis from Ritchie et al [7]. The basic stages of this method are:

1. Generate a research question
2. Identify 'best fit' conceptual framework from the literature
3. Code evidence to identify themes
4. Map themes against the selected framework
5. Identify and interpret relationships between data and framework

The first two stages of this method have already been described in Chapter 3 and in the introduction to this chapter. That is, the research question has already been stated (RQ2.3), and the most appropriate conceptual models have already been identified in the literature, namely Diffusion of Innovations and Hoff and Bashir's trust framework. In the third stage of this framework analysis, data from interviews

were coded to identify prominent themes in how users interact with the RCT classifier (RQ2.2) and what processes inform their trust or lack thereof in the classifier (RQ2.3). These themes were then mapped against both the trust framework and the personas framework in the fourth stage, allowing for interpretation in the fifth and final stage of analysis with respect to the combination of the two frameworks. This combination was used to assess common and divergent themes in how and why each adopter persona trusts in the classifier.

Ethical considerations

In relation to ethical considerations for this study, there was no risk associated with participants' safety through participation in this study. All external communications of the survey's results are completely anonymised in this report and all future dissemination. This study was prospectively approved in accordance with UCL Institute of Education Research Ethics policies.

Results

Phase 1

The survey was completed by 24 individuals; given one information specialist for each of the 54 Cochrane Review Groups, this provides a participation rate of approximately 44%. Descriptive characteristics of participants (i.e., from first topic section of the survey) are presented in

Table 5.1. After mapping the responses against the Diffusion of Innovations adopter personas framework (i.e., the responses from the second topic section of the survey), distribution was similar, but not identical to, the distribution expected by the framework. Figure 5.1 shows the observed distribution among participants and the expected distribution of adopter personas. Of the 24 participants, there were 3 innovators, 4 early adopters, 6 early majorities, 9 late majorities, and 2 laggards.

Table 5.1. Survey participant characteristics

Characteristic / response	n
Years of experience in Information Science	
Less than a year	0
1-5 years	1
5-10 years	2
10-20 years	12
20+ years	8
Prefer not to say	1
Years of experience as Cochrane Information Specialist	
Less than a year	0
1-5 years	6
5-10 years	5
10-20 years	10
20+ years	2
Prefer not to say	1
Aware of Cochrane RCT Classifier	
Yes	24
No	0
Previous use of Cochrane RCT Classifier	
Yes	17
No	7
Frequency of use of Cochrane RCT Classifier*	
Tried once or twice	2
Sometimes	7
Frequent	8
Cochrane RCT Classifier rating (out of 5)*	
5	6
4	6
3	5
Use of CRS Web for study management	
Yes	9
Sometimes	3
No	12

* questions presented only to participants with previous use of Cochrane RCT Classifier (n = 17)

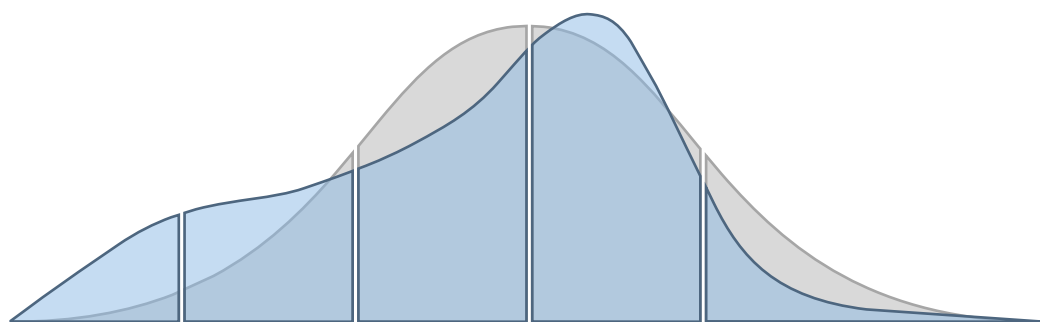
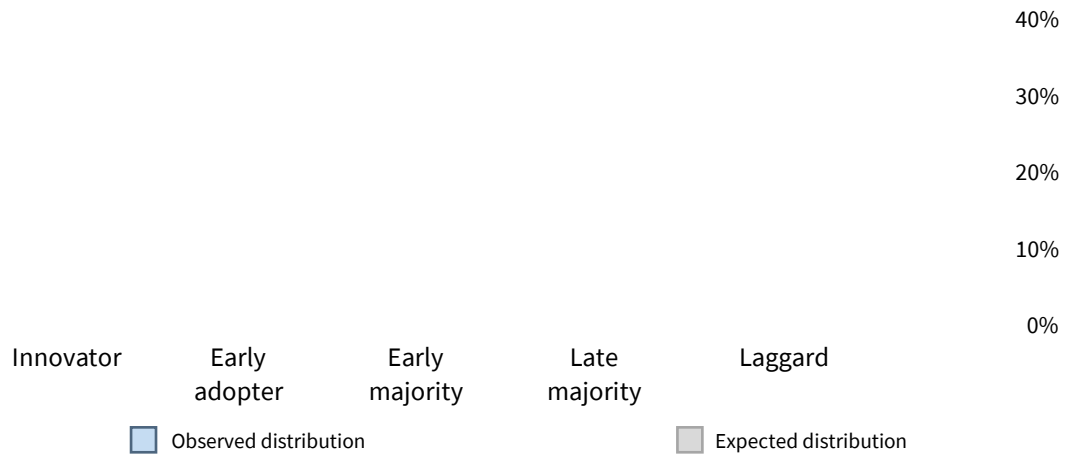


Figure 5.1. Observed and expected distributions among Cochrane Information Specialists



Several notable results emerged apart from the distribution of adopter personas.

First, in response to the question “what are your main considerations in selecting a study or reference management workflow / tool?”, a majority of every persona sub-group responded in a way that included the innovation characteristic of complexity (e.g., “ease of use”). Innovators were the exception; of the three identified innovator respondents, none mentioned complexity in their response to this question. Complexity was mentioned by 75% of early adopters, 66.6% of early majorities, 55.6% of late majorities, and 100% of laggards.

Second, users’ interactions with the RCT classifier correlated in the expected manner according to adopter persona. In describing their use of the RCT classifier, two out of the three identified innovators moved studies in bulk according to classifier results: that is, they trusted the results and did not double-check them. On the same question, 50% of respondents from the early adopter and early majority categories, respectively, behaved the same. Only 22.2% of late majority respondents used the classifier in this way, and neither of the two identified laggard respondents moved studies in bulk according to judgements from the RCT classifier.

Finally, when asked to rate the RCT classifier on a scale of 1 to 5 (with 5 as the positive rating; see Appendix D), the responses aligned with adopter persona. All innovator respondents rated the classifier 5 out of 5, while laggards rated the classifier as 3 out of 5. Results are shown in Figure 5.2.

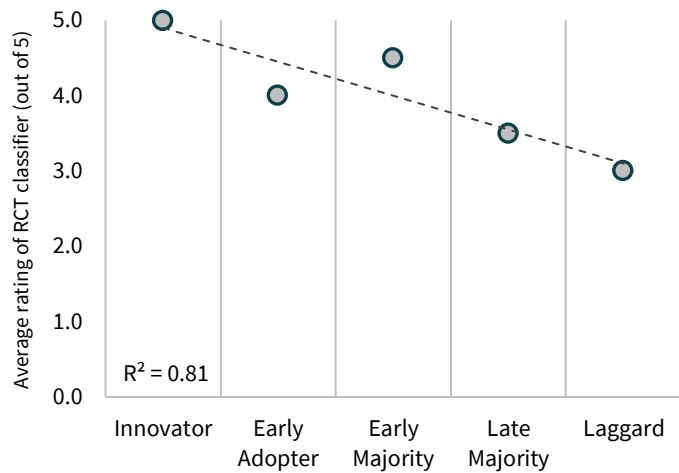


Figure 5.2. Average rating of RCT classifier versus adopter category

Phase 2

Four individuals were interviewed: one from each adopter category except for laggard, as none of the individuals who were coded as a laggard persona indicated availability for a follow-up interview.

Innovator

“I love a bit of new technology!”

The innovator interviewed demonstrated typical innovator persona characteristics. They repeatedly indicated their excitement about trying any new piece of technology, and that attitude extended into their interactions with the RCT classifier. Such enthusiasm was not dampened by their perception of a tool being complex; this aligned with the phase 1 survey results in that complexity was not mentioned as part of their main considerations in choosing a reference or study management tool. In fact, the participant indicated enjoying the process of experimentation with less obvious or intuitive features of new tools. The interviewee also indicated that they had shared their experience with the RCT classifier with several colleagues and believed they had likely persuaded a few to try it out, confirming the expectation that those earliest on the adopter curve may play a role in the next stages of innovation diffusion.

The innovator’s journey into deciding how and why to trust the RCT classifier was largely guided by institutional guidance. They used the threshold score

recommended by the Cochrane documentation and did not do any testing of their own. The interviewee indicated their initial attraction to the tool was driven by a large backlog of work, and they saw the classifier as an opportunity to finally get on top of that backlog. They also mentioned initially hearing about the classifier from specific individuals involved in its development, and consequently knew relatively more about its previous validation than they did about how it works. Cochrane's promotion of the tool was described by the interviewee as sufficient to feel that it had been validated and approved, and therefore was trustworthy.

The innovator interviewed emphasised their own documentation throughout their process, to a greater degree than others mentioned. In describing their process in initially adopting the classifier, they had not only shared information about the process with their Managing Editor and Coordinating Editor, but they had also uploaded documentation to their group website.

Overall, the innovator stated that they preferred more automation in their workflow. However, they also noted that while the RCT classifier used to display scores for each of the studies it processed, it no longer does this, and they would prefer that it display the scores once again. This request fitted in with the rest of their comments indicating a preference to be able to try out new things with new technologies, once again typical of the innovator category.

Early Adopter

“It was available, so I started using it.”

Of the four individuals interviewed, the early adopter was the only one who had experimented themselves with the threshold of the RCT classifier. They initially tried a very cautious cut off and trialled the tool with several systematic reviews before adopting in more generally in their standard workflow.

Like several of the other individuals interviewed, the early adopter indicated feeling overwhelmed by their workflow, and that their adoption decision was informed by this feeling. The early adopter felt that the classifier had not necessarily objectively reduced their workload but had subjectively reduced the pressure they

felt in keeping on top of the number of citations they were managing. They indicated that using the classifier “takes the pressure off of needing to catch everything.”

One of the more notable results from the interview with the early adopter was their rationale in trying the RCT classifier in the first place. While they indicated that it does help with workload, more prominent in their discussion was the simple fact that the tool was available. This fits in with what is expected from early adopters: they are already convinced of the need to change, and do not rely on others to prompt them to do so.

Another striking feature of the early adopter interview data was the discussion around audit trails. While it could be assumed that this is generally important to all information specialists, it was discussed much more prominently by the innovator and by the early adopter. As a secondary reason behind their adoption decision, the early adopter indicated that as the RCT classifier is signed off by Cochrane, it establishes a framework for its use, and they therefore incorporated it to their workflow. On a related note, they talked about the increasing methodological standards from the Cochrane community and felt that using the classifier would improve their transparency for methodological scrutiny.

Early Majority

“I still follow the instructions”

Like the innovator interviewed, the early majority individual mentioned several individuals within the Cochrane community who had influenced their decision to start using the classifier. Their approach in incorporating this influence into their user journey, however, differed in that they “followed instructions” of these individuals in order to start using the RCT classifier rather than learning through their own experimental use as the innovator had; they did not feel capable of figuring it out on their own and therefore relied on external guidance. This is in line with what is expected from the early majority persona, as they typically rely on success stories from peers to inform their adoption decision. Early adopters and early majorities alike are expected to find documentation helpful, and this held true with the early majority individual interviewed for this project. They indicated that they

continue to use the documentation for using the classifier to this day, despite now being a regular user of the tool.

In describing their experience in adopting the classifier, the early majority participant indicated a need to “feel safe” that the tool was working correctly. When asked to describe this in more detail, they indicated that they wanted to be sure it was not losing RCTs, but also indicated an openness to a workflow trade-off: “not that it’s perfect, but that it’s pretty good.”

In another similar decision point to the innovator interviewed, the early majority individual described being moved towards adoption of the classifier out of “desperation” in the face of their ever-growing workload. Continuing through their usage today, they only apply the RCT classifier to the results of specific databases according to the relatively high number of search results they tend to retrieve.

Overall, it was clear through the interview that the early majority individual relied heavily on the information they were provided from their peers, both in their understanding of how the classifier works, and in instructing their day-to-day usage of it.

Late Majority

“More online trainings please!”

Like others, the late majority individual interviewed mentioned first hearing about the classifier from individuals in the Cochrane community who had been involved in its development. Initially they thought it was the ‘holy grail’ of study management, but they have ended up mostly double-checking the results of the classifier. Unlike previous interviewees, the late majority individual did not indicate feeling overwhelmed by their previous workload and did not cite this as a reason for the transition. Instead, they described their current workflow as having the same set of records but in a different order. They indicated a desire to adopt more technology to improve their workflow but cited time constraints as holding them back in this regard; the time they perceive as required to adopt a given technology is greater than the time they currently have available. Though they have already adopted the

classifier, they expressed interest in more ongoing training sessions. They saw the lockdowns from the Coronavirus pandemic as a good opportunity to conduct regular online training sessions but were disappointed in the lack of these. They were unsure whether there might be new elements of the classifier or in CRS Web of which they were unaware, and they hoped that regular training sessions would help with this.

The late majority individual went into comparatively more detail compared to other participants in relation to the interface of tools they had tried. They described a tool similar to the classifier they had previously trialled as “clunky” and “a bit fiddly” and went into further detail about the lengthy process of opening each record to retrieve results. Continuing on this theme, they described their decision to rate the RCT classifier as a 4 out of 5 as informed by the annoyance of having to open separate tabs for each result, resulting in many “extra clicks”. This was the only discussion from any participant that provided data for the internal variability of situational trust in the trust framework.

The late majority individual described their author team as “quite sceptical” of adopting the RCT classifier. This differed from other participants, who generally indicated indifference from their author teams.

Discussion

In addressing research question 2.1, results of the survey indicate that the distribution of adopter categories among Cochrane Information Specialists maps closely to the distribution predicted by Rogers’ Diffusion of Innovations framework. While there were slightly more innovators than predicted, and slightly fewer laggards, the overall trend of the distribution indicates that the adoption curve and adopter categories fit into the context of CISs use of the RCT classifier well. The data from phase 1 therefore indicate that the Diffusion of Innovations framework is applicable in the context of Cochrane Information Specialists’ adoption of the RCT classifier.

This result was strengthened by the results of the unstructured interviews. Each of the individuals interviewed displayed characteristics of their coded category from the phase 1 results, especially with respect to their initial adoption decision.

More specifically, the innovator and early adopter indicated an immediate willingness, if not eagerness, to switch to a new tool and workflow. This aligns with the description of those adopter personas as being already aware of the need to change practices [8]. The early majority and the late majority individuals instead relied on information from their peers in order to inform this decision, once again aligning with the expected and predicted behaviour of their respective adopter categories.

In relation to research question 2.2, data showed trust in the RCT classifier is inversely correlated to adopter persona: those earlier on the adoption curve indicated higher levels of trust in the classifier than did those later on the curve. Earlier adopters were less likely to double-check the results of the classifier and more likely to bulk-approve automated decisions.

When considering the results within the levels of automation framework, those earlier on the adopter curve tended to use the RCT classifier in a way that fell higher on the levels of automation framework [9, 10]. Bulk approval of RCT classifier decisions fits into a level 5 automation (computer executes a suggestion if the human approves; see Table 3.1), whereas double-checking of results reduces the RCT classifier to a level 2 automation (computer provides a full set of decision alternatives). These results show that not only is the Diffusion of Innovations framework applicable in this context, but it is also useful in predicting adopter behaviour, and in drawing novel insights to user behaviour when combined with the levels of automation framework.

Negative data produced in this study should be noted here as well. The survey requested information from respondents about their years of career experience, both as a Cochrane Information Specialist and in the field at large. It might be reasonably assumed that more experience would correlate with more expertise, and according to the selected trust framework, greater subject matter expertise is a strong influence on internally variable trust. However, in this study I observed no relationship between years of experience, trust tendencies, and adopter persona. Though it was expected to play a strong role in my results (as noted in Chapter 3: Trust in automation: Situational trust - Internal variability), internally

variable situational trust and subject matter expertise does not appear to play a strong role in adoption decisions among CISs.

Finally, in relation to research question 2.3, interview participants indicated differences in their adoption processes, in their ongoing use of the RCT classifier, and in the factors affecting their trust in the tool. The early adopter was the only participant to indicate that they had run their own validation tests. The innovator experimented with the RCT classifier, but not specifically with the threshold nor with piloting the tool before fully adopting it, as the early adopter had. The innovator did, however, indicate a preference for being able to see the underlying scores assigned by the classifier. In contrast, both the early majority and late majority interview participants indicated a high level of reliance on the institutional guidance in relation to their initial adoption and their continued use of the RCT classifier. The late majority individual appeared to have a continuing reliance on this guidance, as indicated by their request for regular online training sessions, whereas the early majority individual spoke more about the documentation to assist in using the tool. In addition, the late majority individual was the only interview participant to speak significantly about the tool's user interface. Taken together, the results of this study show that user interface and experience is more influential on trust for later adopter personas. Earlier adopters instead place relatively more importance on the ability to inspect results; in the case of the early adopter this also included direct experimentation with the technical aspects of the tool.

This result connects into some of the results found in the previous chapter. Guideline developers repeatedly raised the importance of the transparency of any applied machine learning. Recall that there were two variations on this idea: *transparency as accountability (compatibility)* and *personal need for double-checking (observability)*. Though this study was primarily focused on the adopter personas of Diffusion of Innovations framework, the themes found by applying the innovation characteristics in the previous chapter's study have appeared again in this study, and therefore warrant further attention. Innovators and early adopter participants appear to focus relatively more on the *personal need for double-checking*; they have either experimented with or examined the threshold scores generated by the classifier themselves. In contrast, the early majority and late

majority participants cited the validation others had performed, thus aligning more closely with the theme of *transparency as accountability*. Given this finding, two questions for future research arise: what is the distribution of adopter personas among guideline developers, and do adopter personas in guideline developers and in CISs show the same behavioural tendencies when using an automation tool? Overall, the results of the survey combined with the results from the interviews showed that users on the first half of the adoption curve – innovators, early adopters, and early majorities – entirely trust in the results of the classifier, while late majorities and laggards prefer to double-check its results.

The evidence presented in this chapter suggests that all adopter personas rely on external variability in situational trust to inform their trust in automation. All adopter categories considered institutional guidance (e.g., from formal Cochrane documentation and presentations) in their adoption decision, and early majority and late majority individuals further used this guidance in shaping their ongoing interactions with the classifier. In an additional layer of externally variable trust in automation, individuals later on the adoption curve placed more importance on their user experience than did earlier adopters. Innovators and early adopters did also cite externally variable situational trust, but also demonstrated more dispositional trust in their adoption decisions and in their trust of the RCT classifier.

Implications for practice

Initial uptake

An overwhelming workload and/or backlog was cited in three of the four interviews conducted, with the single exception being the late majority interview. Considering this result, it should be expected that larger workloads and/or stricter deadlines might result in stronger motivation for uptake of potentially time-saving tools, including automation. Coinciding with the time of writing this thesis, in April 2021 a funder of Cochrane noted the tendency of review groups to deliver projects late [11]. This may contribute to an understanding of the primary importance of review production efficiency in the Cochrane community, and it will be interesting to see if this affects attitudes to or uptake of automation. The results of this study indicate a climate of urgency could be an opportune moment to advocate for further trialling among review groups of the automation tools that are already available.

In addition to overwhelming workloads informing the initial trialling of automation tools, data collected from the interviews indicated the importance of institutional guidance (externally variable situational trust). This influenced all personas' adoption decisions, and furthermore continued to influence the ongoing user experience for later adopters. Cochrane and organisations that wish to pursue methodological innovation should therefore invest in institutional level evaluation, approval pathways and communications that continue even after the initial approval of a tool.

Complexity becomes increasingly important over time

All adopter personas, except for innovators, indicated a high prioritisation of the complexity innovation characteristic in informing their technology selection. Evidence from this study therefore suggests that complexity is the most heavily weighted innovation characteristic among CISs, but that innovators are an exception to this observation. This result could be highly useful in designing not only automation tools for information specialist users, but also for optimising communications. In other words, these results indicate that a majority of information specialists are not going to use something that is not easy to use, even if other innovation characteristics are positive (e.g., relative advantage in timesaving). The late majority participant reinforced this observation with an explanatory barrier to uptake; a lack of time to learn a new technology was cited as a reason for continuing a known workflow over a new one, even if it offered a relative advantage. Similarly, they also provided more detail on the relative difficulty of their user experience both with the RCT classifier and with other tools they had trialled; this point further reinforces that later adopters are more influenced by user experience than are earlier ones.

Automation advocates might use this result to inform research and development prioritisation. In early stages, while innovators are being targeted for the initial rollout of a tool, user experience can be temporarily sacrificed in favour of the tool's performance, assuming limited resources available for technology development. Further, given the innovator participant indicated enjoyment of playing with the technical details of the classifier, other innovators might be persuaded to adopt by including this ability in the initial version of a tool, in contradiction to the intuitive tendency to simplify a tool as much as possible. This sacrifice and technical

detail become more and more risky over time, however, according to these results; late majority and laggards, already resistant to change by disposition, are quite unwilling to take up use of a tool which is not simple to use.

The observed result becomes even more insightful when contrasted against the results of Chapter 4: where guideline developers placed cultural expectations (compatibility) at the forefront, information specialists instead raised complexity. The expectation that different populations involved at different points of the evidence pipeline would return different results from a framework analysis using Diffusion of Innovations proved true, at least in this case. The different results could also be due to the different current stages on the adoption curve of the respective populations. It is nevertheless clear that application of these two frameworks in combination yields insightful results. Future work will benefit from continuing to do so, filling in the remaining gaps of stakeholder populations in addition to CISs and guideline developers.

Study limitations

While 24 respondents to the initial survey delivered some meaningful results, a higher proportion of respondents of the pool of potential participants would have provided more confidence that my results are representative of the full CIS population. The response rate indicates approximately half of the Cochrane Review Groups' Information Specialists completed the survey, and it is difficult to estimate what the distribution of adopter personas among non-respondents might be. The results might be representative, and the distribution does indeed map closely to the Diffusion of Innovations framework, or the remaining non-respondents might fall disproportionately into particular categories (most likely laggards or non-adopters, given the topic of the survey).

Most likely as a consequence of the final sample size of the survey, the availability of potential participants for the second phase of this project was limited. Specifically, no laggards indicated availability for follow-up in the final segment of the survey, and therefore no interview was conducted with an individual who had been coded into the laggard adoption persona. This study therefore is unable to draw any strong conclusions in relation to how and why laggards do or do not trust an

automation tool. Furthermore, the practical decision to interview only one participant of each persona category limits the generalisability of this study's findings. While the qualitative results provided by participants in their interviews provided rich individual data, the limited sample means that it cannot be assumed that such results would translate across additional contexts. Further research in more depth is warranted to determine the generalisability of these findings. Given that the results of this study did confirm the applicability of the adopter categories framework, such research should continue to apply this framework in order to confirm and expand the conclusions presented here.

Conclusion

The results of this survey demonstrated that the Cochrane Information Specialists' distribution of Diffusion of Innovations adopter personas maps quite closely to the predicted distribution, indicating a good fit of this framework to this context. Adopter personas' behaviour towards automation correlated well with their predicted characteristics. This behaviour, examined through the Hoff and Bashir trust framework, showed strong influences of externally variable situational trust in all personas, and dispositional trust in innovators and early adopters. Later adopter categories maintained this reliance on situational trust even after adopting the RCT classifier, and also identified user experience (dynamic learned trust) as influential on their interactions with the tool. Of the Diffusion of Innovations innovation characteristics, complexity was identified as significantly influential for CISs in their selection of workflow management tools. Organisations which wish to advocate for adoption of automation for health evidence synthesis should initially focus on technological abilities to bring innovators and early adopters on board, and over time shift focus to simplification of the user experience and to providing ongoing training and support.

Chapter references

1. Higgins, J., et al. Methodological Expectations of Cochrane Intervention Reviews (MECIR). *Cochrane: London* 2021.
2. O'Mara-Eves, A., et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 2015; 4(1):5.
3. Thomas, J., et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *Journal of Clinical Epidemiology*.
4. 2019. <https://github.com/alan-turing-institute/DSSG19-Cochrane-PUBLIC> (accessed 15 August 2021).
5. Kelley, K., et al. Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care* 2003; 15(3):261-266.
6. Sharma, A., et al. A consensus-based Checklist for Reporting of Survey Studies (CROSS). *Journal of general internal medicine* 2021:1-9.
7. Ritchie, J., et al. *Qualitative research practice: A guide for social science students and researchers*: sage; 2013.
8. Rogers, E.M. *Diffusion of Innovations*. 5th ed: Simon and Schuster; 2003.
9. Sheridan, T.B. *Telerobotics, automation, and human supervisory control*: MIT press; 1992.
10. Sheridan, T.B. and W.L. Verplank, *Human and computer control of undersea teleoperators*. 1978, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
11. Stein, K. Cochrane and NIHR. *Virtually Cochrane*; 2021; Online; 2021.

Chapter 6. Validity

A clustered non-inferiority randomised trial examining the effect of combined human effort and automation on Risk of Bias assessments

Chapter overview

This chapter presents the results of a randomised trial investigating the effect of integrating RobotReviewer into an automation-augmented Risk of Bias workflow. The results of the trial will be contextualised in the available literature, and in particular will be contrasted with prior trials of RobotReviewer. The analytical frameworks selected for this PhD will then be used to interpret the conclusions of the trial. The chapter concludes with a discussion of the trial's weaknesses and recommendations for practice.

Introduction

The first two research chapters of this thesis presented qualitative studies investigating the first theme of my PhD: the adoption of automation in health evidence synthesis. I will now shift to the second theme: the effectiveness of automation for health evidence synthesis. As outlined in the introduction and detailed in the literature review of Chapter 2, much of the existing automation literature focuses on screening tasks. Study quality evaluation and data synthesis, though some of the most time-consuming stages of a systematic review, have relatively less automation evidence currently available. Furthermore, the majority of existing research focuses on efficacy – performance under ideal and controlled circumstances – rather than on effectiveness – performance under ‘real-world’ conditions. While efficacy trials have high internal validity and are important in identifying interventions with observable effects, they can also overestimate an intervention’s effect compared to implementation in practice [1]. Effectiveness trials, in contrast, are less standardised and can account for other factors which may moderate an intervention’s effect. With the trial presented in this chapter, I improve upon the evidence base on both fronts: this trial examines real-world effectiveness of a data synthesis automation tool, namely RobotReviewer [2].

The RobotReviewer tool is an open-access platform which partially automates several elements of data extraction, including Risk of Bias (RoB) assessments that use the Cochrane RoB template, using machine learning (ML) and natural language processing (NLP) [3]; an example assessment is shown in Figure **6.1**.

trial	design	Random sequence generation	Allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment
Wouters H, 2017	RCT	+	+	?	+
Moskowitz JT, 2017	RCT	+	+	?	?
Somerville V, 2019	RCT	+	+	+	+
Conroy MB, 2019	RCT	+	+	?	+

Figure 6.1. An example RobotReviewer assessment

The literature provides some excellent examples of previous research into RobotReviewer’s performance, strengths, and weaknesses. In a 2016 study it was shown that RoB judgements produced by RobotReviewer were only modestly inferior in quality to human-produced judgements [4]. In this study, 12,808 PDFs were annotated and used to train a ML model. These were compared against assessments published previously in the Cochrane Database of Systematic Reviews (CDSR) by a panel of 20 blinded experts who rated judgement accuracy and the relevance of supporting annotations. Judgements provided by RobotReviewer were found to be less accurate, but the absolute difference from judgements in the CDSR were generally less than 10%.

Two previous publications examining RobotReviewer found slightly different results. A 2018 publication used a cross-sectional evaluation to evaluate reliability of RobotReviewer using Cohen’s Kappa coefficient and comparing with human-produced assessments [5]. This evaluation also examined differences by domain and outcome type. RobotReviewer was used to evaluate 1,180 studies, and these RoB assessments were directly compared against assessments completed by human reviewers. They found that RobotReviewer reliability compared to humans was similar for most domains, and superior for selected domains. This is in contrast to a 2020 paper [6] which compared assessments between RobotReviewer and human reviewers for 372 studies. The 2020 study found that accuracy was highly variable across assessment domains, and the study authors therefore concluded that RobotReviewer should not replace evaluations performed by human experts, which aligns with the developers’ guidance.

Studies with similar designs continue to be conducted and published as of the writing of this thesis. Hirt et al (2021) compared the performance of RobotReviewer against human reviewers again, and similarly found variable quality across the domains assessed by the tool [7]. The observed agreement between the automated RoB reports and the human reviewers' assessments ranged from 50% for the 'blinding of outcome assessors' domain to 87% for the 'blinding of participants and personnel' domain. Like Armijo-Olivo et al (2020), the authors here concluded that while RobotReviewer might be helpful, "human reviewer should supervise the semi-automated process."

Despite multiple instances in the literature which concluded that RobotReviewer should be used under human supervision, and direct recommendations from the developers, only one previous publication is available examining this workflow, and is authored by the RobotReviewer developers themselves. Soboczenski et al (2019) [8] recruited 41 participants and assigned four RCTs to each to assess for Risk of Bias. Of these four, two were assessed in a fully manual manner, and the other two in a semi-automated manner with RobotReviewer suggestions. Rather than strictly examining the accuracy of the final assessment, the primary outcome of this study was the time taken to complete RoB assessments. The study showed that the semi-automated workflow reduced time spent on Risk of Bias assessments by 25% (Figure 6.2).

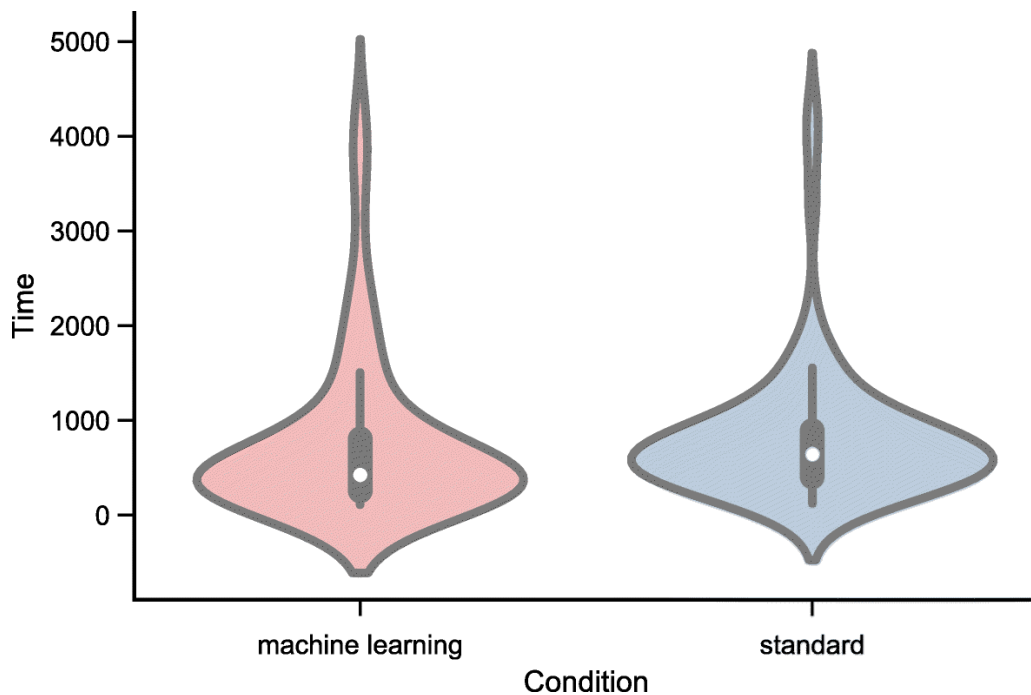


Figure 6.2. Time in seconds to complete Risk of Bias using machine learning versus standard (fully manual), from Soboczenski et al (2019)

With only a single evaluation of RobotReviewer using a semi-automated method rather than a direct comparison of human-only versus fully automated results, the literature largely leaves open the question of the effect of a semi-automated Risk of Bias workflow using RobotReviewer. While these previous evaluations of RobotReviewer accuracy are useful, the lack of published evidence investigating the effect of combining this automation system with human processes is a significant shortcoming. As a semi-automated method is likely to be the manner in which automation is incorporated into evidence synthesis initially, and moreover is the intention of the team creating RobotReviewer, further evidence is needed on the effects of RobotReviewer when used in combination with human systematic reviewers [9].

Finally, all of this previous evidence on RobotReviewer takes a constructed and controlled approach; specific reviewers are given specific and selected RCTs, and their assessments directly compared to those produced by RobotReviewer. Literature with a more ‘real-world’ focus, in which a larger number of reviewers is given a larger number of trials to assess, which have not been pre-selected by the research team, would be useful.

As previously discussed in detail, adoption of automation in health evidence production has been low [10]. In Chapter 4, guideline developers, key gatekeepers in

translating health evidence to practice, expressed the need for robust evidence in this area in order to inform automation adoption decisions, and moreover expressed a preference for augmenting human resources with automation tools over replacement [11]. Current literature falls short on both needs. First, it has focused on laboratory settings (i.e., on the efficacy of automation tools rather than the effectiveness). Second, it has mostly offered simplistic and direct comparisons of automation performance against human performance rather than evaluated the performance of a combination of human and automation effort.

Hesitancy to trust automation in systematic reviews has also contributed to the slow adoption of automation [12]. The well-established foundation of trust in systematic review results is crucial to the evidence ecosystem and to knowledge translation, and thus new methodologies may result in concern around the quality of the results. O'Connor et al (2019) conclude that “automation... must not be perceived as an erosion of current practice standards.” The foundation of trust in evidence synthesis methods must now be renewed to include automation to encourage its adoption, and the evidence necessary is currently lacking.

To address this gap in the currently available literature, this study sought to determine whether, in the context of ‘real-world’ systematic reviews, Risk of Bias (RoB) assessments conducted with machine learning assistance from RobotReviewer were non-inferior in accuracy and in person-time to assessments conducted with human effort alone.

Methods

Given that this trial sought to mimic real-world systematic review conditions as closely as possible, several challenges arose: The design had to account for two distinct individuals conducting assessments within one review, as well as the similarity of studies within a review. A simpler study design was previously considered in which one reviewer was consistently assigned to the RobotReviewer assisted arm, while the other was consistently assigned to the control arm. This presented a number issues in design. First, after the initial assessment, both reviewers would be un-blinded to their allocation. Second, consistent interaction with the intervention also increases the opportunity of a learning effect; with

relatively more interactions with the intervention (i.e., RobotReviewer), there are more opportunities for a reviewer to learn how it performs and pre-judge its assessments. Third, I judged that this also introduced a higher risk that the consensus reviewer – whose assessment was crucial to the primary outcome of accuracy – would be able to determine intervention allocation. Like the individual reviewers knowing their allocation after the initial assessment, the consensus reviewer was likely to learn quickly which reviewer had been assigned to which arm, and therefore potentially have bias before ever examining their individual assessment for that study. That is, a design in which reviewers were randomised rather than studies risked un-blinding of an outcome assessor (the consensus reviewer). Finally, allocation by study rather than by reviewer lessened the impact of an individual reviewer’s ‘skill’ in RoB assessments; if a senior reviewer, or a reviewer otherwise naturally more likely to complete an assessment accurately, were assigned to the RobotReviewer-assisted arm, this would skew the results in favour of the intervention. A paired trial design was therefore selected instead to minimise these risks.

Design

A two-arm, clustered, randomised, single-blind, parallel group, non-inferiority trial was conducted between February 2018 and March 2020. A clustered design was selected to account for similarities of studies and reviewers within each review. Teams of individuals conducting systematic reviews (hereafter ‘reviewers’ and ‘review’, respectively) using Covidence, an online systematic review platform [13], were recruited. Each review acted as a cluster, and each study served as the unit of randomisation.

This trial is reported in line with the Consolidated Standards of Reporting Trials (CONSORT) checklist (included as Appendix E).

Objectives

The aim of the trial was to assess the effectiveness of machine learning assistance for systematic review RoB assessments. This was assessed using two co-primary study objectives:

RQ3.1) Is the accuracy of RobotReviewer-assisted RoB assessments non-inferior to human-only RoB assessments?

RQ3.2) Is the person-time required for RobotReviewer-assisted RoB assessments less than the person-time required for human-only RoB assessments?

Participants

Eligibility

Review teams were eligible for inclusion if their reviews met the following criteria:

1. The systematic review was health related.
2. Risk of Bias assessment was planned to be undertaken in Covidence using the Cochrane Risk of Bias 1.0 template.
3. Risk of Bias assessment would be completed by two independent reviewers, and a consensus assessment by a third reviewer.
4. The systematic review could contribute a minimum of four controlled trials to the analysis.
5. The review authors had not used RobotReviewer previously.

Within eligible reviews, individual studies were eligible for trial inclusion if:

1. A readable PDF could be retrieved.
2. RobotReviewer was able to complete an assessment of the study's PDF.

Recruitment

Recruitment emails were sent to Covidence users who:

1. Were leading a health-related systematic review.
2. Had elected to use the Cochrane RoB 1.0 template.
3. Were leading a review with at least four included studies for which RoB assessment had not been initiated.

The list of potentially eligible users was generated in collaboration with the engineering team from Covidence. Specifically, they assisted in writing a SQL query of the Covidence database to identify users who (1) had at least three collaborators

on their review, (2) had at least one study in the included stage with a blank Cochrane RoB template, (3) had not yet completed full text review, and (4) were active within the previous month on Covidence. This query was originally slightly narrower, seeking users who had at least ten studies with blank Cochrane RoB templates in their included list; a second iteration targeted users with at least five studies meeting this criterion. However, after the low recruitment rate became clear (discussed in the next paragraph), the SQL query was broadened to gain more potential participants. Twitter was also used as a potential recruitment tool to make use of personal networks; this did not result in any successfully recruited participants.

Recruitment emails were sent fortnightly for the duration of the trial. Recipients of these emails were documented; if an individual had previously been contacted and they appeared on a subsequent SQL query result as a potential participant, they were not re-contacted for four months.

Individual video chat sessions were offered to respondents to clarify trial details. Where these were not arranged, all details in relation to trial participation were established via email. The research team was available to participants throughout trial participation if they had any questions or concerns. Potential participants were offered a free Covidence review (value US\$240) and a US\$150 Amazon gift card as compensation for trial participation.

Risk of Bias assessment

The established approach to Risk of Bias assessment is for two reviewers to complete separate, blinded assessments, followed by adjudication by a third reviewer to resolve any discrepancies between the two initial assessments [14]. Various assessment tools and templates are available which describe criteria for quality evaluation [15]. Cochrane Risk of Bias 1.0 is a domain-based evaluation tool and uses seven domains to assess study quality [16]. RobotReviewer performs assessments for four of these domains: Sequence Generation, Allocation Concealment, Blinding of Participants and Personnel, and Blinding of Outcome Assessment. The remaining three domains are not assessed by RobotReviewer partly due to current technological constraints; Incomplete Outcome Data and Selective

Outcome Reporting both require calculations and consultation of a review protocol current outside of RobotReviewer’s capabilities, while Other Sources of Bias is by its design a domain to record non-standardised human judgements on study bias.

When completing a systematic review using Covidence, users may select to use customised assessment domains or to use a pre-formatted Cochrane Risk of Bias 1.0 template. Users may then select a high, low, or unclear risk of bias judgement, record a justification for their decision, usually by recording a relevant section of the study text, and provide additional supporting comments. A blank Covidence Cochrane Risk of Bias 1.0 assessment form is shown in Figure 6.3.

The screenshot shows the 'Cochrane Risk of Bias' form in Covidence. At the top right is a 'Complete' button. Below the title is a 'Show all' button. The main content is divided into two columns. The left column has a purple header for 'SEQUENCE GENERATION' with a dropdown arrow. Below it is a text box with the instruction: 'Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups.' Below the text box is a 'Make judgement' button. Underneath are several grey boxes with right-pointing arrows, each representing a bias domain: 'ALLOCATION CONCEALMENT', 'BLINDING OF PARTICIPANTS AND PERSONNEL', 'BLINDING OF OUTCOME ASSESSORS', 'INCOMPLETE OUTCOME DATA', 'SELECTIVE OUTCOME REPORTING', and 'OTHER SOURCES OF BIAS'. The right column has a header 'Sequence Generation' with a 'CANCEL' button. Below this is the 'RISK OF BIAS' section with three buttons: 'High', 'Low', and 'Unclear'. The 'ANNOTATIONS' section shows 'No annotations'. The 'JUDGEMENT COMMENT' section has a large empty text area. At the bottom right of the form is a 'Save' button.

Figure 6.3. Blank Cochrane Risk of Bias template in Covidence

Trial Procedure

Random sequence generation

For reviews that met the inclusion criteria, I assigned participating reviewers alphabetically to be defined as either Reviewer 1 or as Reviewer 2 for the purposes of intervention allocation. I then used a computer random number generator [17] and simple randomisation to assign the review’s included studies in a 1:1 ratio to have either Reviewer 1 or Reviewer 2 receive RobotReviewer assistance.

RobotReviewer assistance

Study PDFs, previously uploaded to Covidence by review authors, were downloaded from Covidence and uploaded to RobotReviewer for assessment. Judgements and supporting text generated by RobotReviewer were entered into the Covidence RoB form for the reviewer randomised to the intervention arm.

Allocation concealment

Individual reviewers were unaware of each study's allocation until they opened their individual RoB form. At this time, they were presented either with a blank form if the reviewer was allocated to the comparison arm for that study, or a form pre-populated with RobotReviewer suggestions if the reviewer was allocated to the intervention arm for that study.

Blinding of outcome assessment

The consensus reviewer was blinded to individual reviewer allocation on each study. Reviewers were instructed not to discuss study allocation with each other, nor with the individual completing consensus. Reviewer names and study IDs were replaced with numeric identifiers during data analysis.

Intervention

In the intervention arm, Risk of Bias forms in Covidence were pre-populated with suggested judgements and annotations generated by RobotReviewer. Reviewers were advised that they were free to retain all, none, or some RobotReviewer suggestions as they saw appropriate. An example of how a pre-populated RoB form might appear in an intervention-arm study is shown in Figure 6.4.

The screenshot displays the Cochrane Risk of Bias form. On the left, the 'SEQUENCE GENERATION' domain is highlighted in purple and marked as 'LOW'. It contains three annotations: 1) 'A total of 222 (46.1% of those contacted) consented to participate in the trial and were randomly allocated into the intervention or control group.' 2) 'The random numbers were generated and packaged into sealed envelopes by a researcher external to the study.' 3) 'A set of 250 random numbers was generated using Microsoft Excel and the Statistical Package for Interactive Data Analysis (SPIDA) was used to group these numbers by blocks of four to ensure a balanced sample size across both study groups.' Below the annotations is a 'Make judgement' button. Other domains listed include 'ALLOCATION CONCEALMENT' (LOW), 'BLINDING OF PARTICIPANTS AND PERSONNEL' (HIGH), 'BLINDING OF OUTCOME ASSESSORS' (HIGH), 'INCOMPLETE OUTCOME DATA', 'SELECTIVE OUTCOME REPORTING', and 'OTHER SOURCES OF BIAS'. On the right, the 'RISK OF BIAS' summary shows 'Low' selected, with three annotations and a 'JUDGEMENT COMMENT' text area. 'Save' and 'Delete' buttons are at the bottom right.

Figure 6.4. Risk of Bias form pre-populated with RobotReviewer assistance

In the comparison arm, empty individual Cochrane RoB 1.0 forms (Figure 6.3) were presented to participants and completed without RobotReviewer assistance.

Figure 6.5 provides an overview of the trial study design.

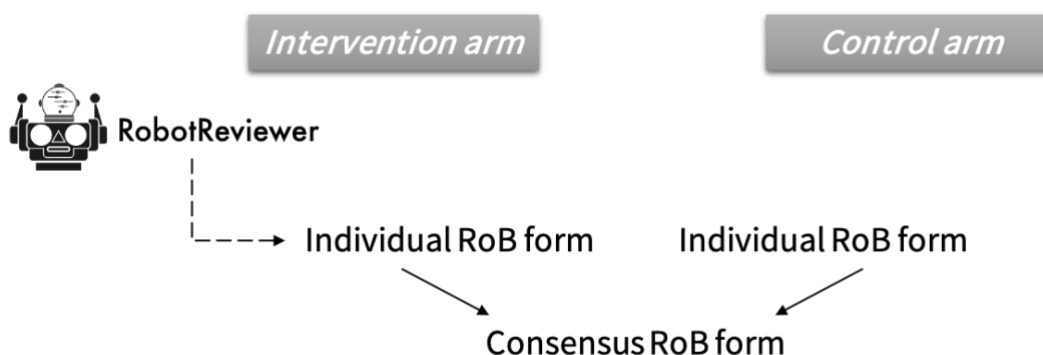


Figure 6.5. Trial study design

Outcome measures

The trial had two co-primary outcomes: accuracy of individual assessments, and person-time required to complete individual assessments.

Accuracy

Accuracy was defined as agreement between an individual RoB assessment and the consensus RoB assessment completed by a third reviewer. Covidence records judgements from both individual forms and the consensus form. Once the consensus assessment was marked as complete, all included studies' judgements were downloaded. Because RobotReviewer combines 'high' and 'unclear' judgements, these were treated as the same for the purposes of assessing accuracy.

Overall accuracy was assessed as a primary outcome. Accuracy at the domain level was assessed as a secondary outcome.

Person-time

Person-time was measured as the total time spent on the RoB form for each study, from beginning an assessment until it was marked as complete. Covidence records all actions with a timestamp, providing an approximation of session time on the RoB form. The total time spent in all sessions, excluding any idling time, which was defined as a period of inaction greater than 20 minutes, was used in the analysis.

Statistical analysis

Statistical methods and analysis for this trial were designed and conducted by Dr Joanne McKenzie. Analyses were conducted in Stata [18].

Sample size calculation

The sample size was calculated prior to trial commencement for the accuracy co-primary outcome. The potential accuracy outcomes of a single RoB assessment can be described in a 2x2 table (Table 6.1).

Table 6.1. Potential accuracy outcomes of an individual Risk of Bias assessment

		RobotReviewer-assisted		
		Correct (1)	Incorrect (0)	
Human-only	Correct (1)	p11	p10	p1+
	Incorrect (0)	p01	p00	
		p+1		

The null and alternative hypotheses are as follows:

- H_0 : The accuracy of RobotReviewer-assisted RoB assessments are inferior to human-only RoB assessments
- H_1 : The accuracy of RobotReviewer-assisted RoB assessments are non-inferior to human-only RoB assessments

Previous studies found 78.3% accuracy of two reviewers measured against a previously published expert RoB judgement, while the accuracy of an entirely automated assessment was 71.0% [4]. In this slightly altered scenario of combining human and ML-generated assessments, we conservatively assumed the higher rate of approximately 30% inaccuracy. A type I error rate of 2.5% [19] was used, and an intra-review correlation of 0.10 and an average of 10 studies per review were assumed.

Based on these assumptions, a series of potential scenarios were modelled to detect a risk difference between arms; detection of a smaller risk difference requires a greater sample size to have sufficient power for analysis. While aiming to set the non-inferiority limit as narrow as possible to maximise the strength of the trial results, it was also necessary to ensure that a practical and achievable sample size was used. I therefore selected a non-inferiority limit -0.10 , which required 26 reviews to detect a difference between p_{01} and p_{10} with 90% power.

With a non-inferiority limit set at -0.1 , this is equivalent to testing:

- $H_0: p_{+1} - p_{1+} \leq -0.1$
- $H_1: p_{+1} - p_{1+} > -0.1$

Figure 6.8 depicts the regions of inferiority and non-inferiority based on the non-inferiority limit.

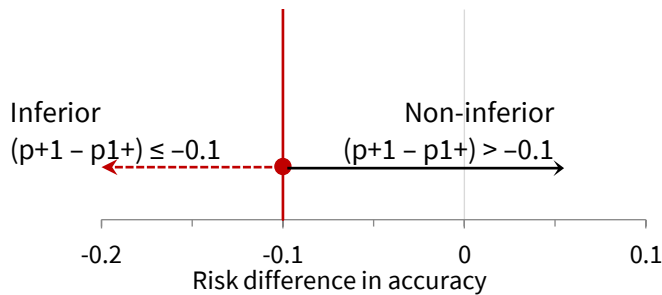


Figure 6.6. Regions of inferiority and non-inferiority

Accuracy

Differences in the marginal proportions of studies with accurate RoB assessments with and without RobotReviewer assistance were calculated across all RoB domains for the primary accuracy outcome, and for each domain individually for the secondary accuracy outcome. An adjusted McNemar test, which adjusts the variance of the difference to account for the clustered and matched-paired design of this trial [20], was used to calculate 95% confidence intervals for accuracy.

Clustering was at the level of the review, with each cluster including multiple studies and also including four RoB judgements for each study in calculation of the primary outcome (i.e. overall accuracy). The adjusted McNemar statistic and its associated Wald confidence interval have been shown to perform well for a small number of clusters [21].

Person-time

A mean difference between RobotReviewer-assisted person-time and human-only person-time required to complete an individual assessment was calculated. The variance of the difference was adjusted for clustering of observations within the same review, arising from the same reviewers undertaking multiple assessments within each review (cluster) it was also then adjusted to account for the effect of clustering.

Ethical considerations

The study protocol was registered with the Monash University Human Research Ethics Committee (MUHREC) for approval as Project 11256, and additionally received approved according to UCL Institute of Education Guidelines for postgraduate research. Funding was provided by a joint PhD studentship from University College London and Monash University.

There was no risk associated with participants' safety from participation in this study. No names or identifying information are included in any published data, including presentations and publications. Names of participants and their systematic reviews were stored in a separate log on a password-protected file, stored securely on a University College London network server. This identifying information was available only to research team members.

Participants were free to withdraw at any time during the trial.

Results

Under the filters for recruitment emails described in the methods, an average of 68.32 potential participants were contacted with each recruitment round, for a total of 1,708 review teams directly contacted. On average, 4.2% responded to the initial email; some of these forwarded information about the trial to colleagues, resulting in 86 potential leads. Of these, 27 were immediately lost to follow-up and were never assessed for eligibility. For reviews that were assessed for eligibility in the trial, 19 were eventually deemed ineligible after discussion of the review method requirements with the potential participants; a further six declined to participate after learning more details about the trial, and 19 were lost to follow-up after their eligibility assessment.

Fifteen review teams were recruited between February 2018 and March 2020, all of which received RobotReviewer assistance. The decision in March 2020 to close the trial stemmed two practical decisions: first, to ensure analysis from this trial could be included in this PhD thesis; second, to exclude the possibility of any longitudinal effects (e.g., changes in practice, uptake of a new Cochrane RoB template) affecting the findings. Eight reviews had not completed RoB assessment as

of May 2020; seven of these indicated these assessments had been indefinitely postponed due to other work commitments or lack of funding, and one changed the search criteria resulting in the exclusion of all previously included eligible studies. Topic areas examined in these excluded reviews were: airways, brain injury, exercise, heart, occupational health, pain, and stroke. The remaining seven reviews were included in the analysis, and included 145 studies, 290 individual assessment forms, and 1160 Risk of Bias judgements across the four Cochrane Risk of Bias domains for which RobotReviewer provides assistance. Topic areas examined in these reviews were: airways, anaesthesia, heart, hepato-biliary, musculoskeletal, and public health.

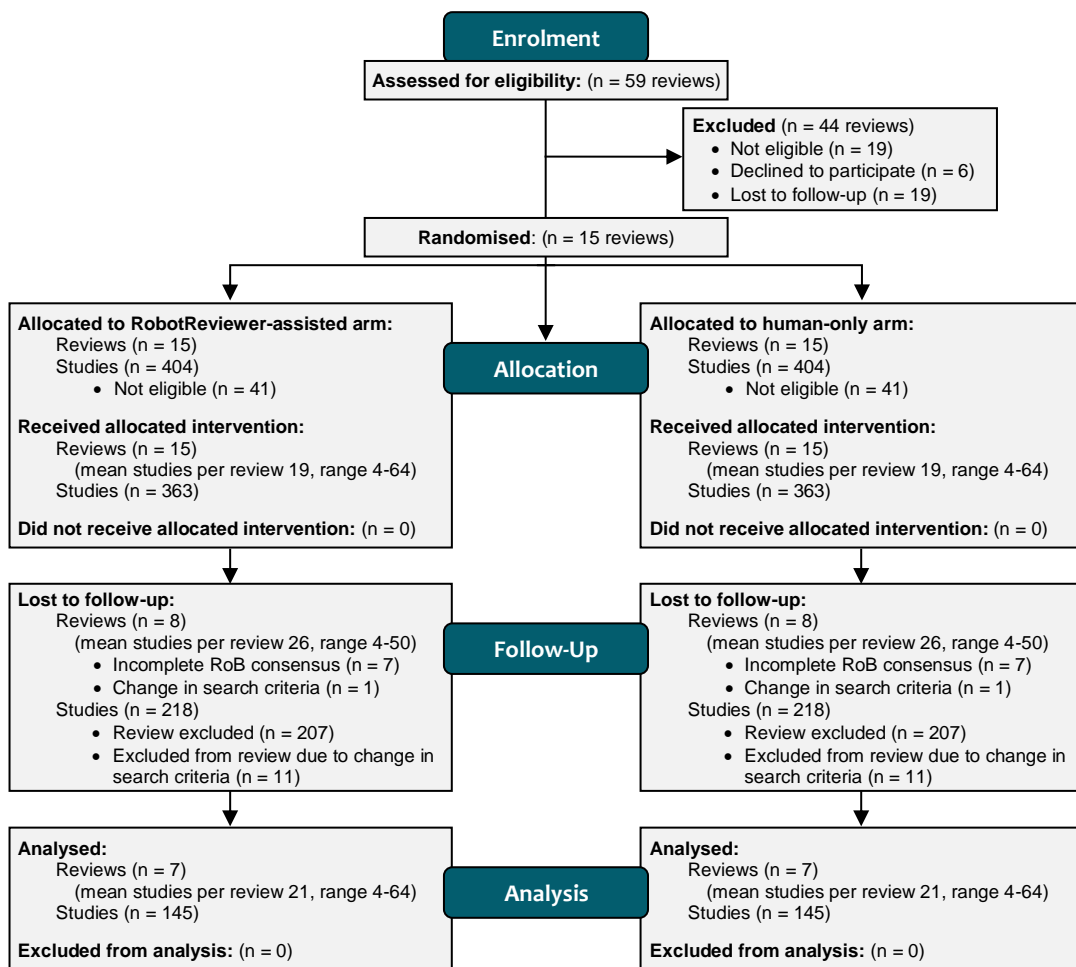


Figure 6.7. Trial flow diagram

Note: Because each study is assessed twice in this paired design, every study is allocated to both trial arms.

Accuracy

The risk difference between RobotReviewer-assisted RoB judgements versus human-only RoB judgements was -0.014 (95% confidence interval: $-0.093, 0.065$). As both the result and inferior bound of the 95% confidence interval are above the pre-defined inferiority threshold of -0.10 , these results support the conclusion that the accuracy of RobotReviewer-assisted RoB judgements are not inferior to human-only RoB.

For the secondary outcome of domain-level accuracy, the risk differences (and 95% confidence intervals) were -0.014 ($-0.091, 0.063$) for Sequence Generation, -0.055 ($-0.211, 0.101$) for Allocation Concealment, -0.021 ($-0.213, 0.172$) for Blinding of Participants and Personnel, and 0.034 ($-0.142, 0.211$) for Blinding of Outcome Assessors. Results are illustrated in Figure 6.8, and detailed data are presented in Table 6.2.

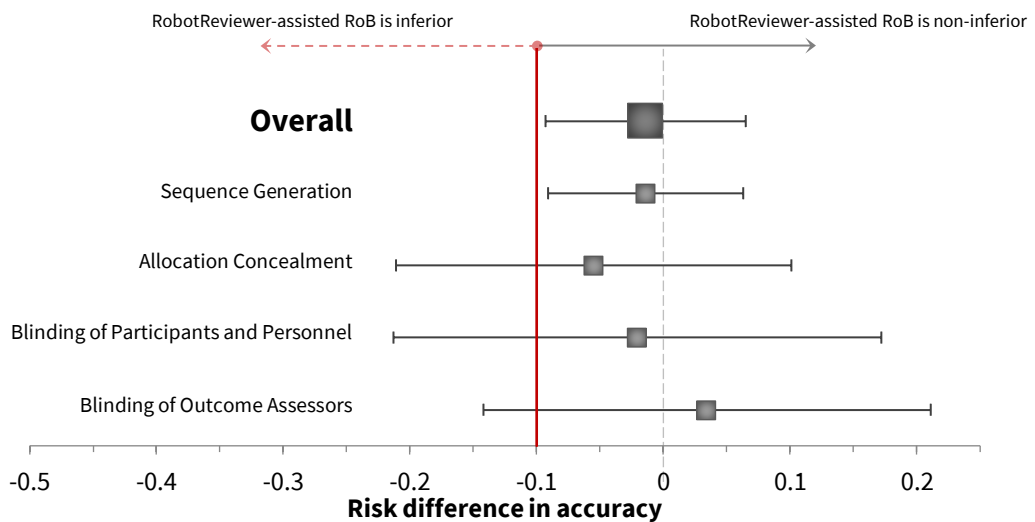


Figure 6.8. Effect of RobotReviewer assistance on RoB accuracy, overall and by domain

Table 6.2. Accuracy data for Risk of Bias

	N	RobotReviewer-assisted		Human-only		Risk Difference [95% CI]
		Accurate	Inaccurate	Accurate	Inaccurate	
Overall	580	515	65	523	57	-0.014 [-0.093, 0.065]
Sequence Generation	145	131	14	133	12	-0.014 [-0.091, 0.063]
Allocation Concealment	145	122	23	130	15	-0.055 [-0.211, 0.101]
Blinding of Participants and Personnel	145	126	19	129	16	-0.021 [-0.213, 0.172]
Blinding of Outcome Assessors	145	136	9	131	14	0.034 [-0.142, 0.211]

Person-time

The weighted mean difference in person-time per individual assessment was 1.40 minutes faster for RobotReviewer-assisted assessments (95% confidence interval: -5.20, 2.41). Values across all reviews ranged from 51 seconds to 49 minutes, with a mean of 9 minutes and 40 seconds. These results are insufficient to conclude that the person-time required for RobotReviewer-assisted RoB assessments is less than person-time required for human-only RoB assessments. Results are illustrated in Figure 6.9, and individual review results are presented in Figure 6.10.

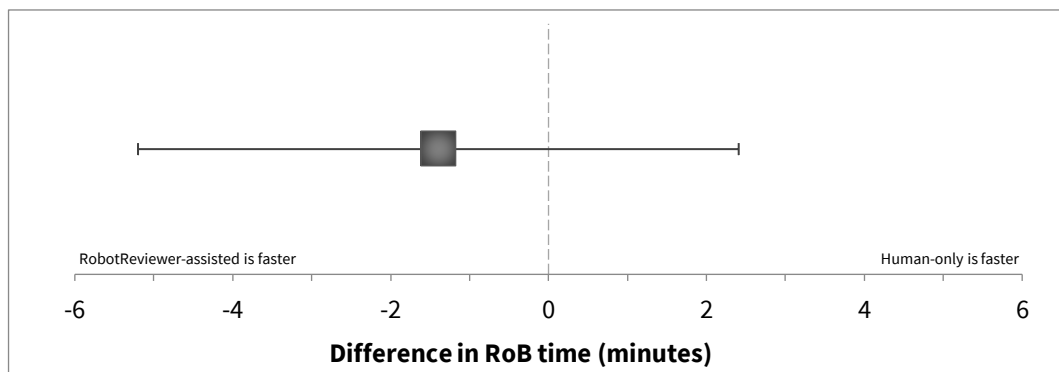


Figure 6.9. Effect of RobotReviewer assistance on time to complete RoB

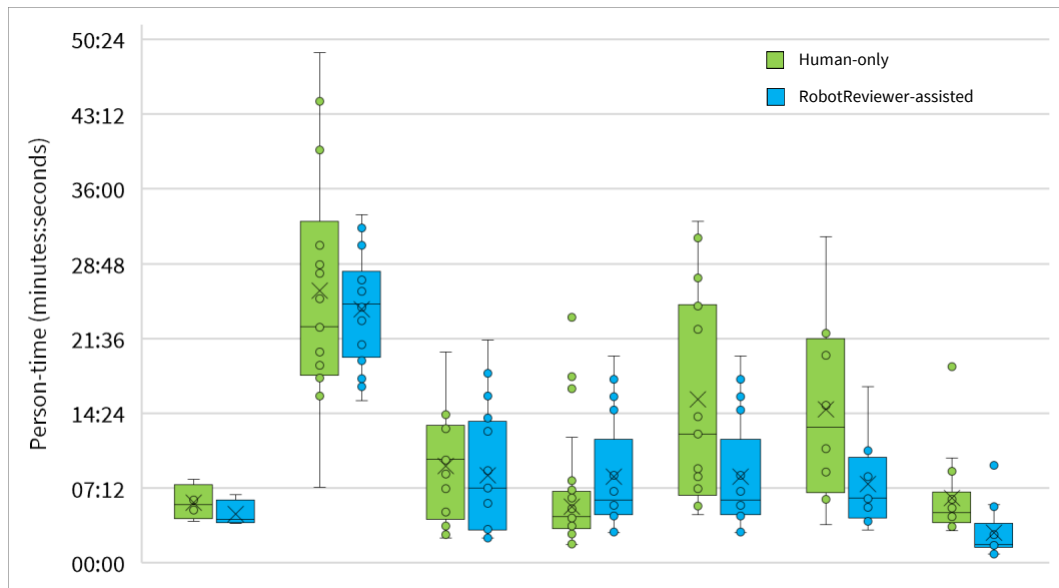


Figure 6.10. Person-time results grouped by review

Discussion

Accuracy of automation-assisted Risk of Bias is non-inferior to human-only Risk of Bias

Establishing the safety of automated or partially automated workflows in systematic reviews is key to building the trust necessary to facilitate adoption of these systems [11, 12]. In this randomised trial in ‘real-world’ systematic reviews, results support the conclusion that machine learning assistance does not negatively impact the accuracy of Risk of Bias assessments. Effect on person-time, while suggesting a potential benefit, remained uncertain. In summary, this trial found a partially automated process resulted in no erosion of systematic review quality; systematic reviewers can therefore incorporate suggestions from RobotReviewer in their methods with confidence the high quality of their evidence will be maintained.

In the face of increasing volume of research output [22], rising methodological expectations [14], and an increasing demand for review currency [23], improving the efficiency of health evidence synthesis is a critical challenge for the field. Machine learning for systematic reviews has the potential to make a significant contribution to these challenges, but despite many years of active research there has been limited uptake of automation innovations [10].

Two novel aspects of the design of this trial substantially strengthen the evidentiary foundations for automation in health-focused systematic reviews: a focus on combining human and machine effort and conduct in real-world conditions rather than artificial or constructed laboratory comparisons.

This trial is the first randomised trial of machine-assistance, rather than machine-only Risk of Bias assessments. As discussed in this chapter's introduction, previous studies have examined the reliability and accuracy of RobotReviewer assessments alone in comparison to human assessments alone. Building on these previous results, this trial now shows that RobotReviewer can be combined with human effort while maintaining a similar quality of output in RoB accuracy.

Although studies that examine the combination of machine and human effort are more difficult to conduct than direct comparisons, they are critical to the identification of automation tools that meet the standards of this community and enable widespread use [24, 25]. As in many other fields of machine learning research, this augmentation of human effort in systematic reviews may be a more feasible short- or medium-term aim than replacement. Current literature also suggests it is the preference of evidence synthesis professionals that automation is used in this assistive manner. The previous qualitative work, presented in Chapter 4, showed that guideline developers are wary of automation that removes human judgement in evidence synthesis [11]. This point has also been highlighted by the International Collaboration for the Automation of Systematic Reviews (ICASR), who have noted that external stakeholders may be concerned that automation will completely erase valuable human judgement and nuance [26].

In addition to previous literature's somewhat narrow focus on direct comparisons of ML versus human-only RoB assessments, the vast majority of research into the use of machine learning for systematic reviews, and all previous research into RobotReviewer, has been conducted in artificial conditions, testing efficacy in highly selected reviews and reviewer subjects. These idealised conditions provide invaluable foundations upon which this trial was built, but they are not easily generalisable to the variable skill-levels and approaches of systematic reviewers, and to the highly heterogeneous topics of ongoing real-world reviews. As such, their findings do not provide information regarding 'effectiveness' and fall

short of supporting real-world adoption. This is also supported by findings from my qualitative study of health guideline developers; though substantial ‘efficacy’ research is available, they still raised the need for further evidence to support automation adoption [11], which suggests that effectiveness research may be a critical contribution towards adoption.

The challenges of conducting an ‘effectiveness’ trial of this type should be noted, however. The recruitment rate was consistently low and for feasibility reasons the trial was stopped before the planned sample size was reached. While the final dataset proved sufficient for analysis of the primary accuracy outcome, it likely impacted on this study’s ability to assess the effect of machine assistance on the person-time outcome and on domain-level accuracy. The selected statistical approach, while methodologically strong, also meant that clustering effects were taken into account, reducing the power of the study to detect a true effect. It might be wise to learn from this for future research and select an analytical approach that is less vulnerable to these effects. Future studies should also endeavour to develop novel methods of promotion and recruitment to address these feasibility challenges.

In summary, this study and its results should encourage well-controlled effectiveness studies, and I suggest they are essential for the development and adoption of machine learning in systematic reviews. Furthermore, these results support a research focus on ‘hybrid’ systems which “combine the strengths of human volunteers and AI” [27].

Contribution of analytical frameworks

It is useful to restate the results of this trial as examined through the lenses of the selected methodological frameworks, as described in Chapter 3. First, this trial tested the non-inferiority of a level 4 or a level 5 automation tool [28, 29]. Because of the flexibility in the choice given to the human researcher, RobotReviewer cannot be entirely fitted into either of these levels, at least not in the context of this trial nor in typical use and under its current guidance. However, given that the trial’s results showed the quality of automation-augmented RoB were non-inferior to human-only RoB, it can also be concluded that allowing human researchers to choose to some

extent between levels of automation according to their situational preference is feasible methodologically defensible.

The results were not sufficient to draw confident conclusions around the effect of automation on person-time use for Risk of Bias, and this trial has left this question open. Regarding automation accuracy, this trial has contributed to observability for potential future adopters by making its methods and data available for review. As for relative advantage in the accuracy of partially automated Risk of Bias, the results of this trial can be used to allay any concerns that incorporating machine learning into judgement tasks would negatively impact review quality or negatively influence reviewer behaviour to become negligent in their assessments. That is, while this trial does not show a relative advantage per se, it does instead address existing concerns that automation might introduce a relative disadvantage in the quality of health evidence.

Trial limitations

Several limitations of this trial should be noted. Most importantly, the methods of time data collection proved to be insufficiently precise to draw confident conclusions. Data on the time taken to complete each Risk of Bias assessment was captured as keystrokes or clicks while participants were interacting with the RoB form in Covidence. Mean time to complete an assessment varied between reviews from seconds to many minutes, but was generally consistent within a review (i.e., cluster). This suggests heterogeneity in the way individual reviewers interacted with the Covidence Risk of Bias form. Some reviewers seemed to have worked through each assessment, inputting data as they went, whereas others completed all assessments within seconds, suggested they completed their assessments and subsequently recorded their decisions in Covidence.

The choice of this method of data collection was intended to avoid the bias and workflow disruption of self-reported outcomes. These results suggest that in the context of variation in human-computer interaction, self-reported outcomes might provide better data, but further work is needed to determine the best approach given the risks of self-reported measurements. On the other hand, the average time to complete an assessment (9 minutes 40 seconds) was generally similar to that found

in previous literature [30], suggesting the measurement was not entirely inaccurate. Finally, a more precise measurement would be preferable in order to exclude the time spent on domains other than those assisted by RobotReviewer. Including total time increases the signal to noise ratio in observing the effect of the intervention, and therefore reduced the ability of this statistical analysis to detect any effect. While both arms included the same number of extraneous assessment domains, and therefore should be inflated by the same amount of time on average, more selectively measuring the time spent on RobotReviewer-assisted domains would strengthen these conclusions. The limitations of this approach are an important methodological finding, and future studies in this field should include multiple methods of data collection in order to triangulate a more precise measure of person-time.

An additional limitation was the failure to meet the intended sample size. While the data proved to be sufficient for the statistical analysis of the primary accuracy endpoint, a larger sample size would undoubtedly have been preferable. Persistent delays in recruitment made reaching the target sample size infeasible and risked the successful completion of the trial. I therefore decided to halt recruitment and disseminate the study results in order to contribute to the evidence base in the use of automation in health evidence. A larger sample size might also have mitigated another risk to the generalisability of these results which must be acknowledged: given that each individual participant interacted with RobotReviewer multiple times in the course of their review, it is possible that a learning effect influenced one or both of the primary outcomes examined in this trial. For example, a reviewer might feel sceptical at first of the RobotReviewer suggestions, but after several interactions determine that they tend to be accurate. This would influence the extent to which they retain the RobotReviewer suggested judgements and annotations, and consequently impact the time required to complete their assessments. However, it is also possible that a learning effect could have the opposite effect: a reviewer might determine they tend to disagree with the suggestions, creating additional work for the reviewer to remove the suggestions and create new judgements.

Lastly, certain characteristics of RobotReviewer and of Cochrane Risk of Bias 1.0 could weaken the sensitivity and consequent impact of these findings. First, the combination of the 'high' and 'unclear' judgements, while consistent with

Cochrane methods, could conceal possible differences between study arms, resulting in a more positive result than is merited. Similarly, Cochrane's method of applying an overall high/unclear rating for a study with any domain-level high/unclear rating might mask inaccuracies in RobotReviewer assessments. Finally, RobotReviewer currently only offers judgements on four of the seven Cochrane RoB 1.0 domains and does not support Cochrane's RoB 2.0 tool, leaving some of the work of RoB assessment unassisted by automation and diminishing potential gains of adopting the tool.

Conclusion

This randomised controlled trial found that the accuracy of combined machine learning and human effort in Risk of Bias assessments is non-inferior to human effort alone. Study results indicated that use of RobotReviewer suggestions did not negatively impact on assessment accuracy overall and for the sequence generation domain; results for other domains were not conclusive. Systematic reviewers can safely adopt RobotReviewer to assist their individual RoB assessments. The study did not provide strong evidence in either direction regarding the impact of ML on the person-time of individual RoB assessments. Further research is warranted to better understand the effect of RobotReviewer-assistance on this outcome. Further studies are also appropriate to replicate these results for both person-time and for accuracy.

In the context of health evidence synthesis methods, this randomised controlled trial has demonstrated the feasibility and benefits of randomised controlled effectiveness trials, and of focusing on combined human and machine learning effort. It is hoped these results and novel methodology will encourage others to pursue these methods and will facilitate wider adoption of ML systems for health evidence synthesis.

Chapter references

1. Singal, A.G., P.D.R. Higgins, and A.K. Waljee. A primer on effectiveness and efficacy trials. *Clinical and Translational Gastroenterology* 2014; 5(1):e45-e45.
2. RobotReviewer. 2018. www.robotreviewer.net.
3. Higgins, J., D. Altman, and J. Sterne. Chapter 8: Assessing risk of bias in included studies. *Cochrane Handbook for Systematic Reviews of Interventions*. 2011.
4. Marshall, I.J., J. Kuiper, and B.C. Wallace. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association* 2016; 23(1):193-201.
5. Gates, A., B. Vandermeer, and L. Hartling. Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the RobotReviewer machine learning tool. *Journal of Clinical Epidemiology* 2018; 96:54-62.
6. Armijo - Olivo, S., R. Craig, and S. Campbell. Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. *Research Synthesis Methods* 2020; 11(3):484-493.
7. Hirt, J., et al. Agreement in Risk of Bias Assessment Between RobotReviewer and Human Reviewers: An Evaluation Study on Randomised Controlled Trials in Nursing-Related Cochrane Reviews. *Journal of Nursing Scholarship* 2021; 53(2):246-254.
8. Soboczenski, F., et al. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. *BMC Medical Informatics and Decision Making* 2019; 19(1):96.
9. Wallace, B. Automating Biomedical Evidence Synthesis: Recent Work and Directions Forward. *BIRNDL@ SIGIR*; 2018; 2018. p. 6-9.
10. Thomas, J. Diffusion of innovation in systematic review methodology: why is study selection not yet assisted by automation. *OA Evidence-Based Medicine* 2013; 1(2):1-6.
11. Arno, A.D., et al. The views of health guideline developers on the use of automation in health evidence synthesis. *Systematic Reviews* 2021; 10(1):16.
12. O'Connor, A.M., et al. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Systematic Reviews* 2019; 8(1):1-8.
13. *Covidence*. 2020, Veritas Health Innovation: Melbourne, Australia.
14. Higgins, J.P., et al. *Cochrane Handbook for Systematic Reviews of Interventions*: John Wiley & Sons; 2019.
15. Critical Appraisal Tools. 2020. <https://www.unisa.edu.au/research/Health-Research/Research/Allied-Health-Evidence/Resources/CAT/>.

16. Higgins, J., D. Altman, and J. Sterne. Chapter 8: Assessing risk of bias in included studies. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. *The Cochrane Collaboration* 2011.
17. Haahr, M. Random.org: True random number service. 2020.
18. StataCorp, *Stata Statistical Software: Release 17*. 2021, StataCorp LLC: College Station, TX.
19. Flight, L. and S.A. Julious. Practical guide to sample size calculations: non - inferiority and equivalence trials. *Pharmaceutical statistics* 2016; 15(1):80-89.
20. Eliasziw, M. and A. Donner. Application of the McNemar test to non - independent matched pair data. *Statistics in medicine* 1991; 10(12):1981-1991.
21. Yang, Z., X. Sun, and J.W. Hardin. Confidence intervals for the difference of marginal probabilities in clustered matched - pair binary data. *Pharmaceutical statistics* 2012; 11(5):386-393.
22. Bastian, H., P. Glasziou, and I. Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Medicine* 2010; 7(9):e1000326.
23. Garner, P., et al. When and how to update systematic reviews: consensus and checklist. *bmj* 2016; 354.
24. de Bie, K., et al. Using AI to help healthcare professionals stay up-to-date with medical research. *AI for Social Good Workshop*.
25. Thomas, J., et al. Living systematic reviews: 2. Combining human and machine effort. *Journal of clinical epidemiology* 2017; 91:31-37.
26. O'Connor, A.M., et al. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Systematic Reviews* 2019; 8(1):57.
27. de Bie, K., et al. Using AI to help healthcare professionals stay up-to-date with medical research.
28. Sheridan, T.B. and W.L. Verplank, *Human and computer control of undersea teleoperators*. 1978, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
29. Sheridan, T.B. *Telerobotics, automation, and human supervisory control*: MIT press; 1992.
30. Hartling, L., et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ* 2009; 339:b4012.

Chapter 7. Economic evaluation

The cost-effectiveness of a semi-automated workflow to maintain a living evidence map

Chapter overview

This chapter will report on the results of a case study economic evaluation of a partially automated workflow applied to an ongoing living map of COVID-19 evidence; this analysis has since been published. A background of the living COVID-19 evidence map will first be provided, followed by the methods used in this cost-effectiveness analysis. The three automation tools tested, along with two modifying methodological choices concerning screening options, will be described, resulting in the comparison of eight study arms. These study arms were evaluated for their incremental effectiveness in terms of recall compared against the baseline workflow, as well as for their cost over a four-week time horizon. The results presented will demonstrate that the partially automated workflow dominated the baseline workflow; that is, use of automation proved less costly and more effective than the manual comparator. These results will then be examined through several sensitivity analyses, and finally placed within the context of the broader literature and within the analytical frameworks selected for this thesis.

Introduction

The early months of 2020 saw the beginning of what would become the global COVID-19 pandemic. While economies and societies struggled with the implications, the serious and rapidly evolving situation led to an unprecedented growth in health evidence synthesis effort.

COVID-19 literature grew exponentially. The United States National Institute of Health (NIH) had indexed more than 28,000 articles as of early June 2020 [1]; as of 20 July 2021, this has grown to a dizzying 168,000 articles. As described in earlier chapters, keeping up with peer-reviewed literature has become near impossible, and this was repeated in the context of the COVID-19 pandemic. This created a unique opportunity and need for artificial intelligence or automation systems to be put into current use, and many took advantage of the opportunity.

Several academic and clinical centres began creating living evidence summaries, systematic reviews, evidence maps, and guidelines, in addition to some private sector efforts towards automation-curated search results and crowd-sourced evidence development. A thorough, though non-systematic and non-exhaustive, list published by the EPPI-Centre identified more than 250 COVID-19 maps, auto-searches, and databases as of 19 June 2020 [2]. Each had slightly differing aims, scope, and deliverables.

In addition to the list of COVID-19 resources, the EPPI-Centre started producing a living systematic map of COVID-19 research evidence, commissioned by the Department of Health and Social Care England (DHSC), via the Evidence Reviews Facility, a team drawn from University College London, the University of York, and the London School of Tropical Hygiene and Medicine [3]. The first search (used as the starting point for subsequent numbering of searches) for this map was published online on 4 March 2020; the initial few searches varied slightly in their interval, but searches were consistently published weekly following search #3 on 24 March 2020.

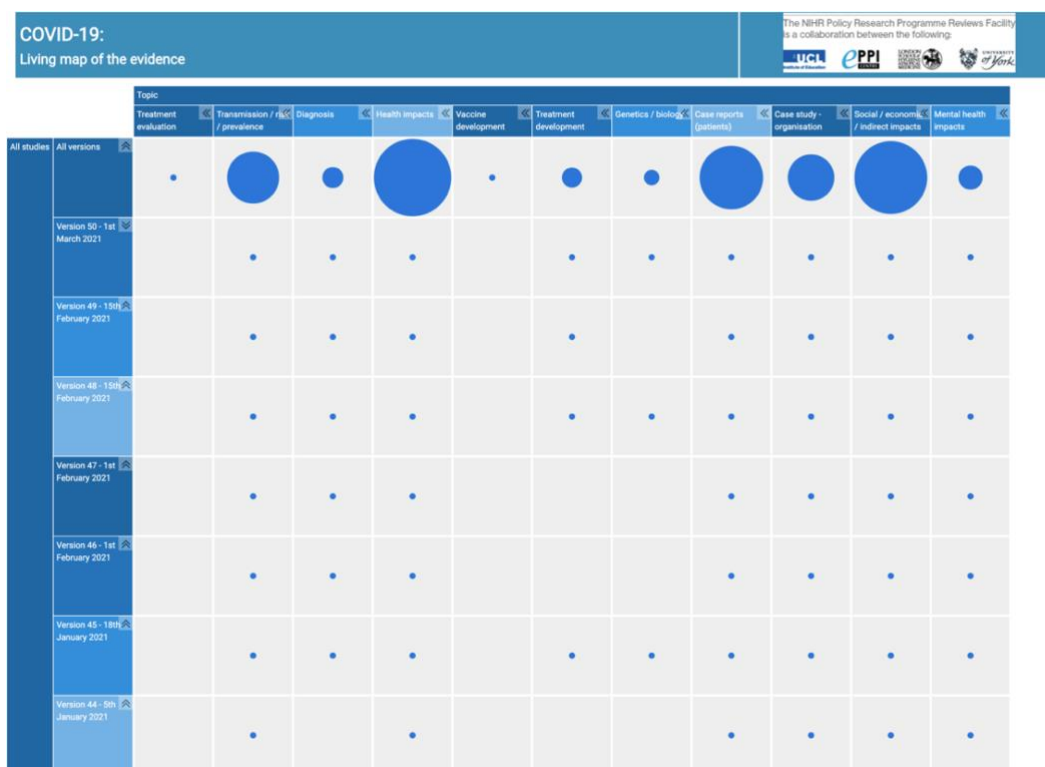


Figure 7.1. Living COVID-19 evidence map published by the EPPI-Centre.

The top row (“All versions”) shows the total studies found in each topic area, while subsequent rows show the relative number of studies found in each version.

A team from the DHSC Reviews Facility (‘the map team’) continues to maintain this living systematic map of COVID-19 evidence (‘the evidence map’ or ‘the COVID-19 map’). The COVID-19 map was originally created, updated, and maintained using a largely manual workflow, underpinned by conventional search and study selection methods and coding tools, hosted in EPPI-Reviewer Web (ER-Web), and repeated on a weekly cycle. However, the rapidly increasing rate of publication of articles with COVID-19-related terms included in their titles or abstracts meant the manual workload involved in maintaining the COVID-19 map was predicted to exceed the capacity of resources available for screening and coding articles. Screeners found themselves pressed for time to complete their share of work by the end of each week, and search specialists found the task of maintaining their study list and deduplication increasingly time-consuming as the number of previous included and excluded records continuously grew. Search yields steadily rose from a few hundred each week to between two and three thousand records per week [4].

Therefore, in parallel with continued maintenance of the COVID-19 map using the current workflow, the map team decided to develop, pilot, and evaluate a

new, semi-automated workflow, enabled by Microsoft Academic Graph (MAG) and MAG Browser (also hosted in EPPI-Reviewer Web).

Microsoft Academic Graph (MAG) is an open access dataset comprising more than 245 million bibliographic records of research articles from across all sciences, with the aim of creating a single, comprehensive dataset of citation information [5]. These are then connected in a large network graph of concept and citation relationships. MAG represents a significant innovation in search in several ways. First, it eliminates the need to search multiple databases by consolidating all information into one database, eliminating the need for complex, bespoke Boolean searches which require significant human expertise and time. Second, MAG automatically indexes records rather than the standard model of relying on publishers for manual updates. Two key questions arise in relation to these innovations. First, does MAG match the recall of multi-database Boolean searches? Second, if MAG is substantially larger than standard sources, does its size negatively impact its precision?

Because of its design, MAG not only offers an opportunity to streamline the searching process to a single source, but also facilitates the integration of machine learning (ML) tools. Put another way, under the Diffusion of Innovations framework, MAG enables expanded compatibility in the pragmatic sense of interoperability with existing systems and automation strategies. Finally, the map team sought to test several other automation tools, namely a binary (include or exclude) ML classifier and active learning priority screening, in conjunction with other methodological variations explained further in the following Methods section.

The intention of this new MAG-enabled, semi-automated workflow was to make the process of maintaining the COVID-19 map more efficient and therefore more sustainable. If the simulated performance of the new semi-automated, MAG-enabled workflow reached acceptable levels during an initial trial period, the map team planned to adopt this new workflow for maintaining the COVID-19 map, instead of the original manual workflow.

Typical concerns regarding automation in systematic reviews were raised within the team. Most importantly, given the risk-aversion of EBM communities and

consequent focus on exhaustive searches which aim to not miss any relevant information, the map team needed to know the effectiveness of the new MAG-enabled workflow in terms of sensitivity (recall). Should the new workflow prove less costly but also less effective, requiring less time and human resources but not identifying as many eligible includes, a specific value of cost per eligible record missed would be helpful in informing the team's decision regarding their preferred workflow.

This study therefore aimed to conduct an economic evaluation to assess the cost-effectiveness of this new semi-automated, MAG-enabled workflow, compared with the conventional manual workflow (current practice), for maintaining the COVID-19 map.

Methods

This study was designed in collaboration with Dr Ian Shemilt. Dr Shemilt contributed simulation data, while I performed the data collection, analysis, and presentation.

Objective

The purpose of this project was to inform the decision-making process of the map team in determining whether to adopt the semi-automated workflow or to continue with the manual workflow (both described in further detail in the following sections). More specifically, the research question was:

RQ4.1) What is the performance and cost-effectiveness in terms of recall, costs, and precision of a semi-automated search and screening method for a living evidence map of COVID-19 evidence?

To answer this question, a cost-effectiveness analysis was conducted using a basic decision-analytic modelling framework comparing the semi-automated workflow (treated as the intervention) with the manual workflow (treated as the comparator). This model-based economic evaluation framework had previously been used to assess the cost-effectiveness of using different screening methods in a case-study systematic review of the effects of undergraduate medical education [6]. This

cost-effectiveness analysis is reported in line with the Consolidated Health Economics Evaluation Reporting Standards (CHEERS) statement (included as Appendix F) [7].

Study arms

The incremental costs and effects of using eight different potential workflows were simulated. The study took a pragmatic approach; these eight arms were defined by the adoption decisions made in practice in the live workflow by the map team. These included three comparator (manual) arms and five intervention (semi-automated) arms. Costs and effects were simulated over a four-week time period in June and July 2020, representing searches 16 through 19, immediately preceding the incremental adoption of automation tools in the live map workflow.

The eight study arms differed across five domains, described in more detail in the following sections. All arms included manual screening of studies and deduplication of search results against all previous searches' included and excluded studies, but differed in their search method, use of a binary machine learning classifier, use of prioritised screening, use of a fixed screening target, and target recall.

Table 7.1. Characteristics of study arms

	Arm	Search	Binary ML classifier	Priority screening	Fixed screening target	Target recall
Manual	Comparator A	MEDLINE & Embase				1.0
	Comparator B	MEDLINE & Embase				0.95
	Comparator C	MEDLINE & Embase			•	0.95
Semi-automated	Intervention A	MEDLINE & Embase	•			0.95
	Intervention B	MEDLINE & Embase	•	•	•	0.95
	Intervention C	MAG				1.0
	Intervention D	MAG	•			0.95
	Intervention E	MAG	•	•	•	0.95

Search: MAG versus MEDLINE & Embase

MAG

Prior to the commencement of this study, members of the EPPI-Reviewer team, in collaboration with Microsoft [5] had developed a novel machine learning

recommender model to automatically search each update of the MAG dataset. Recall from this chapter's introduction that MAG searches represent a substantial methodological shift for search in several ways: it is a single source rather than multiple databases; it automatically updates rather than relying on publishers to index new records; and it can be searched with machine learning rather than with human-developed bespoke (and therefore costly) Boolean search strategies. This model was used to score new records published in each MAG update; records above an optimised threshold are then imported into the map project in EPPI-Reviewer Web (ER-Web) [8]. Once imported to ER-Web, duplicates were semi-automatically identified against known included studies and excluded studies and discarded.

MAG was adopted as the single source of potentially eligible records by the map team from search 35 onward. To simulate the use of MAG in study arms 6 to 8, the records uniquely identified in the MAG dataset during the evaluation period were screened.

MEDLINE & Embase

Conventional Boolean searches were run on MEDLINE and Embase by information specialists on a weekly basis. Each week, the search results were downloaded to an EndNote library and duplicate records discarded, assisted by EndNote's semi-automated duplicate detection. This process leaves the final decision entirely up to the information specialist but suggests duplicates that should be removed. In other words, de-duplication could be viewed as a level 5 automated task in the framework presented in Chapter 3. Once the weekly search was deduplicated, the results were also deduplicated against all previously identified included studies and excluded studies. It is worth noting here that due to the progressively larger library of known includes and excludes, this second round of deduplication steadily increased in time demand. The search results were then provided to the screen-coding team for screening in ER-Web [8]. Once the citations were imported into the ER-Web tool, the screening tasks (i.e., study records) were randomly allocated among the map team. The screen-coders of the map team then completed their screening, including retrieval of any full-texts as they judged necessary.

The MEDLINE and Embase workflow was used from search 1 through search 34, when the team adopted the MAG-enabled workflow. Screening data from

the live workflow between searches 16 through 19 were used to simulate comparators A, B, and C and interventions A and B.

Binary machine learning classifier versus none

A binary machine learning classifier was designed and developed to determine study eligibility. The classifier scores new records according to the likelihood, according to the ML classifier model, that they are eligible for inclusion in the living map. The threshold for inclusion was calibrated during development of the classifier by members of the map team with a target 0.95 recall. The model was trained on known included and excluded studies resulting from the MEDLINE and Embase searches for searches 1 through 15. Once the model was trained on these data, it was applied to the search results from either MAG or MEDLINE/Embase (depending on the study arm) for the evaluation period (searches 16 through 19).

Priority screening versus none

Priority screening used an active learning model to build upon the results of the binary ML classifier. When studies for screening are initially displayed in ER-Web, they are ordered according to the results of the binary ML classifier; higher scoring studies, which are more likely to be included according to the classifier, are listed at the top, and lower scoring studies, which are less likely to be included according to the classifier, are listed at the bottom. As screeners begin to record their decisions, the priority screening mode observes these decisions and periodically updates the order of the study list such that studies more likely to be included according to previous decisions are now listed towards the top. This approach of observing a user's decisions and updating a machine learning model accordingly is called active learning, as was described in Chapter 2. This 'priority screening mode' was adopted by the map team from search 30 onwards.

Fixed screening target versus none

The final two variable decision points of the variant workflows are not forms of automation, but rather other methodological choices which may or may not impact the effects of automation on the map workflow. Given that the map team – like any group of people with a given task – had a finite amount of person-time to spend on the task of screening, they selected an overall screening target of 1,500 study records to be screened each week. For the purposes of data collection in this

study, this fixed screening target was simulated in searches 16 through 19 to produce evaluation data. For study arms not using a fixed screening target, simulated manual screening continued until either 1.0 recall or 0.95 recall was reached, depending on the study arm. The map team adopted a fixed screening target in their live workflow from search 30 onwards.

Target recall = 1.0 versus target recall = 0.95

The target recall, in theory, of most systematic reviews should be 1.0. That is, author teams aim to exhaust all available literature, and to identify every single eligible study for inclusion in their review (or in this case, evidence map). Thus, the baseline workflow was working to a target recall of 1.0, and this target was evaluated in the first of the MAG-enabled study arms (intervention C). When the target recall is lowered to 0.95, this signals a slightly relaxed standard, which is willing to lose or sacrifice 5% of the available literature. Because the binary ML classifier was calibrated to 0.95, any workflow including this tool implicitly works to a target recall of 0.95. For this reason, comparators B and C were included to ensure the inclusion a fair comparison for comparators A, B, D, and E. When the map team adopted the binary ML classifier from search 20 onward, the target recall of 0.95 was also implicitly adopted.

Comparator arms

A summary of study arm information from the previous section and from Table 7.1 is as follows. Each of the three comparator arms used MEDLINE/Embase as a search method. The baseline workflow, comparator A, used a target recall of 1.0, while comparators B and C used a target recall of 0.95 and were included in this study to ensure fair comparisons for intervention study arms which implicitly adopted a target recall of 0.95 via the binary ML classifier. Comparator C was also included to ensure the presence of a fair comparison for intervention arms using a fixed screening target of 1,500 records.

Intervention arms

Each of the intervention arms used one or more novel automation methods. The resulting five intervention arms are the result of distinct combinations of three automation tools, the inclusion or exclusion of a fixed screening target, and either a

1.0 or 0.95 target recall. The three automation tools tested were (1) an automated MAG update search, (2) a binary machine learning classifier determining inclusion or exclusion, and/or (3) active learning prioritisation of the screening list. Final inclusion/exclusion decisions and study categorisation were conducted by human screen-coders in all study arms. Note that any arms using the binary ML classifier implicitly adopt a target recall of 0.95.

Intervention A maintained the MEDLINE and Embase search strategy and simulated the use of the binary ML classifier. Intervention B also used MEDLINE and Embase for search, the binary ML classifier, in addition to active learning priority screening and a fixed screening target of 1,500 records. Interventions C, D, and E all used MAG as a single source for searching. Intervention C used no other automation tools, and therefore had a target recall of 1.0. Interventions D and E used the binary ML classifier, and finally intervention E added priority screening and a fixed target, thus testing all the potential automation tools examined in this study.

Analytical perspective and time horizon

The analytic perspective of the cost-effectiveness analysis was a single employer, specifically a university employer. With this perspective, the main drivers of cost differences between study arms were staff (screen-coder) time required to complete the review of each week's identified potentially eligible studies, and the staff (information specialist) time required to complete search and de-duplication. The time horizon for the study was four weeks; all costs and effects included in the analysis occurred within this isolated time horizon and therefore no discount rate was applied.

Outcomes and measurements

The outcome of interest was the incremental cost-effectiveness of each variant workflow. Effectiveness was defined as the recall or sensitivity of each workflow; that is, perfect effectiveness would be retrieval and identification of 100% of the records eligible for map inclusion. Combined with the cost estimates, cost-effectiveness was defined as the incremental cost per eligible study 'saved' from inappropriate exclusion. Selection of this outcome assumed a preference for avoiding exclusion of eligible studies at the lowest possible cost of doing so.

To calculate this ratio, the following outcomes were measured: recall (sensitivity), precision (specificity), screening workload (number of records), percent of records requiring full-text examination, number of records included, resource use (screen-coder time-on-task and information specialist time-on-task), and costs (unit cost of staff time). To provide for the possibility that either the MEDLINE-Embase search strategy and/or the MAG-enabled search would retrieve unique records, and/or that either method might inappropriately exclude an eligible record, a 'gold standard' recall was constructed which included the combined set of final inclusions from both workflows. Overall recall is reported against this gold standard, while incremental recall is compared against comparator A, i.e., the previous workflow in place prior to the commencement of this study. Precision, screening workload, and the number of records included were automatically collected by ER-Web.

Costs were calculated through the most recently available published university pay scales for University College London (United Kingdom, GBP £) [9] and Monash University (Melbourne, Australia, AUD \$) [10]. Specifically, for the former, spine point 46, grade 9 on the UCL non-clinical grade salary schedule, inclusive of London allowance, was selected. For the latter, the mid-point (salary step 4) of academic level B was selected. These were £30.13 per hour and \$53.94 per hour, respectively. For information specialists, this was calculated as a constant weekly time-use based on their self-reported estimated average. For the comparator workflows and interventions A and B, i.e., those that did not use MAG, interviews were conducted with the two information specialists from the map team to estimate the time required weekly, on average, to complete study retrieval and de-duplication (both between MEDLINE and Embase and against prior searches). For the MAG-enabled workflows, i.e., interventions C, D, and E, the time required each week to de-duplicate was recorded. Self-reported time-use data was collected weekly for the duration of the study using a reporting template (see 0 for supplementary data) supplied to each screen-coder. These were used the average time per 100 records screened, which was then combined with the number of studies screened in each arm to produce the total screening time for each variant workflow. Screening time and either search/deduplication of MEDLINE/Embase or deduplication of MAG then combined with the unit costs to calculate the total cost of each arm. Incremental cost

was calculated for each study arm compared against comparator A, the baseline manual workflow.

Analytical assumptions

In addition to the previously mentioned value assumption of a preference for avoiding inappropriate study exclusion, the base case analysis of this study incorporated three additional assumptions. First, it was assumed that the precision of comparator C was equal to the precision of comparator A (baseline workflow); in other words, that studies are screened in a random order and that the pool of studies not screened contained the same proportion of truly eligible studies as the pool of studies screened. Second, it was assumed that the studies which were identified both by searching MAG and by searching MEDLINE and Embase had the same inclusion rate (i.e., precision) of 0.50. This level of precision is equal to the precision observed while screening the records uniquely identified in the MAG dataset. Finally, of the hypothetical additional studies added to intervention D and intervention E, the distribution of studies that would have been included or excluded was assumed to be the same as the distribution of the known included or excluded studies in the evaluation data (searches 16 through 19).

Sensitivity analyses

Two deterministic, univariate sensitivity analyses were performed. In the first sensitivity analysis, time-on-task needed to screen 100 records was held constant between the study arms. This first analysis was defined prospectively due to the use of self-reported measurements of reviewer efficiency, and moreover because in practice it could be reasonably expected that this measurement could be influenced by factors outside of control of this study. The second sensitivity analysis was conducted post-hoc after observation of variations in precision in the 'live' workflow used to identify studies for the evidence map after adoption of a MAG-enabled workflow from search 35 onwards. In this analysis, precision was varied between plausible lower- and upper-limit values. Because the semi-automated elements of intervention E were effectively the workflow adopted in practice by the map team, this sensitivity analysis was only conducted on that study arm. Results of this analysis address the impact of the second analytical assumption relating to assumed precision, as described in the previous section.

Results

Overall cost-effectiveness

Workflows which used Microsoft Academic Graph (intervention arms C, D, and E) in place of the conventional MEDLINE-Embase search method (comparators A, B, and C) resulted in a higher recall and a lower cost; these results therefore showed that MAG-enabled workflows dominated MEDLINE/Embase workflows. Of the remaining two automation-enabled study arms, intervention arm A (binary ML classifier) resulted in a cost of £15.16 or AU\$27.15 per record saved from inappropriate exclusion, while intervention arm B (binary ML classifier and active learning) resulted in a cost of £13.57 or AU\$24.29 per record saved from inappropriate exclusion. While resulting in substantially different effectiveness, comparator arms B and C each resulted in nearly identical cost per record saved from inappropriate exclusion. Comparator B showed a cost of £1.82 or AU\$3.26 per record saved from inappropriate exclusion, while comparator C showed a cost of £1.83 or AU\$3.28 per record saved from inappropriate exclusion.

Overall cost-effectiveness results are illustrated in a cost plane in Figure 7.2. Note that any data points in the lower-right quadrant represent a workflow that dominates the comparator workflow. The raw data contributing to these calculations are available as supplementary material (see 0).

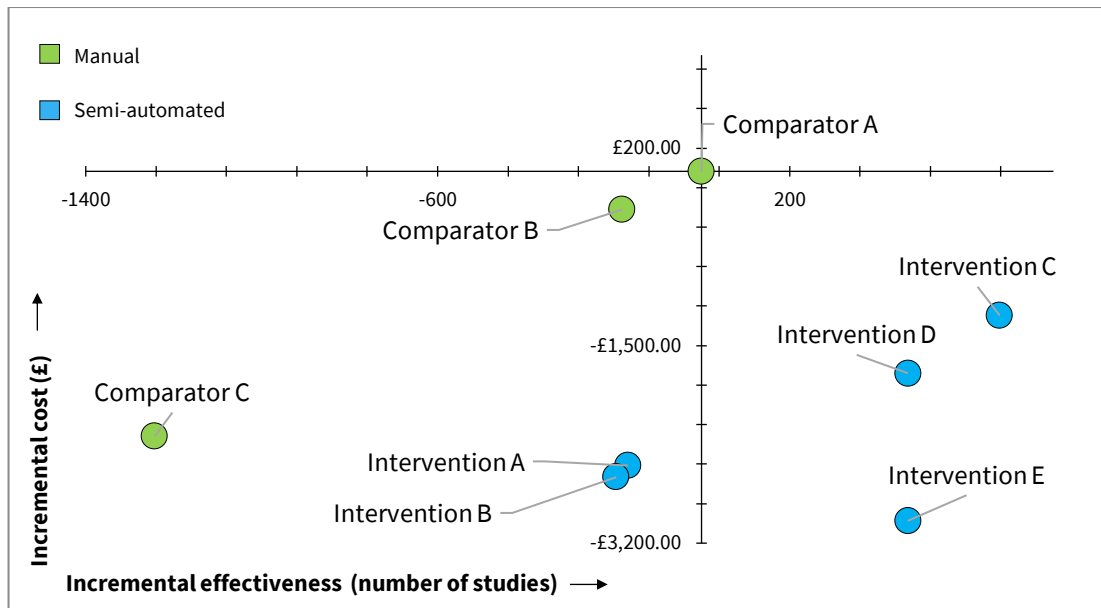


Figure 7.2. Results of cost-effectiveness analysis

Effectiveness

Ranked strictly in terms of recall, the most effective workflow was intervention C (semi-automated searches of MAG with no other ML tools), while the least effective was comparator C (manual searches of MEDLINE/Embase with a fixed screening target).

Table 7.2. Effectiveness results by study arm

	Arm	Recall *	Precision †	Incremental effectiveness ‡
Manual	Comparator A	0.83	0.40	-
	Comparator B	0.79	0.40	-180
	Comparator C	0.55	0.40	-1243
Semi-automated	Intervention A	0.79	0.55	-167
	Intervention B	0.79	0.57	-194
	Intervention C	0.99	0.50	678
	Intervention D	0.94	0.52	469
	Intervention E	0.94	0.86	469

* Recall is the number of eligible records identified divided by the total eligible records from the constructed 'gold standard' recall, which included both MAG-identified and MEDLINE-Embase identified eligible records

† Precision is the number of records included divided by the number of records screen-coded

‡ Incremental effectiveness refers to the number of eligible records identified compared to baseline workflow

Effectiveness in terms of recall (i.e., sensitivity, or the total proportion of all eligible includes identified) was higher in all of the MAG-enabled workflows (interventions C, D, and E). Of the 4378 included studies in the gold standard included studies set, 65 studies were uniquely identified by manual search methods, while 743 were uniquely identified by MAG. Recall was highest in intervention C,

which used a MAG-enabled search strategy but none of the other automation tools tested. In this workflow, 678 additional eligible studies were identified compared to the base case in comparator A, providing a 0.99 recall rate compared against the gold standard recall. Interventions D and E both found an additional 469 eligible studies compared to the base case, or a 0.94 recall. In contrast, the base case (fully manual workflow) resulted in a 0.83 recall; the other comparator arms were even lower at 0.79 recall and 0.55 recall for comparator B and comparator C, respectively.

Among non-MAG-enabled (i.e., MEDLINE/Embase) search workflows that used automation – interventions A and B – both resulted in a modest number of missed eligible studies. Intervention A used the binary ML classifier and identified 167 fewer included studies compared to the base case workflow (comparator A). Intervention B used the binary ML classifier in addition to active learning prioritised screening and found 194 fewer eligible studies than the baseline workflow. These both resulted in a 0.79 recall rate compared to the gold standard recall.

Cost

Two factors influence the total cost of each workflow: the number of studies screened, and the time on task (calculated as hours per 100 records). These data are presented in **Table 7.3**.

Table 7.3. Cost data by study arm

	Arm	Time on task (hours per 100 records [SD])	Number of studies screened	Resource use (hours) *
Manual	Comparator A	2.38 (0.95)	9180	234.08
	Comparator B	"	8722	223.18
	Comparator C	"	6000	158.45
Semi-automated	Intervention A	2.13 (0.56)	6315	150.03
	Intervention B	2.18 (0.47)	6000	146.71
	Intervention C	2.22 (0.13)	8639	184.67
	Intervention D	"	7898	168.92
	Intervention E	"	6000	128.57

* Resource use includes 15.75 hours for information specialist time in arms which used manual search, and 1 hour for information specialist time in arms which use MAG search

Ranked strictly in terms of cost, the least costly workflow was intervention E, which cost £3,179.01 or AU\$5,692.21 less than the comparator workflow. The other

two MAG-enabled workflows showed £1,488.48 or AU\$ 2,664.74 cost saving with intervention C, and £1,963.16 or AU\$3,514.54 cost saving with intervention D.

Table 7.4. Cost results by study arm

	Arm	Total cost (£)	Incremental cost (£)	Total Cost (au\$)	Incremental cost (au\$)
Manual	Comparator A	£7,052.72	-	\$12,626.08	-
	Comparator B	£6,724.53	-£328.19	\$12,038.54	-\$587.54
	Comparator C	£4,774.01	-£2,278.71	\$8,546.63	-\$4,079.45
Semi-automated	Intervention A	£4,520.48	-£2,532.25	\$8,092.75	-\$4,533.34
	Intervention B	£4,420.50	-£2,632.22	\$7,913.76	-\$4,712.32
	Intervention C	£5,564.24	-£1,488.48	\$9,961.34	-\$2,664.74
	Intervention D	£5,089.56	-£1,963.16	\$9,111.55	-\$3,514.54
	Intervention E	£3,873.71	-£3,179.01	\$6,934.88	-\$5,691.21

Among the intervention arms which did not use a MAG-enabled search, results also demonstrated cost savings but decreased effectiveness or recall as previously discussed. Intervention A used the binary ML classifier but no other automation tools and cost £2,532.25 or AU\$4,533.34 less than the comparator. Intervention B saved £2,632.22 or \$4,712.32 compared to comparator A (baseline workflow).

Sensitivity analyses

Time-on-task

To account for the possibility of external factors influencing coders' time required to screen 100 records, and therefore influencing the resulting cost, a deterministic univariate sensitivity analysis was conducted in which time-on-task was held constant between study arms. In this calculation, the overall average time-on-task among all arms was used. The overall results were the same when this scenario was examined: interventions C, D, and E dominated the base case. The cost per inappropriate exclusion avoided in intervention arms A and B was similar but slightly decreased compared to the base case calculations. Intervention A showed a cost of £15.16 or AU\$27.15 per record saved from inappropriate exclusion, while intervention B showed a cost of £13.57 or AU\$24.29 per record saved from inappropriate exclusion.

The results of this first sensitivity analysis are illustrated in Figure 7.3.

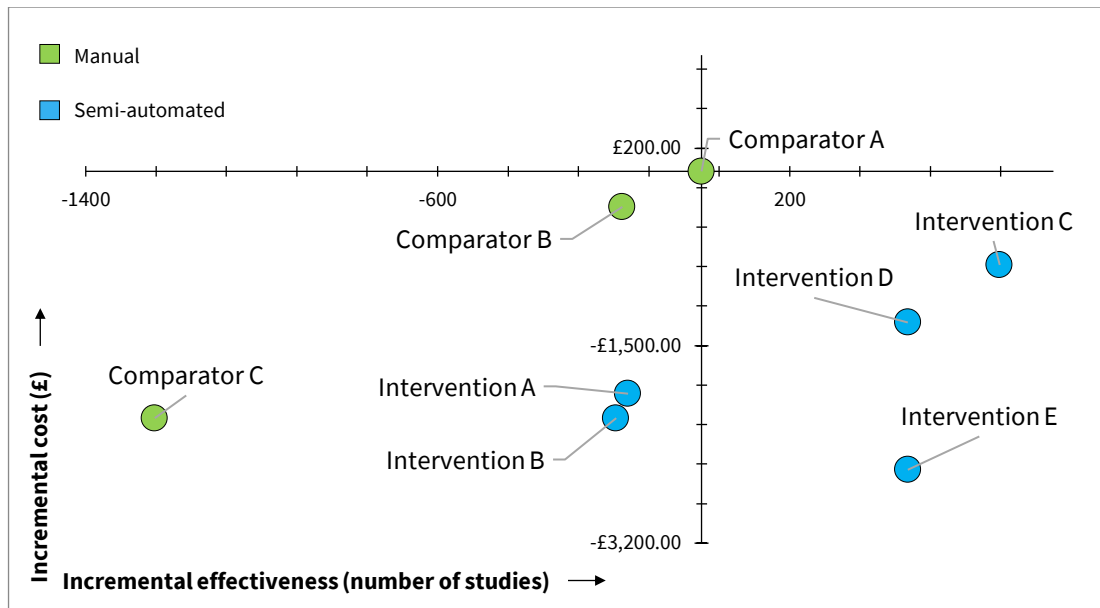


Figure 7.3. Results of cost-effectiveness sensitivity analysis for time-on-task

Precision

An additional sensitivity analysis was conducted post-hoc due to an observed fluctuation in precision observed in practice after adopting a MAG-enabled workflow. Because intervention E was closest to the workflow adopted in practice, the impact of varying precision in this study arm was investigated. The 95% confidence interval of precision observed in practice was 0.55 and 0.72; these values were therefore used in this sensitivity analysis. Though a decrease in precision is typically associated with an increase in recall in traditional screening workflows, lowering the precision significantly impacted the recall (effectiveness) of Intervention E due to the fixed screening target of 1,500 studies. In this scenario it identified 335 fewer records compared to comparator A. The decrease in recall shifted its outcome from one of dominance over the baseline workflow to one of an incremental cost of £9.29 or AU\$ 16.63 per inappropriate exclusion saved. The results of the lower and upper limits of precision sensitivity analysis are illustrated in Figure 7.4 and Figure 7.5.

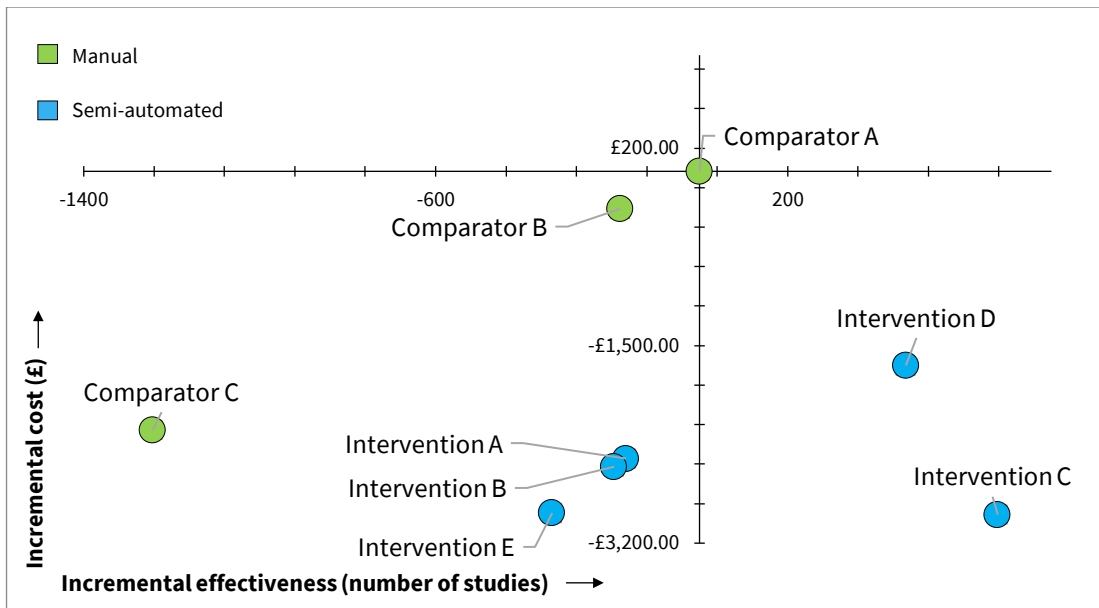


Figure 7.4. Results of cost-effectiveness sensitivity analysis for precision, lower limit

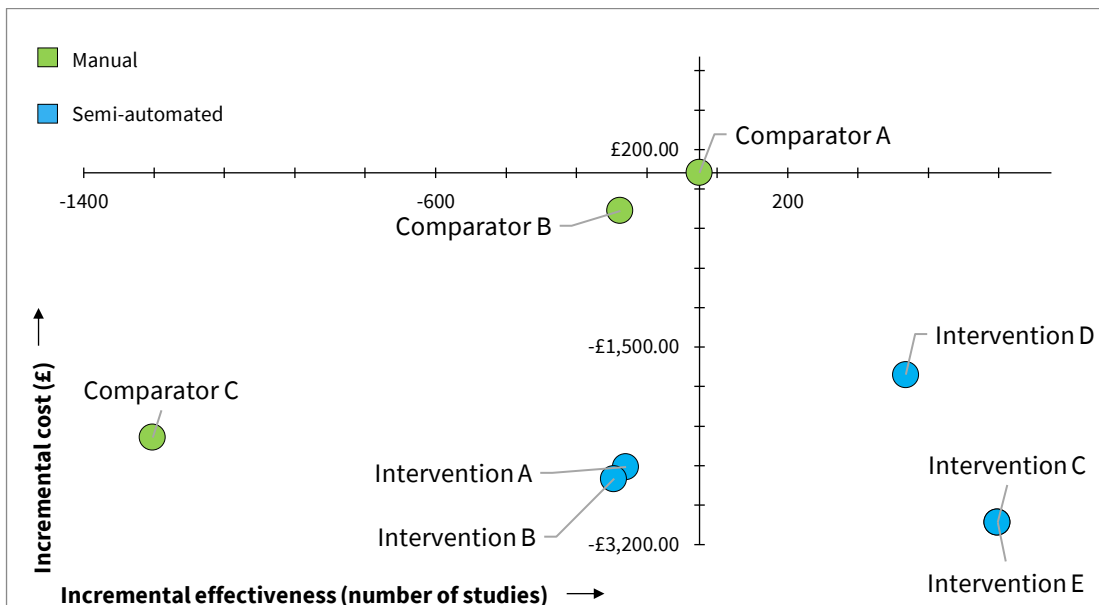


Figure 7.5. Results of cost-effectiveness sensitivity analysis for precision, upper limit

Discussion

The main key finding of this cost-effectiveness analysis was that the semi-automated MAG-enabled workflows dominated the baseline comparator workflow, meaning it was both better in terms of effectiveness and cheaper in terms of cost. This conclusion provides a noteworthy advancement in the evidence relevant to evidence search, given that much of the discussion about applying automation to systematic reviews has centred on the concern of sacrifice in systematic review quality, particularly in terms of recall, a concern which has been reinforced by the

prior research reported in this thesis. While this is a case-specific conclusion, it nonetheless should be considered evidence that automated searches of Microsoft Academic Graph outperform manual database searches.

The semi-automated but non-MAG-enabled arms (interventions A and B), however, showed a decreased recall but cost savings, more in line with the expectations at the outset of this project. This result gives a case study estimate of the cost per inappropriate exclusion avoided that systematic reviewers may now use to inform their adoption decisions. Further work is needed to determine a typical willingness-to-pay with respect to an incremental cost-effectiveness ratio (ICER) in systematic reviews; this is discussed further in the following sections relating to the sensitivity analyses and suggestions for future research. Integrating Microsoft Academic Graph into a live workflow is likely to be a daunting task for the majority of individual researchers, so it might be reasonably expected that they will take a greater interest in intervention arms A and B, which simulate a more probable workflow to be adopted. However, the results of interventions C, D, and E should be noted by the EBM field as encouragement to more seriously consider integration of MAG into their workflows, or more broadly to consider comprehensive single-source bibliographic databases in their systematic review search strategies (such as those published by Semantic Scholar, OpenAlex and The Lens). Perhaps more likely, those working on systematic review tool development might begin exploration of integration of MAG, or its alternatives, to their respective software tools.

The data collected in the course of this analysis reveals further noteworthy time-on-task observations. As shown in the results, all five semi-automated workflows were lower in cost. When examined more closely, it appears that reviewers gained efficiency when using a semi-automated workflow. The average person-time required to screen 100 records under the baseline workflow was 2.37 hours (95% confidence interval: 1.67, 3.08), while the adopted workflow (intervention E) showed 2.13 hours per 100 records screened (95% confidence interval: 1.61, 2.64). Combined with the observed self-reported percentage of records requiring full-text examination, it seems likely that this efficiency gain was partly due to a lower proportion of records requiring full-text examination: the baseline

workflow showed 63.29% records requiring full-text examination (95% confidence interval: 46.35%, 80.24%), while intervention E showed 37.09% (95% confidence interval: 21.29%, 52.90%). From reviewer observations during screening, this is potentially due to a decreased number of records imported without an abstract. Even among those that required full text examination, this was often a quicker process under MAG-enabled workflows due to direct linking to full-text resources; reviewers did not need to put in as much manual work to retrieve the full-text. In any case, the increased efficiency under MAG and other semi-automated workflows should be highlighted in these conclusions.

An additional observation of MAG-enabled screening should be highlighted: the increase in non-English-language records. Further research is required to determine the proportional contribution of these records to the higher recall rate of MAG compared to MEDLINE and Embase, however it was consistently noted during screening that MAG seemed to retrieve a more language-inclusive set of studies for screening, including many studies which met the criteria for map inclusion. While this conclusion is purely observational in nature, if it is reproduced in future research, it could improve equity in research synthesis.

Additional potential comparisons

This analysis elected to use the conventional approach in cost-effectiveness analyses and compared all study arms against current practice, i.e., comparator A. The choice to include several additional comparators to ensure inclusion of a fair comparison, in addition to the pragmatic adoption approach resulting in five intervention arms, enabled observation of the isolated effects of each of the automation tools examined in this study.

Specifically, the incremental effect of the binary ML classifier alone can be examined by comparing intervention A against comparator A (already included in the base analysis), and by comparing intervention D against intervention C; each of these comparisons differ only in the inclusion of the binary ML classifier in the former study arm compared to the latter study arm, respectively. In these comparisons, somewhat similar incremental recall was observed: 167 fewer studies and 209 fewer studies, respectively, in each workflow including the binary ML

classifier. The similar values found in each of these comparisons could be indicative of the overall impact that the classifier tends to have on recall; further investigation is needed in other scenarios, however, to explore and confirm this hypothesis.

Using a similar approach, it is possible to isolate the effect of MAG integration alone using three comparisons: intervention C versus comparator A (already included in the base analysis), intervention D versus intervention A, and intervention E versus intervention B. Each of these comparisons hold all workflow variables equal, apart from the use of MAG as a single source. These comparisons result in incremental effectiveness of 678 studies, 636 studies, and 663 studies, respectively, in the intervention arms which included MAG. Once again, very similar incremental inclusion is observed, suggesting this is attributable to MAG alone. Further research should seek to replicate these results in other contexts.

Sensitivity analyses

The first sensitivity analysis included in this study strengthens the initial conclusions, that a semi-automated workflow is both more effective and less costly than the baseline manual workflow. When time-on-task was held constant between arms, the results were the same, indicating that variations in time-on-task – which is to say, variations in reviewer efficiency – are unlikely to impact the conclusions drawn by the initial analysis.

The second sensitivity analysis, however, provides reason for reflection on these results. The precision observed in practice fluctuated significantly on a weekly basis; as indicated in the results section, the lower 95% confidence interval of these fluctuations showed a 0.55 precision, while the upper confidence interval showed a 0.72 precision. This lower limit changed the initial conclusion of semi-automated workflow dominance to a scenario in which the semi-automated workflow was less costly but also less effective. In such a scenario, an adoption decision would be determined by the maximum acceptable incremental cost. That is, in order to determine the adoption outcome, the cost a given review team is willing to pay for each inappropriate study exclusion avoided would need to be known. Such a cost estimate is highly difficult to estimate. For instance, as discussed in Chapter 4, guideline developers tended to indicate an unwillingness to sacrifice any

effectiveness at all. Those discussions were, however, hypothetical; perhaps in this more concrete scenario, attitudes towards a cost and effectiveness trade-off could be determined.

Finally, it is useful to note that the point at which this sensitivity analysis shifts away from dominance occurs when precision is equal to 0.61. Future efforts in continuing to semi-automate this workflow might aim to raise the lower precision bound to this level, as well as observing the live workflow precision to see if, over time, fluctuations might decrease.

Study limitations

First and foremost, this cost-effectiveness analysis relates to a case study of several semi-automated workflows in the maintenance of a specific living evidence map. The generalisability of these results is therefore limited. However, given the limited research into the cost-effectiveness of semi-automated search methods, in particular taking the approach taken here of calculating ICERs with recall as the designated measure of effectiveness, the results reported from this study still represent a significant contribution to the understanding of the outcomes of adopting automation in health evidence synthesis.

The generalisability of the findings of this analysis are also complicated by the fact that a main driver of the semi-automated workflows' dominance was the substantially higher effectiveness of MAG in terms of recall compared to MEDLINE/Embase searches. This complication is lessened, however, by the presence of fair comparison arms: when intervention arms examined automation tools without the use of MAG searches, they still performed better than comparison arms that adjusted for a fair comparison in target recall and for the use of a fixed screening target.

As described in the methods section, both estimates of unit cost were taken from self-reported measures: first, from interviews with information specialists, and second, from self-reported weekly templates on time-use and number of studies screened. Much like the preceding chapter, the development of an unobtrusive application which observes reviewers' time-use would be very useful for this type of research. The variation in time-use per 100 records screened was minimal,

suggesting that the results are fairly precise and accurate; self-reported outcomes remain, however, open to bias. The first sensitivity analysis aims to address this, and the results of that analysis strengthen the conclusions found in this study.

Suggestions for further research

In addition to the previously mentioned more isolated examination of each of the automation and semi-automation tools considered in this study, several other aspects would benefit from future research.

As has been described in these results and in this discussion, the MAG-enabled workflows (interventions C, D, and E) demonstrated an increased recall compared to the baseline workflow, while the semi-automated but non-MAG-enabled arms (interventions A and B) showed cost savings but decreased recall. This analysis did not examine the more granular aspects of this increased or decreased recall, and instead treated the value of each included or excluded study as equal. That is, I did not examine whether the recall differed among arms with respect to specific study types or designs. Whether the true value of each included study is equal is certainly an open question; the NHMRC, for example, specifies a “hierarchy of evidence” in which certain types of evidence (e.g., RCTs) are more weighted than others (e.g., case reports). It would be highly informative, therefore, to perform a value of information analysis to determine the effect of the presence or absence of particular records on the final results of the review.

Similarly, this analysis did not examine whether one method was superior to another in terms of topic-specific recall. There are nine inclusion topic codes in the living evidence map, and these results did not address the respective distribution of topic codes in the studies uniquely identified by either MAG or by MEDLINE and Embase searches. As above, further information around this distribution would be useful in further informing adoption decisions, assuming certain topics are more valuable than others.

Finally, because this study was focused on informing an adoption decision of already existing and integrated automation tools, the costs of developing each tool were omitted. This study looked primarily at the ongoing cost and effectiveness (recall), rather than development costs. It would, however, be highly informative to

have this information for future groups to use in their own assessments and adoption decisions. Therefore, I suggest to future researchers to examine the relationships between up-front costs of development of automation or semi-automation tools for health evidence synthesis, ongoing costs, and ongoing effectiveness. In particular, given the results of this study have shown dominance of the adopted semi-automated workflow, it opens the door to determine the future point at which the semi-automated workflow is not only dominantly cost-effective, but retroactively dominant when research and development costs are included as well. In other words, at what point do the ongoing savings from a tool pay for its own initial development?

Contribution of analytical frameworks

As in previous chapters, it is helpful to examine the results of this study through the lenses of the three analytical frameworks laid out in Chapter 3. All three frameworks provide insight into the design and analysis of this study, though to varying degrees. The qualitative chapters of this thesis – Chapters 4 and 5 – relied more heavily on the trust framework than Chapters 6 and 7. From its inception, the design and analysis of this study was slanted more towards levels of automation and Diffusion of Innovations.

Trust can be viewed in an observational manner in this project. The map team sought to inform their adoption decision by seeking information about costs and effectiveness of semi-automated workflows. Viewed through the trust framework, the map team positioned their decision as one informed by external variability in situational trust. Recall from Chapter 3 that external variability includes workload and perceived risks and benefits. It could be argued that this project was established largely because of an increasing workload – with more and more evidence being produced related to COVID-19, the map team found themselves hard pressed to keep up with the literature. In addition, information about the risks and benefits was gained through this study; and, as it turned out, the risks were minimal and the benefits substantial, hence the adoption decision to move to the semi-automated workflow.

The applications of automation used in this study covered several levels of the automation taxonomy from Sheridan and Verplank [11, 12], generally at higher

levels than approaches covered in previous chapters. While the development of the automation tools used in this study, namely the MAG search algorithm, the binary ML classifier, and priority screening, were developed prior to its commencement, they each represent a level 10 automation: full computer control of a process. In the case of MAG searching, studies are retrieved from the MAG dataset and imported to ER-Web entirely automatically. Likewise with the ML classifier: this classification takes place without any human input. Finally, priority screening in ER-Web applies active learning to continuously re-order the studies as they are being screened, all without input from a human reviewer, nor informing the human reviewer that it has done so.

Finally, this study contributes significantly to the concept of relative advantage within the Diffusion of Innovations framework. When developing the protocol, it was expected that there would be some sacrifice in recall, or perhaps even a sacrifice in person-time if the automation tools tested proved to be poorly performing; this expectation was informed by previous academic literature [6, 13]. It was thus also expected that the map team would be making their adoption decision by weighing up the benefits against the costs (monetary or simply workload frustration) of the semi-automated workflow. The results of the study, however, indicating semi-automated approaches dominated: it was both more effective, and cheaper. Put in terms of Diffusion of Innovations, in this particular case there was no ‘relative’ advantage, but simply multiple advantages, and consequently an easy decision for the team to make.

Conclusion

This study is the first reported cost-effectiveness analysis of a semi-automated workflow used to maintain a living evidence map. The results demonstrate that a workflow using automated searches of Microsoft Academic Graph, a binary machine learning classifier, and an active learning prioritised screening list dominate the baseline manual workflow. More precisely, an automated, single-source database (in this case, MAG) showed higher recall than multi-database manual searches, and in addition machine learning tools improved search efficiency (i.e., saved costs). Further analyses demonstrated that this result is

not sensitive to changes in reviewer time-on-task efficiency. However, variations in precision did show a weakening of the initial conclusions, and MAG-enabled workflows did not dominate manual workflows when precision dipped below 0.61. These results support the use of MAG-enabled search methods, especially in combination with other ML tools.

Chapter references

1. Hutson, M. Artificial-Intelligence Tools Aim to Tame the Coronavirus Literature. *Nature* 2020:d41586-020.
2. Resources relating to Covid-19. 2021. <http://eppi.ioe.ac.uk/cms/Projects/DepartmentofHealthandSocialCare/Publishedreviews/COVID-19Livingssystematicmapofthevidence/COVID-19Resources/tabid/3767/Default.aspx>.
3. Lorenc, T., et al. COVID-19: living map of the evidence. 2020. http://eppi.ioe.ac.uk/COVID19_MAP/covid_map_v50.html.
4. Shemilt, I., et al. Using automation to produce a ‘living map’ of the COVID-19 research literature. *JEAHIL* 2021; 17(2):11-15.
5. Sinha, A., et al. An overview of microsoft academic service (mas) and applications. *Proceedings of the 24th international conference on world wide web*; 2015; 2015. p. 243-246.
6. Shemilt, I., et al. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews* 2016; 5(1):1-13.
7. Husereau, D., et al. Consolidated health economic evaluation reporting standards (CHEERS) statement. *Cost Effectiveness and Resource Allocation* 2013; 11(1):6.
8. Thomas, J., et al. EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. 2020.
9. 2020/21 UCL Non-clinical grade structure with spinal points. 2021. https://www.ucl.ac.uk/human-resources/sites/human-resources/files/2020-21_ucl_non-clinical_grade_structure_with_spinal_points.pdf (accessed 15 March 2020).
10. Academic Staff Annual Salary Rates. 2021. <https://www.monash.edu/enterprise-agreements/staff-salary-rates/academic> (accessed 15 March 2020).
11. Sheridan, T.B. *Telerobotics, automation, and human supervisory control*: MIT press; 1992.
12. Sheridan, T.B. and W.L. Verplank, *Human and computer control of undersea teleoperators*. 1978, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
13. Gartlehner, G., et al. Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. *Systematic Reviews* 2019; 8(1):277.

Chapter 8. Discussion

The new knowledge gained from this PhD

This thesis aimed to understand why automation is or is not adopted in health evidence synthesis, and what happens in practice when it is adopted. In addition to examining these questions, I also sought to test the applicability of several analytical frameworks, to assess their utility in the context of health evidence synthesis, and to recommend how they might be used in future research and practice.

As described in the introduction of this thesis, these results and analytical frameworks connect to form a novel, evidence-based roadmap for the adoption and implementation of automation in health evidence synthesis. My chosen frameworks interact in new and unique ways to establish this roadmap, which will be discussed in the following sections. The key results from my research will be used in structuring this process and described throughout the following sections. Because of the application of these frameworks and their role as a foundation for my conclusions, the results of this PhD can be translated more easily into other contexts, connected to existing academic literature, and further reinforced by applying and testing them in future studies.

The following sections include the main findings of my research, but it is helpful to summarise each of the individual conclusions from the preceding chapters first.

Health outcomes rely on high-quality health evidence, and this ‘high quality’ should require timely evidence synthesis using rigorous methodological standards. In surveying the current context of health evidence synthesis, however, it is clear that publication of research is outstripping the ability of present resources to keep up. Automation, as a possible solution to this situation, is being developed for many systematic review tasks, but is not widely used.

To begin examining potential barriers to the uptake of automation, I collected qualitative data from guideline developers. They perceived themselves as unfamiliar with automation tools and were conservative in their approach to applying automation. Most importantly, they identified cultural compatibility with the methodological standards of their field as the predominant factor influencing the acceptability of automation in health evidence synthesis.

Information specialists fitted well into the Diffusion of Innovations adoption curve, suggesting this framework is a useful approach in this context. Cochrane Information Specialists generally behaved in alignment with the behaviour expected from their identified adopter personas, and because of this I was able to identify specific strategies for appealing to each. More specifically, externally variable situational trust appeared to influence all personas, while dispositional trust was stronger in innovators and early adopters. Ease of use (i.e., complexity) increased in importance for personas on the latter half of the adoption curve.

In a clustered randomised trial of RobotReviewer assistance for Risk of Bias assessments, I demonstrated that machine learning assistance does not negatively impact the overall quality of RoB assessments. The results of the trial were insufficient to draw confident conclusions about the impact of ML assistance on the time required to complete an assessment.

Finally, the cost-effectiveness analysis showed that use of a single database created using automated web searches had higher recall than searches of multiple, publisher-driven databases. Moreover, the analysis showed that use of this single database in combination with other automation strategies resulted in a more effective (i.e., higher recall) search at a lower cost.

An evidence-based road map of automation in health evidence synthesis

Part 1: Values-based prerequisites

The first entry point on this pathway is that alignment with the values of the field of evidence-based medicine comes before all other priorities. Viewed through the lens of the Diffusion of Innovations framework [1], this is described as the innovation attribute of compatibility. Both of my qualitative chapters support this conclusion, and furthermore draw on the other thematic frameworks which tie this thesis together.

Chapter 4 presented the results of the thematic analysis of guideline developers' opinions towards automation. Above all else, they emphasised

compatibility with the values of the field and alignment with the values which underpin existing practices. When discussing compatibility, they concentrated more on the cultural aspects of this theme rather than on the practical ones. This represents a departure from the ideas put forth in previous literature. ICASR identified interoperability as an “urgent need” in 2019 [2]; my findings do not directly disprove this, but they do present empirical evidence of higher priorities. While interoperable systems are undoubtedly important to the eventual diffusion of automation in health evidence synthesis, the focus on compatibility is associated with a focus on professional values and not on the ability to translate automation data between existing infrastructures.

That guideline developers also highlighted a lack of familiarity with automation tools (whether perceived or real) reinforces the point that alignment with cultural values is a non-negotiable prerequisite for any further discussion of automation in the field of health evidence synthesis. The guideline developers who participated in my first research chapter presented their perspectives outside of an immediate adoption decision and offered up the broad contextual influence of a (perceived) lack of knowledge of automation. Because they view themselves in this way, their insights should be considered as existing around and throughout a hypothetical adoption curve (Figure 3.1), rather than at any particular point. Their concerns are both contextual, influencing the entire process of the diffusion of automation, should it occur, as well as a precondition for the process of diffusion to be considered at all.

When analysing the opinions of the participating guideline developers, Rogers’ innovation characteristics were applied as the deductive analytical framework, but these conclusions can also be situated within the Hoff and Bashir trust framework. Recall the three layers of trust in human-automation interactions: dispositional, situational, and learned [3]. The layer of trust that relates to cultural compatibility is situational trust, and more specifically in this case externally variable situational trust. This aspect includes the cultural expectations and factors outside of the potential user of automation that will influence that specific user’s trust. This description aligns well with the discussion of cultural compatibility: in

relation to a user of automation, the relevant cultural values are organisational (external), fluid (variable), and contextual (situational).

The results presented in Chapter 5 reinforce this argument. As highlighted in that chapter, Cochrane Information Specialists (CISs) are a useful case study because of their potentially advanced stage on the adoption curve for the Cochrane RCT classifier. The data collected from Cochrane Information Specialists (CISs) thus provided insights primarily in relation to the adoption curve, rather than information relating to conditions that precede it. However, interviewees repeatedly raised points which align with the conclusion that engaging with cultural expectations are a prerequisite for the adoption of automation for health evidence synthesis. Several participants described initially hearing about the RCT classifier from specific individuals within the Cochrane community. Such organisational influence can be categorised as another instance of externally variable situational trust, in which individuals are influenced to trust in automation due to the influence of peers and institutional structures. Furthermore, participants' descriptions of these organisational sources were not merely anecdotal recommendations. Rather, they described hearing from these colleagues the steps that had been taken to ensure the RCT classifier met the methodological standards of the CISs. Once again, proof of the cultural expectations of the population of interest, i.e., the methodological standards of CISs, was a threshold for any consideration of adoption of the automation tool.

To summarise these two aspects of the frameworks for this thesis, when considering adoption of automation for health evidence synthesis, cultural compatibility in terms of the Diffusion of Innovations framework and externally variable situational trust in the form of cultural values, norms, and expectations are minimum and seemingly non-negotiable entry requirements for any adoption or diffusion.

On the roadmap being drawn with my PhD results, cultural compatibility and situational trust are the gatekeepers to starting down the road.

Part 2: Strategies for innovators

Once the necessary conditions, as described above, for the beginning of the diffusion of automation are established, more targeted strategies can be taken, informed by the results of this thesis. Location on a potential adoption curve should first be determined or estimated, and then the personalised strategies described in the following sections can be used. These strategies represent the insights I found in the intersections of my chosen frameworks: adopter personas, innovation characteristics, and trust levels each play a role in the adoption curve, and each interact with one another.

Innovators are the first on the adoption curve and anticipated to account for 2.5% of adopters. They are a crucial first step for several reasons. First, they tend to provide useful feedback and refinement for research and development. Second, and more importantly, they act as influencers for later adopter categories; they are therefore an important population to onboard properly. They are also, according to my results, unique among the adopter personas.

First, from the perspective of the trust framework, innovators are most influenced by dispositional trust. That is, the way they approach trust in automation is inherent to their person, and not very likely to be influenced by external factors. An innovator is an innovator, and individuals (generally) are not made into innovators. The dispositional characteristics found in innovators include being non-risk averse and attracted to novelty. They are also self-confident with technological experimentation, which relates to the next aspect from the thematic frameworks which is important to innovators: complexity.

Innovators were unique among the participants of the project presented in Chapter 5 in that they did not identify ease of use (i.e., complexity) as a high priority in software selection. That every other adopter persona identified this innovation characteristic, but innovators did not, demands attention, and should be translated into practice in automation priorities. Given unlimited time and resources every software or automation tools would work perfectly and would be intuitive to use for every possible person; in reality, priorities in research and development must be set. The results of Chapter 5 show that initial priorities do not need to include a simple-

to-use tool and can instead rely on the confidence of innovators with new technologies.

In fact, they expressed a preference for the ability to personally test and manipulate a new tool, and to be given details in relation to how it is working. These preferences could be examined through the Diffusion of Innovations characteristics either as *observability* or as *trialability*. They wanted to see how the machine learning (ML) algorithm was working, fitting more closely with *observability*, but they also wanted to experiment with this algorithm, fitting more closely with *trialability*. Unlike the innovator participants in Chapter 5, *trialability* did not significantly appear in the results of guideline developer participants in Chapter 4. However, when combined with the results of the CISs, further conclusions about the Chapter 4 participants might be drawn. Recall that guideline developers indicated two layers of ‘double-checking’: *ability to double-check (compatibility)*, and *personal need for double-checking (observability)*. CISs who indicated that they personally double-checked the results of the classifier were in the innovator category, so it is possible that the guideline developers who indicated this need might also be innovators. In addition, it seems likely that guideline developers would be encouraged by the hands-on approach of innovators given that the former felt more positively about automation when someone else (in this case, innovator CISs) had the ability to transparently examine ML-produced results. In short, innovators may play a role not only in encouraging other adopters on the benefits of automation, but in demonstrating automation’s trustworthiness to external stakeholders such as guideline developers.

To summarise, when targeting innovators, automation proponents and/or developers should provide hands-on technical functionality for users, even if this comes at a cost of ease of use. Innovators’ use of these transparent features also impacts on the perception of non-user stakeholders. As diffusion continues, however, lowering the prioritisation of ease of use will not always be the best route to wider adoption.

Part 3: Situational trust returns

Innovators are unique in that they do not need to be convinced of a need to change. This likely contributed to the empirical data observed in their automation tool preferences. Other adopter personas, however, do not share this characteristic. Situational trust has already been established as a contextual prerequisite for the consideration of automation, but it becomes more specific as a road marker for the other adopter personas at this stage.

Participants in Chapter 5 highlighted organisational validation, as has already been discussed. However, in addition to it influencing the initial adoption decision, participants in the middle of the adoption curve indicated it also influences their trust levels after they have adopted a tool. They favour continued support from organisational structures to inform their use of a tool (e.g., in the case of my research, the Cochrane RCT classifier as integrated into the Screen4Me workflow, a Cochrane-endorsed workflow which uses ML as well as crowdsourcing to complete study selection).

The participants in my research cited at least two ways that this situational trust is structured. First, they cited the validation that organisations, and individuals within an organisation, perform. Rather than personally test threshold scores, for example, as the innovators did, middle-curve adopters founded their trust on the recommendations and documentation from their networks (both peers and organisational). As it is derived externally from the individual user, this aligns with externally variable situational trust in the Hoff and Bashir three-layered trust model.

A second aspect of externally variable situational trust was also cited: a demanding workload. Participants in my research indicated feeling a necessity to adopt automation due to an increasingly overwhelming set of tasks accruing in their backlog. That my final PhD year coincided with the COVID-19 pandemic only reinforced this observation. After many in the field spent years being reluctant to integrate automation systems to any great extent, a significant number of automated or partially automated health evidence syntheses suddenly appeared throughout 2020 [4]. Having already observed this in my own results when CISs were prompted to adopt Screen4Me due to an overwhelming pace of work, it seems that this sudden switch in adoption attitudes may have been influenced by the urgent nature of the

pandemic. Once again, necessity dictated an adoption decision with respect to automation in health evidence syntheses. Such a circumstance, like organisational endorsement of a tool, is situational, and it is external to the individual user.

To summarise, at the transition from innovators into the middle parts of the adoption curve, externally variable situational trust holds the greatest influence. On the constructed 'road map', the best way to continue a path after initial adoption by innovators is to focus on two signposts: organisational endorsement and ongoing validation, and on establishing a context in which automation is seen as a necessity due to workload or an otherwise pressing situation.

Part 4: Complexity

Thus far, I have combined my analytical frameworks to provide insights as to how to initiate diffusion of automation in health evidence synthesis, how to target innovators, and how to reach the middle of the adoption curve (Figure 3.1). Focus will now shift to the latter half of the adoption curve, and into the post-adoption realm to examine the effects of automation in practice. The previous sections and strategies would set the conditions for adoption by convincing later adopters of the need to change. To best support them in actioning this change, complexity becomes the dominant theme.

The previous sections established that innovators appreciate technological advancement and transparency and do not share the focus on ease of use observed in all other adopter categories. This focus was notable because it was shared across every adopter persona except for innovators; clearly, it is of significant reach and importance. Later categories not only contrasted against innovators in this focus, but their results also contrast against those provided by guideline developers, who provided very little qualitative data related to complexity. These results as a whole indicate that complexity is significant, but as a precondition of automation adoption, and not to innovators; complexity is therefore most suitably targeted to the middle and later portions of the adoption curve.

When examined through the trust framework, the results from middle and later adopters continue to differ from those described previously. Perceptions of a tool's complexity can be classified as learned trust; recall from Chapter 3 that

specific characteristics of a tool are categorised under learned trust. In this case, the specific characteristic which receives attention is the complexity, or lack of it, in an automation tool.

Middle and late-curve participants were very clear that a tool must be simple to learn and to use in order to be considered. However, the late majority participant reinforced this further by speaking at length about the impact of a poor user experience with a tool. Because these personas focus on complexity, but the later personas focused on it even more, it can be inferred that complexity increases in importance over the course of an automation's diffusion.

To summarise, once the previous strategies have succeeded, middle and later adopters are primed to be supported in changing practice, having now been convinced of the need to do so. To best support them in this change, learned trust and lack of complexity are the highest priorities.

Part 5: Post-adoption

With the process of how to most effectively encourage and support adoption of automation in health evidence synthesis described above, focus can now shift to the effects of automation in practice.

Previous academic commentary has focused on semi-automation as a strategy to encourage adoption. This appears to have multiple assumptions: first, that automation will be more acceptable to users if they retain some control, and second, that automation will perform better when paired with human effort. Empirically, according to the results of my thesis, the first assumption is true. The second assumption should be adjusted according to the evidence found in this thesis, however. Every application researched in this thesis represents a semi-automated workflow, and yet they differed greatly in their positions along the levels of automation (LOA) framework [5, 6]. Higher levels of automation, as defined by the levels of automation framework from Sheridan and Verplank (Table 3.1), seemed to have greater success, while still maintaining an overall semi-automated workflow in the broader context of the relevant systematic review. The most successful semi-automation, according to my results, uses high levels of automation but limits this use to well-defined and discrete tasks.

The first example of this in my results is the Screen4Me workflow experienced by the Cochrane Information Specialists. Initially, as highlighted by the innovator participants, the underlying algorithms of this workflow were more transparent and subject to human intervention – approximately a level 2 automation. However, Screen4Me no longer displays the ML-applied scores, and now undertakes bulk actions with the permission of the user – approximately a level 5 automation. This level 5 approach is relatively successful among its target users (CISs) but supports a very well-defined task in the screening stage of a review. In addition, individuals earlier on the adoption curve appeared to implement automation, namely the RCT classifier, in a more automated way, accepting its decisions in bulk and tending not to cross-check results. This was also observed in correlation with trust in the classifier; individuals using the classifier in a way that was higher on the LOA spectrum tended to also trust in the classifier more.

The results of Chapter 7 provide further examples of higher automation providing more benefits. As noted in the discussion of that chapter, Microsoft Academic Graph (MAG), the binary ML classifier, and priority screening mode are all level 10 automations. Their benefits were undeniable in the results of the cost-effectiveness analysis: these highly automated and targeted tools dominated the comparator. MAG identified more eligible studies than the baseline MEDLINE-Embase search strategies and did so at a lower resource cost. In combination with the binary ML classifier and with active learning, the automated workflow continued to improve upon effectiveness while decreasing costs. Automation was more effective than the non-automated workflow, and in this case, the automation of interest was level 10 on the Sheridan and Verplank classification scale.

Though broadly my results support the recommendations described in the previous paragraphs, it should be noted that the results of the RobotReviewer randomised trial presented in Chapter 6 neither support nor refute this proposition. As described elsewhere in this thesis, RobotReviewer's level of automation is dependent on user action: it can be either a level 4 or a level 5 under its current design (i.e., it requires some user input), but from a technical capabilities perspective there is no hindrance to its application at higher levels of automation. With the context of the other results found in my research, the question should be raised in

future research as to whether RobotReviewer would be better implemented at these higher levels.

To summarise this final part of the new roadmap, when implementing automation, the evidence presented in this thesis suggests automation is most effective when a relatively higher level of automation is targeted at a well-defined, discrete task. This is a significant departure from previous literature, and consequently a significant contribution of the findings of this PhD.

Implications for research and practice

Chapter 3 discussed several reasons for the use of my chosen analytical frameworks. These reasons were: in order to better describe and communicate my findings, to locate them within a wider field of academic research, to test the effectiveness of my approach, and most importantly to maximise the ability of my findings to translate across multiple contexts. By using connections between my chosen frameworks to construct the adoption and implementation roadmap described in this chapter, I have accomplished each of these goals. The frameworks structured discussion of my results, connected them to existing literature, and created an evidence-driven guide that can be picked up by future researchers and stakeholders. I have therefore shown the effectiveness of this approach and encourage its use in future research in this area.

Several of the projects presented in this thesis, in addition to testing the analytical approach, used novel study designs. The cost-effectiveness analysis presented in Chapter 7 was based on a single prior publication [7], and now provides a further example, which also follows standard reporting guidelines, of how to conduct these analyses when examining automation in systematic reviews or evidence maps. Cost-effectiveness is an important metric in relation to the relative advantage of an innovation, and it is hoped that by disseminating these methods and results that more will be conducted in the future. The findings of this study are further strengthened by examining a real-world application of the tools examined; finding cost-effective dominance in constructed and controlled settings would have lacked the impact and generalisability of these findings. However, future research would also do well to use my research as the basis for improvement to cost-

effectiveness analyses. Self-reported outcomes could be improved with other methods of measurement, as discussed in Chapter 7, and empirical data could be used to improve the analytical assumptions incorporated into future cost-effectiveness analyses.

The randomised trial of RobotReviewer in combination with human effort presents several methodological innovations. It used a novel methodology which allowed for single blinding of the outcome assessor (the consensus reviewer) while maintaining the standard workflow of two individual, blinded reviewers. In addition, it was one of the first effectiveness (as opposed to efficacy) studies for automation of quality appraisal, and also one of the first to look at semi-automation using RobotReviewer rather than a direct comparison of automation and human results. Like the cost-effectiveness results described above, using this approach offers significant advantages for generalisability and impact of the results as compared with an efficacy study. This approach was not without its challenges, and future researchers can and should draw lessons from this trial. The results of this trial found that providing reviewers with suggestions from RobotReviewer did not negatively impact the quality of their finalised Risk of Bias assessments. Given that this was a previous concern [8], this is a significant finding.

In contributing to the evidence base for automation in health evidence synthesis, my research also raised new questions which should be addressed with future research. The evidence-based roadmap presented in the preceding sections is largely informed by observational and correlational data. While these support my interpretations, causation cannot be concluded from the majority of my findings. Causation studies should be pursued to determine if the hypotheses drawn from these results withstand additional scrutiny.

As could be inferred from the evidence-based roadmap detailed in the previous sections, the findings of my research should suggest to automation tool developers that they should begin their research and development with two primary questions: first, they should establish the cultural compatibility of their tool with the standards of evidence-based medicine professionals. It is clear from my findings that without such compatibility, and other demonstrations of relative advantage, interoperability, or (lack of) complexity are likely to be insufficient to encourage

user adoption. Second, they should seek to determine where on the adoption curve the field currently resides in relation to the specific piece of the evidence pipeline they are aiming to automate. Strategies to encourage adoption are shown in my results to shift over the course of an innovation's diffusion; what works well at one stage and for one adopter category might not be the best strategy at another stage and for another adopter category.

As noted in the previous section, more highly automated processes appeared to be more successful than less automated processes, as defined on the Sheridan and Verplank levels of automation framework. These results are potentially context-specific, however, and merit further investigation in other settings.

Finally, results from this thesis also provide insight for practice in relation to the specific tools analysed in Chapter 6 and Chapter 7. The randomised trial of RobotReviewer, the first of its kind, demonstrated the non-inferiority of integrating RobotReviewer suggestions to Risk of Bias assessments. It therefore supports the adoption of this tool in practice because it demonstrates a lack of risk to the resulting systematic review's quality. The cost-effectiveness analysis looked at the use of Microsoft Academic Graph (MAG), as well as a machine learning classifier to determine eligibility, and an active learning algorithm to prioritise screening; it found the workflows using MAG in combination with ML tools dominated the comparator of manual searches of Medline and Embase. Perhaps as significantly, it found that MAG identified more eligible studies for the living Covid-19 evidence map. This not only supports the adoption of MAG, but also aligns with the cultural expectations that were found throughout the first half of this thesis. That is, use of MAG not only has a demonstrated relative advantage, but it aligns with the compatibility theme (i.e., cultural situational trust) expectations found in this PhD which intensely prioritise recall. These results therefore support wider adoption of MAG, or similar tools, and are encouraging of the shift away from publisher-reliant databases towards single-source, automated databases.

Conclusions

My PhD aimed to explore the adoption and the effectiveness of automation technologies in health evidence synthesis. I sought to determine why individuals, teams, or organisations choose to adopt automation technologies, and what happens if they do adopt automation. To achieve this aim, I applied a novel combination of several analytical frameworks: Rogers' Diffusion of Innovations [1], Hoff and Bashir's human-automation trust [3], and Sheridan and Verplank's levels of automation [5, 6]. These frameworks were used to collect qualitative information which then informed an evidence-based roadmap of the necessary conditions for the acceptance and uptake of automation in health evidence synthesis. Next, I used quantitative methods to evaluate the effectiveness of automation in two specific contexts and found automation in both cases to either maintain the same quality as human effort, or to improve upon it. With these findings, my approach was successful and has contributed to the academic literature an evidence-based and structured guide to encouraging the adoption and supporting the use of automation in health evidence synthesis.

This thesis has shown that values-based professional expectations are the first prerequisite in initiating the diffusion of automation among health evidence professionals. Once automation is able to demonstrate this alignment, innovators can be targeted by focusing on hands-on technical transparency and capabilities. With innovators on board, situational trust contributes to middle-curve adoption by way of organisational endorsement and support. Complexity, or lack thereof, in automation tools becomes increasingly important over time, and is therefore of significant importance to late-curve adopters. Finally, when implemented, higher levels of automation are more successful, but work best when targeted at a well-defined and discrete task.

On the theme of effectiveness, my results endorse the adoption of automation for specific tasks as they do not negatively impact health evidence quality. With this structured guide to the promotion of automation, as well as novel examples of its effectiveness, it is hoped that this PhD will support the field's ability to develop, evaluate and implement automation in health evidence synthesis.

Chapter references

1. Rogers, E.M. *Diffusion of Innovations*. 5th ed: Simon and Schuster; 2003.
2. O'Connor, A.M., et al. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Systematic Reviews* 2019; 8(1):57.
3. Hoff, K.A. and M. Bashir. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 2014; 57(3):407-434.
4. Resources relating to Covid-19. 2021.
<http://eppi.ioe.ac.uk/cms/Projects/DepartmentofHealthandSocialCare/Publishedreviews/COVID-19Livingssystematicmapofthevidence/COVID-19Resources/tabid/3767/Default.aspx>.
5. Sheridan, T.B. *Telerobotics, automation, and human supervisory control*: MIT press; 1992.
6. Sheridan, T.B. and W.L. Verplank, *Human and computer control of undersea teleoperators*. 1978, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
7. Shemilt, I., et al. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews* 2016; 5(1):1-13.
8. Soboczenski, F., et al. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. *BMC Medical Informatics and Decision Making* 2019; 19(1):96.

Appendix A. COREQ checklist

COREQ (Consolidated criteria for REporting Qualitative research) checklist

Topic	Item No.	Guide Questions / Description	Reported on Page No.
Domain 1: Research team and reflexivity			
<i>Personal characteristics</i>			
Interviewer / facilitator	1	Which author/s conducted the interview or focus group?	78
Credentials	2	What were the researcher's credentials? E.g. PhD, MD	78
Occupation	3	What was their occupation at the time of the study?	78
Gender	4	Was the researcher male or female?	18
Experience and training	5	What experience or training did the researcher have?	78, 19-20
<i>Relationship with participants</i>			
Relationship established	6	Was a relationship established prior to study commencement?	78
Participant knowledge of the interviewer	7	What did the participants know about the researcher? e.g. personal goals, reasons for doing the research	78
Interviewer characteristics	8	What characteristics were reported about the interviewer/facilitator? e.g. Bias, assumptions, reasons and interests in the research topic	19-20
Domain 2: Study design			
<i>Theoretical framework</i>			
Methodological orientation and Theory	9	What methodological orientation was stated to underpin the study? e.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis	79-81
<i>Participant selection</i>			
Sampling	10	How were participants selected? e.g. purposive, convenience, consecutive, snowball	78
Method of approach	11	How were participants approached? e.g. face-to-face, telephone, mail, email	78
Sample size	12	How many participants were in the study?	81
Non-participation	13	How many people refused to participate or dropped out? Reasons?	81
<i>Setting</i>			
Setting of data collection	14	Where was the data collected? e.g. home, clinic, workplace	78
Presence of non-participants	15	Was anyone else present besides the participants and researchers?	78
Description of sample	16	What are the important characteristics of the sample? e.g. demographic data, date	81
<i>Data collection</i>			
Interview guide	17	Were questions, prompts, guides provided by the authors? Was it pilot tested?	78
Repeat interviews	18	Were repeat interviews carried out? If yes, how many?	81
Audio / visual recording	19	Did the research use audio or visual recording to collect the data?	78
Field notes	20	Were field notes made during and/or after the interview or focus group?	n/a

Duration	21	What was the duration of the inter views or focus group?	81
Data saturation	22	Was data saturation discussed?	n/a
Transcripts returned	23	Were transcripts returned to participants for comment and/or correction?	79
Domain 3: analysis and findings			
<i>Data analysis</i>			
Number of data coders	24	How many data coders coded the data?	79
Description of the coding tree	25	Did authors provide a description of the coding tree?	79-81
Derivation of themes	26	Were themes identified in advance or derived from the data?	79-81
Software	27	What software, if applicable, was used to manage the data?	79
Participant checking	28	Did participants provide feedback on the findings?	n/a
<i>Reporting</i>			
Quotations presented	29	Were participant quotations presented to illustrate the themes/findings? Was each quotation identified? e.g. participant number	82-91
Data and findings consistent	30	Was there consistency between the data presented and the findings?	82-91
Clarity of major themes	31	Were major themes clearly presented in the findings?	82-91
Clarity of minor themes	32	Is there a description of diverse cases or discussion of minor themes?	82-91

Developed from: Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*. 2007. Volume 19, Number 6: pp. 349 – 357

Appendix B. Interview instrument

Screening: Can you describe your experience and role in guideline development?

- Must be involved in the decision-making process of whether a piece of evidence is accepted into the guideline or not. If participant is determined early in the interview to not fit this criterion, interview will be wrapped up.
1. What do you view as the goal of guideline development?
 2. What do you view as the end goal of evidence synthesis (including systematic reviews)?
 3. What method do you or does your organisation use to collect evidence?
 4. Are there any limitations on the type of evidence you will accept?
 5. What do you think are the best methods for production of evidence?
 6. What is your opinion of machine learning in relation to evidence synthesis?
 7. What is your opinion of crowdsourcing in relation to evidence synthesis?
 8. What do you feel the opinion of the guideline developer community as a whole is, in relation to machine learning and crowd?
 9. Would you be more likely or less likely to accept the conclusions of a systematic review which had used automation in its protocol?
 10. What factors might influence this decision?
 11. In your ideal world, what would the future look like in terms of use of automation for evidence synthesis? What sorts of capabilities would these technologies have, and what would they not have?

Appendix C. CROSS checklist

Checklist for Reporting of Survey Studies

Section/topic	Item	Item description	Reported on page #
Title and abstract			
Title and abstract	1a	State the word “survey” along with a commonly used term in title or abstract to introduce the study’s design.	108, 111
	1b	Provide an informative summary in the abstract, covering background, objectives, methods, findings/results, interpretation/discussion, and conclusions.	108
Introduction			
Background	2	Provide a background about the rationale of study, what has been previously done, and why this survey is needed.	109-111
Purpose/aim	3	Identify specific purposes, aims, goals, or objectives of the study.	112
Methods			
Study design	4	Specify the study design in the methods section with a commonly used term (e.g., cross-sectional or longitudinal).	113
	5a	Describe the questionnaire (e.g., number of sections, number of questions, number and names of instruments used).	113
Data collection methods	5b	Describe all questionnaire instruments that were used in the survey to measure particular concepts. Report target population, reported validity and reliability information, scoring/classification procedure, and reference links (if any).	112-114
	5c	Provide information on pretesting of the questionnaire, if performed (in the article or in an online supplement). Report the method of pretesting, number of times questionnaire was pre-tested, number and demographics of participants used for pretesting, and the level of similarity of demographics between pre-testing participants and sample population.	114
	5d	Questionnaire if possible, should be fully provided (in the article, or as appendices or as an online supplement).	208
Sample characteristics	6a	Describe the study population (i.e., background, locations, eligibility criteria for participant inclusion in survey, exclusion criteria).	108, 114
	6b	Describe the sampling techniques used (e.g., single stage or multistage sampling, simple random sampling, stratified sampling, cluster sampling, convenience sampling). Specify the locations of sample participants whenever clustered sampling was applied.	114
	6c	Provide information on sample size, along with details of sample size calculation.	114
	6d	Describe how representative the sample is of the study population (or target population if possible), particularly for population-based surveys.	114, 116
Survey administration	7a	Provide information on modes of questionnaire administration, including the type and number of contacts, the location where the survey was conducted (e.g., outpatient room or by use of online tools, such as SurveyMonkey).	113, 116
	7b	Provide information of survey’s time frame, such as periods of recruitment, exposure, and follow-up days.	115

	7c	Provide information on the entry process: ->For non-web-based surveys, provide approaches to minimize human error in data entry. ->For web-based surveys, provide approaches to prevent “multiple participation” of participants.	113
Study preparation	8	Describe any preparation process before conducting the survey (e.g., interviewers’ training process, advertising the survey).	
Ethical considerations	9a	Provide information on ethical approval for the survey if obtained, including informed consent, institutional review board [IRB] approval, Helsinki declaration, and good clinical practice [GCP] declaration (as appropriate).	117
	9b	Provide information about survey anonymity and confidentiality and describe what mechanisms were used to protect unauthorized access.	114
Statistical analysis	10a	Describe statistical methods and analytical approach. Report the statistical software that was used for data analysis.	n/a
	10b	Report any modification of variables used in the analysis, along with reference (if available).	n/a
	10c	Report details about how missing data was handled. Include rate of missing items, missing data mechanism (i.e., missing completely at random [MCAR], missing at random [MAR] or missing not at random [MNAR]) and methods used to deal with missing data (e.g., multiple imputation).	113
	10d	State how non-response error was addressed.	113
	10e	For longitudinal surveys, state how loss to follow-up was addressed.	n/a
	10f	Indicate whether any methods such as weighting of items or propensity scores have been used to adjust for non-representativeness of the sample.	n/a
	10g	Describe any sensitivity analysis conducted.	n/a

Results

Respondent characteristics	11a	Report numbers of individuals at each stage of the study. Consider using a flow diagram, if possible.	117
	11b	Provide reasons for non-participation at each stage, if possible.	n/a
	11c	Report response rate, present the definition of response rate or the formula used to calculate response rate.	117
	11d	Provide information to define how unique visitors are determined. Report number of unique visitors along with relevant proportions (e.g., view proportion, participation proportion, completion proportion).	113
Descriptive results	12	Provide characteristics of study participants, as well as information on potential confounders and assessed outcomes.	118
Main findings	13a	Give unadjusted estimates and, if applicable, confounder-adjusted estimates along with 95% confidence intervals and p-values.	n/a
	13b	For multivariable analysis, provide information on the model building process, model fit statistics, and model assumptions (as appropriate).	n/a
	13c	Provide details about any sensitivity analysis performed. If there are considerable amount of missing data, report sensitivity analyses comparing the results of complete cases with that of the imputed dataset (if possible).	n/a

Discussion			
Limitations	14	Discuss the limitations of the study, considering sources of potential biases and imprecisions, such as non-representativeness of sample, study design, important uncontrolled confounders.	129
Interpretations	15	Give a cautious overall interpretation of results, based on potential biases and imprecisions and suggest areas for future research.	124-127
Generalisability	16	Discuss the external validity of the results.	129
Other sections			
Role of funding source	17	State whether any funding organization has had any roles in the survey's design, implementation, and analysis.	n/a
Conflict of interest	18	Declare any potential conflict of interest.	n/a
Acknowledgements	19	Provide names of organisations/persons that are acknowledged along with their contribution to the research.	114

Appendix D. Survey instrument

Note: this survey was originally conducted online via Google Forms; this print-out version is for reference purposes only

Technology adoption among Cochrane Information Specialists

Thank you for responding to our call for participants in this survey. The purpose of this survey is to describe and to understand the current adoption of the Cochrane RCT Classifier, and technology decision-making processes in general among CIS's. There are no right or wrong answers! Results of this survey will be disseminated to participants, and will be submitted for peer-reviewed publication.

This survey may be completed anonymously if preferred. Any identifiable information will be anonymised for all communications external to the research teams. Data will be retained in accordance with University College London Guidance, including General Data Protection Regulation (GDPR) Compliance.

The complete survey will take less than 10 minutes to complete. You are free to withdraw at any time.

If you wish to withdraw after you've submitted your answers, or for any other queries, please contact anneliese.arno.17@ucl.ac.uk.

* Required

1. I have read the above information and would like to complete the survey. *

Mark only one oval.

- Yes *Skip to question 2*
 No *Skip to question 23*

**Starter
questions**

This section will ask you a few questions to streamline the rest of your survey.

2. Approximately how long have you worked in information science? *

Mark only one oval.

- Prefer not to say
- Less than a year
- Between 1 and 5 years
- Between 5 and 10 years
- Between 10 and 20 years
- More than 20 years

3. Approximately how long have you worked as a Cochrane Information Specialist? *

Mark only one oval.

- Prefer not to say
- Less than a year
- Between 1 and 5 years
- Between 5 and 10 years
- Between 10 and 20 years
- More than 20 years

4. With which Cochrane Review Group do you primarily work? *

Mark only one oval.

- Prefer not to say
- Acute Respiratory Infections
- Airways
- Anaesthesia
- Back and Neck
- Bone, Joint and Muscle Trauma
- Breast Cancer
- Childhood Cancer
- Colorectal
- Common Mental Disorders
- Consumers and Communication
- Cystic Fibrosis and Genetic Disorders
- Dementia and Cognitive Improvement
- Developmental, Psychosocial and Learning Problems
- Drugs and Alcohol
- Effective Practice and Organisation (EPOC)
- Emergency and Critical Care
- Ear, Nose and Throat
- Epilepsy
- Eyes and Vision
- Fertility Regulation
- Gynaecological, Neuro-oncology and Orphan Cancers
- Gynaecology and Fertility
- Haematologica
- Heart
- Hepato-Biliary
- HIV/AIDS
- Hypertension
- Incontinence
- Infectious Diseases
- Inflammatory Bowel Disease
- Injuries

- Kidney and Transplant
- Lung Cancer
- Metabolic and Endocrine Disorders
- Methodology Review
- Movement Disorders
- Multiple Sclerosis and Rare Diseases of the CNS
- Musculoskeletal
- Neonatal
- Neuromuscular
- Oral Health
- Pain, Palliative and Supportive Care
- Pregnancy and Childbirth
- Public Health
- Schizophrenia
- Sexually Transmitted Infections
- Skin
- Stroke
- Tobacco Addiction
- Upper GI and Pancreatic Diseases
- Urology
- Vascular
- Work
- Wounds
- Other

5. Are you aware of the Cochrane RCT Classifier? *

Mark only one oval.

- Yes
- No

6. Have you used the Cochrane RCT Classifier? *

Mark only one oval.

Yes *Skip to question 7*

No *Skip to question 14*

RCT Classifier Users

7. Around when did you first try the RCT Classifier? *

8. Which statement most closely aligns with your use of the RCT Classifier? note: we are interested in all usage contexts (eg. review production, register maintenance, or any others) *

Mark only one oval.

I tried it once or twice, and did not use it again

I use it occasionally

I use it sometimes

I use it frequently

9. Do you use CRS Web to manage your references and/or your studies? *

Mark only one oval.

Yes

Sometimes

No

10. When (approximately) did you first start using CRS Web regularly? (please leave blank if you do not use CRS Web)

11. What are your main considerations in selecting a study or reference management workflow / tool? *

12. Which statement most closely aligns with your use of the RCT Classifier? *

Mark only one oval.

- I use it, but don't really pay attention to its results
- I assess the study/trial first, and then check with the Classifier
- I filter by the Classifier results, and then assess manually as a double-check
- I filter by the Classifier results and move studies in bulk based on the Classifier results
- Other: _____

13. What is your overall opinion of the RCT Classifier? *

Mark only one oval.

	1	2	3	4	5	
Hate it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Love it

[Skip to question 17](#)

RCT Classifier non-Users

14. Do you use CRS Web to manage your references and/or your studies? *

Mark only one oval.

- Yes
- Sometimes
- No

15. When (approximately) did you first start using CRS Web regularly? (please leave blank if you do not use CRS Web)

16. What are your main considerations in selecting a study management workflow / tool?

Skip to question 17

Technology preferences

17. *Mark only one oval per row.*

	Agree	Disagree
I know people who work in software development	<input type="radio"/>	<input type="radio"/>
Colleagues often come to me for advice about technology	<input type="radio"/>	<input type="radio"/>
When I try out a new technology, I like to tell my peers about my experience	<input type="radio"/>	<input type="radio"/>
I don't usually have informal conversations with colleagues about technology or software.	<input type="radio"/>	<input type="radio"/>
I am open to taking risks with my workflow, and feel these can result in benefits	<input type="radio"/>	<input type="radio"/>
I see myself as an opinion leader among Information Specialists	<input type="radio"/>	<input type="radio"/>
I tend to adopt new technologies around the same time as my peers	<input type="radio"/>	<input type="radio"/>
I tend to adopt new technologies before my peers	<input type="radio"/>	<input type="radio"/>
I tend to adopt new technologies far before my peers	<input type="radio"/>	<input type="radio"/>
I tend to adopt new technologies once most of my peers have changed over	<input type="radio"/>	<input type="radio"/>
I tend to think time-tested workflows are preferable to new technologies	<input type="radio"/>	<input type="radio"/>
I think people tend to rush in to adopting a new technology	<input type="radio"/>	<input type="radio"/>
I think people are excessively hesitant about trying machine-learning	<input type="radio"/>	<input type="radio"/>

18. When adopting a new technology, which statement sounds most like you?

Mark only one oval.

- I like to learn through using it
- I like to learn through a taught workshop
- Other: _____

19. When adopting a new technology, which statement sounds most like you?

Mark only one oval.

- I like to try things out -- if it doesn't work then I'll ditch it later
- I like to be very cautious -- it's important to carefully consider all angles
- Other: _____

20. When adopting a new technology, which statement sounds more like you?

Mark only one oval.

- I like to adopt improvements as soon as possible, and share my experience with my peers
- I like to stick to what I know until most people I know have tried it and can tell me about their experiences
- Other: _____

21. Which of the following best describes your current opinion of machine-learning (ML) for systematic reviews?

Mark only one oval.

- Excited: ML is the future of systematic reviewing
- Optimistic: ML is promising, and I'd like to start using it (if I'm not already)
- Open-minded: I haven't formed an opinion yet, and want to see what the evidence says
- Cautious: I'm sceptical of ML, and think more people need to try it out before I can form an opinion
- Concerned: I'm worried about the adoption of ML, and think it might be ill-advised

22. My decision to adopt a new technology is influenced by:

Mark only one oval per row.

	Not at all	Somewhat	A lot
What my peers are using	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My previous experience	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Recommendations made by organisations such as Cochrane, NICE, etc.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
What people I manage like to use	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My "gut" feeling	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Previous evidence I've seen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Recommendations from peers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
What's new on the scene	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
What was used in my training (eg. university modules, workshops, etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Availability of information about how it works	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Availability of instructional information	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
What is most established or vetted	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Thank
you!

Thank you for your participation in this survey.

If you'd like to hear more about this research, or just have a chat, please contact anneliese.arno.17@ucl.ac.uk.

23. If you are comfortable to be contacted for follow-up questions, please provide your email below.

This content is neither created nor endorsed by Google.

Google Forms

Appendix E. CONSORT checklists

CONSORT 2010 extension

Checklist of information to include when reporting a cluster randomised trial

Section/Topic	Item No	Standard Checklist item	Extension for cluster designs	Page No.
Title and abstract				
	1a	Identification as a randomised trial in the title	Identification as a cluster randomised trial in the title	132
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	See table 2	n/a
Introduction				
Background and objectives	2a	Scientific background and explanation of rationale	Rationale for using a cluster design	133-137
	2b	Specific objectives or hypotheses	Whether objectives pertain to the cluster level, the individual participant level or both	138-139
Methods				
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	Definition of cluster and description of how the design features apply to the clusters	138
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons		140
Participants	4a	Eligibility criteria for participants	Eligibility criteria for clusters	139
	4b	Settings and locations where the data were collected		141
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	Whether interventions pertain to the cluster level, the individual participant level or both	138, 141-142
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	Whether outcome measures pertain to the cluster level, the individual participant level or both	143
	6b	Any changes to trial outcomes after the trial commenced, with reasons		n/a
Sample size	7a	How sample size was determined	Method of calculation, number of clusters(s) (and whether equal or unequal cluster sizes are assumed), cluster size, a coefficient of intracluster correlation (ICC or k), and an indication of its uncertainty	144

	7b	When applicable, explanation of any interim analyses and stopping guidelines		n/a
Randomisation:				
Sequence generation	8a	Method used to generate the random allocation sequence		141
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	Details of stratification or matching if used	141
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	Specification that allocation was based on clusters rather than individuals and whether allocation concealment (if any) was at the cluster level, the individual participant level or both	142
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	Replace by 10a, 10b and 10c	141
	10a		Who generated the random allocation sequence, who enrolled clusters, and who assigned clusters to interventions	141
	10b		Mechanism by which individual participants were included in clusters for the purposes of the trial (such as complete enumeration, random sampling)	n/a; clusters contained two individual reviewers and one consensus reviewer in accordance with the eligibility criteria
	10c		From whom consent was sought (representatives of the cluster, or individual cluster members, or both), and whether consent was sought before or after randomisation	140
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how		142
	11b	If relevant, description of the similarity of interventions		142
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	How clustering was taken into account	146

	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses		n/a
Results				
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	For each group, the numbers of clusters that were randomly assigned, received intended treatment, and were analysed for the primary outcome	148
	13b	For each group, losses and exclusions after randomisation, together with reasons	For each group, losses and exclusions for both clusters and individual cluster members	148
Recruitment	14a	Dates defining the periods of recruitment and follow-up		147
	14b	Why the trial ended or was stopped		
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	Baseline characteristics for the individual and cluster levels as applicable for each group	147; topic areas of reviews are included
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	For each group, number of clusters included in each analysis	148
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	Results at the individual or cluster level as applicable and a coefficient of intracluster correlation (ICC or k) for each primary outcome	149, 151
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended		150
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory		n/a
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)		147
Discussion				
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses		154
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	Generalisability to clusters and/or individual participants (as relevant)	152

Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	151-156
Other information			
Registration	23	Registration number and name of trial registry	147
Protocol	24	Where the full trial protocol can be accessed, if available	227
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	147

CONSORT 2006 extension

Checklist for Non-inferiority and Equivalence Trials

PAPER SECTION And topic	Item	Descriptor	Reported on Page No.
TITLE & ABSTRACT	1	How participants were allocated to interventions (e.g., "random allocation", "randomised", or "randomly assigned"), <i>Specifying that the trial is a non-inferiority or equivalence trial.</i>	141
INTRODUCTION Background	2	Scientific background and explanation of rationale, <i>Including the rationale for using a non-inferiority or equivalence design.</i>	133-137
METHODS Participants	3	Eligibility criteria for participants (<i>detailing whether participants in the non-inferiority or equivalence trial are similar to those in any trial(s) that established efficacy of the reference treatment</i>) and the settings and locations where the data were collected.	139, 141
Interventions	4	Precise details of the interventions intended for each group <i>detailing whether the reference treatment in the non-inferiority or equivalence trial is identical (or very similar) to that in any trial(s) that established efficacy</i> , and how and when they were actually administered.	142
Objectives	5	Specific objectives and hypotheses, <i>including the hypothesis concerning non-inferiority or equivalence.</i>	138
Outcomes	6	Clearly defined primary and secondary outcome measures <i>detailing whether the outcomes in the non-inferiority or equivalence trial are identical (or very similar) to those in any trial(s) that established efficacy of the reference treatment</i> and, when applicable, any methods used to enhance the quality of measurements (e.g., multiple observations, training of assessors).	143
Sample size	7	How sample size was determined <i>detailing whether it was calculated using a non-inferiority or equivalence criterion and specifying the margin of equivalence with the rationale for its choice</i> . When applicable, explanation of any interim analyses and stopping rules (<i>and whether related to a non-inferiority or equivalence hypothesis</i>).	144
Randomisation -- Sequence generation	8	Method used to generate the random allocation sequence, including details of any restrictions (e.g., blocking, stratification)	141

Randomisation -- Allocation concealment	9	Method used to implement the random allocation sequence (e.g., numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned.	141
Randomisation -- Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups.	141
Blinding (masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated.	142
Statistical methods	12	Statistical methods used to compare groups for primary outcome(s), <i>specifying whether a one or two-sided confidence interval approach was used</i> . Methods for additional analyses, such as subgroup analyses and adjusted analyses.	144
<i>RESULTS</i> Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analysed for the primary outcome. Describe protocol deviations from study as planned, together with reasons.	148
Recruitment	14	Dates defining the periods of recruitment and follow-up.	147
Baseline data	15	Baseline demographic and clinical characteristics of each group.	n/a
Numbers analysed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was <i>"intention-to-treat"</i> and/or <i>alternative analyses were conducted</i> . State the results in absolute numbers when feasible (e.g., 10/20, not 50%).	148
Outcomes and estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (e.g., 95% confidence interval). <i>For the outcome(s) for which non-inferiority or equivalence is hypothesized, a figure showing confidence intervals and margins of equivalence may be useful.</i>	147
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory.	n/a
Adverse events	19	All important adverse events or side effects in each intervention group.	n/a
<i>DISCUSSION</i> Interpretation	20	Interpretation of the results, taking into account the <i>non-inferiority or equivalence hypothesis and any other study hypotheses</i> , sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.	151-156
Generalisability	21	Generalisability (external validity) of the trial findings.	152
Overall evidence	22	General interpretation of the results in the context of current evidence.	151-156

Appendix F. CHEERS checklist

Consolidated Health Economics Evaluation Reporting Standards (CHEERS)

Section/item	Item number	Recommendation	Reported on page number
Title and abstract			
Title	1	Identify the study as an economic evaluation or use more specific terms such as “cost-effectiveness analysis”, and describe the interventions compared.	159
Abstract	2	Provide a structured summary of objectives, perspective, setting, methods (including study design and inputs), results (including base case and uncertainty analyses), and conclusions.	n/a
Introduction			
Background and objectives	3	Provide an explicit statement of the broader context for the study.	160
		Present the study question and its relevance for health policy or practice decisions.	163
Methods			
Target population and subgroups	4	Describe characteristics of the base case population and subgroups analysed, including why they were chosen.	164
Setting and location	5	State relevant aspects of the system(s) in which the decision(s) need(s) to be made.	164
Study perspective	6	Describe the perspective of the study and relate this to the costs being evaluated.	168
Comparators	7	Describe the interventions or strategies being compared and state why they were chosen.	164-167
Time horizon	8	State the time horizon(s) over which costs and consequences are being evaluated and say why appropriate.	168
Discount rate	9	Report the choice of discount rate(s) used for costs and outcomes and say why appropriate.	168
Choice of health outcomes	10	Describe what outcomes were used as the measure(s) of benefit in the evaluation and their relevance for the type of analysis performed.	168
Measurement of effectiveness	11a	<i>Single study-based estimates</i> : Describe fully the design features of the single effectiveness study and why the single study was a sufficient source of clinical effectiveness data.	168
	11b	<i>Synthesis-based estimates</i> : Describe fully the methods used for identification of included studies and synthesis of clinical effectiveness data.	Not applicable
Measurement and valuation of preference-based outcomes	12	If applicable, describe the population and methods used to elicit preferences for outcomes.	168
Estimating resources and costs	13a	<i>Single study-based economic evaluation</i> : Describe approaches used to estimate resource use associated with the alternative interventions. Describe primary or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.	Not applicable

Section/item	Item number	Recommendation	Reported on page number
	13b	<i>Model-based economic evaluation:</i> Describe approaches and data sources used to estimate resource use associated with model health states. Describe primary or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.	168
Currency, price date, and conversion	14	Report the dates of the estimated resource quantities and unit costs. Describe methods for adjusting estimated unit costs to the year of reported costs if necessary. Describe methods for converting costs into a common currency base and the exchange rate.	168
Choice of model	15	Describe and give reasons for the specific type of decision-analytical model used. Providing a figure to show model structure is strongly recommended.	168
Assumptions	16	Describe all structural or other assumptions underpinning the decision-analytical model.	168
Analytical methods	17	Describe all analytical methods supporting the evaluation. This could include methods for dealing with skewed, missing, or censored data; extrapolation methods; methods for pooling data; approaches to validate or make adjustments (such as half cycle corrections) to a model; and methods for handling population heterogeneity and uncertainty.	168-170
Results			
Study parameters	18	Report the values, ranges, references, and, if used, probability distributions for all parameters. Report reasons or sources for distributions used to represent uncertainty where appropriate. Providing a table to show the input values is strongly recommended.	171
Incremental costs and outcomes	19	For each intervention, report mean values for the main categories of estimated costs and outcomes of interest, as well as mean differences between the comparator groups. If applicable, report incremental cost-effectiveness ratios.	172-174
Characterising uncertainty	20a	<i>Single study-based economic evaluation:</i> Describe the effects of sampling uncertainty for the estimated incremental cost and incremental effectiveness parameters, together with the impact of methodological assumptions (such as discount rate, study perspective).	Not applicable
	20b	<i>Model-based economic evaluation:</i> Describe the effects on the results of uncertainty for all input parameters, and uncertainty related to the structure of the model and assumptions.	174
Characterising heterogeneity	21	If applicable, report differences in costs, outcomes, or cost-effectiveness that can be explained by variations between subgroups of patients with different baseline characteristics or other observed variability in effects that are not reducible by more information.	Not applicable
Discussion			
Study findings, limitations, generalisability, and current knowledge	22	Summarise key study findings and describe how they support the conclusions reached. Discuss limitations and the generalisability of the findings and how the findings fit with current knowledge.	176-183

Section/item	Item number	Recommendation	Reported on page number
Other			
Source of funding	23	Describe how the study was funded and the role of the funder in the identification, design, conduct, and reporting of the analysis. Describe other non-monetary sources of support.	163
Conflicts of interest	24	Describe any potential for conflict of interest of study contributors in accordance with journal policy. In the absence of a journal policy, we recommend authors comply with International Committee of Medical Journal Editors recommendations.	Not applicable

Appendix G. Data availability statements

All files listed below are available via Open Science Framework and are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CCo 1.0 Public domain dedication).

Chapter 6. Validity

File name	Description	DOI
RobotReviewer Covidence Protocol.docx	Trial protocol as registered with Monash	doi.org/10.17605/OSF.IO/H8AN9
Accuracy data.xlsx	Anonymised trial data for overall accuracy and for domain-specific accuracy	doi.org/10.17605/OSF.IO/H8AN9
Time data.xlsx	Anonymised trial data for time spent on RoB assessments	doi.org/10.17605/OSF.IO/H8AN9

Chapter 7. Economic evaluation

File name	Description	DOI
Search strategies.docx	Links to MEDLINE and Embase search strategies; custom search strategy	doi.org/10.17605/OSF.IO/24W53
Time log workbook.xlsx	Workbook provided to screen-coders to collect time-on-task information	doi.org/10.17605/OSF.IO/24W53
Time on task data.csv	De-identified time-on-task data collected from the screen-coders	doi.org/10.17605/OSF.IO/24W53
Base case analysis data.csv	All values used in cost-effectiveness calculations	doi.org/10.17605/OSF.IO/24W53
Sensitivity analysis – Precision.csv	All values used in calculations assessing the impact of precision on the cost-effectiveness analysis	doi.org/10.17605/OSF.IO/24W53
Sensitivity analysis – Time on task.csv	All values used in the calculations assessing the impact of time-on-task on the cost-effectiveness analysis	doi.org/10.17605/OSF.IO/24W53