



UvA-DARE (Digital Academic Repository)

Big data in medical research

From understanding to improving automation

van Altena, A.J.

Publication date

2022

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

van Altena, A. J. (2022). *Big data in medical research: From understanding to improving automation*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

BIG DATA IN MEDICAL RESEARCH

FROM UNDERSTANDING TO IMPROVING AUTOMATION



ALLARD J. VAN ALTENA

Big Data in medical research: from understanding to improving automation

Allard J. van Altena

Big Data in medical research: from understanding to improving automation
Allard J. van Altena
PhD Thesis, University of Amsterdam (UvA), The Netherlands
ISBN: 978-94-6458-561-2

Layout: Shayan Shahand with the help of F. Maggi, V. Gayevskiy, and the T_EX
community. Font and layout tweaks by Allard J. van Altena.
Cover design: Rodney Lichtveld
Printed by: Ridderprint | ridderprint.nl

© 2022 Allard J. van Altena

Big Data in medical research: from understanding to improving automation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. P.P.C.C. Verbeek

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen in de Agnietenkapel

op vrijdag 9 december 2022, te 13:00 uur

door **Allard Jan-Jaap van Altena**

geboren te Rotterdam

Promotiecommissie:

Promotor:	Prof. dr. A.H. Zwinderman	AMC-UvA
Co-promotor:	Dr. S.D. Olabarriaga	AMC-UvA
Overige leden:	Prof. dr. D. Krefting	Universitätsmedizin Göttingen
	Prof. dr. ing. A.H.C. van Kampen	AMC-UvA
	Prof. dr. A. Abu-Hanna	AMC-UvA
	Prof. dr. H.A. Marquering	AMC-UvA
	Prof. dr. M.C. Schut	AMC-UvA
	Prof. dr. M. Hoogendoorn	Vrije Universiteit Amsterdam

Faculteit der Geneeskunde

The research described in this thesis was carried out at the department of epidemiology and data science of the Amsterdam university medical centers.

Parent: "If all your friends jumped of a bridge would you follow them?"

Machine learning algorithm: "Yes"

Contents

List of Figures	ii
List of Tables	iii
Preface	v
Chapter 1 Introduction	1
Part I Understanding Big Data	5
Chapter 2 Understanding Big Data themes	7
Chapter 3 Usage of the term Big Data	29
Part II Adoption of automation tools	45
Chapter 4 Usage of automation tools	47
Part III Improving automation tools	67
Chapter 5 Training sample selection	69
Chapter 6 Improving active learning performance	93
Chapter 7 Discussion	107
Appendices	113

List of Figures

2.1	Plate notation of topic modelling	11
2.2	Corpus generation: documents extracted per literature database, documents removed from the corpus, and total number of included documents	13
2.3	Frequency of the top 2000 unique words in the corpus	14
2.4	Word cloud of the top-100 unique words in the corpus	14
2.5	AIC curves	15
2.6	Convergence of the log-likelihood for the chosen model M_{25}	15
2.7	Distribution of topics over documents	18
3.1	Results of searching, fetching, and cleaning two corpora	34
3.2	Word cloud for corpora	36
3.3	Pipeline from sampling the data to training the models and retrieving performance metrics	37
3.4	ROC curves	39
3.5	Word cloud of selected features	40
3.6	False Omission Rate over time	40
4.1	Technology Acceptance Model 2, adapted for this study	50
4.2	Reported tool usage	55
4.3	Summary of the System Usability Scale questions	57
5.1	Overview of workflow for the approaches using different training data	78
5.2	Boxplot of model performance stratified by the training set size	81
5.3	Boxplot of SIMILAR performance stratified by the training set size	81
5.4	Cosine similarities between all pairs of reviews in our dataset	87
5.5	WSS@95 performance for SIMILAR and ALL models for individual reviews	89
5.6	Training time, measured in seconds, for training set sizes 1 and 49	90
6.1	Representation of the active learning cycle	97
6.2	Overview of the process for the experiments using different approaches	99
6.3	Boxplots of the Inverse Burden (IB) at 95% yield for each initial pool size	100

List of Tables

2.1	Description of themes identified in Big Data definitions from literature	17
2.3	Summed annotations per topic and theme, and overall theme per topic	19
2.4	Top 20 words for the 25-topic model identified with TM	24
2.5	Raw annotation results per observer	27
3.1	Characteristics of the corpora after cleaning	37
4.1	Number of responses per tool for S2	56
4.2	Number of emails sent for S1, failed to deliver emails, and survey responses per group	58
4.3	Questions for S1	61
4.4	Generic questions for S2	62
4.5	Specific questions for S2	63
4.6	Usability questions for S2	64
5.1	Document characteristics after cleaning	73
5.2	Reasons for missing abstract	73
5.3	Review groups according to disease	73
5.4	Metadata collected for each review	74
5.5	P-values for SIMILAR versus ALL performance and SIMILAR versus RANDOM performance	80
5.6	P-values for other versus disease groups 1-7 performance, both are stratified by the training set size	80
5.7	Pearson correlation between the performance and cosine similarity for each training set size in the SIMILAR approach	80
5.8	Pearson correlations between review metadata and WSS@95 performance	88
6.1	Inverse burden results for each of the initial pool sizes	101

αύχένα τε στιβαρόν καὶ στήθεα λαχνήεντα,
δῦ δὲ χιτῶν', ἔλε δὲ σκῆπτρον παχύ, βῆ δὲ θύραζε
χωλεύων: ὑπὸ δ' ἀμφίπολοι ῥώνοντο ἄνακτι
χρύσειαι ζωῆσι νεήνισιν εἰοικυῖται.
τῆς ἐν μὲν νόος ἐστὶ μετὰ φρεσίν, ἐν δὲ καὶ αὐδὴ
καὶ σθένος, ἀθανάτων δὲ θεῶν ἅπο ἔργα ἴσασιν.

[...] [Hephaistos] took up a heavy stick in his hand, and went to the doorway limping.
And in support of their master moved his attendants.
These are golden, and in appearance like living young women.
There is intelligence in their hearts, and there is speech in them and strength,
and from the immortal gods they have learned how to do things.

Golden attendants of Hephaistos
Homer, Iliad 18. 415 ff (trans. Lattimore)

Preface

Inventing ways to make ones life easier and more pleasant transcends modern times. The idea of *automata* that aid humans in various tasks dates back to ancient Greek myths. On more than one occasion the greek eposes described Hephaistos creating automata, either to help him personally, to protect islands, or because Zeus told him to do so.

The ancient Greek lacked the imagination for a design such as R2-D2, automata were therefore often beautiful women or 70 meter tall men. In modern times automata take many forms and fulfill many tasks, from the simplest Excel macro to robot vacuums and massive industrial machines.

In this thesis I take a peek into the pithos of Pandora and describe the automata I have lovingly worked on. I sincerely hope that my work improves the lives of others.

Allard J. van Altena
Utrecht
March 2022



Introduction

1.1 General introduction

In the current field of medical science, and I dare say *almost any* modern research field, it is impossible to imagine a world without data. Many thousands of researchers contributed and refined methods for the creation, use, and reuse of data. Here the phrase “standing on the shoulders of giants” really comes into its right. The foundations for this type of research were laid ages ago. There are some examples of famous early-day researchers that first wandered onto the path of data science, one of which is John Graunt.

In 1662, Graunt was one of the first to document his observational research, a type of research where data is gathered first and hypotheses and conclusion are derived from that data [1]. What is interesting from the viewpoint of a data scientist is that Graunt gathered data from weekly published “Bills of Mortality”, containing data about the number of burials in London. Using this data he discovered that more boys than girls are born; made the first somewhat accurate estimate of the population of London; and made time trends for many diseases. These are, possibly, the first examples of using available data for research purposes.

Graunt’s method of observing the data and generating hypotheses is often frowned upon in modern science [1]. Applying this style of research without the proper methods can lead to incorrect conclusions that are ‘supported’ by data. In short, throwing every known method at data will most likely yield *some* result, even when they do not hold any water. However, modern science provides many tools that, when applied in good faith, enable observing data and generating hypotheses from those observations.

Another noteworthy founder of research methodology is John Snow. In his publication of “1854 Broad Street cholera outbreak” he combined cholera occurrence data with geo-spatial data into a map [2]. From this map he derived the source of the cholera infection and persuaded the authorities to remove it. The truthfulness of this story and specifics of his methods have been questioned [3]. However, the outlines of mixing data types to create evidence, a common practice in modern research, are clearly present.

The examples of Graunt and Snow show that data collection, aggregation, and processing are connected to the medical research field since its inception. Since those early years, science fields have grown into using data and advances in technologies are still actively changing the culture of science. Communication and exchange of information has never been easier, opening doors for collaborations unhindered by distance and time. This is reflected in a shift from local science with small research teams to widely distributed teams of researchers from many institutions working together. The terms ‘little’ science and ‘big’ science have been coined in the 60’s to describe different research cultures [4].

A field of science is described as ‘little’ when locally managed small communities use heterogeneous methods and data [5]. On the other hand, ‘big’ science concerns distributed and often international teams that jointly collect, analyse, preserve, and share data using shared facilities [4].

Using data in a ‘big’ manner brings challenges and opportunities. The commonly accepted term to describe data to which these challenges apply is ‘Big Data’, a term introduced by Gartner in 2011 [6]. Although the term Big Data is widely used, studies show that its meaning is much debated and many different definitions exist [7, 8].

Big data is mostly understood as the manipulation of large data volumes [9–11]. However, it is well recognized that, apart from its volume, data may have other characteristics that make it Big. Possibly the most famous definition of Big Data is captured in three “V’s” introduced by Laney in 2001: volume, velocity, and variety [12]. But now a wide range of V’s has been added to describe Big Data, such as: veracity, value, and variability [13]. Other studies [14–16] ignore these labels and simply consider any data that calls for methods beyond the conventional¹ Big Data. In Chapters 2 and 3 of this thesis I will delve deeper into the meaning and usage of the term Big Data.

Many research fields constantly push the boundaries of what is conventional to achieve their purposes. One such field is that of (bio)medical systematic reviews. Systematic reviews are a cornerstone of medical decision making [17]. They bring together the findings from multiple studies in a structured, reliable, and preferably unbiased way. As such, reviews provide a good reflection of the current scientific understanding of a certain topic.

A systematic review consists of retrieval, appraisal, and synthesis of evidence. The whole process is mostly manual and time consuming. For a full-time researcher, depending on expertise and review complexity, a review can take from 6 months to several years. Most of this time is spent on retrieving studies and determining their relevancy for the research question at hand.

In many reviews the number of studies to appraise, also called *screening*, is very large. Searches returning 10,000 items are not uncommon, and in some extreme cases 800,000 to 1 million items need to be screened [18]. With an ever-growing body of published literature [19] and a multitude of questions that need to be answered, the current practice is unsustainable [20].

One of the proposed solutions to this problem is computerised support of the screening process. The resulting tools are often called ‘automation tools’. Automation tools come in many forms and a well researched form is *text mining*. Text mining is a machine learning method that attempts to find structure in text data. Using this structure the method can then (partially) automate the appraisal of studies in a systematic review, by labelling the studies yielded by the systematic search as relevant or irrelevant to the research question. The degree of success depends on many factors, among which data quality and similarity of the labelled and unlabelled studies are prime factors.

Automation tools bring many challenges. Adoption among reviewers is often low because methods are unobtainable or difficult to use, a topic I explore further in Chapter 4. Furthermore, the performance of relevancy prediction is often lacking, an issue to which I contribute a possible solution in Chapters 5 and 6.

While the fields of Big Data and systematic review automation have both been studied extensively, there are still many unanswered questions. In this thesis I aim to (1) uncover a common understanding of Big Data in the (bio)medical field; (2) aid in improving the adoption of automation tools among systematic reviewers; and (3) contribute to the effectiveness of automation tools.

¹Data whose characteristics call for methods beyond the tried-and-true; necessity of scalable systems for storage, processing, manipulation, analysis, visualisation.

1.2 Outline

My research project started with an open aim of contributing to Big Data in the field of (bio)medical science. This led me to a specific Big Data challenge that systematic reviewers face. This thesis follows this path by starting at exploring a common understanding of the term Big Data in the scope of (bio)medical science. The focus then shifts to systematic reviewers and their use of automation tools to deal with Big Data. Lastly, I developed a method that improves the performance of existing automation tools.

1.2.1 Understanding Big Data

While Big Data is a key component of many (bio)medical studies, it has yet to receive a formal definition. We have observed that different research fields have many different interpretations for Big Data, some of which have a negative connotation. We believe this wide spread of understanding hampers communication and results in missed opportunities. Chapter 2 pursues a better understanding of the topics covered by the term Big Data through a data-driven systematic approach using text analysis of scientific (bio)medical literature.

Skeptics argue that Big Data is just a hype term, representing nothing new or at best just an extension of what has been done for decades [21]. Therefore, in Chapter 3 we assess the value of the term Big Data when used by researchers in their publications.

1.2.2 Solutions to a data deluge

As stated above, systematic reviews are a cornerstone of evidence-informed decision making, but the process is very time-consuming. With the rapid expansion of scientific information produced and research questions to be addressed, there is a growing workload on reviewers, making the current practice unsustainable without the aid of automation tools. In Chapter 4 we investigate why the adoption of automation tools among systematic reviewers seems to be lagging and identify potential barriers and facilitators for adoption.

1.2.3 Applying solutions in practice

In Chapter 5 we introduce an approach to improve the performance of systematic review automation tools. We focussed on a subset of automation tools that support the screening process of a systematic review by using text mining to predict the relevancy of each study that needs to be screened. Using the predictions the reading order can be adjusted so that the reviewer sees the studies that are most likely to be relevant first.

To predict how relevant a study is a prediction method first needs to learn from previous systematic reviews where the relevant studies were appraised by researchers, a process called *training*. For this to work we assume that there were transferable characteristics between the previous reviews and the current review. However, systematic reviews mostly have very specific and unique research questions.

Prediction methods often use all available data to train on. However, in the case of systematic reviews we hypothesise that this waters down the transferable characteristics leading to less precision in prediction. Our proposed approach chooses which data to use during training based on a metric that measures similarity between reviews. This approach leads to less data to train, but the selection has a high similarity to the data from the review that we are predicting, improving the performance of the automation tool. Chapter 5 introduces our approach and Chapter 6 builds upon our first insights and applies the proposed approach in an active learning setting.

Finally, in Chapter 7 all results presented in this thesis are discussed and ideas for future research are proposed.



Understanding Big Data



Understanding Big Data themes from scientific biomedical literature through topic modeling

In Journal of Big Data, volume 3, pages 1-21, 2016

A.J. van Altena, P.D. Moerland,
A.H. Zwinderman, S.D. Olabarriaga

Abstract

Nowadays, Big Data is a key component in (bio)medical research. However, the meaning of the term is subject to a wide array of opinions, without a formal definition. This hampers communication and leads to missed opportunities. For example, in the (bio)medical field we have observed many different interpretations, some of which have a negative connotation, impeding exploitation of Big Data approaches.

In this paper we pursue a better understanding of the term Big Data through a data-driven systematic approach using text analysis of scientific (bio)medical literature. We attempt to find how existing Big Data definitions are expressed within the chosen application domain. We build upon findings of previous qualitative research by De Mauro et al., which analysed fifteen definitions and identified four key Big Data themes (i.e., information, methods, technology, and impact). We have revisited these and other definitions of Big Data, and consolidated them into eight additional themes, resulting in a total of twelve themes.

The corpus was composed of paper abstracts extracted from (bio)medical literature databases, searching for 'big data'. After text pre-processing and parameter selection, topic modelling was applied with 25 topics. The resulting top-20 words per topic were annotated with the twelve Big Data themes by seven observers. The analysis of these annotations show that the themes proposed by De Mauro et al. are strongly expressed in the corpus. Furthermore, several of the most popular Big Data V's (i.e., Volume, Velocity, and Value) also have a relatively high presence. Other V's introduced more recently (e.g. Variability) were however hardly found in the 25 topics. These findings show that the current understanding of Big Data within the (bio)medical domain is in agreement with more general definitions of the term.

2.1 Introduction

The usage of the term ‘Big Data’ has picked up since 2011. This was the year that Gartner introduced “Big Data and Extreme Information Processing and Management” in its hype cycle [6]. Furthermore, increased interest is visible in the ever growing search traffic shown by Google Trends [22]. Scientific publications in (bio)medicine, which are our main interest in this study, also show a massive increase in the number of papers published yearly that mention Big Data [13].

Still, in spite of the popularity of this term, there is much debate about the definition of Big Data. In 2001 Gartner (called “META Group” at the time [23]) published a report that in hindsight is often referred to as the first description of Big Data. It defines the term through Information Technology (IT) challenges described by three Big Data aspects (V’s): volume, velocity, and variety [12].

Over the years this has evolved into many interpretations. Mostly, companies define Big Data in the light of their prime business, meaning that Google will mention analysis (e.g., Google Flu), while Oracle emphasises volume and storage [24], and IBM or Microsoft focus on computation and usability [25]. In a web-blog, posted on the data science sub-domain of the Berkeley school of information, 43 ‘thought leaders’ from the industry were asked for their definition of Big Data [26]. Not many of these leaders agreed with each other and definitions range from “data that cannot fit easily into a standard relational database” to “Big data is not all about volume, it is more about combining different data sets and to analyze it in real-time to get insights for your organisation”. On a governmental level, the US National Institute of Standards and Technology (NIST) defined Big Data in 2014 as the need for scalable technology and four V’s: Volume, Velocity, Variety, and Variability. Finally, in the scientific domain, Big Data is mostly understood as the challenges of working with large volumes of data [9–11].

Possibly due to this great variety of definitions, in practice we have observed many different interpretations of the term Big Data among (bio)medical scientists. Some understand Big Data as a positive development, and actively pursue usage of new methods and technology associated with the term [13]. Others, however, view it as a harmful influence on, for example, the strength of research evidence, preferring classical statistical methods [27]. A better understanding of Big Data would facilitate communication and clarify expectations regarding this overloaded term [28].

Some researchers have attempted to capture comprehensive definitions of Big Data, such as De Mauro et al. [8], Ward and Barker [7], and Andreu-Perez et al. [13]. The first two focus on no domain in particular, whereas Andreu-Perez et al. [13] focuses on health-oriented applications. Of particular interest is the work by De Mauro et al., which analysed various Big Data definitions and from these distil their own. Their proposed definition is based on four themes found in the underlying definitions that were gathered, namely Information, Methods, Technology, and Impact. Note that all the cases mentioned above are based on qualitative literature studies. Hansmann and Niemeyer [29], however, used text mining to understand the themes included in Big Data literature. They combined automatic and manual approaches to identify three themes: IT infrastructure, methods, and data. While these efforts have been valuable for a better understanding of the term Big Data, they do not present systematic evidence of the actual themes used in the scientific literature, in particular for the (bio)medical research domain.

In this paper we present our efforts to answer the following research question: Which themes from various existing Big Data definitions are expressed in (bio)medical scientific publications? For this purpose, we adopted a data-driven systematic approach. First, Big Data definitions were revised and 12 themes were identified. Then, (bio)medical literature was systematically gathered from two scientific databases (i.e., PubMed and PubMed Central) and analysed automatically with text mining. While there are many text mining and clustering methods, we chose Topic Modelling (TM) [30, 31] because this method captures two aspects that are important for this dataset: words may have multiple meanings or

interpretations and documents may contain one or more topics. The topics identified through TM were annotated with the 12 themes by a small group of observers. In the following sections we detail the methods, present the results and discuss our findings.

2.2 Methods

In this section the construction of the corpus is described, followed by an explanation of the concepts behind TM. Then the application of TM to the corpus is presented in three steps: pre-processing, model fitting, and post-processing. Finally we present the gathering and summary of existing Big Data definitions, and the process used to identify them in the topics determined by TM.

2.2.1 Corpus

The corpus of documents was created by querying two literature databases focused on (bio)medical publications: PubMed and PubMed Central (PMC). The search queries were as follows:

- *PubMed*: “big data”[TIAB] OR (big[TIAB] AND “health data”[TIAB]) OR “large data”[TI];
- *PMC*: “big data”[TI] OR “big data”[AB] OR (big[TI] AND “health data”[TI]) OR (big[AB] AND “health data”[AB]) OR “large data”[TI].

Each query was built to search for literal use of the term ‘big data’, therefore selecting documents that were self-identified with Big Data. No word spacing was allowed to minimise the amount of irrelevant results. The terms ‘big health data’ and ‘large data’ were added because they also retrieved relevant literature, especially for publications before 2011, when the term Big Data was not popular yet.

Titles and abstracts were exported from the databases and merged into a local repository for further processing. Based on the title (stripped of all special characters and spaces) or the Digital Object Identifier (DOI), if available, duplicates were removed from the corpus. Lastly, any record with an empty abstract (i.e., not provided in the database) was also removed from the corpus.

2.2.2 Topic modelling concepts

A specific type of TM was chosen, namely Latent Dirichlet Allocation (LDA) [30]. Throughout this paper the abbreviations TM and LDA are used interchangeably to indicate topic modelling through the application of LDA. The concept of TM is captured in Figure 2.1 using the plate notation [30–32]. Plate D denotes the set of documents, while $\theta^{(d)}$ is the multinomial distribution over topics for document d . Plate $N_{(d)}$ denotes the set of words w for a specific document d , while z is the topic to which word w is assigned. Lastly, plate T denotes the set of topics where $\phi^{(z)}$ is the multinomial distribution over words for topic z .

In TM, θ , ϕ , and z are the latent variables that have to be estimated. Together with the Dirichlet distributed hyperparameters α and β , the model is called Latent Dirichlet Allocation [30, 32]. The hyperparameters α and β should be interpreted as smoothing factors for respectively topic-to-document (θ) and word-to-topic (ϕ) assignments.

2.2.3 Topic modelling implementation

The statistical software R [33] was used to implement the pre-processing, TM fitting, model selection, and post-processing steps.

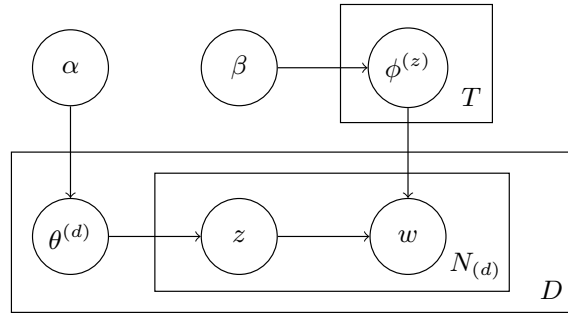


Figure 2.1: Plate notation of topic modelling, plates are shown as rectangles and the arrows indicate conditional dependencies. Shows the relations between known variables (documents D , number of words $N_{(d)}$, and words w), latent variables (multinomial distributions $\theta^{(d)}$ and $\phi^{(z)}$, and word to topic assignment z), and hyperparameters (α and β).

Pre-processing We used the R `tm` and `quantda` packages [34, 35] to execute the pre-processing steps. Processing consisted of removing stop words taken from the SMART list [36, 37] (e.g., about, the, which)¹. Extra stop words were added, they were either junk words resulting from processing steps, or terms that appeared very often and diluted the TM outcome, such as ‘big data’, ‘introduction’ and ‘discussion’². From the remaining words, bi-grams were created with function `dfm`: two words that occur next to each other at least fifteen times in the whole corpus are joined by an underscore (e.g., health_care). Furthermore, words were stemmed with function `stemDocument`; e.g., ‘develop’, ‘developed’, and ‘development’ were all stemmed to ‘develop’. Lastly, words longer than 26 characters were removed.

Fitting We fitted the model by estimating the latent variables θ , ϕ and z , which was done with the R `topicmodels` package [39]. Directly calculating θ and ϕ was shown to be suboptimal [32], therefore we used a Bayesian approach from the `topicmodels` package using Gibbs iterative sampling to approximate the distribution z . In this sampling process the probability of a word occurring in a topic is estimated. This probability of a given word-to-topic assignment is calculated from how often the word already occurs in the topic and how dominant the topic is for the document from which the word was sampled. Once the model fitting converges, θ and ϕ can be derived from the approximated distribution z with the posterior function.

Multiple models were fitted to determine the best TM parameters. We first conducted experiments to find adequate values for α and β . These influence the model as follows: with a small α (i.e., with many topics $\alpha = 50/T$ becomes smaller) it is likely for documents to contain only a few topics, whereas a bigger α (i.e., few topics) results in more topics per document. A small β similarly makes it likely for a topic to contain a mixture of a few words, thereby pushing the model to select highly specific words per topic. A range of values was fitted for both α and β and model outcomes were compared. Within a reasonable range (i.e., $0.1 < \alpha < 1$) we observed only minor differences between topics. Ultimately, fixed values were chosen for α and β , respectively $50/T$ and 0.01 as suggested in the literature [32, 40].

¹The full list can be found at [38]

²The complete list is: big, data, ieee, discussion, conclusion, introduction, methods, psycinfo database, rights reserved, record apa, journal abstract, apa rights, psycinfo, reserved journal

Model selection Modes were selected by analysing the likelihood for varying numbered of topics in the range $T \in \{5, 10, 15, \dots, 100, 150, 200, \dots, 500\}$. However, likelihood alone cannot be used to find the best model. A penalising factor has to be added for the model's complexity (i.e., the number of variables that have to be estimated). Two information criteria were considered, namely the Bayesian Information Criterion (BIC) [41] and the Akaike Information Criterion (AIC) [42]. When increasing the number of topics in a model, each topic becomes more specific and, therefore, easier to interpret. BIC puts more emphasis on the simplicity (in terms of the number of free parameters) of the model, resulting in a smaller number of topics as compared to AIC. We therefore chose to perform model selection using the AIC. In the case of TM, the variables to be estimated are the latent variables ϕ and θ , which grow with the number of topics. The model where the AIC reached its minimum was considered the optimal model. Equation (2.1) defines the AIC, where T is the number of topics in model M_T , L is the likelihood of model M_T , and W is the number of unique words in the corpus:

$$AIC(M_T) = -2 \log(L) + 2 ((T - 1) + T(W - 1)) \quad (2.1)$$

Post-processing θ and ϕ were retrieved for the optimal model. We then calculated the relevance of words within a topic according to the method described by Sievert et al. [43]. Equation (2.2) defines how relevance r was calculated for word w in topic t given λ :

$$r(t, w | \lambda) = \lambda \log(\phi_{tw}) + (1 - \lambda) \log\left(\frac{\phi_{tw}}{p_w}\right) \quad (2.2)$$

The relevance is a convex combination of two measures: the topic-specific distribution (ϕ_{tw}) and 'lift' (ϕ_{tw}/p_w), that is a ratio between topic-specific and corpus-wide distributions. These measures can be balanced with $0 \leq \lambda \leq 1$, by giving more weight to ϕ ($\lambda = 1$) or to the lift ($\lambda = 0$). In our experiments a value of 0.6 was chosen for λ , as suggested in Sievert et al. [43]. $T \times W$ relevancies were calculated (i.e., each word had one relevance score per topic) and used to sort the most relevant words per topic.

2.2.4 Big Data definitions

The definition proposed by De Mauro et al. was used as a starting point for this study. Furthermore, the underlying definitions gathered in De Mauro et al. were reassessed and where necessary updated (e.g., updates in white papers published by industry). Lastly, a publication by Andreu-Perez et al. [13] was added because it defined six Big Data V's in the context of (bio)medical research.

All the definitions were analysed. If the definition was given in free text, the major themes were extracted. Themes were then grouped on similarity, for example, Volume and Size were merged into one theme. For various reasons a few definitions were discarded, as discussed in Section 2.3.3.

2.2.5 Topic analysis

Topic model results were analysed manually by inspecting the top relevant words (i.e., 20 per topic). The observers received a list of topics and a description of each theme. They were instructed to read all the words in each topic, then consult the Big Data definition themes, and finally provide their opinion about which themes are associated with that set of words. Each of the topics was assigned zero, one, or more themes by each observer individually. In total seven persons performed the analysis independently: each of the authors and three external health data scientists.

2.3 Results

This section reports the results of corpus extraction, TM model fitting and selection, gathering and consolidation of Big Data definitions, and annotation of topics with the themes.

2.3.1 Corpus

A total of 1,659 documents were extracted from Pubmed and 543 from PubMed Central (see Section 2.2.1). After removing duplicates and records with an empty abstract, 1,308 documents were included in the corpus as shown in Figure 2.2.

After pre-processing (see Section 2.2.3) 136,339 words remained in the corpus, of which 7,849 were unique. A large portion (7,081 words) had a low frequency (< 40 occurrences). Figures 2.3 and 2.4 give an impression of the corpus's contents, showing a frequency plot of the top 2000 words, that seems to be in accordance with Zipf's law [44]. To create the word cloud the top 100 most frequent words were extracted (as marked with the vertical line in the frequency plot).

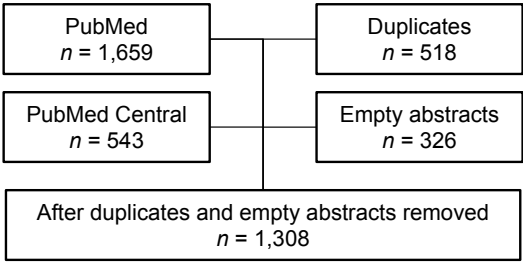


Figure 2.2: Corpus generation: documents extracted per literature database, documents removed from the corpus, and total number of included documents.

2.3.2 Topic modelling and model selection

In total 49 models M_T were fitted with T ranging between 5 and 500. The AIC curve for all fitted models M is shown in Figure 2.5. The minimum of the AIC curve lies at $T = 14$, however the differences are small until $T = 25$. We also calculated the distances between topics from diverse models ($T \in \{14 - 25\}$), showing that topics are fairly stable (data not shown). When increasing the number of topics, changes observed include one topic splitting into two topics or a new topic appearing. We saw no major reorganisation of topics or words within topics. We also observed that increasing the number of topics in the model makes the terms in each individual topic more specific. For example, one topic covering both application and Big Data themes might be split into two separate topics in a larger model. We therefore selected M_{25} for annotation, as this model has a better interpretability compared to M_{14} (more specific topics), with comparable quality of model fit (similar AIC).

To assess the robustness of the model M_{25} , the log-likelihood was tracked for each iteration of Gibbs sampling. This model was fitted three times with fixed input, but with different starting seeds for the sampling. The outcome of these fits is presented in Figure 2.6. It shows that the log-likelihood reaches its approximate maximum after 100 to 150 iterations. Models run with a higher number of iterations (up to 4000, data not shown) showed no major difference in log-likelihood convergence, therefore, final models such as M_{14} and M_{25} were run for 500 iterations. The top-20 most relevant words per topic of the M_{25} model are shown in Table 2.4.

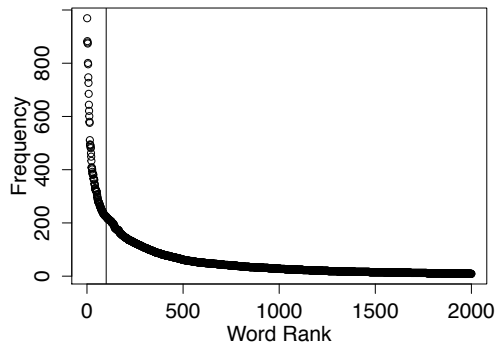


Figure 2.3: Frequency of the top 2000 unique words in the corpus. The vertical line is the cut-off point ($n = 100$) used for the word cloud.

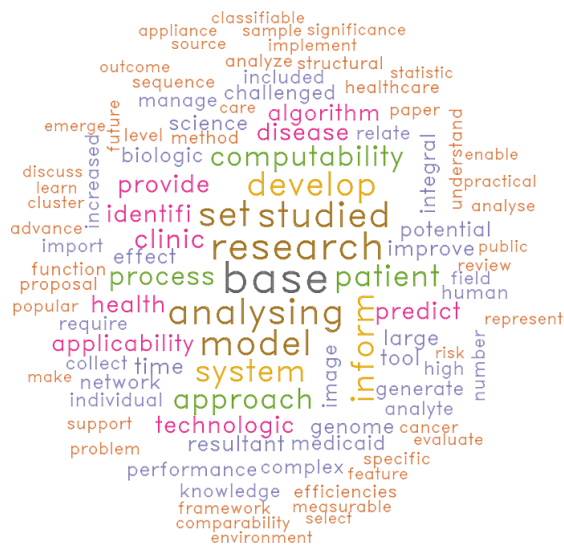


Figure 2.4: Word cloud of the top-100 unique words in the corpus.

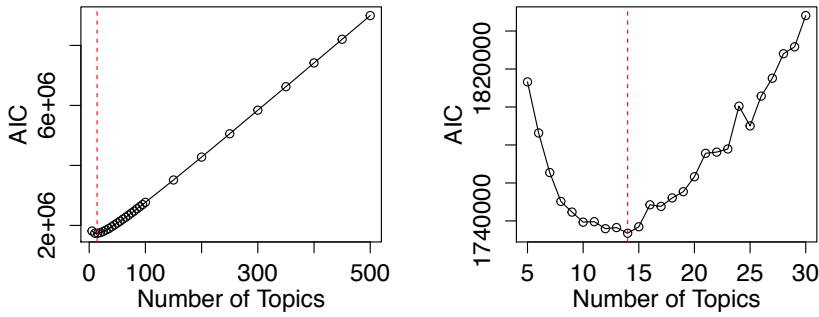


Figure 2.5: *Left*: AIC curve of the 49 fitted TM models (20 models between $T = 5$ and $T = 30$ not plotted, see right). The minimum is marked by the dotted line ($T = 14$). *Right*: Close-up of the AIC curve between $T = 5$ and $T = 30$, showing 26 fitted TM models. The minimum is marked by the dotted line ($T = 14$).

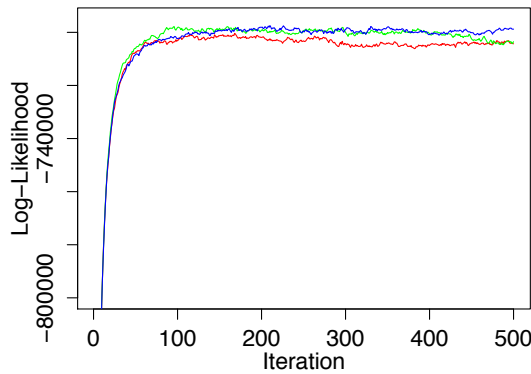


Figure 2.6: Convergence of the log-likelihood for the chosen model M_{25} for three runs starting from different seeds.

2.3.3 Big Data definitions

In total 17 definitions of Big Data were considered from the following sources [7, 8, 12–16, 24, 45–53]. Table 2.1 presents the results of our analysis listing the found themes, their description, and respective sources. Note that we have not attempted to consolidate the names of the themes, leaving the complete description as found in the sources. The definitions can be divided into three groups, with each group containing multiple themes.

The first group (I) correspond to the Big Data V's that occur in various forms in many of the analysed definitions. Some words were merged into one theme because they are essentially pseudonyms of each other. For example: volume, size, voluminous, and cardinality were found in ten of the definitions and, from their descriptions, refer to the amount of data. Also note that velocity and continuity, and complexity and variety were combined.

The second group (II) correspond to the aggregated themes proposed by De Mauro et al. that represent concepts of a higher level of abstraction than the previous group.

The third group (III) includes a theme identified in three definitions, that describe Big Data as data that is *beyond conventional* processing and analysis. The V's describe data by many different aspects, but none of those define a hard limit beyond which data becomes big. The theme 'beyond conventional' therefore describes Big Data as something that needs novel specialised and scalable solutions. This also means that the types of problems and applications that are assigned to the scope of Big Data change over time, as technology and methods evolve and improve.

The fourth group (IV) was not found in the studied definitions, but was added to cope with the reality of our data. Because the body of literature used in this study was obtained from (bio)medical literature databases, we expected to see application-related themes to be strongly represented in the resulting topics. We therefore included the Application theme to classify those topics that do not fall under Big Data.

Note that some definitions considered by De Mauro et al. were not used here:

- the definition by Microsoft [50] was a web-blogpost from 2013, therefore possibly outdated;
- Shneiderman et al. [51] does not specifically mention Big Data, as it was a publication from 2008 when this term was not in use yet;
- the definition by Manyika et al. [53] was only described in the executive summary;
- Mayer-Schönberger et al. [52] propose an abstract definition that was considered too difficult to convert into interpretable themes for topic analysis.

2.3.4 Topic analysis

The list of topics and words and Big Data themes were analysed by the seven observers. The observers all worked at the local department of epidemiology, biostatistics and bioinformatics, therefore they were extremely suitable for the annotation task. The Big Data themes (Table 2.1) and topic words (Table 2.4) were well understood and the task could be finished without further help in a reasonable amount of time (30 minutes to an hour).

The raw annotation results are displayed per observer and per topic in Table 2.5. Note that some observers did not assign any theme to some topics, and that in many cases more than one theme was assigned to the topics. Table 2.3 presents the frequency of themes assigned per topic, highlighting high or unanimous agreement among the observers (shown underlined and bold). It also shows the *overall* themes, i.e., those that were assigned to a topic by at least four observers.

In four topics less than four observers assigned the same theme to it (i.e., 3, 17, 19, and 25). Out of the remaining 21 topics, five had unanimous agreement between the observers for some theme (i.e., 6, 7, 8, 20, and 21). The remaining 16 topics could be split into topics

Table 2.1: Description of themes identified in Big Data definitions from literature.

	Theme Name	Theme Description	Definition Sources
I	Volume, Size, Voluminous, Cardinality	Large quantities of data in number of bytes; size of available data (e.g. all records instead of a sample); beyond conventional storage techniques; number of records at a particular instance.	[7, 12, 13, 15, 16, 24, 45–47, 49]
	Velocity, Continuity	Flow rate at which data is created, stored, analysed, and visualised; increased through invention of new data streams such as social media; beyond conventional means of processing, needing new techniques such as streaming; growth of data over time.	[12, 13, 16, 24, 45–47]
	Variety, Complexity	Many different types of data; not bound to a traditional data format; format changes over time; heterogeneous and unstructured data.	[7, 12, 13, 15, 16, 24, 45–47, 49]
	Veracity	Trustworthiness of data; reliability of data quality and gathering environment.	[13, 45]
	Value	Worth/relevancy of data (e.g. economic, individual/privacy, societal, humanity value).	[13, 24, 48]
	Variability	Consistency of data over time; influences which systematically change data measures over time.	[13, 47]
II	Information	Where signals are turned into data (e.g. book digitalisation, or gathering from personal device measurements).	[8]
	Technology	Tools, systems, and software (e.g. scalable processing and transmission systems such as Hadoop).	[7, 8, 14, 15, 47, 48]
	Methods	Procedures and their application (e.g. clustering, natural language processing, machine learning, neural networks, visualisation).	[8, 14, 48]
	Impact	Ethical, business, societal.	[8]
III	Beyond conventional	Data whose size call for methods beyond the tried-and-true; necessity of scalable systems for storage, processing, manipulation, analysis, visualisation.	[14–16]
IV	Application	About the application domain treated in the papers.	-

with a single overall theme (i.e., 2, 4, 9, 10, 11, 13, 14, 15, 16, 18, 22, 24) and topics with two overall themes (i.e., 1, 5, 12, 23).

Note that the most frequently assigned theme was Application (66 times), followed by the themes in the second group, proposed by de Mauro et al.. From the themes in the first group, Volume and Velocity occurred more often than the others. Notably, Variability was hardly identified among these topics.

Figure 2.7 presents the distribution of topics over documents based on the probability of each topic to each document (i.e., θ). The large majority of topics (in black) have a strong presence in only a few hundred documents. However, there are four topics (in red and blue) that deviate from this pattern. The two red topics (topic 1 and 2, see Table 2.4) have a stronger presence in more documents as compared to the topics pictured in black. The blue topics (topic 3 and 5, see Table 2.4) have a stronger presence in nearly all documents.

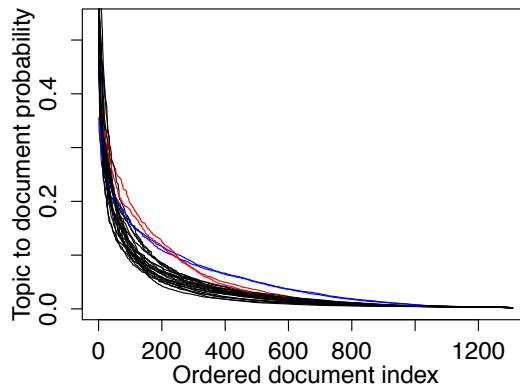


Figure 2.7: Distribution of topics over documents (i.e., θ , y-axis). The documents are sorted on topic-to-document relevance within each topic. The x-axis represents the order of the sorted documents. Each line represents one topic, in black. Exceptions are topics 1 and 2, plotted in red, and topic 3 and 5, plotted in blue.

2.4 Discussion

In this paper we attempted to identify themes related to Big Data definitions in a large corpus of (bio)medical literature through topic modelling. We have followed a structured and objective approach as much as possible. This process delivered novel and interesting results, that however need to be carefully interpreted due to remaining limitations in our study.

2.4.1 Identification of themes in Big Data definitions

Due to the lack of a consolidated and widely accepted definition of Big Data, it was necessary to consult a large number of scientific papers. This work is limited to scientific literature, but obviously there are many other definitions of Big Data that have not been considered in our work, such as the Berkeley blog mentioned in the introduction [26]. Nevertheless, most of the definitions in [26] can be mapped to the themes identified in

Table 2.3: Summed annotations per topic and theme, and overall theme per topic (≥ 4 counts).

Topic	Themes											Overall
	Volume	Velocity	Variety	Veracity	Value	Variability	Information	Technology	Methods	Impact	Beyond con.	
1				2	<u>5</u>					4	2	value, impact
2		1	1	3			1	1		1	4	application
3										1	3	-
4	1	1						2	<u>6</u>			methods
5	<u>5</u>	2	1					3			4	volume, beyond conventional
6		1						<u>7</u>			2	technology
7				1			1	1	<u>7</u>			1 methods
8			1				2					<u>7</u> application
9					1					4		2 impact
10	1		2				2	1	3		1	4 application
11					1					2		<u>6</u> application
12	<u>6</u>	<u>5</u>			1			2	1		1	volume, velocity
13				1		1	1		<u>5</u>			1 methods
14					1		4	1		1		2 information
15				1	1		1			3		4 application
16	1						2		1		1	<u>5</u> application
17		1			1	1	2	3			1	-
18				1	1		3	1	2			4 application
19		1	1		1		1	1	1	1		3 -
20							1		<u>7</u>			methods
21			1	1			1			1		<u>7</u> application
22		2		1			<u>6</u>					3 information
23	1			1			<u>6</u>				1	4 application, information
24	1	1	1		1		3			1		4 application
25	1	2					2	1	3			-
total	17	17	8	12	14	2	39	24	36	19	11	66

this study. Interestingly, the word cloud in [26] highlights words such as size, complex, and techniques, that are also found in the descriptions of the themes consolidated in Table 2.1. Furthermore, there are qualitative approaches to describing the Big Data field in publications such as Chen et al. [28] and Tsai et al. [54]. Note that, although these works do not strive to deliver a formal definition, the description of the Big Data field in both these publications include the same aspects found in the definition themes (see Section 2.3.3).

We have observed a large overlap among the Big Data definition literature considered in this study, nevertheless with variations in the focus applied by each author. Furthermore, certain themes occur more often than others in the definitions (Table 2.1). The original three V's (Volume, Velocity, Variety) occur in many definitions compared to the relatively 'newer' V's (Veracity, Value, Variability), that are present in only a few. This is also the case with Technology and Methods that are found in definitions more often than Information and Impact.

Finally, as the corpus was gathered from (bio)medical literature databases, we expected to find topics describing this domain. Therefore, the theme 'Application' has been introduced, that is obviously not found in the published Big Data definitions. Indeed, the annotation results presented in Table 2.3 show that 10 out of 25 topics have been annotated with Application by the majority of the observers. Note that the large fraction of application-related words might have overshadowed others that are related to Big Data themes. Scrubbing the corpus of application-related words could be used to circumvent this problem. This opens the possibility for fitting highly granular models that would be more easily interpretable and better reflect Big Data instead of the research field topics.

2.4.2 Corpus gathering

By design, in this study we only considered papers that were self-annotated with Big Data, whatever definition the authors might have used. This led to an interesting observation by one observer who could not find his research domain in any of the topics. However, the searched databases certainly included this domain and many of the Big Data themes could potentially be assigned to its papers. The domain could be missing due to various reasons, such as a low frequency of this research domain in the corpus. However, this observer acknowledged to consider his domain as 'conventional', therefore, papers published about this research domain most likely do not mention Big Data and were therefore not captured in the search performed in this study.

Note also that we only considered two databases, whereas many others could be included as well (e.g., Scopus or Ovid). Nevertheless, PubMed and PMC are important sources in medical research and therefore have been considered sufficiently representative for the purposes of our study.

Finally, a potential limitation of our study is that only abstracts were included in the corpus instead of full-text papers. Our assumption is that the abstracts contain the essence of a paper and are therefore representative of the actual themes found in a full paper. Moreover, it is currently still difficult to retrieve and parse full papers in an automated fashion, which would have severely limited the number of papers considered in our study.

2.4.3 Automatic identification of topics

In the progress of this research various text mining approaches were attempted to identify relevant topics to characterise the publications. First, we attempted to use AlchemyAPI [55], a natural language processing service that is accessible through the web. However, in a pilot experiment of 100 documents we observed that the number of results produced would be too big for effective analysis (i.e., 3,774 results, of which 3,006 were unique). Moreover, AlchemyAPI's method is implemented by proprietary code, so relations between documents and results were difficult to interpret.

We continued searching for a text mining method and considered document clustering to find the definition themes in literature. In principle, document clustering could capture themes but results are often limited to one theme per document. Furthermore, analysing document clusters to find definition themes would be a non-trivial (if not impossible) task.

A seemingly more suitable method was topic modelling, a method that can discover latent semantics in text. The main purpose of topic models is described as “discovering main themes that pervade large unstructured collections of documents” [31]. Furthermore, TM captures multiple meanings of words, but most importantly, it can identify multiple topics for each observed document. The LDA approach is perhaps the most popular and common topic model. The R package implementing the algorithm `topicmodels` had 22,576 downloads in 2015³. Moreover, the paper describing the underlying model by Blei et al. [30] has been cited over 16,000 times⁴. We therefore chose to use the LDA implementation of TM because of its appropriateness for our data, the relative ease of use of this approach (i.e., ready to use implementations in R), and extensive use in the literature by our peers.

Various TM approaches were tried to find a model with a manageable number of topics that allowed for manual annotation. The largest challenges were encountered during model selection. Two model evaluation methods (i.e., perplexity and harmonic mean) are often used in TM literature [29, 32, 56, 57]. The harmonic mean method calculates an approximation of the marginal likelihood of a fitted model, while perplexity measures how well a fitted model can predict unseen data. These criteria were calculated for multiple models with varying parameters expecting that the model decision boundary lay at some optimum of the response curve. For both criteria we were looking for a sudden decrease in marginal difference between two consecutive data points (i.e., models). Unfortunately, in our case, even when fitting models with up to 1,500 topics (data not shown), the curves did not show an optimum.

Finally we opted for TM with model selection through AIC, a method based on likelihood and model complexity (see Section 2.2.3). The AIC curve shows an optimum at M_{14} , however M_{25} was chosen for further analysis. While experimenting with the parameter T we noticed that quantitatively measuring model fit did not relate to the interpretability of the topics, as also noted in [43, 58]. Comparison between models showed that there was no major reorganisation of topics (data not shown), but increasing the number of topics made them more specific and therefore more interpretable.

2.4.4 Manual annotation of topics

Subjectivity of the manual annotation is one of the limitations of this study. Some research has been done in objectifying the analysis of TM results [40, 43, 59, 60]. However, so far, the results of TM cannot be quantitatively evaluated [29, 58]. For the purpose of this study, a group of seven observers was deemed enough for the topic analysis. We also present all the data in the paper, such that the reader can assess the topics themselves to confirm or dispute our results.

We took great effort to objectify the interpretation of TM results, but seven is a small number of observers. Ideally more persons should be involved in the assessment of theme assignment. For example, crowd sourcing services such as Mechanical Turk could be used [61]. However, this particular annotation task requires sufficient background knowledge in health data science, which significantly reduces the pool of suitable observers.

All the observers in this study were trained in health data science, therefore they are familiar with the terms and concepts that appeared in the topics and the Big Data themes. Nevertheless, no baseline assessment was performed to more precisely understand their own interpretations, which might have introduced some noise in our results.

³<http://cran-logs.rstudio.com/> on 9 June 2016

⁴<https://scholar.google.com/> on 20 October 2016

In general, the observers reported some difficulty to associate words with a theme. They also noted that their annotation decisions were mostly based on words that stood out in the topic, meaning that not all words were considered equally. This possibly led to the discrepancy between annotators displayed by the results (Tables 2.3 and 2.5). For example, when asked, annotator F noted that he chose Technology for topic 4 because of the specific word ‘cluster’, while all others chose Methods. Note that cluster could be interpreted as a computer cluster (i.e., Technology) or a cluster used in unsupervised machine learning (i.e., Methods). Furthermore, note that Information is often co-annotated or interchanged with Application. For example, neuroimaging, neuroscience, image, and signal are present in topic 23. The first two words can be associated with Application, and the latter with Information. Also, topics containing words referring to data (e.g., images and age) have been annotated as Information and/or Application by some observers. For such reasons some observers said that it was possible that their annotation might change slightly if they would analyse the topics again.

2.4.5 Big Data themes in biomedical literature

Despite annotation subjectivity we consider to have found sufficient agreement between the observers to support our findings, which show how Big Data themes are identified in biomedical literature (see Table 2.3).

Technology and Methods are found fairly often in topics. Note that the identification of these themes is facilitated because they can be associated to concrete terms such as device, cloud, and platform for Technology, or model, infer, and simulate for Methods. From the V's, Volume and Velocity were the most identified themes, which are also easily associated with terms such as large scale, performance, and computability. These terms are frequently used in practice, explaining why they have been so strongly identified in topics 4, 5, 6, 7, 12, 13, and 20.

Impact, Variety, Veracity, Value, and Beyond Conventional were annotated less often. Because these are more abstract concepts it is likely that they are more difficult to discover within topics. For example, Value was annotated to topic 1, containing words such as secure, challenged, and protect. Compared to concrete themes (e.g., Technology and Volume), it was more difficult for the annotators to find a fitting theme. Variability was annotated only twice, however we do believe that it is an integral part of Big Data. Variability not being recognised could mean that the observers could not identify the theme properly (due to poor theme description or understanding), or that the topics in the selected model could not capture this theme (due to insufficient representation in the corpus).

Each of the themes from the definition by De Mauro et al. (Information, Methods, Technology, Impact) was annotated more often than any other (apart from Application). Note that by design these themes are defined in a broader manner, meaning that they include the others. For example, Methods includes a few V's such as Volume and Velocity as well as Beyond Conventional. Perhaps due to their broadness, the themes from De Mauro et al. were chosen more easily, indicating that their definition covers the understanding of Big Data in a better way. However, one might wonder whether these themes are exclusively related to Big Data or whether they will also pop-out in other types of papers. The set-up of our study is not able to answer this question.

2.5 Related work

Other studies have been performed to discern a definition of Big Data [7, 8, 13]. These have provided an overview of Big Data research in different research fields [13]; a literature analysis to discover Big Data themes and a proposal for their consolidation into one definition [8]; and an analysis of industry statements on Big Data [7]. Each of these studies used qualitative methods, whereas our work builds upon their findings with a quantitative

method. In particular, our study provides evidence that supports the definition proposed by De Mauro et al. [8] and an aggregation of its underlying definitions (see Table 2.1).

Many researchers have applied TM for text analysis in various fields [62]. Most similar to our approach is a study by Hansmann and Niemeyer [29], that applied TM to a Big Data corpus to discover its characteristics. Their research identified three themes, namely IT infrastructure, methods, and data, and applied TM in two stages. The first stage separated the corpus of 248 manually selected papers into the three themes mentioned above. Then, in the second stage, TM was applied to the papers that had been grouped by theme. An in-depth word-by-word analysis of Big Data characteristics was performed on the second stage TM results. The meaning of each word was assessed, finding the important concepts for each of the themes and where research focus lies in the corpus. Our work differs from [29] in three ways. First, their analysis was based on only three Big Data themes, whereas we used multiple definitions leading to twelve themes. Secondly, we collected a larger corpus resulting from a systematic review of the literature. Lastly, the research goals differ: instead of finding the defining concepts for each of the themes, our approach identifies existing definitions in a biomedical Big Data corpus.

There are also more sophisticated (and complex) text analysis approaches such as the method described by Hurtado et al. [63]. Whereas we applied a bag-of-words principle, where each word is considered independently, the method by Hurtado et al. processes whole sentences and preserves context information. In [63] text mining was applied to find trends in topics over time and predict topic popularity in the future. While this is not applicable in our current case it might be interesting for further research (e.g., finding trends of Big Data over time within scientific literature). Lastly, their method to generate topics also gives them a concise label built from the topic's keywords. This would partially remove subjectivity from annotation, however interpretation of the results is still bound to human interpretation.

2.6 Conclusion

In this work we describe a systematic study that attempted to answer the question: 'Which themes from various existing Big Data definitions are expressed in (bio)medical scientific publications?'. A large number of existing definitions were analysed and consolidated into twelve themes. A large corpus of representative biomedical scientific publications was collected and automatically analysed with text mining to identify the 25 most relevant topics based on title and abstract. Manual annotation was performed by seven observers to identify Big Data themes in the topics. In spite of the limitations of our study, the results show that these themes can be identified in this corpus. Volume, Velocity and Value are recognized frequently, but in particular results show strong presence of the themes defined by De Mauro et al. (i.e., Information, Methods, Technology, and Impact). This finding indicates that their definition of Big Data is supported by the current understanding expressed by authors when they use the term Big Data in their own (bio)medical publications in this corpus. To our knowledge this is the first time that this is shown in a systematic manner for literature in an application field.

Acknowledgements

This work was carried out on the High Performance Computing Cloud resources of the Dutch national e-infrastructure with the support of SURF Foundation. Furthermore, we would like to thank the observers for their work on annotating the results. This publication was supported by the Dutch national program COMMIT/.

Appendices

Table 2.4: Top 20 words for the 25-topic model identified with TM

Topics				
1	2	3	4	5
health	patient	article	algorithm	challenged
research	clinic	review	cluster	analyte
healthcare	hospital	discuss	learn	tool
policies	electron	field	method	amount
health_care	care	recent	feature	technologic
privacies	outcome	issue	efficiencies	computability
nation	medicaid	aspect	approximate	analysing
ethic	record	focus	tree	require
protect	ehr	emerge	represent	advance
govern	clinical_research	future	fast	varieties
inform	health_record	highlight	matrix	solution
secure	clinician	current	accuracies	growth
challenged	treatment	context	problem	large_amount
share	improve	overview	distance	massive
concern	assess	paper	hierarchical	generate
access	healthcare	paradigm	computability	dataset
communities	qualities	confer	faster	vast
fund	potential	natural	calculate	process
health_informatics	patient_care	technologic	graph	handle
health_system	routine	literature	outperform	infrastructural
6	7	8	9	10
system	model	age	change	network
process	predict	risk	nurse	molecular
device	infer	influenza	innovated	structural
framework	statistic	indicating	science	biomarker
cloud	regress	exposure	social	complex
architectural	simulate	cohort	question	heterogeneities
hadoop	predictor	rate	historian	integral
applicability	bayesian	symptom	influence	systems_biology
service	fit	month	practical	mechanical
manage	good	yearbook	insight	omic
platform	optimal	variable	cultural	approach
design	prior	life	turn	character
mapreducible	base	death	product	dynamomics
computability	variable	diabetes	food	function
base	machine_learning	adjust	societies	biologic
support	high_dimensional	geographic	understand	transit
implement	tradition	condition	drive	edge
task	rank	factor	evolution	topological
deploy	parameter	demographic	scientific	protein
cloud_computing	feature	incidence	principle	organ
11	12	13	14	15
disease	dataset	effect	search	biomedical
prevent	time	group	social_media	informatic

Table 2.4 Continued from previous page

epidemiologic	sample	measurable	language	science
vaccination	large_scale	testable	google	medicinal
progress	computability	estimate	word	medicaid
immune	speed	analysing	public	educate
leverage	performance	studied	relate	research
popular	increased	statistic	psychological	learn
initial	approach	bias	trend	personalized_medicine
develop	thousand	large	emoticon	era
heart	step	random	twitter	ontological
administration	rate	valuable	message	disciplinary
intervention	implement	power	online	translate
generate	full	method	relationship	student
blood	memorial	sample_size	social	scientist
advance	scale	marker	visit	train
public_health	hundred	find	content	impact
reported	block	large_set	caseness	workshop
consensus	applicability	import	posit	discoveries
earlier	multiple	error	investigacin	knowledge
16	17	18	19	20
genet	web	sequence	mine	classifiable
gene	resource	genome	knowledge	set
associating	code	bioinformatic	extract	object
phenotype	file	proteome	inform	large_set
pathway	laboratories	high_throughput	chemical	class
disease	public	dna	specialised	noise
genotype	compress	transcriptome	plant	general
factor	semantic	protein	biologic	pair
enrich	software	composite	concept	performance
trait	retrievable	ngs	develop	abilities
genome_wide	access	metagenome	toxic	neural_network
metabolic	share	virus	construct	similar
genome	format	analysing	note	train
mutated	inform	host	curate	dimension
number	interface	biologic	rich	machine
identifi	source	assemble	gap	categorical
polymorphism	platform	cell	preservation	appliance
individual	metadata	microbiome	ecological	formula
regular	storage	align	diverse	encounter
unification	exchange	human	abstract	coefficient
21	22	23	24	25
drug	visual	image	cancer	low
target	activated	brain	studied	reduce
cell	human	disorder	tumor	time
event	behavior	signal	valid	base
screen	mobile	subject	research	reduction
response	environment	resolution	registries	digital
experiment	interact	neuroimaging	therapeutic	node
detected	exploration	function	database	energies
analyse	user	neuron	injuries	deep
adversary	collect	segment	oncologist	small
multiple	sensor	psychiatric	clinical_trials	cost
compound	tool	connectome	claim	size
profile	wearable	neuroscience	therapies	numerator
miss	quantifiable	mode	efficacies	operability

Table 2.4 Continued from previous page

type	track	mri	diagnostic	combina
potential	movement	scan	heterogeneities	peak
combina	physical	quantitation	set	spectral
meta	display	analysing	specific	structural
complete	smartphone	microscopic	ongoing	locate
point	interest	multi	consortium	qualities

Table 2.5: Raw annotation results per observer. The following coding is used to represent the themes described in Table 2.1: vol = Volume, velo = Velocity, vera = Veracity, info = Information, met = Methods, tech = Technology, imp = Impact, app = Application, beyond = beyond conventional.

Topic	Theme assignment grouped by observer						
	A	B	C	D	E	F	G
1	imp, value		value	app, imp, value	vera, value	imp, app, vera	imp, value
2	vera, app		imp, app	info, app	vera, velo	app	tech, variety, vera
3					imp, app	app	app
4	met	met	vol, met	met	tech, met	tech, velo	met
5	vol, velo, beyond	tech	vol, tech, beyond	beyond, vol, velo	tech, complex, beyond	vol	vol
6	tech	tech	tech, velo	tech, beyond	tech, beyond	tech	tech, variety, vera
7	met	met	vera, met	met	tech, met, info, app	met	met
8	app	app	info, app	app, info	app	app	variety, app
9	app			imp	imp	imp	value, imp, app
10	app	met, tech	variety, info, met	app, met	app	app, variety, info	vol, beyond
11	app	app	app	app, imp	app	app	imp, value
12	tech, vol, velo	vol	vol, velo	vol, velo, beyond	tech, vol, velo	vol, velo	met, vol
13	variability, vera	met	met	met	app, info	met	met
14	info	info	tech, app	app, info	imp	info	value, imp, app
15	imp	app	imp	app	info, app	app, imp	value, vera
16	app	met	app	info, app	info, app	app	beyond, vol
17	value	info	tech, beyond	info	continuity, variability	tech	value, tech
18	app	met	info	app, info	met, app, tech, info	app	vol, vera
19	value	app	met, app	info	continuity, app	variety	tech, imp
20	met	met	met	met	met, info	met	met
21	app	app	app	app, imp	info, app	app	variety, app, vera
22	info, velo	info	info, app	info, vera	velo, continuity, app	app, info	info
23	info, app	app	info, app	info	info	app, info	beyond, vol, vera, info
24	value	app	info, app	info, app	continuity, info, imp	app	vol, variety
25	met	met	info	info, app	info, met, tech	vol, velo	velo
total	33	22	39	40	53	35	49



Usage of the term Big Data in biomedical publications: a text mining approach

In Big Data and Cognitive Computing, volume 3, article 13, 2019

A.J. van Altena, P.D. Moerland,
A.H. Zwinderman, S.D. Olabarriaga

Abstract

In this study we attempt to assess the value of the term Big Data when used by researchers in their publications. For this purpose, we systematically collected a corpus of biomedical publications that use and do not use the term Big Data. These documents were used as input to a machine learning classifier to determine how well they can be separated into two groups and to determine the most distinguishing classification features.

We generated 100 classifiers that could correctly distinguish between Big Data and non-Big Data documents with an area under the Receiver Operating Characteristic (ROC) curve of 0.96. The differences between the two groups were characterised by terms specific to Big Data themes – such as ‘computational’, ‘mining’, and ‘challenges’ – and also by terms that indicate the research field, such as ‘genomics’. The ROC curves when plotted for various time intervals showed no difference over time.

We conclude that there is a detectable and stable difference between publications that use the term Big Data and those that do not. Furthermore, the use of the term Big Data within a publication seems to indicate a distinct type of research in the biomedical field. Therefore, we conclude that value can be attributed to the term Big Data when used in a publication and this value has not changed over time.

3.1 Introduction

With approximately 3,700 documents mentioning Big Data in the PubMed library between 2011 and the time of writing, it can be said that the term Big Data is widely used in biomedical research. This, however, does not mean that a clear-cut meaning of the term is being applied, as can be attested from the many publications – both formal and informal – written on the subject. This sentiment is underwritten in publications such as Tian et al. [64] and Mayer-Schonberger et al. [65], which state that there is no rigorous definition of Big Data and it still remains something of a work-in-progress. The reasons above, in conjunction with a massive increase in use in the last few years [66], raises the question of what value the term holds when used in a scientific document.

By comparing documents that use the term with those that do not, one can find out what distinguishes these two groups of documents from each other and determine how well they can be separated [67]. We further refer to these two groups respectively as Big Data (BD) and non-Big Data (NBD) documents. The degree to which BD can be separated from NBD documents gives insight in the value of the Big Data term, and inspecting the distinctive features tells us something about its meaning. The influence of some hype effect can be measured through the change of value of the term over time.

In our work we are particularly interested in discovering differences between BD and NBD documents in the scope of biomedical research literature. Our hypothesis is that the term Big Data describes research with common characteristics that are distinguishable from those found in other biomedical research. Also, we hypothesise that, through overuse or hype, the meaning of the term has become diluted over time. In this study we therefore investigate the following questions:

1. How well can documents that use the term Big Data be distinguished from documents that do not use the term in a comparable corpus?
2. What are the distinguishing features between BD and NBD documents?
3. Does the distinguishability of BD and NBD documents change over time?

The large number of published literature makes it nigh impossible for a researcher to keep up with the status quo [68]. Therefore, we seek answers to these questions through text mining on a corpus of BD and NBD documents from two biomedical literature databases. The label BD or NBD was given based on the presence or absence of the term Big Data in the title or abstract. BD and NBD documents were cleaned and preprocessed to be used as input to a machine learning classifier that trains a model to determine the most distinguishing features. To assess the stability of the applied methods multiple datasets were created and tested, each with a different random mix of documents. Features that were selected consistently were used in further analysis.

This work builds on previous research published in a conference proceedings [69]. In this previous work we also investigate whether BD and NBD documents are distinguishable using text mining tools. There we concluded that Big Data biomedical research articles can be reliably identified. Here we extend that work, the BD corpus in the current study has nearly doubled in size and we analyse a larger portion of the available scientific documents. Furthermore, the analysis methods were adapted and simplified.

3.2 Related work

The meaning of Big Data is being discussed at various levels. In 2001 Gartner published a report which in hindsight is often referred to as the first description of Big Data. It defines the term through information technology challenges described by three Big Data aspects (V's): volume, velocity, and variety [12]. This definition has had many additions and adaptations over time and a relatively stable six V's (volume, velocity, variety, veracity, value, and variability) are in common use nowadays [13].

At an informal level, various blogs have debated about the usage of the term Big Data, and the hype that surrounds it. These blogs cover a wide spectrum of opinions expressed by members of the scientific community and industry. At one end of the spectrum we have 'The emperors cloths', by Levi [27], which states that Big Data appears to be a fad with many potential downfalls in the medical field. On the other end, some state that Big Data has become the 'new normal' in information processing. Gartner describes that the aspects of Big Data have evolved into various other areas such as data science [70]. IBM states that Big Data techniques are no longer an option, but a necessity [71]. The large majority, however, adopt definitions of Big Data that often focus on technological aspects such as the storing and processing of data [72, 73].

While blogs are informal and subjective sources, there are also many researchers investigating the meaning of Big Data more systematically. Some approached this in a qualitative manner by analysing existing definitions, describing similarities and differences, and merging them into an overarching definition. For example, De Mauro et al. [8] looked at fifteen existing definitions and derived four overarching aspects that define Big Data: *Information* describes the aspects directly related to data such as its volume and variety; *Technology* and *Methods* describe the techniques to make use of data; and lastly *Impact* describes the value – either scientific or economic – that data may generate. Others are, for example, Ward et al. [7] and Gandomi et al. [74] which assess existing (industry) definitions to find and describe common aspects between them. There is also research focused on definitions within a specific research area, such as: Kudva et al. [75] for smart cities, Wolfert et al. [76] for smart farming, and Hashem et al. [77] for cloud computing. These studies are aimed at helping researchers identifying the intersections between the research area that they know and Big Data.

Other researchers applied quantitative methods and extracted common features from research publications. Hansmann et al. [29] identified topics in a corpus of Big Data publications and described them in the light of existing definitions. They concluded that Big Data is described by data, information technology infrastructure, and methods of data analysis. Similarly, our previous work [78] mined the topics of Big Data publications and matched them against the six Big Data V's and the definition posed by [8]. We concluded that, while some V's are often identified (volume, velocity, value), the presence of aspects from the definition of de Mauro et al. is especially strong.

More closely related to the research in this paper is the work of Hahn et al. [66] who analysed the changes in popularity of specific areas in bioinformatics over time. They gathered a set of scientific literature and applied a keyword and topic modelling based analysis. Their results show that the term Big Data has a massive increase in popularity over time and several research areas of bioinformatics are shifting to Big Data techniques.

The previously mentioned studies attempted to understand and define Big Data in the broad scope of a research field, including methodological aspects. The meaning of Big Data, however, has also been derived from the characteristics of datasets alone. By applying a taxonomy [79] of potential Big Data aspects to 26 datasets, Kitchin et al. [80] investigated which aspects are common in 'Big' datasets. They concluded that velocity and exhaustivity (i.e., the dataset is a sample or $n = all$) are the most distinguishing aspects. Moreover, they stated that volume and variety, which are traditionally related with Big Data, do not qualify as meaningful aspects without velocity or exhaustivity.

As it can be seen from the above, the term Big Data may be used to describe different aspects. Defining Big Data only through dataset characteristics, as proposed by Kitchin et al., provides a narrow perspective, excluding aspects such as methods and technology that are included by many others. Depending on the point of view, Big Data definitions may overlap but are often not fully in agreement with each other. Therefore, when the term Big Data is used in a scientific document, it is unclear what it really means and whether this is a marker of unique characteristics.

3.3 Data and methods

3.3.1 Corpus collection

The corpora were obtained through querying and cleaning of BD publications, and then matching these to NBD publications through new querying and cleaning steps. Overviews are shown in Figures 3.1a and 3.1b respectively. The implementation of the methods described in this section can be found on GitHub [81].

3.3.1.1 Big Data corpus

BD documents were collected from PubMed and PubMed Central (PMC) using the Entrez Programming Utilities API [82]. We searched for the literal use of the term “Big Data” in either the title or abstract. The following search queries were used:

- *PubMed* (“big data”[TIAB] OR (big[TI] AND data[TI]))
AND (“2011/01/01”[PDAT] : “3000/12/31”[PDAT]) AND english[Language]
- *PMC* (“big data”[TI] OR “big data”[AB]) AND (“2011/01/01”[PDAT] : “3000/12/31”[PDAT])

The query did not allow distance between the words ‘big’ and ‘data’ to minimise the number of irrelevant results. For the same reason we limited the search to publications after 2011. Note that 3000/12/31 is the default value that PubMed uses when no limit is given for the end date. We noticed that documents containing the term “Big Data” between single quotes were not returned by the PubMed search, therefore the sub-query (big[TI] AND data[TI]) was added and the gathering was repeated.

The search used the `esearch` function of the Entrez API, which yielded 3,679 PubMed and 1,387 PMC results. With the `efetch` function the following information was retrieved and stored in a local database: titles, abstracts, and metadata (i.e., publication date, publication type, DOI, journal, journal ISSN, and journal ISO).

An overview of the cleaning process is shown in Figure 3.1a, and the steps are described in order below.

(1) Some documents had to be removed as they could not be retrieved by the `efetch` function. (2) Documents with empty abstracts were removed as they did not contain enough data to be useful in the classification. (3) In our previous study [69] we observed that documents such as comments and letters to the editor have different structure and content, therefore documents other than research papers were removed¹. The document type was determined with the `PublicationTypeList` field in the Entrez API output. (4) We observed that not all journals in the corpus primarily covered biomedical research, so these had to be removed manually. All journals with three or more documents in the corpus were inspected by one of the authors (AA), as we assumed that journals with less documents did not have a big impact on the corpus overall. The titles of the documents were scanned to estimate the research field of the journal, and where the field did not become clear the abstracts were analysed as well (see Dataset S1 for the complete list of journals). (5) Lastly, any duplicates were removed based on title or DOI.

The search was performed on 2018/05/13 and yielded 5,066 documents, and through cleaning 2,554 were removed, resulting in a BD corpus of 2,512 documents.

3.3.1.2 Non-Big Data corpus

NBD documents were collected through the Entrez API similarly to the BD corpus. To make the NBD documents comparable with BD documents, the PubMed and PMC databases were

¹Full list of removed document types: Addresses, Bibliography, Biography, Book, Clinical Conference, Comment, Congresses, Consensus Development Conference, Consensus Development Conference, NIH, Dataset, Directory, Editorial, Guideline, Interview, Lectures, Letter, News, Published Erratum.

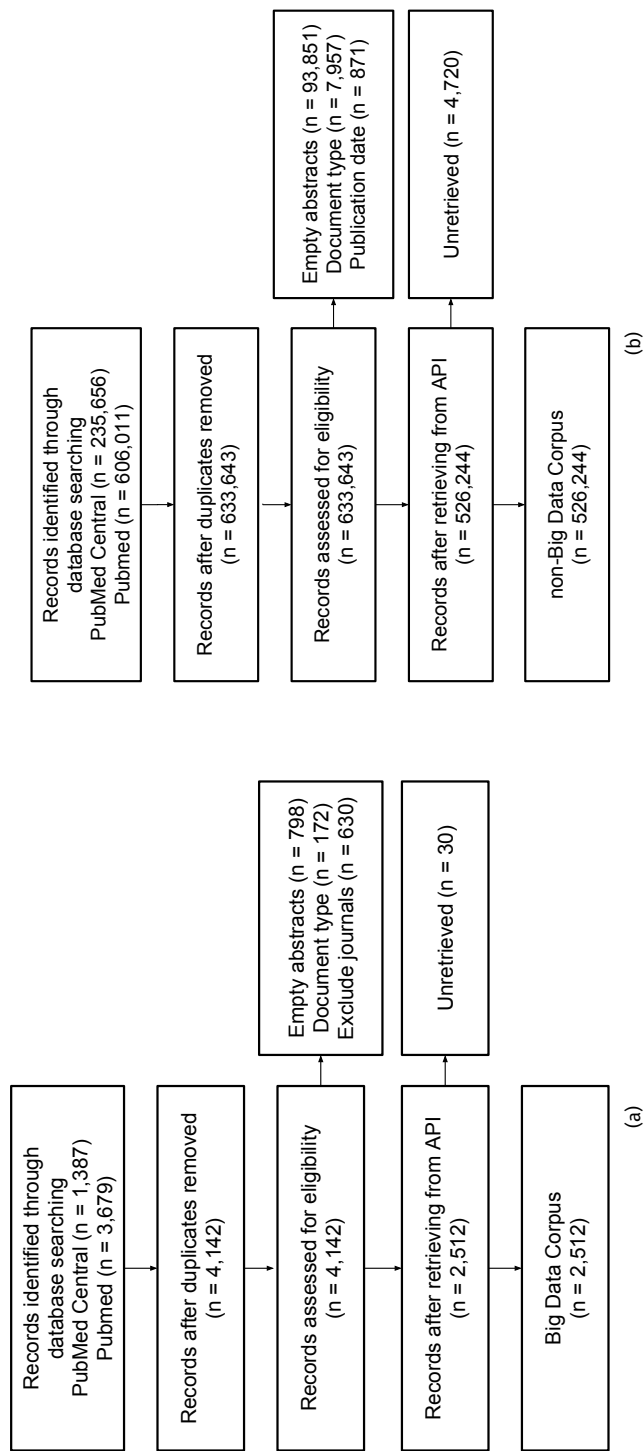


Figure 3.1: Respectively the results of the search, fetch, and cleaning of: (a) the Big Data corpus; and (b) the non-Big Data corpus. Diagram has been adapted from the PRISMA guideline [83] to fit our use case.

queried for each journal in the BD corpus. Furthermore, the publication date range was set to the minimum and maximum publication year of the BD documents in each journal. For example, the following query would match BD publications between 2012 and 2016 in the journal *Nature Communications*: “*Nature Communications*”[Journal] AND (“2015”[PDAT] : “2018”[PDAT]).

An overview of the cleaning process is shown in Figure 3.1b. The process was similar to the BD corpus cleaning described above, with two differences: (1) no journals had to be removed; (2) some documents were pre-published and had a publication date in the future, so these were removed.

The match was performed on 2018/05/13 and yielded 841,667 documents. Through cleaning 315,423 were removed, resulting in a NBD corpus of 526,244 documents.

3.3.2 Dataset preparation

In this section we describe the preprocessing of the individual documents from the BD and NBD corpora and their characteristics. Furthermore, we describe the sampling of the datasets used as input to the classification method as described in Section 3.3.3. The implementation of the methods described in this section can be found on GitHub [81].

We cleaned all documents so that they contained only unaccented alphabetical letters. The following items were removed: HTML tags², special characters (e.g., &, %), and numbers. Stopwords were removed using the `english` list from the NLTK python library [84] in addition to ‘big data’ and ‘big’. Lastly, the documents were tokenised and any too short (< 2 characters) or too long (> 34 characters) tokens were removed, as they were unlikely to be real words.

The characteristics of the corpora after document cleaning are shown in Table 3.1. Word clouds of the top-100 most frequent terms in both the BD and NBD corpora are shown in respectively Figures 3.2a and 3.2b. When normalised, the corpora showed a similar trend in documents per year and tokens per document (shown respectively in Table S2 and Figure S3). Note that the minimum number of tokens in the NBD corpus was zero for both the title and abstract, indicating empty fields. Later inspection showed that this was due to three malformed documents in PubMed (PubMed IDs: 27529366, 27529367, and 27529368).

We sampled datasets so that they consisted of an equal number of BD and NBD documents. The sampling process is shown in Figure 3.3. For each dataset the whole BD corpus was included and paired with a random sample of the NBD corpus, resulting in sets of 5,024 documents. Datasets were split into 90% training and 10% validation data. To cover a larger part of the NBD corpus and test the stability of the classifier, 100 datasets were created. We did not apply a cross validation, as each dataset was randomly sampled from the NBD corpus. While this approach does not guarantee coverage of all NBD documents we assume that the random sampling ensures a fair spread of the variety in the NBD documents.

²PubMed data may contain the following tags: <i>, <u>, , <sup>, and <sub>



Figure 3.2: Word cloud of the: (a) Big Data corpus and (b) non-Big Data corpus. The top-100 words are shown, their size is proportional to their frequency in the respective corpus.

Table 3.1: Characteristics of the corpora after cleaning. *: mean [minimum - maximum]. Docs: Documents. Note that 2018 only covers 2018/01/01 to 2018/05/13.

	Big Data	non-Big Data
# docs	2,512	526,244
# journals	1,189	1,144
# docs per journal*	2 [1-73]	460 [1-10,298]
# docs per year		
2011	5	839
2012	18	5668
2013	100	23825
2014	271	53307
2015	411	87220
2016	631	134876
2017	728	175590
2018	348	44919
# tokens		
all*	133 [13-516]	139 [0-1,210]
title*	9 [1-28]	10 [0-57]
abstract*	125 [10-511]	129 [0-1,205]
# unique tokens		
all*	94 [12-287]	91 [0-425]
title*	9 [1-24]	10 [0-48]
abstract*	92 [10-287]	89 [0-424]

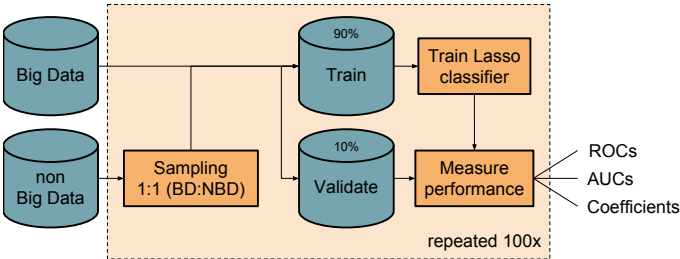


Figure 3.3: Pipeline from sampling the data to training the models and retrieving performance metrics.

3.3.3 Classification

In this section we describe how the classifiers were trained and how performance measures were calculated. The input to the classifiers were the 100 datasets constructed as described in Section 3.3.2. Furthermore, we describe how the influence of time (i.e., publication date) was evaluated. The implementation of the methods described in this subsection can be found on GitHub [85].

The process of classification is shown in Figure 3.3. We implemented a logistic regression with LASSO penalty using the `glmnet` R package [33, 86]. This method was used because of its ability to discard features and limit the size of the final model, thereby identifying the most relevant features.

For each dataset the training data was used to fit a model with `cv.glmnet`. A range of lambda values was tested using 10 folds. Predictions and coefficients were extracted using `lambda.1se` on the validation data. `lambda.1se` was chosen instead of `lambda.min` because it gives the simplest model within one standard error of the minimal misclassification rate, limiting the number of selected features. We extracted the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) to assess the performance of the classification models. Furthermore, the coefficient values of the selected features were retrieved.

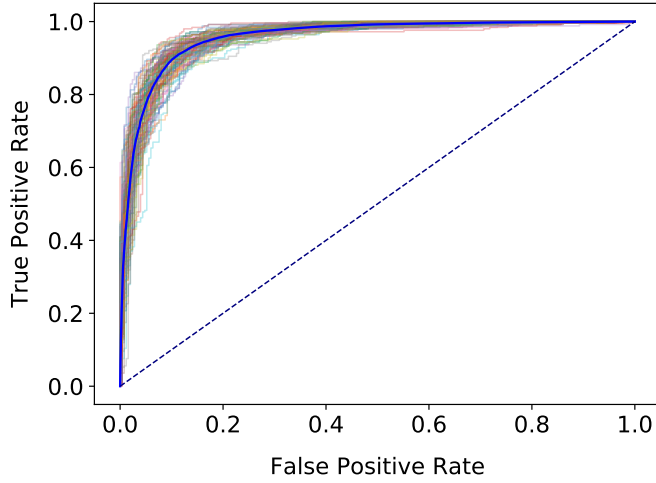
Trend analysis To answer research question 3 a stratified analysis by year of publication was performed. We used the BD and NBD corpora, but split them into the following bins: 2011-2014, 2015, 2016, 2017, 2018. Note that documents from 2011 to 2014 were combined because each year included a relatively small number of documents, which could result in unreliable results. For each bin we sampled twenty datasets with the same approach as described in Section 3.3.2.

Each dataset was classified using the same process as described in Section 3.3.3, and additionally a confusion matrix was retrieved. The matrix was used to calculate the False Omission Rate (FOR). This metric was chosen because it reflects the chance of a negatively classified document to be false negative. We hypothesized that over time this metric would increase caused by a dilution of the value of the term Big Data. When the value of 'Big Data' becomes diluted, documents without Big Data characteristics might carry the BD label, and be included in our BD corpus. If no Big data characteristics are present, the classifier should label them as NBD, resulting in a false negative. This situation is captured by the FOR.

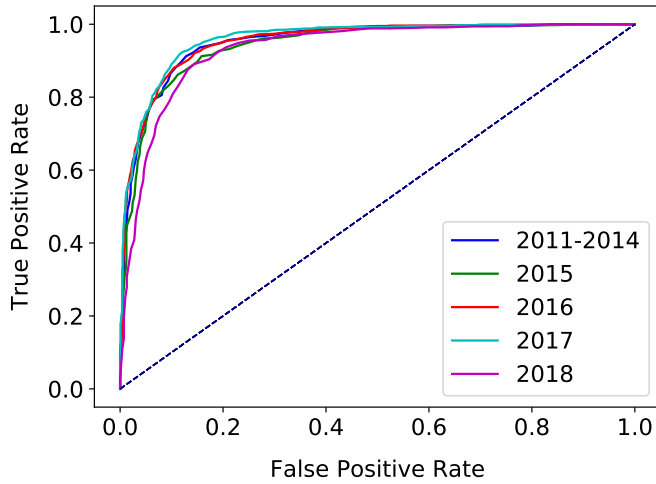
3.4 Results

The mean AUC over all 100 datasets was 0.96 with a standard deviation of 0.009. ROC curves are shown in Figure 3.4a and were relatively stable over the 100 datasets. Lastly, we retrieved the frequency of each unique feature. Logistic regression with LASSO penalty creates a model with a subset of the input features, therefore the features may differ between each model. As described above a model was trained for every dataset, we counted the frequency of each unique feature. Then, all features that occurred at least fifty times were used to create a word cloud, which is shown in Figure 3.5. The results were used to answer research questions 1 and 2.

The outcomes of the analysis are shown in Figures 3.4b and 3.6. Note that the ROC curves and FOR curve do not show trends along time.



(a)



(b)

Figure 3.4: (a): ROC curves for all 100 datasets with average curve highlighted (blue). (b): ROC curve for each period of time. Each period of time consists of twenty datasets, the mean curve is plotted.



Figure 3.5: Word cloud of the selected features for all 100 datasets. Their size is proportional to the number of times they were selected.

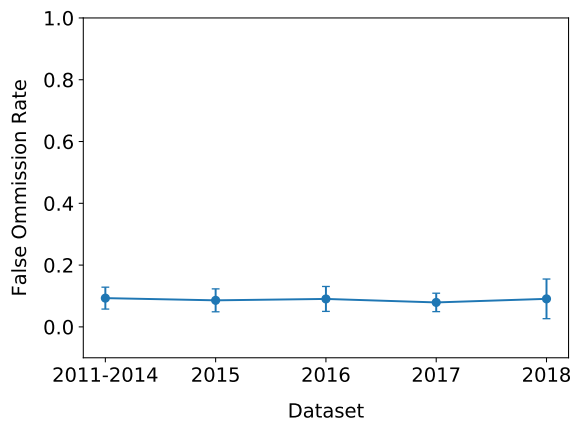


Figure 3.6: False Omission Rate over time, each year consists of twenty datasets, the mean and standard deviation are plotted.

3.5 Discussion

In this study we set out to answer the question whether and how a corpus of Big Data documents can be separated from non-Big Data documents published in the biomedical literature. Furthermore, we looked into the distinguishing features and whether the date of publication of a document has an effect on its distinguishability. Below we analyse and discuss the results and limitations concerning the creation of the datasets and classifiers, distinguishing features and trends over time.

Corpus, datasets, and classification We created two corpora – BD and NBD – and randomly sampled 100 datasets from them, each with a one-to-one ratio of BD and NBD documents. For each dataset a classifier was trained and the classification performance was tested. The generated models have a high performance with an average AUC of 0.96. Analysis of the ROC curves showed that the model performance remains stable over the 100 datasets. These results answer research question 1: a classifier based on bag-of-words approach can reliably and with a high performance separate BD and NBD documents.

Distinguishing features To give an impression of the BD and NBD corpora two word clouds were created, respectively Figures 3.2a and 3.2b. While the most frequent word (respectively data and patients) differs between the corpora, there is much overlap between the word clouds. Most of the differences may be found in the research fields that are covered. For example, the BD word cloud contains genetic and genome, most likely stemming from the genomics field, which has a high interest in Big Data applications. These insights, however, do not give a complete story about the differences between the BD and NBD corpora. Therefore, research question 2 was answered by extracting the words that were selected as most distinguishing features between the two sets of BD and NBD documents – see Figure 3.5. We apply below the Big Data definition proposed by de Mauro et al. [8] to interpret these words.

Under the *Information* aspect, words such as massive and large are the most noticeable. More interestingly, many words can be associated with *Technology* and *Methods*, for example: computational, mining, and machine (possibly from “machine learning”). *Value* aspects can be identified in words such as future, era, and challenges. Note that a word like era, as in “the era of big data”, can also be associated to hype. Lastly, there are words that do not fit in the definition as proposed by de Mauro et al., but instead identify a specific research fields such as omics and genomics. These words are related to the areas that tend to handle large datasets.

Note that other Big Data word clouds have been published, for example at the Gartner blog [73] and the United Kingdom parliament website [72]. Note that these word clouds include more words that are associated with the size of data – petabytes, volume, size – as compared to our word cloud in Figure 3.5. However, many words are similar, therefore supporting our findings.

Trends over time There is a clear increase in the usage of the term Big Data in biomedical literature along time. Here however our focus is in changes over time regarding distinguishability between papers that use the term and that do not.

In our previous work [69] we found a trend over time in the False Discovery Rate (FDR). This indicated that more papers were incorrectly classified as BD in more recent years. From this we concluded that, whilst Big Data concepts are still being discussed, researchers used the term Big Data less often in later years, although their content includes Big Data aspects. In the current study we found no trend for the FDR (data not shown) neither for the FOR. While the current work does not differ in methodology from the previous one, it uses better data. In the study presented here we improved the document searching and sampling approach by restricting the types of included BD documents while increasing the

number of matched NBD documents. We believe that the current datasets better represent the published works in biomedical literature that are relevant for this study.

The ROC curves for various time intervals (Figure 3.4b) show no trends in distinguishability. The same conclusion can be drawn for the FOR (Figure 3.6). These results answer research question 3, rejecting our hypothesis that the term Big Data became diluted over time.

Value of the term Big Data In Section 3.2 we showed that there is a wide spectrum of opinions on the value and the definition of the term Big Data. Concerning the definition, our findings show that the term is consistently used to identify a distinct field of research within a biomedical scope. Moreover, the characteristics of this field align with existing formal and informal definitions of the term.

With respect to the value of the term, our findings do not support the opinions that Big Data is a fad or the 'new normal'. A fad would die out over time, and a 'new normal' would permeate the literature. In both cases one would expect to see less distinguishability as time progresses. However, we did not find such a trend over time, which suggests that these opinions are not valid in the context of biomedical scientific literature.

3.5.1 Limitations

There were several limitations to our approach. Firstly, we restricted the corpus to biomedical documents, therefore the BD and NBD corpora were collected from two biomedical online libraries, PubMed and PMC. There are other libraries available such as Scopus and Ovid, however they do not provide a public API, which would make this study impractical.

Another limitation is the use of only titles and abstracts in the analysis, because the full-text is not directly available in the used libraries. We assume that the main message of each document is represented in their title and abstract, but it is possible that more complex concepts are only expressed in the full text. PMC contains open-access articles and therefore often, but not always, includes full-text in the API results. These would represent only a small portion of the BD corpus and were therefore not used in this study.

To ensure a certain quality in our corpora we had to remove documents, for example because they lacked an abstract. Note that, while some documents had to be discarded due to quality criteria, all eligible BD documents were included in our analysis. The corpus was also restricted on represented journals because we noticed that some journals included in the PubMed or PMC are not specific to biomedical research. We manually curated a list of journals to be removed from the corpus, however this was non-exhaustive and partly subjective. Therefore, some of the documents included in the corpus might be from other research fields.

Finally, we used all BD documents in each set and matched them with an equal amount of NBD documents. Because there were 2,512 BD documents, a theoretical maximum of 251,200 unique NBD could be included. While this is about half of the total amount of NBD documents, we assume that (even considering repeats) the random sampling ensures a fair spread of the variety in the NBD documents. The ROC curves show little variation between the models, supporting this assumption.

3.6 Conclusion

In this research we investigated the question whether Big Data literature in the biomedical field can be distinguished from literature that does not use the term. To our best knowledge, this is the first study to analyse this question using quantitative methods in this research field. From our results, we conclude that there is indeed a detectable and stable distinction between BD and NBD documents in the biomedical field. Furthermore,

we found no trends over time that indicate a change in the distinguishability between BD and NBD documents. This suggests that the value of the term remains the same, in spite of its increased usage in the biomedical literature.

The differences between the BD and NBD documents are mostly captured in terms that are associated with Big Data themes previously described by others. Furthermore, the distinguishing features seem to be sensitive to words that indicate data types belonging to certain research fields, such as 'omics'. These words suggest that certain research fields tend to use the term Big Data in their publications more often. This is probably due to the affinity of some areas of biomedical research with large datasets and computational methods, such as bioinformatics. Therefore, even when taking possible hype into account, the use of the term Big Data within a publication seems to indicate a distinct type of scientific publication in the biomedical field. Recognising this may help biomedical researchers to identify themselves with this new field, increasing participation in this growing community and taking more benefit from it.

Acknowledgements

This work was carried out on the High Performance Computing Cloud resources of the Dutch national e-infrastructure with the support of the SURF Foundation.



Adoption of automation tools



Usage of automation tools in systematic reviews

In Research Synthesis Methods, volume 10, pages 72-82, 2019

A.J. van Altena, R. Spijker, S.D. Olabarriaga

Abstract

Systematic reviews are a cornerstone of today's evidence-informed decision making. With the rapid expansion of questions to be addressed and scientific information produced, there is a growing workload on reviewers, making the current practice unsustainable without the aid of automation tools. While many automation tools have been developed and are available, uptake seems to be lagging. For this reason, we set out to investigate the current level of uptake and what the potential barriers and facilitators are for the adoption of automation tools in systematic reviews.

We deployed surveys among systematic reviewers that gathered information on tool uptake, demographics, systematic review characteristics, and barriers and facilitators for uptake. Systematic reviewers from multiple domains were targeted during recruitment, however, responders were predominantly from the biomedical sciences.

We found that automation tools are currently not widely used among the participants. When tools are used, participants mostly learn about them from their environment, for example through colleagues, peers, or organisation. Tools are often chosen on the basis of user experience, either by own experience or from colleagues or peers. Lastly, licensing, steep learning curve, lack of support, and mismatch to workflow are often reported by participants as relevant barriers.

While conclusions can only be drawn for the biomedical field, our work provides evidence and confirms the conclusions and recommendations of previous work, which was based on expert opinions. Furthermore, our study highlights the importance that organisations and best practices in a field can have for the uptake of automation tools for systematic reviews.

4.1 Introduction

Systematic reviews are a cornerstone of today's evidence-informed decision making [17, 87]. These can be decisions around medical tests or treatment, food safety, environmental questions, or avoidable research waste [88], just to name a few. By synthesising all relevant evidence regarding a certain topic, systematic reviews provide a good reflection of the current scientific knowledge.

A systematic review, as its name suggests, is a highly structured process which can be divided into 15 tasks [20] concerning the retrieval, appraisal, and synthesis of evidence. The whole process is mostly manual and time consuming. For a full-time researcher, depending on expertise and complexity, a review can take from 6 months to several years. With the ever-growing body of literature being produced [19], and the multitude of questions that need to be answered, the current practice is unsustainable [20].

Looking more closely at the various tasks within a systematic review, it becomes clear that some areas could benefit from automation to speed-up the process whilst maintaining the high standards associated with a systematic review. While there are many definitions of automation, here we understand automation as a repeatable computerised method that performs a task normally executed by the researchers or that aids in their decision making process.

An example of a task eligible for automation is the identification and selection of studies. A systematic review sets out to identify a comprehensive set of relevant studies across multiple sources in order to obtain a reliable sample of studies and minimising bias [17]. This is done by executing highly sensitive searches in multiple literature databases and other resources and retrieving between a thousand and tens of thousands studies, with some examples reaching up to one million studies [89]. Each study needs to be assessed manually by two independent researchers based on title and abstract, which is a very time-consuming process.

An analysis by Tsafnat et al. [20] showed that several automation tools have been developed to assist the screening of titles and abstracts based on machine learning methods. A comprehensive list can be found in the Systematic Review Toolbox [90] and an overview of their performance can be found in O'Mara et al. [18] and Shemilt et al. [89].

Although automation tools have been around for several years, their adoption seems to be low, as reported in the summary of the 2016 meeting by the International Collaboration for Automation of Systematic Reviews (ICASR) [91]. While experts from the field (e.g., ICASR, O'Mara et al. [18]) indicate that uptake is lagging, no studies have been carried out to investigate whether this is indeed the case and what might be causing it. Understanding the level of, and condition for, uptake would help to devise bespoke implementation strategies for the introduction of automation in the systematic review process. For this reason, we set out to investigate the following research questions:

- What is the uptake of automation tools that support the execution of systematic reviews?
- What are the barriers and facilitators that lead to the use of these tool?

We deploy two surveys among a population of systematic reviewers. The first survey is geared towards finding the uptake of automation tools, population demographics, and characteristics of systematic reviews. This survey is designed to be accessible and short so as to reach a high response rate. The second survey is a follow-up that goes in-depth on the specifics of automation tools. To capture all possible barriers and facilitators that lead to, or hamper, their use we applied an existing conceptual model for assessing the acceptance and usability of new technologies.

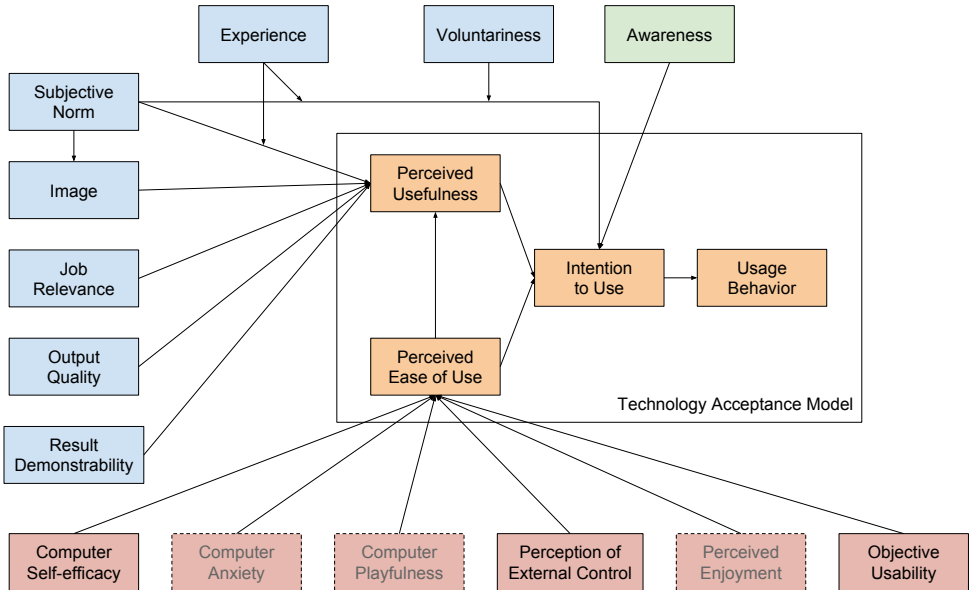


Figure 4.1: Technology Acceptance Model 2, adapted for this study. This model structured the questions for S2, see also Tables 4.4 to 4.6. The original model (orange) was established in 1985 [92]. The model was extended twice in 2000 [93, 94] (blue, red) which added more detailed concepts. Because we hypothesised that a barrier might be that reviewers do not know about the existence of automated tools, we added the concept Awareness to the model. From the concepts describing Perceived Ease of Use, only Computer Self-efficacy, Perception of External Control, and Objective Usability were used.

4.2 Methods

Our study collected data through two surveys among systematic review practitioners. In this section we describe how the surveys were designed, participants were recruited, and the data was analysed.

For the remainder of this paper the systematic review practitioners are referred to as *reviewers* or *participants*. Furthermore, the first survey is referred to as *S1* and the second survey as *S2*. Similarly, to refer to a specific question – for example question 1 in survey 1 – we use the following notation: *S1.1*.

4.2.1 Survey design

In this section we describe the process that led to the formulation of the questions presented to the participants in *S1* and *S2*.

4.2.1.1 First Survey

S1 covered three areas of interest, namely: demographics of the participants, characteristics of the systematic reviews they performed, and the automation tools they were using or considering to use. These areas were captured in ten questions summarised in Table 4.3 and detailed in Section 1 of the supplemental material [95]. The survey was implemented using SurveyMonkey [96]. Below, the rationale for each question is described.

S1.1-3 obtained information about the workload of the participant to assess whether a higher workload had an influence on the usage of tools. S1.4 covered the tools a participant might use - more details about tool selection are described below. S1.5-9 considered the following demographics: age, country, professional position (i.e., seniority), field of research, and computer proficiency. S1.10 was optional and collected an email address for participation in S2.

Tool selection for S1.4 Candidate tools were gathered from the 'Systematic Review Toolbox' website [90], which is a highly comprehensive list of tools compiled by researchers in the field. Based on publicly available information and personal experience, tools that automated any part of the systematic review process were selected¹. The tools were analysed by two researchers (AA and ML). Disagreements between their decisions were resolved by RS.

The Systematic Review Toolbox website listed 111 tools. Of these, 54 were selected by AA and ML, and, after disagreements were resolved, 31 tools remained. Finally, three well known tools (i.e., EndNote, RevMan, and Covidence) were added to the list as controls, although they do not fulfil the inclusion criteria for S1.4. See the complete list of tools in Section 1.4 of the supplemental materials [95]. Note that, to avoid missing out on tools, participants were encouraged to indicate tools that they felt were missing in the *Other* option of S1.4. Answers to this option were analysed by AA to identify additional automation tools.

4.2.1.2 Second Survey

S2 was structured using the Technology Acceptance Model 2 (TAM) [93, 94] illustrated in Figure 4.1. TAM defines the aspects that lead to Intention to Use, which may be used to predict Usage Behaviour of a certain (new) technology. The model was chosen to structure the questions in S2 because it describes the underlying motives, barriers, and facilitators (intention to use) for the uptake of tools (usage behaviour).

Below we describe each concept as used in our study (adapted from Chuttur et al. [97]):

- *Image*: in which way tool usage reflects upon the status of an individual among peers;
- *Experience*: the degree to which an individual has used the tool;
- *Voluntariness*: the degree to which an individual uses the tool out of own free will, as opposed to being obliged to use by, e.g., an organisation;
- *Job Relevance*: the degree to which a tool fits the task and workflow of an individual;
- *Output Quality*: the degree to which the output of the tool matches the individual's job goals;
- *Result Demonstrability*: the degree to which the usage of the tool can be linked with beneficial results (e.g., decreased job completion time);
- *Perception of External Control*: the degree to which an individual can control resources (time, money) that lead to usage of the tool;
- *Computer Self-efficacy*: covered in S1.9;
- *Objective Usability*: measured with the System Usability Scale (see below);
- *Awareness*: the degree to which an individual knows about the existence of the tool.

Note that some of the concepts were not used because they did not fit the goals of the survey. These were: computer anxiety, computer playfulness, and perceived enjoyment.

¹The used definition of 'automation' is stated in Section 4.1

Questions There were three categories of questions, which are summarised below – see Tables 4.4 to 4.6 for the questions that belong to each category.

- *The generic questions (S2.1-13)*: obtained information about how the participant learns about the existence of an automation tool, how they determine whether the tool fits their requirements, and how much effort that takes.
- *The specific questions (S2.14-24)*: obtained similar information as the generic questions, however posed in the context of an individual tool. This enabled assessment of barriers and facilitators comparatively among tools.
- *The usability questions (S2.25-34)*: were covered using the System Usability Scale (more details below).

S2 was implemented through a custom web-based system developed using PHP and MySQL to accommodate for the flexibility and control required for this survey [98]. Firstly, as control, the participants were presented with their previous answers to three questions from S1. They were asked to update the number of systematic reviews they were involved in (S2.0.2), the average number of search results (S2.0.3), and which tools they were using (S2.0.4). Based on their answers to S2.0.4, the participants were then assigned to one of two versions of the S2 survey.

One version included generic questions (S2.1-13), and was shown to participants that did not use, or considered to use, any tool. We refer to this as the ‘non-user’ group. The other version was presented to participants that indicated to be using some tool – the ‘user’ group – and included all questions (S2.1-34). The specific and usability questions (S2.14-34) were posed in the context of a single tool that the participant had indicated to be using in S2.0.4. To limit the burden of the participant, at first, specific information was asked for a single tool determined by the survey system. From all tools that the participant listed (S2.0.4), the tool with the smallest number of responses in the survey so far was selected to maximise the spread of responses. At the end of the survey, participants could choose to answer the specific and usability questions for additional tools of their own preference. The survey could be completed for as many tools as desired.

System Usability Scale TAM describes objective usability as a determining aspect for technology adoption. Furthermore, we hypothesised that usability would have a major influence on the use of tools.

The System Usability Scale (SUS) was chosen for its longevity, being proposed during the mid-eighties and formally published in 1996 [99]. The popularity of SUS has grown since then and, in spite of its pros and cons, the overall layout of the scale has remained stable [100]. The scale consists of ten questions (see Table 4.6), rated with a 1 to 7 Likert scale from *Strongly Disagree* to *Strongly Agree*.

For the usability analysis, answers to the ten questions were added together and interpreted as proposed by Bangor et al. [101]. A tool with score larger than 68 is considered to have good usability.

4.2.2 Participants

A variety of approaches were used to recruit participants for S1. Our aim was to get a mix of participants regarding workload, seniority, research field, country of residence, and tool usage.

Network Emails were sent to the professional network of RS, which includes the following organisations: Cochrane Information Retrieval Methods Group (IRMG) list, Cochrane editorial and methods digest readers, ICASR members and its contact list, and attendees of a meeting on automation tools in systematic reviews at Bristol University (UK).

Cochrane library Cochrane authors were contacted directly through the publicly available email addresses in the Cochrane library [102]. Articles from issues 1, 2, and 3 of the year 2017 were retrieved. The email of contact authors found in the articles was extracted and an invitation was sent.

Own Organisation To contact the reviewers in our organisation (Amsterdam UMC, University of Amsterdam) the library archives of 2015 and 2016 were used. These archives contain a fairly comprehensive list of all publications by AMC authors. The articles were filtered for systematic reviews using PubMed, retrieved, and an invitation was sent.

Own Department The survey was distributed in our department (Clinical Epidemiology, Biostatistics, and Bioinformatics) using an internal mailing list.

A single invitation was sent to all potential participants in S1. Responses were collected between October 2016 and April 2017.

Invitations for S2 were sent to all participants that provided their email address in S1.10. Up to two reminders were sent over the course of the subsequent four weeks. Responses for S2 were collected in the period between October and November 2017.

We also compared characteristics of the groups that answered S1 and the subgroup that answered S2 after the updates in S2.0.2-4, with the goal of identifying possible bias. This was done using the Pearson correlation coefficient implemented with the `rcorr` function from the `Hmisc` R-package [103, 104].

4.2.3 Data analysis

Data was cleaned by removing incomplete responses (i.e., participants that started, but did not answer all questions). Furthermore, some of the questions in S1 and S2 included an option *Other*, where the participant could enter any free text. These answers were inspected by AA and, where possible, mapped to the closest existing option. Answers that could not be mapped to an existing option were used to clarify our observations. Below we describe further processing for each survey.

First Survey The number of selected stages for S1.1 were counted per participant. The number of considered, incidentally used, and regularly used tools selected in S1.4 were counted per participant, as well as the total number of tools selected for each of these categories together. Finally, correlations between questions were analysed using the `hetcor` function from the `polycor` R-package [104, 105]. Question S1.6 was not included as the data was not suitable to test for correlation.

Second Survey There were three types of questions: multiple choice, single choice, and scales. For all question types the answer count and frequencies (fraction of total responses) statistics were determined. For multiple choice questions, in addition to the fraction of total responses, the fraction of participants that had chosen an option was determined. Lastly, for scales, the mean and median of the answers was calculated.

The questions S2.14-34 were posed in the context of a specific tool that the participant had indicated to be using, so the answers were grouped by tool. Tools with less than five completed surveys were pooled together to enable more meaningful statistical analysis.

Each question type was visualised as a table and a bar chart showing total frequency. The tool-specific multiple choice questions (S2.14-34) were grouped by tool and visualised as a grouped bar chart. Scale questions were visualised with box plots.

4.3 Results

In this section we describe the participants reached for both S1 and S2, the number of responses, and a summary of the responses.

4.3.1 Participants

First Survey Participants were recruited through four different approaches. The number of responses and contacts (where possible) for each of these is listed in Table 4.2. Note that the number of emails sent for the network and department are an approximation, as these were sent out to mailing lists. In total 172 responses were given to S1. Three of the responses were incomplete, so 168 responses were analysed.

Second Survey A total of 109 participants (65%) filled out their email address in S1.10 and were contacted for S2. For the initial invitation, two emails bounced and 7 returned an out-of-office, therefore the invitation reached 100 potential participants. S2 had 62 complete and three incomplete responses, therefore a response rate of 62% was obtained.

Six participants belonged to the 'non-user group', and filled out only the generic part of S2. The remaining 56 belonged to the 'user group' and filled out the complete survey.

Characteristics Participant characteristics measured in S1 resembled normal distributions with good spread for research stages (S1.1), age (S1.5), and position (S1.7). Computer proficiency was spread over the scale of 0 to 10 (S1.9). Nevertheless, more participants report their proficiency as 'basic' as opposed to 'advanced', showing a slight trend towards less proficient participants (median of 4, mean of 3.71).

Participants were spread over 22 countries (S1.5). The major countries of residence are the Netherlands with 33% of participants and the United Kingdom with 20%, followed by Australia (8%), Canada (7%), the United States (6%), and Denmark (6%).

Lastly, 95% of all participants (S1.6) performed reviews in the medical sciences field. Note that multiple fields could be chosen, the medical sciences field made up 78% of all answers given to S1.6. The complete distributions of the characteristic are shown in Section 1 of the supplemental materials [95].

To assess the participant bias considering the research domain a subgroup analysis was performed. For privacy reasons, questionnaire results were analysed on an aggregated level. The questionnaire design, therefore, limits the possibilities for an exhaustive analysis. We restricted the subgroup to the 81 participants recruited through the network method. This subgroup contains researchers with a mix of research environments. When comparing this group against the whole population of 168 participants little variance was shown in the research domain distribution. The full analysis is shown in Section 1.8 of the supplemental materials [95].

Correlations between the group of S1 participants and subgroup of S2 participants were calculated for: number of stages, number of reviews, number of papers, number of used tools, age, position, and proficiency. Number of reviews and number of papers show a relatively low correlation, respectively 0.71 and 0.61 (data shown in Section 2 in [95]). The differences indicate no clear increasing or decreasing trend. All other characteristics were highly correlated between the S1 and S2 groups, indicating that there is no difference between the participants. Therefore, the conclusions drawn for any analysis on the S1 group can be applied to the S2 group.

4.3.2 Survey responses

The complete results of both surveys are included as supplemental materials [95]. Below we summarise and highlight the most relevant findings.

4.3.2.1 First Survey

Approximately 36% of the participants are involved in four or more systematic reviews per year (S1.2). Furthermore, approximately 60% of the participants work with reviews with more than 3000 search results (S1.3).

Reported tool usage (S1.4) is summarised in Figure 4.2. Note that in the results presented below we did not include the three validation tools. From the 168 participants, 54 answered that they were incidentally or regularly using one or more of the listed tools. Fourteen participants consider many tools, out of which six consider all and use none. In total 92 participants indicated to not be using nor considering any tool. The three control tools (EndNote, RevMan, and Covidence – leftmost bars in Figure 4.2) are either used or considered by many, respectively 78%, 71%, and 49%. Automation tools regularly used by at least one participant were: Epistemonikos, EPPI-reviewer, GATE, NLM Medical Text Indexer, Rayyan, RevMan HAL, SWIFT-Review, and Systematic Review Assistant.

The option *Other* of S1.4 resulted in 33 unique tool suggestions. Fifteen of them were listed on the Systematic Review Toolbox website, but were not included during our tool selection process because they did not fulfil our criteria. The remaining 18 tools were found to be in one of the following cases: not an automation tool, not available (no public information, or discontinued), or very specific to a research field. Therefore, no additional tools were listed in S2.

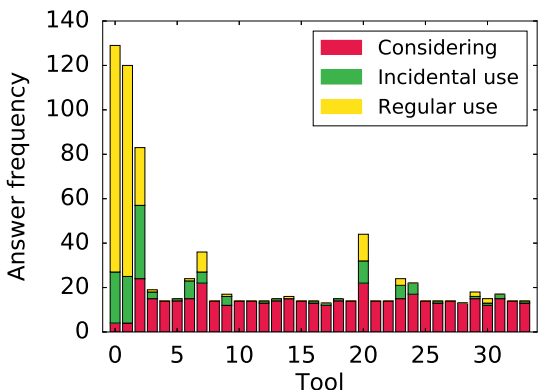


Figure 4.2: Reported tool usage, outcome of S1.4. Red: considering the tool; green: incidentally using the tool; yellow: regularly using the tool. See Section 1.4 in the supplemental materials [95] for the names of the tools included in this figure.

4.3.2.2 Second Survey

The first step of S2 provided information about changes in tool usage between S1 and S2. In total 108 changes were made by 32 participants (see Table 4 in the supplementary materials). We observed two major categories of changes: 1) a tool was not used in S1 but became used in S2 (15 cases); and 2) a tool was used in S1 but was no longer used in S2 (4 cases). These changes did not influence the correlation in participant characteristics between the S1 group and S2 subgroup as shown in Section 4.3.1.

Below we highlight the most relevant results for S2. Note that all mentioned percentages are the fraction of participants that have chosen a given option. The survey was completed for fourteen distinct tools - see counts in Table 4.1. Nine tools fell below the cut-off of five responses and were grouped together for the result analysis.

Table 4.1: Number of responses per tool for S2. The responses contained fourteen distinct tools. Nine tools fell below the cut-off of five responses and were grouped together for the result analysis.

ID	Count	Tool
1	33	EndNote
2	23	Review Manager (RevMan)
3	16	Covidence
21	13	Rayyan
8	6	EPPI-Reviewer
<i>Grouped tools</i>		
99	4	Abstrackr
99	3	Epistemonikos method of searching
99	3	SWIFT-Review
99	2	RevMan HAL
99	2	RobotReviewer
99	1	DoctorEvidence (DOC Data)
99	1	GATE
99	1	NLM Medical Text Indexer (MTI)
99	1	Spa

The responses to questions S2.1, S2.3, S2.5, S2.14, and S2.20 indicate that the participant’s environment has impact on the tools that they know and might use. In S2.1 the majority of participants indicated to hear about tools through their colleagues (77%), peers (71%), or organisation (40%). The same pattern is found in S2.14, which asked the same question for a specific tool and obtained the answers: colleagues (57%), peers (39%), and organisation (42%). Determining a tool’s task, and whether it fits in the workflow of the participant (S2.3), is often based on experience, either personal (84%) or from peers (77%). Moreover, participants indicated personal experience (69%) and peer experience (61%) as the most contributing aspects for the effort during this process (S2.5). The responses to S2.20 also show that experience is often the basis to assess the quality of results that a tool produces: own (24%) and colleagues (28%).

In questions S2.8-11 we collected the aspects and reasons for starting and stopping to use tools. Responses to S2.8 show that about half (48%) of the participants had used another tool which they stopped using. The most important reasons for participants to stop using a tool (S2.9) were poor usability (43%), lacking functionality (37%), and not fitting their workflow (37%). In S2.10, 35% of the participants indicated that they would like to use a tool but do not. The responses to S2.11 show that the most important factors are: cost of licensing (32%), missing the support of colleagues (27%), and the effort to learn a new tool, such as learning curve (23%) and lack of time (9%).

Responses to S2.21 and S2.22 are expressed on a scale of 1 to 7, and indicate that automation tools help to reduce the time a participant’s job takes (median = 3) and improve the quality of their work (median = 5). Lastly, the usability (S2.24-34) is similar for all tools (Figure 4.3), being overall rated as good (i.e., usability score > 68).

4.4 Discussion and conclusions

In this research two surveys were conducted to investigate the uptake of automation tools that support in the conduct of systematic reviews, as well as the barriers and facilitators that lead to their use.

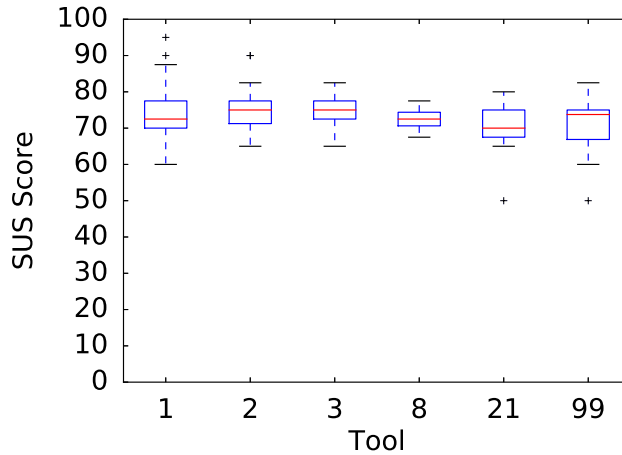


Figure 4.3: Summary of the System Usability Scale (SUS) questions, S2 question 25 to 34. The SUS scores are summed for all participants, the maximum score a tool can achieve is 100 (i.e., 10 questions, 10 points for each question). SUS scores are shown grouped by tool (see Table 4.1) as boxplot.

Our assumption was that researchers with a high interest in automation solutions would respond to our survey. For example, reviewers with a high workload or those that already use a large number of tools. However this has not occurred. Although the number of participants is small relative to the total population of systematic reviewers, a good spread of characteristics was obtained for S1, and similarly for S2. This was true except for two characteristics, namely country and research domain, discussed below.

Firstly, the geographic spread of participants is somewhat limited. Although we collected responses from 22 countries, 53% of those came from the Netherlands or the United Kingdom alone. This can be attributed to our recruitment methods, and it is likely not representative of actual interest in systematic review automation worldwide.

Regarding the research field, note that we attempted to reach out to participants from various backgrounds, but our population, however, mostly covers reviewers in the medical domain. This bias could be explained by our recruiting methods, which reached more researchers from the medical sciences (Cochrane, organisation, and department methods) than from mixed research fields (network method) – see Table 4.2. A subgroup analysis was performed to assess whether the participants recruited through the network method differ from the whole population regarding research field. When comparing this group against the whole participant population the distribution had only slight differences – see full analysis in Section 1.8 of the supplemental materials [95]. The bias, however, could also be present in the field of systematic reviews. It is a fact that the medical research domain has conducted more systematic reviews, and for far longer, compared to the other domains. When looking at an analysis of research domains for systematic review articles in the literature database Web of Science², a large portion of this research is in the biomedical sciences. We therefore consider that our population reflects the dominance of medical research in carrying out systematic reviews, but our results should be carefully interpreted in the context of other fields.

²Searching for the terms 'systematic review' in the title and then using the 'analyze results' function.

Table 4.2: Number of emails sent for S1, failed to deliver emails (i.e., bounced or out-of-office), and survey responses per group. *: estimate.

	Total	Failed	Responses
Network	2500*		81
Cochrane	327	22	48
Organisation	261	27	37
Department	50*		6
Total	3138*	49	172

The results of our survey show no specific trends regarding workload and participant characteristics. Below we summarise our results categorised by the TAM concepts:

- *Image*: The standing of a researcher among their peers was found to not have an effect on tool usage behaviour, although a full assessment of this could not be performed based on our results.
- *Experience*: Either personal experience or the experience of peers with a tool is used to choose, learn, assess and validate tools.
- *Voluntariness*: Tools are often used voluntarily, however in some cases tools are enforced by institutions or organisations. For example, researchers from the medical domain, and especially those working in Cochrane, need to follow relatively strict protocols regarding the execution of reviews, which could favour or inhibit the use of particular tools.
- *Job Relevance*: The fit of a tool in the participant's tasks and workflow are equally important.
- *Output Quality*: Whether the output of a tool matches with the goals of a participant is often not formally assessed, but understood through own experience or the experience of others. Tool documentation and scientific publications are used less often.
- *Result Demonstrability*: Beneficial effects of tools are time reduction and quality improvement. Mostly measured through personal experience or the experience of others.
- *Perception of External Control*: Resources that lie outside of the control of participants are, for example, cost of licensing and lack of time. These resources are often controlled on an institutional level. Therefore, perception of external control seems to be a relevant factor for choosing and using tools.
- *Objective Usability*: Even though some of the tools are more used than others, no difference in usability (measured with the SUS) was found between these tools.
- *Awareness*: Participants mostly learn about the existence of new tools through their environment.

Regarding tool uptake, from S1.4 we can conclude that automation tools are not widely used by the participants (32% use automation tools, discounting the three control tools). Tool uptake was also confirmed with S2.0.4 for a subset of the population – see Section 2 introduction in [95]. When comparing answers between S1 and S2, we noticed that in more cases participants started to use tools, than stopped using tools, which may indicate increase in tool uptake over time. This can be explained by the participants of S2 having larger interest in automation tools, and therefore being more likely to invest time in completing the survey. However, no bias could be detected through correlation of the number of used tools per participant between S1 and S2 groups. Therefore, our observations could also indicate that these participants went through a exploration phase during S1 and stopped before S2.

In spite of low uptake, we observed that all the tools are considered effective because they reduce time (S2.21). Moreover, participants indicated to perform many (and large)

systematic reviews (S1.2-3). Therefore our results indicate that the uptake of automation tools is not directly linked to a large workload or potential benefits.

We also observed that ‘good’ usability was reported for all tools that are in use by the participants of this study (S2.24-34). Although our results show no direct influence of usability in tool uptake, poor usability and a steep learning curve have been reported as barriers to start using – or continue to use – a tool (S2.8-11). It might be that the other tools, not in use, have lower usability (S2.9), which would confirm our hypothesis that usability is a relevant factor in tool uptake.

Regarding barriers, licensing, steep learning curve, lack of support, and mismatch to workflow are often and equally reported as relevant factors (S2.8-11). Lack of time was also indicated by some as a barrier to assess the applicability and to evaluate the quality of a new tool (S2.11).

Regarding facilitators, we observed that the reviewer’s environment (i.e., colleagues, organisation, peers) plays a major role in getting acquainted with, and using, tools (S2.1, S2.3, S2.5, S2.14, and S2.20).

Our results mostly confirm and support many of the conclusions and recommendations of the 2016 ICASR meeting [91]. A barrier mentioned is the shortage of studies showing the benefits of these tools. We indeed observed that the quality of tools results is often not formally validated, and, when it is, this is mostly done by trying the tool or through experience of colleagues (S2.20), demanding additional effort (S2.4, S2.5). Lack of transparency in automation tools was another blocking factor mentioned, however in S2.18 the option ‘the tool does not explain to me how it generates the results’ was chosen only once (1/74) as a reason for not using a tool.

As facilitators, ICASR suggests integration with other tools along the complete workflow, which was also identified as an important factor for tool usage (S2.6). Another facilitator put forward is the joint development of validation of tools and quality criteria that can be used to assess output of automation tools. This necessity is confirmed by our results, which point towards a strong influence of the environment and community for improving awareness, evaluation, and support for tools.

A review by Thomas et al. [106] mostly names the same barriers and facilitators as the ICASR meeting. The two main conclusions from this work are that automation solutions should “have a demonstrative relative advantage and are clearly compatible with the needs of systematic reviewers”. These conclusions are supported by the results of this study, as described above.

Our findings highlight the importance that organisations and best practices in a field can have for the uptake of automation tools for systematic reviews. We argue that organisations and communities should play a leading role in validating the quality of results generated by the tools, raising awareness about them, and supporting their use.

Acknowledgements

We thank Dr. M. Leeftang for her support with selecting the automation tools, developing S1, and testing S2, and also M. Hillen for support in structuring and giving feedback on S1.

Appendices

Table 4.3: Questions for S1, Other: question had a free text input.

#	Question	Answer options	Other
<i>Section 1: Systematic review characteristics</i>			
1	In which stage(s) of the systematic review are you involved?	Preparation, Retrieval, Appraisal, Synthesis, Write-up [20].	
2	In how many systematic reviews were you involved on average in the past two years?	Less than 1 per year, 1 per year, 2 per year, 3 per year, 4 per year, More than 4 per year.	
3	What is the average number of search results in systematic reviews you are involved in?	Less than 500, 500 to 1500, 1501 to 3000, 3001 to 9000, 9001 to 18000, More than 18000.	
4	Which of the following computer tools, programs, or software to automate (parts of) the systematic review process do you use?	List of selected tools from [90] (see Section 4.2.1.1). Options: I'm considering using it, I use it incidentally, I use it regularly. No selection indicates 'not using or considering'	✓
<i>Section 2: Demographics</i>			
5	Current age	Younger than 20, 20 to 30, 31 to 40, 41 to 50, 51 to 60, Older than 60.	
6	Country of residence	List of all countries.	
7	Current position	Student, PhD candidate, Post-doc, Researcher, Senior researcher, Other (specify).	✓
8	Field(s) of systematic reviews	Medical sciences, Education sciences, Social sciences, Computer sciences, Life sciences.	
9	Computer proficiency	Scale of 1 to 10 with three labels: Basic, Intermediate, Advanced.	
10	Email address	Optional field for participation in S2.	

Table 4.4: Generic questions for S2, Other: question had a free text input.
SubQ: sub-question of, i.e., question only had to be answered if the previous was answered with 'yes'.

#	Question	Answer type	Other	SubQ
1	How do you keep yourself informed about available tools?	Multiple choice	✓	
2	How iterate do you consider yourself about technology-assisted systematic reviews?	Scale		
3	How do you determine which task(s) a tool performs and whether it fits your workflow?	Multiple choice	✓	
4	How much effort does it take to determine which task(s) a tool performs and whether it fits in your workflow?	Scale		
5	Which aspects mostly contribute to the effort to determine which task(s) a tool performs and whether it fits in your workflow?	Multiple choice	✓	
6	Which is most important when assessing a tool?	Scale		
7	What is necessary for you to start using a new tool?	Multiple choice	✓	
8	Are there tools that you have used in the past, but that you currently don't use?	Single choice		
9	Why don't you use these tools anymore?	Multiple choice	✓	8
10	Are there tools you would like to use but don't?	Single choice		
11	Why don't you use these tools?	Multiple choice	✓	10
12	In which way do systematic review automation tools impact your systematic reviews?	Scale		
13	Would you like to stay informed about the results of the survey?	Single choice		

Table 4.5: Specific questions for S2, Other: question had a free text input.
SubQ: sub-question of, i.e., question only had to be answered if the previous was answered with 'yes'.

#	Question	Answer type	Other	SubQ
14	How did you hear about this tool?	Multiple choice	✓	
15	Is the tool commonly used by your peers?	Scale		
16	For which task(s) are you currently using this tool?	Multiple choice	✓	
17	Are there other tools available that do the same task?	Single choice		
18	Why have you not chosen for the other tool(s)?	Multiple choice	✓	17
19	Why do you use this tool?	Multiple choice	✓	
20	How did you assess the quality of the results?	Multiple choice	✓	
21	How does the tool impact your productivity?	Scale		
22	How does the tool impact the quality of your work?	Scale		
23	How much experience do you have with this tool?	Scale		
24	Do you use this tool by your own choice?	Single choice	✓	

Table 4.6: Usability questions for S2, all scales run from strongly disagree (1) to strongly agree (7).

#	Question	Answer type
25	I like to use this tool	Scale
26	I find this tool unnecessarily complex	Scale
27	I think this tool is easy to use	Scale
28	I need(ed) the support of a technical person to be able to use this tool	Scale
29	I find that various functions in this tool are well integrated	Scale
30	I think there is too much inconsistency in this tool	Scale
31	I would imagine that most people would learn to use this tool very quickly	Scale
32	I find the tool very cumbersome to use	Scale
33	I feel very confident using this tool	Scale
34	I needed to learn a lot of things before I could get going with this tool	Scale



Improving automation tools



Training sample selection: impact on screening automation in diagnostic test accuracy reviews

In Research Synthesis Methods, volume 12, pages 831-841, 2021

A.J. van Altena, R. Spijker,
M.M.G. Leeflang, S.D. Olabarriaga

Abstract

When performing a systematic review, researchers screen the articles retrieved after a broad search strategy one by one, which is time-consuming. Computerised support of this screening process has been applied with varying success. This is partly due to the dependency on large amounts of data to develop models that predict inclusion.

In this paper, we present an approach to choose which data to use in model training and compare it with established approaches. We used a dataset of fifty Cochrane diagnostic test accuracy reviews, and each was used as a target review. From the remaining 49 reviews, we selected those that most closely resembled the target review's clinical topic using the cosine similarity metric. Included and excluded studies from these selected reviews were then used to develop our prediction models. The performance of models trained on the selected reviews were compared against models trained on studies from all available reviews.

The prediction models performed best with a larger number of reviews in the training set and on target reviews that had a research subject similar to other reviews in the dataset. Our approach using cosine similarity may reduce computational costs for model training and the duration of the screening process.

5.1 Introduction

Even for an experienced review team a single systematic review can take between 6 months and several years [20, 107]. Approximately half of this time is spent on developing the research protocol, performing the search, and assessing the results [108]. Therefore, reducing the time spent on these tasks has a big impact on the efficiency of the review process [20].

Within the medical field, Diagnostic Test Accuracy (DTA) studies often do not follow a standard design and are generally poorly reported [109]. Therefore, search strategies cannot depend on design descriptors or commonly reported terminology. Complex and broad literature searches are needed, resulting in a high number of documents needing screening to find relevant studies. This leads to a relatively large part of the review-time being spent on screening and selection [110, 111].

Machine learning methods can aid the screening process through ranking or classification of relevant documents [18]. Generally, there are two types of machine learning; supervised and unsupervised. Supervised methods use data that have been manually labelled as being relevant or not. Unsupervised methods learn from trends in unlabelled data. Both types have been applied before in systematic reviews with varying levels of success (see for example [112–114]). In this study we focus on a supervised method that needs a training set of labelled data. This method can only be applied under the assumption that the labelled training data share ‘transferable knowledge’ with the unlabelled data on which it is tested [115]. In the case of systematic reviews, transferable knowledge may refer to, for example, the clinical topic or patient population being similar in the relevant studies in the training set and in the relevant studies that the model aims to select.

Typically, as much data as possible is used when building machine learning models, because more examples used during training will usually yield a more robust model. However, because systematic reviews focus on a specific research question, the question arises whether using all available training data indeed results in optimal model robustness. After all, when considering which data to use for a review about Alzheimer, another review about Alzheimer might provide a better training set than a review about cancer. Therefore, excluding the review about cancer from the training data might improve the model’s robustness because the remaining training data is less diluted.

In this study, we use a set of fifty DTA reviews and simulated the screening process for each of those reviews as a target review (i.e., the review for which a model is built). The remaining 49 reviews were used to build the model using three different approaches to select the training data. The first and novel approach used a similarity metric to select a subset of reviews similar to the target review. The second approach used all 49 reviews as a training set. The third approach randomly selected a training set. The models trained with these three approaches were tested on each target review and the resulting performance was compared. We hypothesise that creating a training set specifically for the target review will yield a better prediction performance, because the transferable knowledge is not diluted by non-relevant training data.

5.2 Data preparation

In this study we used the dataset provided by the 2017 CLEF eHealth Lab “Technologically Assisted Reviews in Empirical Medicine Overview” [116]. This dataset consisted of fifty DTA reviews published in the Cochrane Library and contained the following information about each review: its unique identifier (in the form of CD0XXXXX), review title, the search query, and the search results (PubMed IDs of all found documents). Also, for each search result there were two labels indicating whether: 1.) it was included in the systematic review after screening the title and abstract of the document, and 2.) it was included after reading the

full text of the document. We used the latter because they represent the inclusions that need to be found after the review process is completed.

Data gathering The dataset was cleaned by the organisers of the lab, and limited to search results available through the PubMed search engine. The Entrez Programming Utilities API [82] was used to retrieve data about the search results based on their PubMed ID. In total we retrieved 266,966 documents using the `efetch` function of the Entrez API. We used all documents regardless of whether they were inclusions or not. For each document, if available, we stored the following data in a local database: review identifier, document title, document abstract, publication date, publication type, DOI, journal, journal ISSN, journal ISO, inclusion label, and PubMed ID.

Text preprocessing The title and abstract of all in- and excluded studies were used to build the prediction models. We cleaned the text so that it contained only unaccented alphabetical letters. We removed: HTML tags¹, special characters (e.g., &, %), and numbers. Stopwords (e.g., the, what, was) were removed using the `english` list from the NLTK Python library [117]. Lastly, the documents were split into separate words and any short (< 2 characters) or long (> 34 characters) words were removed, as they were unlikely to be real words, or words that distinguish the topic of a review. The Python code implementation is available at [118].

The characteristics of our dataset after preprocessing are shown in Table 5.1. Reviews had an average of 5,339 documents with on average 93 inclusions. The smallest review had 64 documents and the largest 43,363. The review with the fewest inclusions contained 2 inclusions, while the largest number of inclusions was 619. This resulted in a mean inclusion rate of 4% with a minimum of 0.015% (2 on a total of 12,705) and a maximum of 20% (23 on a total of 114). For an overview of all the metadata collected per review see [119].

Abstracts were missing for 45,033 documents (17%). Table 5.2 shows the characteristics of these documents. Three major characteristics were found: 1.) the document was written in a foreign language and not available in English, 2.) the document was published before (approximately) 1975 and was not digitally available, and 3.) the document was not a primary research publication (e.g., comment, case report, etc.). They were kept in the dataset nevertheless, because 359 of them were inclusions.

Review metadata enrichment DTA review questions are usually constructed according to three elements, describing the people suspected of the disease (Patients, P); the diagnostic tests that were evaluated in the review (Index test(s), I); and a definition of the disease (Target condition, T) [120]. Of these elements, the target condition (T) can be mapped to a standardized system and was therefore added to the review dataset. In preparation, one of the authors (AA) read the abstracts of the 50 reviews and identified the International Classification of Diseases, 10th revision (ICD-10) code for the target condition using the ICD-10 browser [121]. Each review was assigned the best fitting code suggested by the auto-complete function of the ICD-10 browser. If more than one code was available, both codes were assigned. Together with another author (ML) the codes were reviewed. Codes were adjusted if both authors agreed that the resulting code would better reflect the research topic of the review at hand.

We categorised diseases into disease groups using the first letter of the ICD-10 code. Twelve reviews could not be grouped based on disease codes, so we created a catch-all group coined 'other'. A total of eight groups were identified, with Alzheimer (G) and dementia (F) combined as one group (see Table 5.3). Table 5.4 shows the metadata collected for each review, including the disease group of the review question.

¹PubMed data may contain the following tags: <i>, <u>, , <sup>, and <sub>

Table 5.1: Document characteristics after cleaning.
*: mean [minimum - maximum]

number of DTA reviews	50
total number of documents	266,966
included documents	4,661
# words per document*	922 [0 - 9,795]
# unique words per document* per review	70 [9 - 529]
# documents*	5,339 [64 - 43,363]
# included documents*	93 [2 - 619]
% included documents*	4% [< 1% - 20%]
missing abstracts	
# all documents	45,033 (17%)
# included documents	359 (7%)

Table 5.2: Reasons for missing abstract. Note that there was overlap between the characteristics, as an document might both be written in a foreign language and be published before 1975.

	All	Inclusions
Foreign language	16075	81
Before 1975	14721	24
Not journal article	23368	142

Table 5.3: Review groups according to disease (target condition).

Group	# reviews	ICD-10	Disease
1	2	A	Tuberculosis
2	4	B	Parasitic
3	8	C	Cancer
4	12	G and F	Dementia & Alzheimer
5	4	K	Liver
6	5	M	Musculoskeletal system
7	3	Q	Down syndrome
8 (other)	12	-	Various

Table 5.4: Metadata collected for each review.

Identifier	# docs.	# incl.	ICD-10	Secondary ICD-10	Disease group
CD007394	2545	95	B44.0		2
CD007427	1521	123	M75.4		6
CD007431	2074	24	M54.3	M54.5	6
CD008054	3217	274	N87.9		Other
CD008081	970	26	H35.81	E14.3	Other
CD008643	15083	11	S32.001A	M54.5	6
CD008686	3966	7	M53.9	M54.5	6
CD008691	1316	73	I25.10	Z94	Other
CD008760	64	12	I85		Other
CD008782	10507	45	G30	F06.7	4
CD008803	5220	99	H44.51		Other
CD009020	1584	162	M75.101	M25.5	6
CD009135	791	77	B55.0		2
CD009185	1615	92	N10		Other
CD009323	3881	122	C25.9	C24.1	3
CD009372	2248	25	I61.9		Other
CD009519	5971	104	C34.90	C80	3
CD009551	1911	46	B44.0		2
CD009579	6455	138	B65		2
CD009591	7991	144	N80		Other
CD009593	14922	78	A15.3	U84.9	1
CD009647	2785	56	E86		Other
CD009786	2065	10	C56	C80	3
CD009925	6531	460	Q90.2		7
CD009944	1181	117	C16.9	C80	3
CD010023	981	52	S92.2		Other
CD010173	5495	23	C06.9	C80	3
CD010276	5495	54	C06.9	C80	3
CD010339	12807	114	K80		5
CD010386	625	2	F03	F06.7	4
CD010409	43363	76	C51	C77.4	3
CD010438	3250	39	D68.9	T14.9	Other
CD010542	348	20	K70		5
CD010632	1504	32	F03	F06.7	4
CD010633	1573	4	G31.8	F02.8	4
CD010653	8002	45	F20		4
CD010705	114	23	A15.3	U84.9	1
CD010771	322	48	F03		4
CD010772	316	47	F03		4
CD010775	241	11	G30	F03	4
CD010783	10905	30	G30	F03	4
CD010860	94	7	G30	F03	4
CD010896	169	6	G31.0	F03	4
CD011134	1953	215	C18	C80	3
CD011145	10872	202	F03		4
CD011548	12708	113	K80		5
CD011549	12705	2	K80		5
CD011975	8201	619	Q90.2		7
CD011984	8192	454	Q90.2		7
CD012019	10317	3	N80		Other

5.3 Methods

5.3.1 Prediction models

A plethora of feature extraction and classification methods were available. We selected representatives of approaches often used in related literature [18, 116].

Feature extraction The input that a prediction model is trained on are called the features. To extract these features from the gathered data we chose the Term Frequency (TF) because of its simplicity. Document frequency weighting was added to the term frequency matrix (TF-IDF) to adjust for words that generally occur more frequently in texts.

Classifiers We chose a Random Forest classifier because it is relatively simple and is much used in systematic review prediction applications [116]. The classifier was implemented using the `RandomForestClassifier` method from the `scikit-learn` library [122].

Each classifier method has a set of parameters that need to be determined before training on a dataset. The Random Forest classifier has parameters for the shape of the trees that will be generated, for example, the maximum depth of one branch on the tree. Parameters have a different optimum for each dataset. To find these optimal values we used grid search, a technique where a range of values is tested with a small portion of the training set. Performance of the resulting models is measured and the parameter settings of the best model are retained to train the model on the complete training set.

First, only a subset of all possible value combinations is tried in a random search. The ranges for the parameters are very wide to find the specific value range where the prediction model approaches its optimal state. Results of the random search were inspected and a smaller set of parameter values was chosen for the full grid search. In the full search all parameter values are tested and only the model with the highest performance is retained. Final prediction models were trained using a full grid search for each systematic review. A detailed description of the selection process can be found in Appendix A.

5.3.2 Model performance metric

Using the `predict_proba` function from the `sklearn` library, the predicted probability of being an inclusion was retrieved for each document in the target review. The reading order of documents was determined by sorting the predicted probability from highest to lowest. Models were judged on their ability of ordering the documents such that inclusions would be encountered earlier during the screening process. This reduced the number of documents needed to be read during the screening process, thus saving work and time.

This concept of performance is captured in the metric Work Saved over Sampling (WSS), introduced by Cohen et al. [113]. For a specified level of recall², WSS measures the fraction of documents that a review author does not need to read as a result of the ranking, as compared to a random ordering. WSS is calculated as follows:

$$WSS = \frac{TN + FN}{n} - (1 - R) \quad (5.1)$$

where TN and FN are the number of true and false negatives respectively, n is the total number of documents, and R is the level of recall. Recall is defined as:

$$R = \frac{TP}{TP + FN} \quad (5.2)$$

²i.e., the fraction of correctly identified inclusions, in statistics the term 'sensitivity' is used.

where TP is the number of true positives.

We adopted the commonly used WSS at a recall level of 95% (Work Saved over Sampling @ 95% (WSS@95)), which is defined as follows by [113, 116]:

$$WSS@95 = \frac{TN + FN}{n} - 0.05 \quad (5.3)$$

WSS@95 ranges between 0.95 and -0.05. Respectively, indicating a perfect classification or a poor classification where all documents have been labelled as inclusion and $TN + FN = 0$.

5.3.3 Similarity metric

The similarity between the potential training data and the target review was measured using titles and abstracts. In our study the documents were mathematically represented as a vector from the Term Frequency Inverse Document Frequency (TF-IDF) matrix, so we adopted the cosine similarity metric [123], which is designed for vectorial representations of documents.

Cosine similarity measures the cosine of the angle between two vectors of an inner product space [124], being defined as follows by Huang et al. [123]:

$$SIM(\vec{d}_a, \vec{d}_b) = \frac{\vec{d}_a \cdot \vec{d}_b}{\|\vec{d}_a\| \|\vec{d}_b\|} \quad (5.4)$$

where \vec{d} is a vector representation of a text document (i.e., a single row of the TF-IDF matrix). The inner product space is calculated as:

$$\vec{d}_a \cdot \vec{d}_b = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \dots + a_n b_n \quad (5.5)$$

where n is the length of the document vectors. Similarity ranges from 0 (not similar, vectors are at an angle of 90°) to 1 (perfectly similar, vectors are at an angle of 0°). Because the word counts of the TF-IDF matrix cannot be negative, similarity cannot be negative either. Note that Equation (5.4) may look similar to bivariate correlations such as the Pearson correlation.

We were interested in the similarity between reviews, and not between individual documents in these reviews. We therefore calculated the similarity based on the mean feature vector for all documents in each review \vec{r}_i :

$$\vec{r}_i = \frac{\sum_{j=1}^{n_i} \vec{d}_{ij}}{n_i} \quad (5.6)$$

where \vec{d}_{ij} is the feature vector for document j from review i , and n_i is the number of documents in review i . The similarity between all pairs of reviews is then defined as:

$$S_{ik} = SIM(\vec{r}_i, \vec{r}_k) \quad (5.7)$$

$$i \in \{1 \dots 50\}, k \begin{cases} k \in \{1 \dots 50\} \\ k \neq i \end{cases}$$

where \vec{r}_i and \vec{r}_k are the mean review vectors for respectively reviews i and k . Cosine similarity was calculated between all pairs of reviews ($50 \times 49 = 2,450$ in total).

5.3.4 Workflow

We refer to the approaches used in this study as selected data (SIMILAR), all data (ALL), and random data (RANDOM). The SIMILAR approach used a similarity metric to select documents for the training set. Training sets were constructed by using the documents from the $n \in \{1, 2, 5, 10\}$ reviews most similar to the target review³. The ALL approach used all of the remaining 49 reviews as the training set. And lastly, the RANDOM approach selected $n \in \{1, 2, 5, 10\}$ random reviews.

The SIMILAR approach was compared with the ALL and RANDOM approaches. ALL was chosen because it is the standard in machine learning, following the rule of thumb that more data equals better models. RANDOM was added as a control.

As described above, prediction models were trained using a Random Forest classifier and the features from the TF-IDF matrix. The models were used to rank the test set (i.e., the documents from the target review) and the WSS@95 was calculated. We repeated this process five times to account for model training variability. For all three approaches each of the fifty DTA reviews were used as test set once. This would train 1,000 models each for the SIMILAR and RANDOM approaches (i.e., $50 \text{ reviews} \times 4 \text{ training set sizes} \times 5 \text{ repeats} = 1,000$), and 250 models for the ALL approach (i.e., $50 \text{ reviews} \times 5 \text{ repeats} = 250$). The workflow is shown in Figure 5.1.

Analysis Performance of the SIMILAR, ALL, and RANDOM approaches were analysed using boxplots. Statistical significance of the results was analysed using a Wilcoxon rank sum test. The Wilcoxon test was executed for each pair of training set sizes. Resulting p-values were adjusted for multiple testing using the Bonferroni method. The significance tests were implemented using the `stats.ranksums` function from the SciPy package [125] and the `stats.multitest.multipletests` function from the statsmodels package [126] respectively. Additionally, the same analysis was applied to the performance results after stratification into disease groups: 1-7 and 'other'.

Lastly, we analysed the correlation between model performance and cosine similarity. To determine the correlation we first retrieved the mean WSS@95 and mean cosine similarity per review for each training set size used in the SIMILAR approach ($n \in \{1, 2, 5, 10\}$). Then, the `corr` function of the Pandas package was used to calculate the correlation [127].

³A maximum of 10 similar reviews was chosen after analysing preliminary data on cosine similarity scores. We observed that reviews are mostly similar to just a few other reviews. Similarity rapidly drops and at the 10th review similarity is mostly equal. Data are shown in Appendix B.

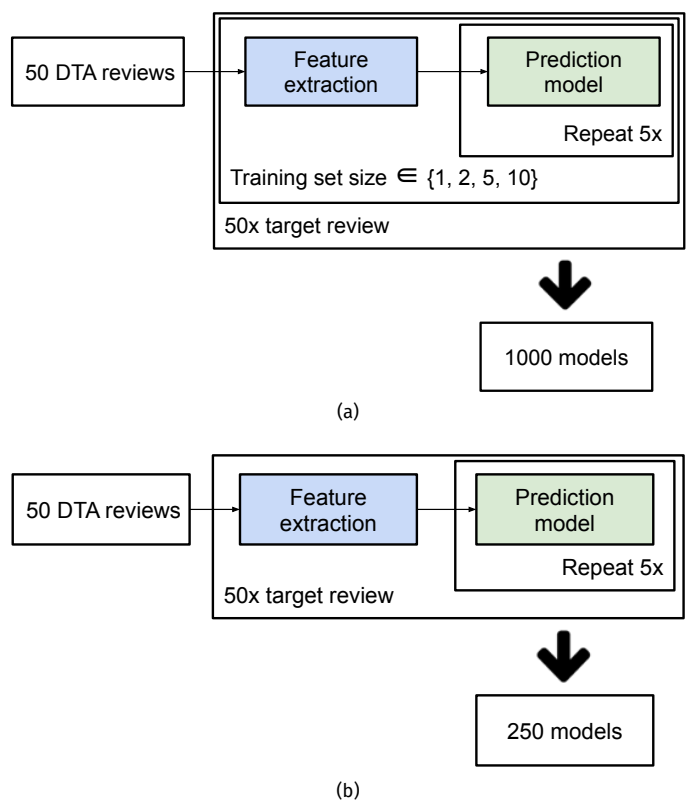


Figure 5.1: Overview of workflow for the approaches using different training data: (a) selected data (SIMILAR) and random data (RANDOM), and (b) all data (ALL). Feature extraction was implemented using TF-IDF. The prediction model was implemented using the Random Forest classifier.

5.4 Results

Approach comparison The overall prediction performance results obtained in the SIMILAR, ALL and RANDOM approaches are shown in Figure 5.2. The SIMILAR and ALL results indicate that on average the best performance is obtained when all training data is used. The ALL significantly outperforms SIMILAR for all training set sizes. Furthermore, the median performance in the SIMILAR approach is higher for larger training sets (Table 5.5).

With smaller training sets ($n \in \{1, 2, 5, 10\}$) the models from the SIMILAR approach outperform those from the RANDOM approach. However, the difference in performance for the training sets with size 5 and 10 is not statistically significant, see Table 5.5.

Influence of the ‘other’ disease group on performance Figure 5.3 presents the overall performance results obtained for all training set sizes ($n \in \{1, 2, 5, 10, 49\}$) stratified by disease group. The stratified results show that in general the prediction performance is higher for reviews that belong to a disease group, as opposed to those that do not. The difference in performance is significant over all training set sizes as seen in Table 5.6.

Correlation between cosine similarity and performance Results for the correlation analysis are shown in Table 5.7. The values in the diagonal show that a moderate correlation (0.32 - 0.47) exists between the performance of a review and its similarity to the training set.

Table 5.5: P-values for SIMILAR versus ALL performance and SIMILAR versus RANDOM performance. SIMILAR is stratified by the training set size. M is the median WSS@95 performance over all models. *: p-value is significant.

SIMILAR	ALL 49 ($M = 0.49$)	RANDOM			
		1 ($M = 0.25$)	2 ($M = 0.25$)	5 ($M = 0.33$)	10 ($M = 0.39$)
1 ($M = 0.36$)	< 0.001*	0.03*			
2 ($M = 0.40$)	< 0.001*		< 0.01*		
5 ($M = 0.39$)	< 0.001*			0.66	
10 ($M = 0.43$)	0.05*				1.00

Table 5.6: P-values for other versus disease groups 1-7 performance, both are stratified by the training set size. All p-values are significant. M is the median WSS@95 performance over all models.

Groups 1-7	Other				
	1 ($M = 0.26$)	2 ($M = 0.28$)	5 ($M = 0.29$)	10 ($M = 0.33$)	49 ($M = 0.42$)
1 ($M = 0.40$)	0.009				
2 ($M = 0.44$)		< 0.001			
5 ($M = 0.43$)			0.005		
10 ($M = 0.47$)				0.002	
49 ($M = 0.51$)					0.038

Table 5.7: Pearson correlation between the performance and cosine similarity for each training set size in the SIMILAR approach.

	Performance	Similarity			
		1	2	5	10
1		0.40			
2			0.47		
5				0.36	
10					0.32

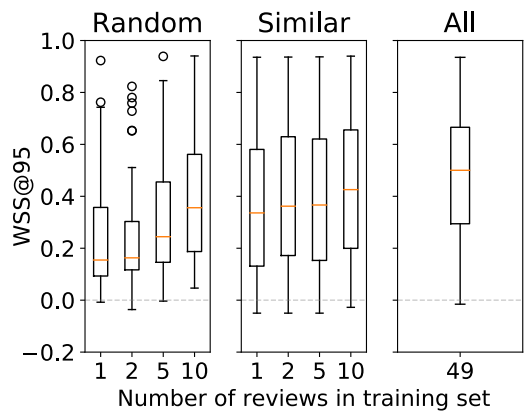


Figure 5.2: Boxplot of model performance stratified by the training set size. Performance is shown separately for the RANDOM, SIMILAR, and ALL approaches.

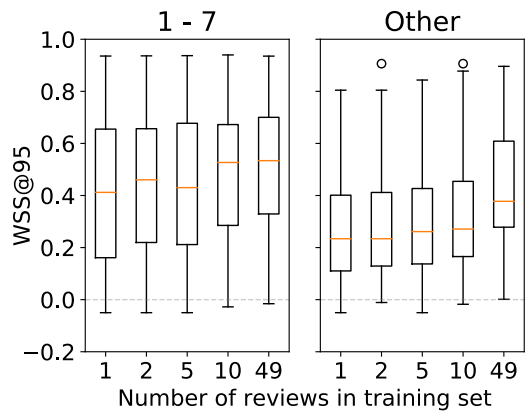


Figure 5.3: Boxplot of SIMILAR performance stratified by the training set size. The results are shown for groups 1-7 and other.

5.5 Discussion

In this study we investigated whether computerised support of the systematic review screening process could be improved. Our hypothesis was that a smaller, but more focused training set would improve performance. We assessed the use of cosine similarity for selecting data for the training set and compared the classification performance of this approach to approaches using all available data and randomly selected data.

Approach comparison Analysis of the SIMILAR and ALL approaches shows that, when considering all fifty reviews in our dataset, the best performance is obtained when all training data are used. This rejects our hypothesis that a more targeted training set is beneficial for prediction. However, when reviews in the ‘other’ group are considered separately from those in disease groups 1-7, we observe that they perform significantly worse at any training set size or approach. This indicates that a training set with topically similar reviews is crucial for prediction model performance.

When review authors start a new systematic review, they may not have a large training set with many previously undertaken systematic reviews at hand. Our findings indicate that, in these situations, it may be worthwhile to gather a training set based on a few systematic reviews on a similar topic. Reviewers who develop many systematic reviews, for example within a guideline committee or a review-developing enterprise, may want to invest into creating a repository of past reviews to use as a training set. The extra investment to select related reviews for a training set can thus be prevented.

The size of the training set is a major factor in the computational cost of a machine learning method (see Appendix D for a comparison between two training set sizes). Building a prediction model is much faster for smaller training sets. However, a comparison of the SIMILAR and RANDOM approaches shows that careful selection is important for classification performance. Note that the median performance of the SIMILAR approach is higher than the RANDOM performance, especially when smaller training sets were used. Furthermore, a moderate correlation was found between performance and cosine similarity. A greater correlation between performance and similarity means that similarity is selecting useful samples from the dataset for training. From this we conclude that, given a target review, cosine similarity can indeed identify transferable knowledge in the available data. For computerised support developers the proposed SIMILAR approach may be useful to reduce the training set size in settings where plenty of data is available and training on all data is infeasible.

The models always performed well for some reviews, regardless of the size of the training set, while for other reviews the models always performed poorly. Note that this is the reason that the performance boxplots (shown in Figure 5.3) cover nearly the whole possible range for the WSS@95 metric. Although other researchers hypothesized that this may be due to the number of included studies in a review, our additional analysis (Appendix C) did not reveal a clear explanation for this effect.

Reproducibility Reproducibility of methods is often problematic in systematic review automation literature because the proposed methods are difficult to reproduce and compare [18, 128]. We attempted to mitigate this problem by using relatively simple methods and a publicly available dataset. The dataset was provided in the 2017 CLEF eHealth Lab [116] and is curated such that it was available through PubMed using automated methods. We also provide the complete repository of code that was used to train the models and analyse the results in [118].

Limitations Because the dataset has a relatively small number of inclusions, a missing abstract on an inclusion has a relatively large influence on the model's performance (data shown in Appendix C). Most are not available at all and cannot be added manually. On

the other hand, using the dataset ‘as is’ enabled us to consider all fifty reviews in the CLEF dataset.

Another limitation of our study is that we only had titles and abstracts available for the similarity metric. We assume that most of the relevant information of each document is represented in their title and abstract, but it is possible that complex concepts were only expressed in the full text. Unfortunately, full-text data for all documents in the CLEF dataset is not available through PubMed in an automated way. The techniques that would enable collection of full-text documents and analysis of the influence of full-text documents on prediction performance are issues for further research.

The disease group is not the only potentially transferable knowledge among DTA reviews, as it refers only to the ‘Target condition’ aspect. The ‘Patients’ and ‘Index test’ aspects remain unexplored in this study. Further analysis of these aspects might therefore identify a different set of reviews in the ‘other’ group. Nevertheless, because cosine similarity takes all words in the documents into account, we hypothesise that it also captures the P and I aspects. Further research is needed to test this hypothesis and to adapt or extend the similarity metric to further increase its ability to detect transferable knowledge.

5.5.1 Other approaches

Although our approach for selecting training samples is novel, training sample selection itself is not a new idea in machine learning. There are numerous examples that attempt to enrich, balance, or create datasets in other domains. The techniques proposed in these papers often stem from the same type of problems: there is little to no data to train or the available data is noisy or unbalanced. Below we compare our approach with some of these other training sample selection approaches.

Cohen et al. [129] conclude that a topically similar training set almost always outperforms a set that is not. However, they also note that finding topically similar data for training is impractical. Our work, however, offers a practical approach to identify relevant training data through cosine similarity.

An example of enriching a training set for natural language models is shown by Moore et al. [130]. They showed that curating the data and selecting only those samples that improve the classifier increases the performance of the final language model. This approach is similar to ours, as we used cosine similarity to select only those reviews that are similar to the target review. In both approaches less training data is used to improve the classifier performance.

Imbalanced datasets, where the negative examples in the dataset massively outweigh the positive examples or vice versa, are often challenging in machine learning. Unlike, for example, the Random Forest classifier used in this study, there are many classifiers that cannot handle unbalanced datasets and yield a bad prediction. Nowadays there are many techniques that address unbalanced datasets. An example is shown in Kubat et al. [131]. They present a simple technique that only removes negative examples while preserving all the positive examples in the dataset. This preservation is important for systematic reviews because generally they have very few positive documents (i.e., inclusions). In this paper we did not apply such a sample selection technique, however for further research it might be interesting to combine the proposed training set selection based on cosine similarity with a technique that tackles dataset imbalance.

Lastly, instead of selection of samples we could also choose to make a sub-selection of the features that are extracted. The data used as input to the machine learning method is represented as a matrix with one sample per row and one column per feature. In the case of systematic reviews we have documents with words. The feature matrix therefore has one row per document and one word per column, and each cell contains the occurrence of a word in a document. The sample selection techniques discussed above will remove rows from this matrix whereas feature selection removes columns. As discussed in Adeva et

al. [132] feature selection reduces the training set size and condenses the important features which has a beneficial effect on the prediction model. Even though feature selection might have increased the overall performance of the prediction models, we chose not to apply it, which made it possible to focus on training set selection.

5.6 Conclusion

We have shown that cosine similarity can be used to select a training set that is relatively similar to the articles one aims to screen for. We have also shown that using all available data outperforms a dataset containing data selected using cosine similarity. Nevertheless, in cases where reviews on a similar topic are available, good prediction performance can be achieved with significantly smaller training sets.

For systematic reviewers it might be worthwhile to gather a few previously undertaken systematic reviews on a similar topic when applying computerised support to the screening of a new systematic review. However, when a large set of systematic reviews is available the extra investment to make a selection can be avoided.

The approach proposed in this work is meant to improve future tools that provide computerised support for systematic reviewers. Further research may investigate the benefits of our approach in a practical setting.

Acknowledgements

This work was carried out on the High Performance Computing Cloud resources of the Dutch national e-infrastructure with the support of the SURF Foundation. We like to thank A.H. Zwinderman and P.D. Moerland for their support designing the methodology, and B.D. Yang and M. Borgers for proofreading the manuscript.

Appendices

A Parameter selection

Each classifier method has parameters that need to be set before training on a dataset. To build the model with the highest performance the optimal value of each parameter needs to be found. For the Random Forest classifier we tested values for: `bootstrap`, `max_depth`, `max_features`, `min_samples_leaf`, `min_samples_split`, and `n_estimators` [133]. Because the number of possible parameter value combinations quickly increases a random search is done first. In the random search a random set of parameter values is tested to get a general sense of the optimal parameter settings.

For the random search we chose parameter values as follows. Parameters with numeric values were mostly chosen with equal steps between two extremes. For example, the `max_depth` test range was [10, 20, 30, ..., 90, 100, 110]. For parameters without clear boundaries online resources such as [134] were used to determine a suitable range. Parameters for which a choice had to be made (boolean or from a list of options) included most, if not all, options. For example, the `bootstrap` range contained the complete set of options: `True`, `False`. If options were dropped it was because they were similar to another option in the set. For example, the `max_features` may contain: `auto`, `sqrt`, and `log2`. We chose to forgo the `auto` option because it is equal to `sqrt`. Because this resulted in a search grid with less options the search time was reduced.

The random search was performed on ten of the fifty systematic reviews in the dataset. The combination of parameters that yielded the best result was kept for each review tested. Using these results the values that would be tested in the full grid search was determined by one of the researchers (AA):

1. if one value gave the best result for all tested reviews it was chosen as a definitive value;
2. if a value had a small range of values the range was used in the full search;
3. if a value had a large range, the extremes of the range were kept but most of the intermediate values were removed.

These choices were made to size the full grid in such a way that training the models on the complete dataset could be run in an acceptable amount of time.

Final prediction models were trained using a full grid search for each systematic review. The difference with random grid searches is that the complete set of parameter value combinations is tested. The model with the best performance is returned from the full grid search. The parameters used for both the random and full searches are described in the code found in [118].

B Cosine similarity analysis

We calculated the cosine similarity between each review and the remaining 49 reviews. The similarity scores were then sorted from highest to lowest and plotted in Figure 5.4. Each line in Figure 5.4 depicts the cosine similarity score of a target review to the remaining 49 reviews. Overall, after ordering, the similarity starts high and drops down rapidly in the first couple of reviews. Reviews tend to be most similar to less than ten other reviews. For this reason, a maximum of ten reviews was chosen for construction of the training sets for the SIMILAR approach as described in Section 5.3.4.

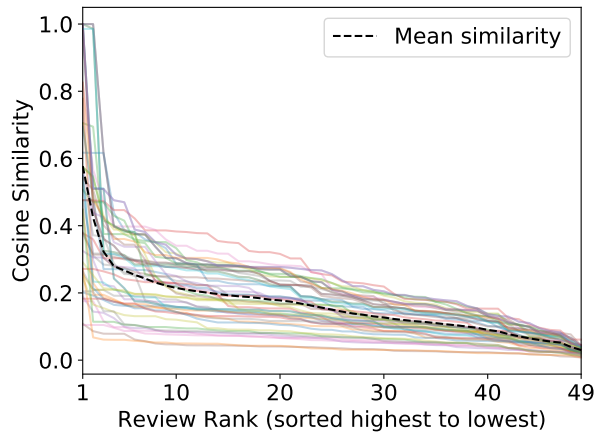


Figure 5.4: Cosine similarities between all pairs of reviews in our dataset (2,450 pairs). Each line in the figure represents a review and its similarity against the other 49 reviews, ranked from most to less similar.

Table 5.8: Pearson correlations between review metadata and WSS@95 performance.

Metadata type	Performance
% empty abstracts	0.08
% empty abstracts in inclusions	-0.28
% inclusions	-0.24
# words in abstracts	0.22
# words in titles	0.05
# of documents	0.05
is update	-0.04
publication year	0.11

C Performance trends

To get insight in the overall trends in the performance results we plotted the WSS@95 of the SIMILAR and ALL approaches. The results for both approaches were split into the fifty reviews and sorted from lowest to highest median WSS@95, the plot is shown in Figure 5.5.

On visual inspection of the plot a couple of observations were made. There is a relatively large spread in the performance of the SIMILAR approach. Often the smaller training sets ($n \in \{1, 2\}$) have a lower performance compared to the larger sets ($n \in \{5, 10\}$). However, there is no clear trend in performance gain or loss.

Some of the reviews perform better on a smaller training set (e.g., CD010772), others have approximately the same performance (e.g., CD011549), and the remaining reviews clearly perform better on more data (e.g., CD009323). A clear trend is found in the overall performance of reviews when put next to each other. For some reviews, performance is always high (e.g., CD011549), or low (e.g., CD009020), irrespectively of training set size. In their work Cohen et al. [129] noted that one of their reviews had a divergent prediction performance (i.e., lower than all others), which they figured was likely due to the low number of inclusions.

To test this observation we investigated the characteristics of the reviews. For each review we collected the following:

- percentage of empty abstracts in the review as a whole;
- percentage of empty abstracts only in the included documents;
- percentage of included documents;
- average number of words in the abstracts and titles;
- number of documents;
- whether the review was an update;
- and, the review's publication year.

A Pearson correlation was calculated between each of the metadata columns and the performance of the ALL approach.

The correlation results are shown in Table 5.8. No strong correlations were found. The percentage of empty abstracts in the included documents was moderately negatively correlated with performance (-0.28). As was the percentage of included documents (-0.24). The number of words in the abstracts was moderately positively correlated with performance (0.22).

Base on these results we hypothesise that the differences in performance are likely due to a combination of review metadata. Or perhaps there are some unobserved influences such as the availability of reviews in the dataset written by the same authors. These reviews will often have similar research topics and search strategies and are therefore more valuable in the training set. However, an in-depth analysis of these effects is outside of the scope of this paper and remains for further research.

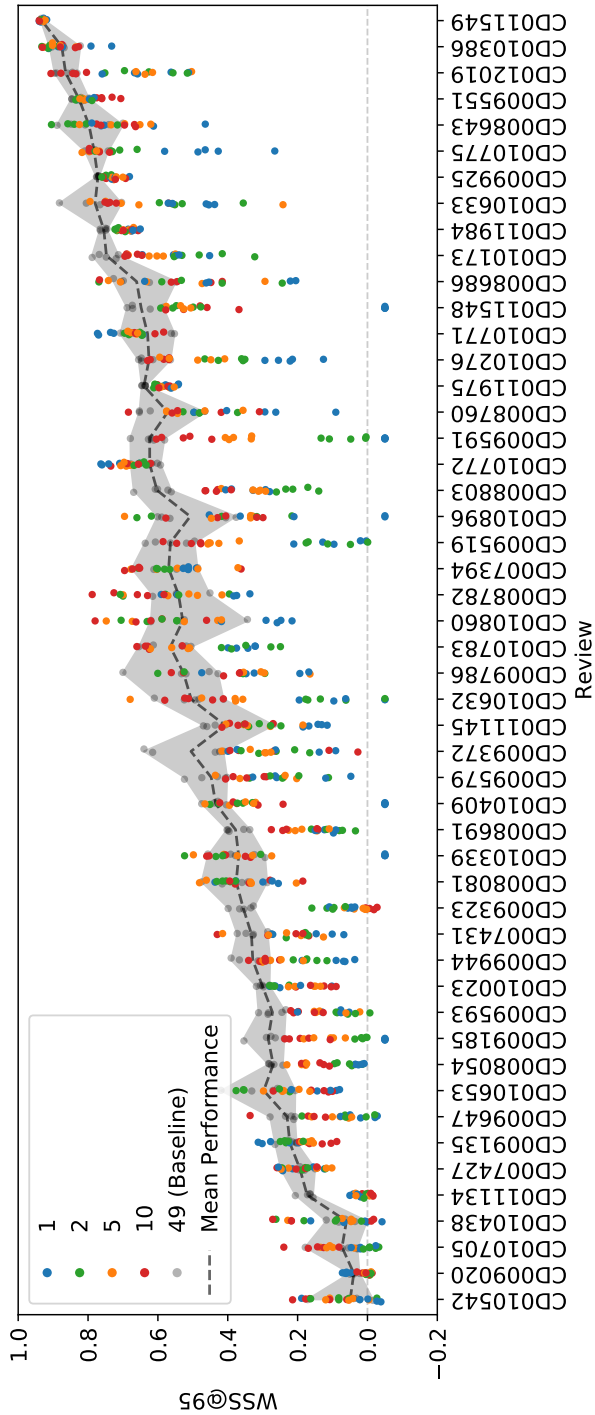


Figure 5.5: WSS@95 performance for SIMILAR and ALL models for individual reviews (total 1,250 models). Each colour represents a training set size used for the SIMILAR approach. Note that, to adjust for variance, five models were trained for each training set size, see Figure 5.1. Therefore, each review has five points per colour. The grey area represents the performance range of the ALL approach. The reviews are ordered by increasing median performance in the ALL approach.

D Computational effort

To illustrate the difference in computational effort we tracked the training time for two training set sizes: 1 and 49. The results are plotted in Figure 5.6. The training set with one review is significantly ($p < 0.001$) faster. The results show that computational effort is much lower for smaller training sets.

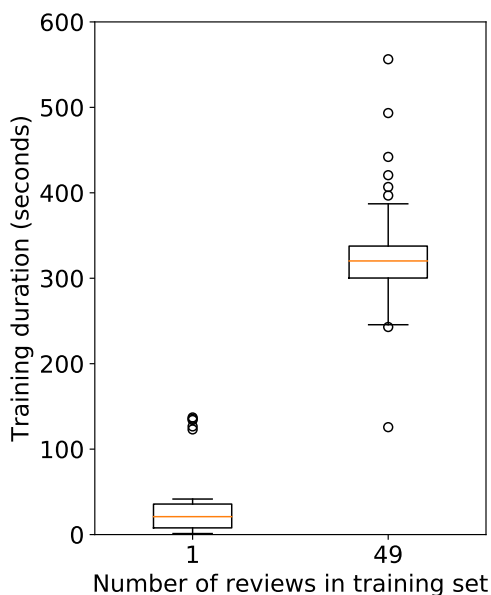


Figure 5.6: Training time, measured in seconds, for training set sizes 1 and 49.



Improving active learning performance through training sample selection

In Submission

A.J. van Altena, A.H. Zwinderman,
S.D. Olabarriaga

Abstract

Finding relevant papers for systematic reviews is a time-consuming process. Computerised support methods may aid researchers and has been applied with varying success. These methods rely on large amounts of data to train a model and predict paper inclusion. In a previous study we introduced a approach that selects training data on relevancy to the review at hand and ultimately improve prediction performance. In this paper we test our approach when applied with an active learning model.

We used a dataset of 38 Cochrane diagnostic test accuracy reviews. Every review was used as a target and the remaining 37 reviews were used to train the model. Multiple models were created with a varying number of selected reviews in the training set and performance was compared.

The performance of a model trained using five reviews was significantly higher compared to a model using all 37 reviews.

We found that, give our dataset, five is the optimal number of reviews to use in model training. Our proposed approach is agnostic to the subject of the data, therefore it is likely that the optimal number changes depending on the overall similarity of the dataset used.

Our approach improves the prediction performance of active learning models by selecting smaller datasets. These results may benefit systematic reviewers by reducing the time needed to determine the relevancy of papers.

6.1 Introduction

6.1.1 Background and significance

Systematic reviews are the cornerstone of today's evidence-based decision making in medicine [17, 87]. By synthesising all relevant evidence regarding a certain topic, systematic reviews provide a good reflection of the current scientific knowledge. However, they require a considerable amount of effort. A single review can take from 6 months to several years, depending on expertise of the research team and review size and complexity [20, 107]. Approximately half of this time is spent on developing the research protocol, performing the search, and assessing (i.e., screening) the results [108]. Therefore, reducing the time spent on these tasks will have a substantial impact on the review process [20].

Machine learning methods can aid the screening process of systematic reviews through ranking or classification of relevant articles. Many studies have been performed investigating both supervised and unsupervised methods with varying levels of success [18]. In this study we focus on a supervised learning method that needs a labelled dataset. Our approach uses the assumption that data from previous reviews shares transferable knowledge with the current systematic review [115] and thus can be used as training data for the supervised learning method. However, because systematic reviews focus on a specific research question, reusing data likely has limitations. Cohen et al. [129] state that this leaves only one opportunity for training: when a systematic review is updated after a few years, the data collected for the previous review can be used as a training set for the updated review. However, another solution is to use active learning, which has been shown to work with systematic reviews in various cases, for example [135–138].

Active learning is an interactive and iterative process where the reviewer provides information to the machine learning method while reviewing documents. The active learning loop can be started without any external training data. This would essentially yield randomly queried documents at first. However, an initial pool of documents may be provided to train the prediction model for the first time. When the initial pool has data with plenty transferable knowledge the relevant documents are found earlier, saving work for the reviewer. Note, however, that a large initial pool may dilute the transferable knowledge. This means that adding a reviewer-labelled document during the active learning loop has a small influence on the model as a whole because, by default, each document is equally important during model training.

6.1.2 Objective

In this study we focus on the question whether it is sensible to use a small number of initial documents to train the supervised learning method while keeping transferable knowledge intact. In a previous study we introduced an approach to select a subset of similar documents into the training set [139]. We found that the performance of models trained on all data slightly outperformed those trained on a selected subset of data. However, because of the characteristics of active learning, we hypothesise that selecting documents and creating a smaller initial pool outperforms a larger initial pool.

In this study we test this hypothesis by performing active learning on a curated set of Diagnostic Test Accuracy (DTA) systematic reviews. To this end, we create prediction models using our approach, of a similarity selected initial pool, and measure the performance. This approach is then compared with two others. The first approach uses all available data in the initial pool and the second uses no data in the initial pool. To our knowledge, our work is the first to use a similarity metric to select a subset of all available data as an initial pool for active learning with the aim of improving prediction performance.

6.2 Materials and methods

6.2.1 Data preparation

The data used in our experiments were taken from a previous study [139]. Below, we summarise the data gathering and enrichment steps which were taken in the previous study (for the full description refer to [139]).

Data gathering and cleaning In the experiments we used a dataset that was prepared for the 2017 CLEF eHealth Lab “Technologically Assisted Reviews in Empirical Medicine Overview” [116]. The dataset provided by the CLEF eHealth Lab consisted of PubMed IDs for all the documents that were found during the search phase of 50 systematic reviews. Each PubMed ID had a label indicating whether the PubMed document was included in the systematic review after screening the title, abstract, and full-text by the reviewers. This label is the outcome variable that is to be predicted by the prediction models.

The dataset contained a total of 266,966 documents. Reviews had an average of 5,339 documents of which, on average, 93 were included. The smallest review had 64 documents and the largest 43,363. The review with the fewest inclusions contained 2 inclusions, while the largest number of inclusions was 619. Overall this resulted in an inclusion rate of 4% with a minimum of 0.015% (2 on a total of 12,705) and a maximum of 20% (23 on a total of 114). For an overview of all the metadata collected per review see [119].

Titles and abstracts were retrieved for all documents. These were then preprocessed to remove any unwanted characters (e.g., numbers, accented letters) and stopwords (e.g., the, what, was). The words in the cleaned documents were counted and stored in a Term Frequency (TF) matrix. The Python code implementation is available at [118].

In our previous study [139] we noted that twelve reviews had low performance, independently from the approach used to build the prediction model. This is probably because these reviews all had a unique research topic. We therefore considered that these reviews would not be good candidates for similarity selection (see Appendix A). Therefore, in the current study, we excluded these twelve reviews from our experiments, leaving 38 reviews in the dataset.

6.2.2 Active learning

In this study we used an active learning method. The active learning process consisted of various components:

- *Labelled pool*: documents that have been assigned a label by the reviewer. Usually this is done after reading the title and abstract of the document. The label denotes whether the document was included or excluded from the review.
- *Unlabelled pool*: documents that have yet to be assigned a label.
- *Query*: a sample of documents pulled from the unlabelled pool.
- *Oracle*: the reviewer who labels the queried documents.
- *Active learning model*: the prediction model trained with labelled data to predict the certainty of inclusion for unlabelled data.
- *Initial pool*: an optional starting set of labelled documents.

These components together formed the active learning loop (see Figure 6.1). Documents were queried from the unlabelled pool and presented to the oracle. The oracle then assigned an *include* or *exclude* label to the document, and the document was then added to the labelled pool. A model was trained using the labelled pool. The model predicted the relevancy of the remaining unlabelled documents and assigned each a relevancy score (between 0 and 1). The documents were then prioritised by the score and the process was repeated. This continued until all documents were labelled by the oracle.

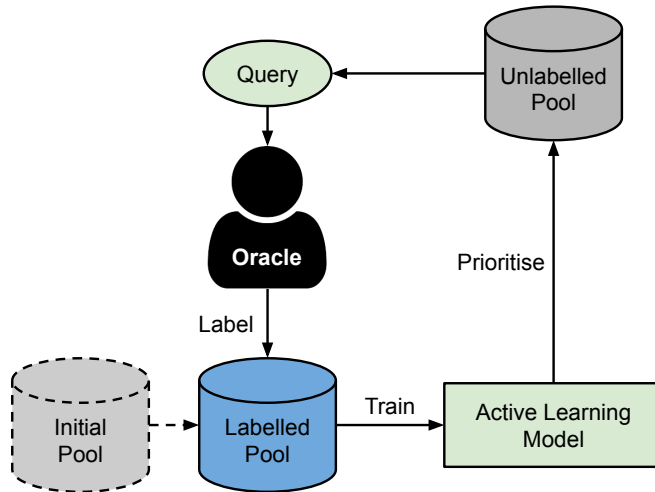


Figure 6.1: Representation of the active learning cycle. The oracle assigns labels to one or more documents from the pool of unlabelled documents. The labelled documents are then used to train or update the active learning model. The model prioritises the unlabelled pool and the cycle starts anew.

6.2.3 Prediction models

For each of the 38 reviews in the dataset we trained a prediction model using active learning. For every model the unlabelled pool was filled with the documents from a single review (i.e., the test set). The labels of the unlabelled pool (i.e., ground truth) were hidden from the model. A training set was compiled from the documents of the remaining reviews, with their corresponding labels, and put into the initial pool. Using the initial pool, a prediction model was trained. The model then predicted the relevancy for each document in the unlabelled pool. We prioritised the documents such that the document with the highest relevancy score was at the top of the list to be presented to the oracle [140].

In the real world, a reviewer would then read the title and abstract for one (or a couple) of documents from the top of the list and assign the inclusion label. For our experiments we used a simulated reviewer as oracle. The oracle took the ground truth labels that were hidden from the model and assigned them to the documents that were at the top of the relevancy list. The labelled documents were then added to the labelled pool and another cycle of the active learning process was started. The process was executed until all included documents from the test set were found by the oracle. The Python code implementation of the process described above is available at [118].

Machine learning methods We chose the Support Vector Machine (SVM) classifier to construct the prediction models. SVMs were reported to work well with high-dimensional text data, as we had in our dataset [136]. Also, this classifier lent itself well to active learning because the prediction model could be updated with one sample instead of having to train a completely new model. Lastly, this method was chosen because it was often used in related literature, for example [116, 136], facilitating reuse and comparison. We implemented SVM using the `SGDClassifier` method from the `sklearn` library [122].

The TF matrix representing the documents in our dataset was used as input to the SVM classifier. We chose to randomly undersample the dataset because it had heavily unbalanced classes (i.e., very few ‘inclusions’ and many ‘exclusions’). Undersampling was implemented using the `RandomUnderSampler` method from the `sklearn` library.

This method undersamples only the majority class, in our case the exclusions, until it resembles the minority class in size. Using a cross-validating grid search, we optimised the tolerance and alpha parameters of the SVM classifier. The values $\{10, 1, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001\}$ were used for both parameters. The best performing classifier was selected from the search and used to execute the active learning loop.

6.2.4 Cosine similarity metric

In our previous paper [139] we introduced an approach using cosine similarity to select data into the training set. Cosine similarity measures the cosine of the angle between two vectors of an inner product space [124], defined by Huang et al. [123] as follows:

$$SIM(\vec{r}_a, \vec{r}_b) = \frac{\vec{r}_a \cdot \vec{r}_b}{\|\vec{r}_a\| \|\vec{r}_b\|} \quad (6.1)$$

where \vec{r} is a vector representation of a review. Similarity ranges from 0 (not similar, vectors are at an angle of 90°) to 1 (perfectly similar, vectors are at an angle of 0°). Because the word counts of the TF matrix cannot be negative, similarity cannot be negative either.

We were interested in the similarity between complete reviews, and not between individual documents in these reviews. We therefore calculated the similarity based on the mean word count vector for all documents in each review \vec{r}_i :

$$\vec{r}_i = \frac{\sum_{j=1}^n \vec{d}_{ij}}{n_i} \quad (6.2)$$

where \vec{d}_{ij} is the word count vector for document j from review i , and n_i is the number of documents in review i . The similarity between all pairs of reviews is then defined as:

$$S_{ik} = SIM(\vec{r}_i, \vec{r}_k) \quad (6.3)$$

$$i \in \{1 \dots 38\}, k \begin{cases} k \in \{1 \dots 38\} \\ k \neq i \end{cases}$$

where \vec{r}_i and \vec{r}_k are the mean vectors for respectively reviews i and k . Similarity was calculated between all pairs of reviews ($38 \times 37 = 1,406$ in total).

6.2.5 Performance metrics

Models were judged on their ability of ordering the documents in such a way that the reviewer (or the oracle in our simulations) encounters the inclusions as early as possible in the screening process. To assess the models performance, we used yield and burden, two metrics for evaluation of active learning models introduced by Wallace et al. [136]. Yield is the fraction of relevant documents that the model correctly identified, defined as:

$$Yield = \frac{TP^L + TP^U}{TP^L + TP^U + FN^U} \quad (6.4)$$

where TP and FN are, respectively, the number of true positives and false negatives, and the superscript L and U denote manually labelled documents and unlabelled documents. Note that yield is the adaptation of *recall* for an active learning setting. Burden measures the number of documents that had to be screened manually, being defined as:

$$Burden = \frac{N^L + TP^U + FP^U}{N} \quad (6.5)$$

where FP is the number of false positives and N the total number of documents.

During the active learning process the yield and burden were measured for each loop. A yield level of 95% is commonly chosen to compare models [113, 116]. Therefore, when the model first reached a yield of 95%, the burden was retained for analysis. To get a more intuitive outcome, we took one minus burden (referred to as: Inverse Burden (IB) or IB at 95% [138]). It measures the fraction of documents that did not have to be read by the reviewer during the screening:

$$1 - Burden = \frac{TN^U + FN^U}{N} \quad (6.6)$$

where TN is the number of true negatives. The IB value ranges from 0 to 1. A higher value is better because it means less burden and therefore less work for the reviewer.

6.2.6 Experiments

We tested prediction model performance using three approaches to create the initial pool:

- Selecting a subset of the dataset using the similarity metric to create the initial pool (coined `similar-pool`);
- Using all the available data in the initial pool (coined `all-in-pool`);
- and using no initial pool (coined `no-pool`).

The process for each approach is shown in Figure 6.2. For the 38 reviews in our dataset prediction models were trained using a SVM and the features from the TF matrix. The active learning loop was repeated until all documents were labelled. Because we applied random undersampling, we repeated this process 50 times, each time with a new random draw to account for variability in the randomly sampled exclusions.

To create the initial pool for the `similar-pool` approach we used the similarity metric to create sets containing the top $s \in \{1, 2, 3, 4, 5, 7, 9, 10, 15, 30\}$ most similar reviews. For the `all-in-pool` approach all 37 reviews were used as the initial pool. Lastly, for the `no-pool` approach we simulated the oracle having to find the first inclusion from which we could start training a model. The oracle labelled documents using the original order, i.e., as returned by the PubMed search, until it encountered a relevant document. The model was then trained for the first time and the active learning loop was started.

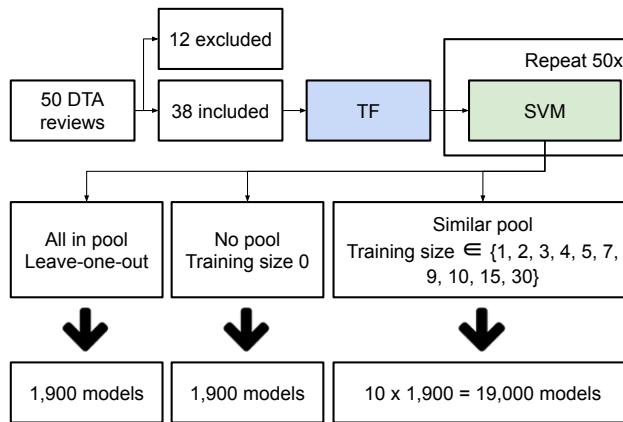


Figure 6.2: Overview of the process for the experiments using different approaches: no data in initial pool (`no-pool`); using all remaining data in initial pool (`all-in-pool`); selecting similar reviews into the initial pool (`similar-pool`).

Analysis Yield was monitored for each loop in the active learning process. For each model, when the yield reached 95%, the IB was collected. A boxplot from the IB was created for each initial pool size (i.e., 0, 1, 2, 3, 4, 5, 7, 9, 10, 15, 30, and 37). To underscore the difference in performance a polynomial regression line was fitted to the median performance values.

6.3 Results

The overall IB at 95% yield performance of all experiments is shown in Figure 6.3. Each boxplot shows the results of the 1,900 models trained and tested for the no-pool, similar-pool, and all-in-pool experiments. The polynomial line was fitted with a residual sum of squares of 0.0059.

A comparison of classifier performance against the all-in-pool baseline is shown in Table 6.1. The highest difference in performance is found at initial pool size five. The median IB for pool size five is 0.15 higher than the all-in-pool performance, a 60% increase. The results indicate that on average the best performance was obtained with smaller initial pool sizes such as those from the similar-pool experiments. Especially the pool sizes of four and five have a higher performance.

6.4 Discussion

In this study we investigated whether it is possible to improve classification performance of a supervised active learning method by using less data in the training set. Using a performance metric specific for active learning methods, we compared results obtained with training sets that were composed using different strategies: a similarity metric (similar-pool); all data (all-in-pool); and no training data (no-pool).

Prediction performance The similar-pool approach performed better compared to the all-in-pool and no-pool approaches. The box plots in Figure 6.3 show that the

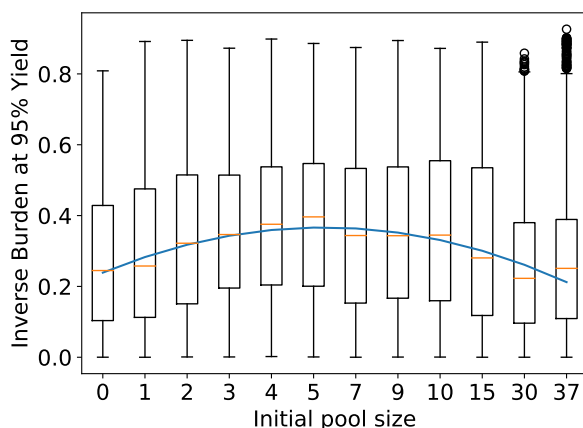


Figure 6.3: Boxplots of the IB at 95% yield for each initial pool size. Each boxplot consists of 1,900 measurements (i.e., 38×50 models). The polynomial line was fitted to the median values of the boxplots with a residual sum of squares of 0.0059.

Table 6.1: Inverse burden results for each of the initial pool sizes.

Initial pool size	Inverse burden at 95% Yield			Difference median versus all-in-pool
	Minimum	Median	Maximum	
0	< 0.01	0.24	0.81	-0.01
1	< 0.01	0.26	0.89	0.01
2	< 0.01	0.32	0.89	0.07
3	< 0.01	0.35	0.87	0.10
4	< 0.01	0.38	0.90	0.13
5	< 0.01	0.40	0.89	0.15
7	< 0.01	0.34	0.87	0.09
9	< 0.01	0.34	0.89	0.09
10	< 0.01	0.34	0.87	0.09
15	< 0.01	0.28	0.89	0.03
30	< 0.01	0.22	0.86	-0.03
37 (all-in-pool)	< 0.01	0.25	0.93	

best overall performance is reached using five reviews in the initial training set. Therefore, training an active learning method with just five selected reviews out of all 37 available reviews saves the most time and effort for the systematic reviewer.

For the largest (30 and 37) and the smallest (0 and 1) training set sizes, performance is nearly equal. In the situation where a choice between a set size of 30 and 37 has to be made, the smaller size should be preferred from a computational point of view, since the size of the training set decreases and therefore processing time as well.

It is surprising that the median performance of active learning without any initial training pool is almost equal to using the single most similar review that is available. However, the upper quadrant of the box plot (see initial pool size 0 and 1 in Figure 6.3) reaches a marginally higher performance. These results show that, in a few select cases, including only one review is better than using no reviews in the initial pool. We hypothesise that these performance results stem from the (few) highly similar reviews in our dataset. Further analysis of this observation, however, is outside the scope of this paper.

Real life application We found that five is the optimal number of reviews in the training set. This number, however, is likely to change given the dataset of available reviews. For example, if the dataset would contain only cancer research, performance might increase when using more than five reviews. We used a metric that is agnostic to the research subject and chose to not manually select reviews into the initial pool. Therefore, the question whether a research specific initial pool improves the performance of classification remains open to further research.

In our experiments no stopping criterion was applied for the active learning. Therefore, the screening continued until all documents were assessed. Ideally, the active learning loop can be stopped when the prediction method is certain it has found some percentage of all the relevant documents (usually 95% or 100%). Because the number of relevant documents is often very small compared to the number of irrelevant documents, this certainty may only occur very late in the reviewing process. Thus, because prediction models have the ability of ordering the documents in a way that those most likely to be relevant are placed at the beginning of the reading order, they are often viewed as a tool to reduce the difficulty of reviewing [135]. Reviewers that know that a given document has a low chance of relevancy may be able to exclude it by only reading a few key pieces of information in the title or abstract. Inversely, a high chance of relevancy will invoke a

more in-depth screening of the document. Prediction methods that quickly give confident relevancy scores save more time in this process.

6.4.1 Other approaches

While technically not an active learning method, we found that Jonnalagadda et al. [141] take an interesting approach to predict the relevance of systematic review documents using cosine similarity. First, they extract semantic vector representations of the documents from the data, using a combination of random indexing and a directional model. These vectors are then used to compare the documents labelled relevant by the reviewer to the remaining documents in the dataset. The most similar documents are then presented to the reviewer for labelling. The process is then repeated until all relevant documents have been identified.

Other research often investigates other ways of optimising the active learning method. For example, Wallace et al. has introduced two novel feature extraction techniques: in [136] they combine multiple feature spaces (e.g., words from the abstract combined with PubMed MeSH terms) into a multi-view active learning strategy and [135] describes a strategy where experts provide domain knowledge to the model before the prediction loop is started. Hashimoto et al. [138] introduces topic detection, which extracts features using knowledge of the words in a document together with knowledge about phrases, sentences, and paragraphs.

Also Miwa et al. [137] researched other approaches to enhance the active learning model. They looked at various methods for: data imbalance (solving the issues with having many more irrelevant than relevant documents); ensemble classifiers (training multiple models and combining them into a single prediction model); covariate shift (a problem where the distribution of the test data differs from the training data); and clustering (creating new features by clustering the data unsupervised before training the active learning model).

Our proposed approach reduces the size of the training data but does not change its form. Therefore, it can be applied together with any of the enhancements listed above and potentially improve the classification performance as shown in this paper.

6.5 Conclusion

We have shown that cosine similarity is a relatively simple and effective tool for selecting relevant reviews into an initial training set for active learning purposes. Performance of active learning methods using a selected dataset was significantly higher when compared against using all the available data or when starting without any data in the training set. We found that five reviews in the training set reached the highest performance. However, it is likely that this number may change when the composition of the dataset changes. Nevertheless, results indicate that smaller training set used as initial pool for active learning methods may greatly reduce the reading time for systematic reviewers.

Acknowledgements

We like to thank P.D. Moerland for his support designing the methodology and experiments. Furthermore, we thank M.M.G. Leeftang for her expertise checking the ICD-10 codes assigned to the CLEF reviews and M. Borgers for proofreading the manuscript. Lastly, we thank Shuxin Zhang and René Spijker for their contributions during the initial conceptualisation of this research.

Appendices

A Research Subject Group Performance

In previous research [139] we grouped the reviews in the dataset based on their research subject. The subject was assigned using the International Classification of Diseases, 10th revision (ICD-10). Reviews with the same ICD-10 code were grouped together. Those that were found to have a unique ICD-10 code in the dataset were assigned to the 'other' group.

Our previous results showed that classification performance was lower for the reviews in the 'other' group. We tested for statistical significance, which showed that the difference in performance is significant over all training set sizes.



Discussion

7.1 General discussion

In this thesis we showed that Big Data is a term that sees plenty of use in (bio)medical research. Because Big Data covers many different subjects its definition can only be captured in broad terms. Nevertheless, a quantifiable difference exists between literature that used the term Big Data and those that did not, showing that the label *Big Data* indicates a separate field of research.

Also, we found that systematic reviewers are a group of researchers who encounter Big Data challenges in many phases of their work. They stand to benefit greatly from applying solutions to these challenges. However, due to a number of barriers, adoption of existing solutions is lacking.

Lastly, we propose and assess a new approach to computerised support of literature appraisal in systematic reviews. Our new approach showed a significant increase in performance when compared to existing approaches, an outcome that could aid systematic reviewers by reducing the time spent on appraising literature.

In this final chapter we will discuss the findings above in the light of current literature.

7.2 Understanding Big Data

In Chapter 2 we described how existing Big Data definitions are expressed within (bio)medical research literature. This study built upon findings of previous qualitative research by De Mauro et al. that was published in 2016. De Mauro et al. analysed fifteen definitions and identified four key Big Data themes [8]. We have revisited these and other definitions of Big Data, and consolidated them into eight additional themes, resulting in a total of twelve themes. Manual annotation of the (bio)medical literature we collected showed a strong presence of the themes proposed by De Mauro et al.. We noted that these themes are defined broadly, thereby capturing many of the other eight themes in them. These results indicate that, at that time, the understanding of Big Data was mostly captured in broad terms.

Since our study, others have researched the definition of Big Data. For example, Sestino et al. proposed a model with three themes: implications, applications, and methods [142]. This model captures the same concepts as the four themes by De Mauro et al. and therefore each theme by Sestino et al. has a broader definition. Also, De Mauro et al. further refined their definition a few years later by assigning subtopics to the four major themes [143]. Note that while these subtopics give a clearer description of the major

themes, they do not narrow them down. Favaretto et al. have studied the understanding of Big Data by interviewing researchers [144]. They note that out of 39 participants only one could pinpoint a formal definition of Big Data. The others rather opted to indirectly describe the term. For example, by telling about the research that they thought was Big Data.

The studies described above all gravitate to more or less the same conclusion: it remains difficult to draw a well-defined boundary around the concepts of Big Data. Nevertheless, I will argue that the term still holds value for researchers. Favaretto et al. found a commonality in the responses of participants, even though they worked in a range of different research fields. This indicates that a presumed gap in understanding does not necessarily exist in practice. Take, for example, a dataset consisting of three small sources, each with their own data types (such as: measurements, text, timestamps). It is not unlikely that a researcher working with this dataset will run into issues due to the variety of data and needs techniques to solve them. The necessary techniques may be labelled as Big Data, either because they are developed for voluminous datasets or because the developer of the technique applied a different definition. Regardless, the technique will be very useful to the researcher trying to solve their data challenge. Therefore, maybe the definition of Big Data is less important than applying it to whatever a researcher thinks is appropriate.

In Chapter 3 we described that there was a detectable difference between publications that use the term Big Data and those that did not. This was achieved by feeding a large number of (bio)medical research papers into a machine learning method. These research papers belonged to two groups: those that contained the term Big Data and those that did not. We found that the machine learning method had a high performance in distinguishing characteristics of papers in each of the groups, thereby differentiating Big Data research from ‘other’ research. Moreover, analysing the most commonly used words in the Big Data papers, we found that the use of the term Big Data within a publication seemed to indicate a distinct type of research in the biomedical field. We concluded that value can be attributed to the term Big Data when used in a publication. This conclusion strengthened our belief that it is useful to attach the term wherever a researcher thinks it is applicable, thereby making the work findable for others.

Others also applied similar techniques to understand Big Data. Hahn et al. used a scope similar to our research while Parlina et al. and Mohammadi et al. looked at all literature published within a specific literature database [66, 145, 146]. All conclude that interest increases for some Big Data topics, such as data mining and parallel computing, while it decreases for others. Mohammadi et al. also state that methodology seems to have caught up with the availability of data and we are now in the implementation stage [146]. This indicates that Big Data has reached the “plateau of productivity” on the Gartner hype cycle¹ and has, or will, become a regular tool.

Thus, understanding Big Data from the published literature seems like a fulfilled field of research. However, Big Data research focus changes over time, which may be of interest to new researchers or businesses. They can get up to speed with the latest trends in Big Data quickly, using the results of studies like those named above.

7.3 Adoption of automation tools

In Chapter 4 we investigated the adoption of automation tools among systematic reviewers. To this end we deployed surveys and found that automation tools were not widely used among the participants. When tools are used, participants mostly learn about them from

¹The Gartner hype cycle places emerging technologies on a course that describes five levels of expectations: technology trigger, peak of inflated expectations, trough of disillusionment, slope of enlightenment, and plateau of productivity. If the technology reaches the end of the hype cycle it has become part of the regular toolkit a user might use.

their environment, for example through colleagues, peers, or organisation. Tools are often chosen on the basis of user experience, either by own experience or from colleagues or peers. Lastly, licensing, steep learning curve, lack of support, and mismatch to workflow are often reported by participants as relevant barriers. These results provide evidence and confirmed the conclusions and recommendations of previous work from others such as O'Connor et al. [91], which was based on expert opinions. Gates et al. performed similar research into tool adoption [147]. They focussed on the users of three specific tools and reached similar conclusions as we did in our work.

Even though adoption is low, automation tools may greatly benefit systematic reviewers. Clark et al. showed this in their work [148]. By applying tools to many of the phases of the review process and optimising the workflow, a team of systematic reviewers completed a review in two weeks². This is an impressive improvement over the median completion time of 41 weeks reported in Borah et al. [149]. In a follow-up study Clark et al. perform another systematic review with two teams: a manual team and an automation team [150]. The time spent on tasks, where automation was possible, was reduced from 42 hours to 12 hours, while the methodological quality of the review was maintained.

All the results above imply that adoption is not hindered by the performance of automation tools, but rather by the users' sentiment towards the tool. Clark et al. state that an increase in associating efficiency with using tools increases the adoption of those tools [148]. The same is implied in the works of Gates et al. and Scott et al. [147, 151]. Both papers indicate that tool developers need to actively process user feedback to improve their product, which will ultimately lead to higher adoption. Specifically, Gates et al. discuss that important contributors to the adoption of automation tools are: (1) *usability*: from three tools investigated only one was deemed usable according to a standardised usability score; (2) *reliability*: trust in tools may be boosted by allowing the user to customise the level of risk they are willing to take (e.g., manually setting relevancy thresholds, thereby reducing/increasing the number of 'hits' with a trade-off of missing relevant publications); (3) *fit with systematic review workflows*: multiple experienced researchers could not retrieve the results they intended to get from the tools.

A great example of development that takes the recommendations above into account is the Dextr tool [152]. Dextr is an automation tool for identifying, extracting, and connecting data entities from environmental health animal studies. During its development end-users were kept in the loop to adjust the usability and interoperability. Meanwhile, reliability was built into the design by making it possible for users to query the tool on its reasoning behind the outcomes. Quality of the work slightly decreased while time spent on the tasks was halved, similar to Clark et al. [150]. It would be interesting to validate whether this approach to development actually overcomes barriers in adoption. The Dextr tool, or any other newly developed tool that uses the same approach, should be compared with similar automation tools that did not follow the given recommendations.

Lastly, the results described in Chapter 4 highlight the importance that organisations and best practices in a field can have for the adoption of automation tools for systematic reviews. A 2021 research by Arno et al. focussed on the opinions of systematic review guideline developers [153]. They found that guideline developers are mostly concerned with the fit of tools with current practices and values, and are less concerned by usability or validation of the tools. This is a valuable insight because guideline developers decide the rules that systematic reviewers should adhere to for scientifically solid work. Not following these guidelines makes it difficult to get published. Therefore, the sentiment of this group directs most of the decisions made in the systematic reviews field. Because guideline developers are crucial in adoption of tools, I argue that this should be the first group to be convinced of the reliability of automation tools.

²Note that the setup was somewhat artificial because the team consisted of experienced reviewers that had blocked off time for the whole duration of the project. In reality it is likely that less experienced members of the team and various other tasks slow down a review.

7.4 Improving automation tools

In Chapter 5 we introduced and assessed a new approach to improve the performance of automation tools. We focussed on automation tools that could be applied in the appraisal stage of systematic reviews to automatically identify the relevant studies from the literature search. To predict relevancy, a prediction algorithm needs to be trained on systematic reviews for which the appraisal of studies was manually completed by a researcher. The algorithm learns how to make correct predictions from each given sample. A rule of thumb in machine learning is to use all available data, because more examples used during training will usually yield better predictions. However, systematic reviews investigate unique research questions. The question arises whether using all available data actually results in the best performing algorithm. After all, when considering which data to use for a review about Alzheimer, another review about Alzheimer might provide a better training set than a review about cancer. Therefore, excluding the review about cancer from the training data might improve the algorithm's performance because the remaining training data is less diluted.

Our approach attempted to select training data by measuring similarity between the available data of completed systematic reviews and the systematic review we were automating. We selected only the most similar data and then trained and tested our prediction algorithm. In contrast to our hypothesis, we found that algorithms trained on more data performed better. However, algorithms trained for reviews that had a research subject similar to other reviews in the dataset got better results. We concluded that our proposed approach had the potential to improve prediction performance in those cases.

To test our conclusion, in Chapter 6 we applied our approach in an active learning setting. In this setting the tool made a few predictions and presented them to the user. The user appraised these predictions and gave feedback to the tool. The tool assessed the feedback and made new predictions. This process was repeated until a stopping criteria was reached (e.g., all relevant items were found). Because the information provided by the user was the most relevant to the review at hand it should weigh heavily in the prediction process. Reducing the number of samples in the initial training data should therefore affect the machine learning performance. In Chapter 6 we found that the performance of an active learning tool trained using less data was significantly higher compared to a tool using all available data. Therefore, we concluded that our approach improved prediction performance when using active learning. Moreover, another benefit of this approach was that the reduction in data also reduced the computational effort of training the machine learning method. This meant that training was faster, leading to less downtime for users and a potential cost reduction for IT infrastructure.

As noted in Chapters 5 and 6, there are many studies that investigate optimisation of training data. To our knowledge, none of these focussed on reducing the size of the training set based on the similarity of the available training data. However, data selection using a similarity metric was applied in at least one other research field. Unnikrishnan et al. describes a method of applying cosine similarity (i.e., the same metric we applied) for selecting training data in a mobile health dataset [154]. Their dataset consists of users, with some agents having little data and others having lots. They solved this data imbalance by selecting similar data for the small agents instead of using all available data. Overall, they concluded that better performance is reached when using less data. This outcome shows that the benefits of similarity selection are applicable in more settings than just systematic review automation and the general rule-of-thumb that *'more data equals better outcomes'* does not always apply.

We did not compare our approach to approaches specific to the systematic review automation field. We had two main reasons for this: (1) our approach is a building block to a complete automation tool; and (2) the lack of consistency in reporting of automation tool development, described in 2015 by O'Mara et al., is still present in 2021 [18, 155] making it extremely difficult to compare automation tools. In Chapters 5 and 6 we strive for

open science by providing the accompanying (code) implementations and data. However, whether this is enough for our approach to be dissected and implemented by others is difficult to tell. There are no guidelines for developers on how to present their methods. In this perspective, the automation tool community could learn a lot from the systematic review guideline developers. In my opinion the community would benefit greatly from a specification of the research field, guidelines to releasing underlying methods, and standardised performance metrics.

7.5 Concluding remarks

In this thesis we found a definition for the term Big Data and that it holds value in (bio)medical science literature. We then described barriers and facilitators for the adoption of automation tools among systematic reviewers. We also introduced a new method that boosts the performance of systematic review automation tools while reducing the computational cost. Furthermore, I strongly argue for developing guidelines and recommendations for automation tool developers because they will benefit all parties involved. Developers could build better tools by learning from their users and by comparing them with competing tools. Authoritative instances, such as guideline developers, would be able to assure that the quality of scientific work will not be affected. Systematic reviewers would have to spend less time on repetitive tasks. And, ultimately, the consumers of systematic reviews will get the latest scientific insights earlier with more frequent updates.



Appendices

References

- [1] K. J. Rothman. 'Lessons from John Graunt'. In: *The Lancet* 347.8993 (1996), pp. 37–39. doi: 10.1016/s0140-6736(96)91562-7.
- [2] R. Frerichs. 'John Snow'. In: *Encyclopædia Britannica*. 2012. URL: <https://www.britannica.com/biography/John-Snow-British-physician> (visited on 29/12/2021).
- [3] K. S. McLeod. 'Our sense of Snow: the myth of John Snow in medical geography'. In: *Social science & medicine* 50.7-8 (2000), pp. 923–935. doi: 10.1016/S0277-9536(99)00345-7.
- [4] A. M. Weinberg. 'Impact of large-scale science on the United States'. In: *Science* 134 (1961), pp. 161–164. doi: 10.1126/science.134.3473.161.
- [5] D. J. de Solla Price. *Little science, big science... and beyond*. Columbia University Press New York, 1986. doi: 10.7312/pric91844.
- [6] J. Fenn and H. LeHong. *Hype cycle for emerging technologies, 2011*. Tech. rep. Gartner, 2011.
- [7] J. S. Ward and A. Barker. 'Undefined By Data: A Survey of Big Data Definitions'. In: *CoRR abs/1309.5821* (2013). URL: <http://arxiv.org/abs/1309.5821>.
- [8] A. De Mauro, M. Greco and M. Grimaldi. 'A formal definition of Big Data based on its essential features'. In: *Library Review* 65.3 (2016), pp. 122–135. doi: 10.1108/LR-06-2015-0061.
- [9] A. Jacobs. 'The Pathologies of Big Data'. In: *Commun. ACM* 52.8 (Aug. 2009), pp. 36–44. ISSN: 0001-0782. doi: 10.1145/1536616.1536632.
- [10] T. DeRouen. 'Promises and pitfalls in the use of "Big Data" for clinical research'. In: *Journal of Dental Research* 94.9 (Sept. 2015), 1075–1095. doi: 10.1177/0022034515587863.
- [11] P. Zikopoulos and C. Eaton. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. 1st ed. Vol. 1. New York, NY, USA: McGraw-Hill Osborne Media, 2011. ISBN: 0071790535.
- [12] D. Laney. '3D data management: Controlling data volume, velocity and variety'. In: *META Group Research Note* 6 (Feb. 2001), p. 70.
- [13] J. Andreu-Perez et al. 'Big Data for Health'. In: *IEEE Journal of Biomedical and Health Informatics* 19.4 (July 2015), pp. 1193–1208. ISSN: 2168-2194. doi: 10.1109/JBHI.2015.2450362.
- [14] D. Fisher et al. 'Interactions with Big Data Analytics'. In: *interactions* 19.3 (May 2012), pp. 50–59. ISSN: 1072-5520. doi: 10.1145/2168931.2168943.
- [15] H. Chen, R. H. Chiang and V. C. Storey. 'Business Intelligence and Analytics: From Big Data to Big Impact'. In: *MIS Q.* 36.4 (Dec. 2012), pp. 1165–1188. ISSN: 0276-7783. doi: 10.2307/41703503.
- [16] E. Dumbill. 'Making sense of big data'. In: *Big Data* 1.1 (2013), pp. 1–2. doi: 10.1089/big.2012.1503.
- [17] D. Gough, S. Oliver and J. Thomas. *An introduction to systematic reviews*. 1st ed. 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP UK: Sage Publications Ltd, 2012. ISBN: 1473929431.

- [18] A. O'Mara-Eves et al. 'Using text mining for study identification in systematic reviews: a systematic review of current approaches'. In: *Systematic reviews* 4.1 (2015), p. 5. doi: 10.1186/2046-4053-4-5.
- [19] H. Bastian, P. Glasziou and I. Chalmers. 'Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up?' In: *PLOS Medicine* 7.9 (Sept. 2010), pp. 1–6. doi: 10.1371/journal.pmed.1000326.
- [20] G. Tsafnat et al. 'Systematic review automation technologies.' In: *Systematic reviews* 3.1 (2014), p. 74. ISSN: 2046-4053. doi: 10.1186/2046-4053-3-74. PMID: 25005128.
- [21] F. X. Diebold. 'On the Origin(s) and Development of the Term 'Big Data''. In: *PIER Working Paper* (2012). doi: 10.2139/ssrn.2152421.
- [22] Google. *Google Trends*. URL: <https://www.google.com/trends/explore#q=big+data> (visited on 28/03/2016).
- [23] Gartner. *Gartner Acquisitions*. URL: http://www.gartner.com/technology/about/acquisition_history.jsp (visited on 27/03/2016).
- [24] J. P. Dijcks. *Oracle: Big data for the enterprise*. Tech. rep. Oracle, Oct. 2012. URL: <http://www.oracle.com/technetwork/database/bi-datawarehousing/wp-big-data-with-oracle-521209.pdf> (visited on 12/09/2016).
- [25] IBM. *IBM - What is Big Data?* Accessed through Google cache. URL: <https://www.ibm.com/software/data/bigdata/what-is-big-data.html> (visited on 17/12/2015).
- [26] J. Dutcher. *What Is Big Data?* 2014. URL: <https://datascience.berkeley.edu/what-is-big-data/> (visited on 12/09/2016).
- [27] M. Levi. *Kleren van de keizer [The emperor's clothes]*. Column, Medisch Contact. Oct. 2015.
- [28] M. Chen, S. Mao and Y. Liu. 'Big Data: A Survey'. In: *Mobile Networks and Applications* 19.2 (2014), pp. 171–209. ISSN: 1572-8153. doi: 10.1007/s11036-013-0489-0.
- [29] T. Hansmann and P. Niemeyer. 'Big Data - Characterizing an Emerging Research Field Using Topic Models'. In: *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01*. WI-IAT '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 43–51. ISBN: 978-1-4799-4143-8. doi: 10.1109/WI-IAT.2014.15.
- [30] D. M. Blei, A. Y. Ng and M. I. Jordan. 'Latent Dirichlet Allocation'. In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022. ISSN: 1532-4435. doi: 10.5555/944919.944937.
- [31] D. M. Blei. 'Probabilistic Topic Models'. In: *Commun. ACM* 55.4 (Apr. 2012), pp. 77–84. ISSN: 0001-0782. doi: 10.1145/2133806.2133826.
- [32] M. Steyvers and T. Griffiths. 'Probabilistic topic models'. In: *Handbook of Latent Semantic Analysis* 427.7 (2007), pp. 424–440. doi: 10.4324/9780203936399.
- [33] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015. URL: <https://www.R-project.org/>.
- [34] I. Feinerer, K. Hornik and D. Meyer. 'Text mining infrastructure in R'. In: *Journal of Statistical Software* 25.5 (2008). R package version 0.6-2, pp. 1–54. URL: <http://www.jstatsoft.org/v25/i05/>.
- [35] K. Benoit and P. Nulty. *quanteda: Quantitative analysis of textual data*. R package version 0.8.5-10. 2015. URL: <http://github.com/kbenoit/quanteda>.
- [36] D. D. Lewis et al. 'RCV1: A New Benchmark Collection for Text Categorization Research'. In: *J. Mach. Learn. Res.* 5 (Dec. 2004), pp. 361–397. ISSN: 1532-4435. doi: 10.5555/1005332.1005345.
- [37] G. Salton. *The SMART Retrieval System-Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1971. ISBN: 0138145253. doi: 10.5555/1102022.

- [38] D. D. Lewis et al. -. URL: <http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-s-mart-stop-list/english.stop> (visited on 20/11/2015).
- [39] B. Grün and K. Hornik. 'topicmodels: An R package for fitting topic models'. In: *Journal of Statistical Software* 40.13 (2011). R package version 0.2-2, pp. 1–30. URL: <http://www.jstatsoft.org/v40/i13/>.
- [40] J. Chuang et al. 'Topic model diagnostics: Assessing domain relevance via topical alignment'. In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 2013, pp. 612–620. doi: 10.5555/3042817.3043005.
- [41] G. Schwarz. 'Estimating the Dimension of a Model'. In: *Ann. Statist.* 6.2 (Mar. 1978), pp. 461–464. doi: 10.1214/aos/1176344136.
- [42] H. Akaike. 'Information Theory and an Extension of the Maximum Likelihood Principle'. In: *Selected Papers of Hirotugu Akaike*. Ed. by E. Parzen, K. Tanabe and G. Kitagawa. New York, NY, USA: Springer New York, 1998, pp. 199–213. ISBN: 978-1-4612-1694-0. doi: 10.1007/978-1-4612-1694-0_15.
- [43] C. Sievert and K. E. Shirley. 'LDAvis: A method for visualizing and interpreting topics'. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 2014, pp. 63–70. doi: 10.13140/2.1.1394.3043.
- [44] G. K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Indianapolis, IN, USA: Addison-Wesley press, 1949. ISBN: 161427312X.
- [45] M. Schroeck et al. 'Analytics: The real-world use of big data'. In: *IBM Global Business Services* (Jan. 2012), pp. 1–20. URL: <https://www.bdvc.nl/images/Rapporten/GBE03519USEN.PDF> (visited on 12/09/2016).
- [46] S. Suthaharan. 'Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning'. In: *SIGMETRICS Perform. Eval. Rev.* 41.4 (Apr. 2014), pp. 70–73. ISSN: 0163-5999. doi: 10.1145/2627534.2627557.
- [47] L. Chang. *NIST Big Data Interoperability Framework: Volume 1, Definitions*. NIST, Sept. 2015, p. 32. doi: 10.6028/NIST.SP.1500-1.
- [48] D. Boyd and K. Crawford. 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon'. In: *Information, Communication & Society* 15.5 (2012), pp. 662–679. doi: 10.1080/1369118X.2012.678878.
- [49] I. Center. *Big Data Analytics*. Tech. rep. Intel IT Center, 2012. URL: <https://www.intel.com/content/dam/www/public/us/en/documents/reports/intel-corp-big-data-policy-position-paper.pdf> (visited on 12/09/2016).
- [50] Microsoft. *The Big Bang: How the Big Data Explosion Is Changing the World*. URL: <https://news.microsoft.com/2013/02/11/the-big-bang-how-the-big-data-explosion-is-changing-the-world/> (visited on 11/02/2013).
- [51] B. Shneiderman. 'Extreme Visualization: Squeezing a Billion Records into a Million Pixels'. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. SIGMOD '08. Vancouver, Canada: ACM, 2008, pp. 3–12. ISBN: 978-1-60558-102-6. doi: 10.1145/1376616.1376618.
- [52] V. Mayer-Schönberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. UK: John Murray Publishers, 2013. ISBN: 1848547927, 9781848547926.
- [53] J. Manyika et al. 'Big data: The next frontier for innovation, competition, and productivity'. In: (June 2011). URL: <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation> (visited on 12/09/2016).
- [54] C.-W. Tsai et al. 'Big data analytics: a survey'. In: *Journal of Big Data* 2.1 (2015), p. 21. ISSN: 2196-1115. doi: 10.1186/s40537-015-0030-3.

-
- [55] AlchemyAPI. *Alchemy*. URL: <http://www.alchemyapi.com> (visited on 15/12/2015).
 - [56] H. M. Wallach et al. 'Evaluation Methods for Topic Models'. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: ACM, 2009, pp. 1105–1112. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553515.
 - [57] C. Sievert. *Finding structure in xkcd comics with Latent Dirichlet Allocation*. URL: <https://cpsievert.github.io/xkcd/> (visited on 20/11/2015).
 - [58] J. Chang et al. 'Reading Tea Leaves: How Humans Interpret Topic Models'. In: *Advances in Neural Information Processing Systems 22*. Ed. by Y. Bengio et al. Red Hook, NY, USA: Curran Associates, Inc., 2009, pp. 288–296. DOI: 10.5555/2984093.2984126.
 - [59] J. H. Lau et al. 'Automatic Labelling of Topic Models'. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Portland, Oregon: Association for Computational Linguistics, 2011, pp. 1536–1545. ISBN: 978-1-932432-87-9. DOI: 10.5555/2002472.2002658.
 - [60] Q. Mei, X. Shen and C. Zhai. 'Automatic Labeling of Multinomial Topic Models'. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '07. San Jose, California, USA: ACM, 2007, pp. 490–499. ISBN: 978-1-59593-609-7. DOI: 10.1145/1281192.1281246.
 - [61] Amazon. *Amazon Mechanical Turk*. URL: <https://www.mturk.com> (visited on 27/02/2016).
 - [62] W. X. Zhao et al. 'Comparing Twitter and Traditional Media Using Topic Models'. In: *Proceedings of the 33rd European Conference on Advances in Information Retrieval*. ECIR'11. Dublin, Ireland: Springer-Verlag, 2011, pp. 338–349. ISBN: 978-3-642-20160-8. DOI: 10.1007/978-3-642-20161-5_34.
 - [63] J. L. Hurtado, A. Agarwal and X. Zhu. 'Topic discovery and future trend forecasting for texts'. In: *Journal of Big Data* 3.1 (2016), pp. 1–21. ISSN: 2196-1115. DOI: 10.1186/s40537-016-0039-2.
 - [64] X. Tian. 'Big data and knowledge management: a case of déjà vu or back to the future?' In: *Journal of Knowledge Management* 21.1 (2017), pp. 113–131. DOI: 10.1108/JKM-07-2015-0277.
 - [65] V. Mayer-Schönberger and K. Cukier. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Mariner Books, Jan. 2013, p. 257. ISBN: 9780544227750.
 - [66] A. Hahn, S. D. Mohanty and P. Manda. 'What's Hot and What's Not? - Exploring Trends in Bioinformatics Literature Using Topic Modeling and Keyword Analysis'. In: *Bioinformatics Research and Applications: 13th International Symposium, ISBRA 2017, Honolulu, HI, USA, May 29 – June 2, 2017, Proceedings*. Ed. by Z. Cai, O. Daescu and M. Li. Springer International Publishing, 2017, pp. 279–290. DOI: 10.1007/978-3-319-59575-7_25.
 - [67] S. Weiss et al. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Jan. 2004. DOI: 10.1007/978-0-387-34555-0.
 - [68] H.-k. Zhou, H.-m. Yu and R. Hu. 'Topic discovery and evolution in scientific literature based on content and citations'. In: *Frontiers of Information Technology & Electronic Engineering* 18.10 (Oct. 2017), pp. 1511–1524. ISSN: 2095-9230. DOI: 10.1631/FITEE.1601125.
 - [69] A. J. van Altena et al. 'Analysis of the term 'big data': Usage in biomedical publications'. In: *2017 IEEE International Conference on Big Data (Big Data)*. Dec. 2017, pp. 1253–1258. DOI: 10.1109/BigData.2017.8258051.

- [70] N. Heudecker. *Big Data Isn't Obsolete. It's Normal*. 2015. URL: <http://blogs.gartner.com/nick-heudecker/big-data-is-now-normal/> (visited on 18/05/2018).
- [71] A. Foo. *Face It, Big Data Is the New Normal*. 2013. URL: <http://www.ibmbigdatahub.com/blog/face-it-big-data-new-normal> (visited on 18/05/2018).
- [72] Anon. *Big Data Series*. 2014. URL: <https://www.parliament.uk/mps-lords-and-offices/offices/bicameral/post/work-programme/big-data/> (visited on 18/05/2018).
- [73] D. Laney. *Big Data's 10 Biggest Vision and Strategy Questions*. 2015. URL: <http://blogs.gartner.com/doug-laney/big-datas-10-biggest-vision-and-strategy-questions/> (visited on 18/05/2018).
- [74] A. Gandomi and M. Haider. 'Beyond the hype: Big data concepts, methods, and analytics'. In: *International Journal of Information Management* 35.2 (2015), pp. 137–144. ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2014.10.007.
- [75] S. Kudva and X. Ye. 'Smart Cities, Big Data, and Sustainability Union'. In: *Big Data and Cognitive Computing* 1.1 (2017). ISSN: 2504-2289. DOI: 10.3390/bdcc1010004.
- [76] S. Wolfert et al. 'Big Data in Smart Farming – A review'. In: *Agricultural Systems* 153 (2017), pp. 69–80. ISSN: 0308-521X. DOI: 10.1016/j.agsy.2017.01.023.
- [77] I. A. T. Hashem et al. 'The rise of “big data” on cloud computing: Review and open research issues'. In: *Information Systems* 47 (2015), pp. 98–115. ISSN: 0306-4379. DOI: 10.1016/j.is.2014.07.006.
- [78] A. J. van Altena et al. 'Understanding big data themes from scientific biomedical literature through topic modeling'. In: *Journal of Big Data* 3.1 (2016), p. 23.
- [79] R. Kitchin. 'Big data and human geography: Opportunities, challenges and risks'. In: *Dialogues in Human Geography* 3.3 (2013), pp. 262–267. DOI: 10.1177/2043820613513388.
- [80] R. Kitchin and G. McArdle. 'What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets'. In: *Big Data & Society* 3.1 (2016), p. 2053951716631130. DOI: 10.1177/2053951716631130.
- [81] A. J. van Altena. *AMCeScience/python-miner-pub*. 2018. URL: <https://github.com/AMCeScience/python-miner-pub/>.
- [82] Bethesda (MD): National Center for Biotechnology Information (US). *Entrez Programming Utilities Help*. 2010. URL: <https://www.ncbi.nlm.nih.gov/books/NBK25501/> (visited on 18/05/2018).
- [83] D. Moher et al. 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement'. In: *Journal of Clinical Epidemiology* 62.10 (2009), pp. 1006–1012. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2009.06.005.
- [84] E. Loper and S. Bird. *NLTK: The Natural Language Toolkit*. 2002.
- [85] A. J. van Altena. *AMCeScience/R-contrast-pub*. 2018. URL: <https://github.com/AMCeScience/R-contrast-pub/>.
- [86] J. Friedman, T. Hastie and R. Tibshirani. 'Regularization Paths for Generalized Linear Models via Coordinate Descent'. In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/>.
- [87] D. Gough and D. Elbourne. 'Systematic Research Synthesis to Inform Policy, Practice and Democratic Debate'. In: *Social Policy and Society* 1.3 (2002), pp. 225–236. DOI: 10.1017/S147474640200307X.
- [88] I. Chalmers and P. Glasziou. 'Avoidable waste in the production and reporting of research evidence'. In: *Lancet* 374 (2009), pp. 86–89. DOI: 10.1016/S0140-6736(09)60329-9.

- [89] I. Shemilt et al. 'Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews'. In: *Research Synthesis Methods* 5.1 (2013), pp. 31–49. doi: 10.1002/jrsm.1093.
- [90] C. Marshall. *Systematic Review Toolbox*. 2017. URL: <http://www.systematicreviewtools.com/> (visited on 14/10/2016).
- [91] A. M. O'Connor et al. 'Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR)'. In: *Systematic Reviews* 7.1 (Jan. 2018), p. 3. ISSN: 2046-4053. doi: 10.1186/s13643-017-0667-4.
- [92] F. D. Davis. 'A technology acceptance model for empirically testing new end-user information systems: Theory and results'. PhD thesis. Massachusetts Institute of Technology, 1986. doi: 1721.1/15192.
- [93] V. Venkatesh. 'Determinants of Perceived Ease of Use : Integrating Control , Intrinsic Motivation , and Emotion into the Technology Acceptance Model'. In: *Information System Research* 11.4 (2000), pp. 342–365. ISSN: 10477047,15265536. doi: 10.1287/isre.11.4.342.11872. PMID: 3961358.
- [94] V. Venkatesh et al. 'A Theoretical Extension of the Technology Acceptance Model : Four Longitudinal Field Studies'. In: *Management science* 46.2 (Feb. 2000), pp. 186–204. doi: 10.1287/mnsc.46.2.186.11926.
- [95] A. J. van Altena, R. Spijker and S. D. Olabarriaga. *Supplementary Material For Paper: Usage of Automation Tools in Systematic Reviews*. 2018. URL: <https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fjrsm.1335&file=supplementary-revision-version-21-09-18.pdf>.
- [96] SurveyMonkey Inc. *SurveyMonkey*. 2018. URL: <http://www.surveymonkey.com>.
- [97] M. Y. Chuttur. 'Overview of the technology acceptance model: Origins, developments and future directions'. In: *Working Papers on Information Systems* 9.37 (2009), pp. 9–37. ISSN: 1535-6078. doi: 10.1021/jf001443p. PMID: 11453748.
- [98] A. J. van Altena. *Release Submission - AMCeScience/survey-system*. 2018. URL: <https://github.com/AMCeScience/survey-system/releases/tag/Submission>.
- [99] J. Brooke. 'SUS-A quick and dirty usability scale'. In: ed. by P. W. Jordan et al. Vol. 189. 194. London: Taylor and Francis, 1996, pp. 4–7. ISBN: 9780429157011.
- [100] J. Brooke. 'SUS: A Retrospective'. In: *J. Usability Studies* 8.2 (Feb. 2013), pp. 29–40. ISSN: 1931-3357. doi: 10.5555/2817912.2817913.
- [101] A. Bangor, P. Kortum and J. Miller. 'Determining what individual SUS scores mean: Adding an adjective rating scale'. In: *Journal of usability studies* 4.3 (2009), pp. 114–123. doi: 10.5555/2835587.2835589.
- [102] *2017 Issue 1 | Cochrane Library*. 2017. URL: <http://www.cochranelibrary.com/cochrane-database-of-systematic-reviews/table-of-contents/2017/issue1/> (visited on 24/03/2017).
- [103] F. E. Harrell Jr, with contributions from Charles Dupont and many others. *Hmisc: Harrell Miscellaneous*. R package version 4.0-3. 2017. URL: <https://CRAN.R-project.org/package=Hmisc>.
- [104] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: <http://www.R-project.org/>.
- [105] J. Fox. *polycor: Polychoric and Polyserial Correlations*. R package version 0.7-9. 2016. URL: <https://CRAN.R-project.org/package=polycor>.
- [106] J. Thomas. 'Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation?' In: *OA Evidence-Based Medicine* 1.2 (Oct. 2013), pp. 1–6. doi: 10.13172/2053-2636-1-2-1109.

- [107] *The Cochrane Collaboration: Cochrane Handbook for Systematic Reviews of Interventions*. 51st edition. 2011. URL: <http://www.cochrane.org/training/cochrane-handbook>.
- [108] I. E. Allen and I. Olkin. 'Estimating Time to Conduct a Meta-analysis From Number of Citations Retrieved'. In: *JAMA* 282.7 (Aug. 1999), pp. 634–635. ISSN: 0098-7484. DOI: 10.1001/jama.282.7.634.
- [109] D. A. Korevaar et al. 'Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD'. In: *BMJ Evidence-Based Medicine* 19.2 (2014), pp. 47–54. ISSN: 1356-5524. DOI: 10.1136/eb-2013-101637.
- [110] *Diagnostic Test Accuracy Working Group Handbook for DTA reviews*. 2013. URL: <http://srdta.cochrane.org/handbook-dta-reviews> (visited on 03/11/2019).
- [111] H. Petersen et al. 'Increased workload for systematic review literature searches of diagnostic tests compared with treatments: Challenges and opportunities'. In: *JMIR medical informatics* 2.1 (2014), e11. DOI: 10.2196/medinform.3037.
- [112] J. Liu, P. Timsina and O. El-Gayar. 'A comparative analysis of semi-supervised learning: the case of article selection for medical systematic reviews'. In: *Information Systems Frontiers* 20.2 (2018), pp. 195–207. DOI: 10.1007/s10796-016-9724-0.
- [113] A. M. Cohen et al. 'Reducing workload in systematic review preparation using automated citation classification'. In: *Journal of the American Medical Informatics Association* 13.2 (2006), pp. 206–219. DOI: 10.1197/jamia.M1929.
- [114] M. Miwa et al. 'Reducing systematic review workload through certainty-based screening'. In: *Journal of biomedical informatics* 51 (2014), pp. 242–253. DOI: 10.1016/j.jbi.2014.06.005.
- [115] K. Weiss, T. M. Khoshgoftaar and D. Wang. 'A survey on transfer learning'. In: *Journal of Big Data* 3.9 (2016). DOI: 10.1186/s40537-016-0043-6.
- [116] E. Kanoulas et al. 'CLEF 2017 technologically assisted reviews in empirical medicine overview'. In: *CEUR Workshop Proceedings*. Vol. 1866. 2017, pp. 1–29.
- [117] S. Bird, E. Klein and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009.
- [118] A. J. van Altena. *AMCeScience/feature-miner*. 2020. URL: <https://github.com/AMCeScience/feature-miner-pub/>.
- [119] A. J. van Altena. *Review metadata*. Mar. 2019. DOI: 10.6084/m9.figshare.7804094.
- [120] M. D. F. McInnes et al. 'Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement'. In: *JAMA* 319.4 (Jan. 2018), pp. 388–396. ISSN: 0098-7484. DOI: 10.1001/jama.2017.19163.
- [121] *ICD-10 Version:2010*. URL: <https://icd.who.int/browse10/2010/en> (visited on 03/12/2018).
- [122] F. Pedregosa et al. 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [123] A. Huang. 'Similarity measures for text document clustering'. In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand. 2008, pp. 49–56.
- [124] W. H. Gomaa and A. A. Fahmy. 'A survey of text similarity approaches'. In: *International Journal of Computer Applications* 68.13 (2013), pp. 13–18. DOI: 10.5120/11638-7118.
- [125] P. Virtanen et al. 'SciPy 1.0—Fundamental Algorithms for Scientific Computing in Python'. In: *arXiv e-prints*, arXiv:1907.10121 (July 2019), arXiv:1907.10121. arXiv: 1907.10121 [cs.LG].

- [126] S. Seabold and J. Perktold. 'statsmodels: Econometric and statistical modeling with python'. In: *9th Python in Science Conference*. 2010. doi: 10.25080/Majora-92bf1922-011.
- [127] T. pandas development team. *pandas-dev/pandas: Pandas*. Version 1.0.3. Feb. 2020. doi: 10.5281/zenodo.3509134.
- [128] B. K. Olorisade, P. Brereton and P. Andras. 'Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist'. In: *Journal of Biomedical Informatics* 73 (2017), pp. 1–13. issn: 1532-0464. doi: 10.1016/j.jbi.2017.07.010.
- [129] A. M. Cohen. 'Optimizing feature representation for automated systematic review work prioritization'. In: *AMIA annual symposium proceedings*. Vol. 2008. American Medical Informatics Association. 2008, p. 121. PMID: 18998798.
- [130] R. C. Moore and W. Lewis. 'Intelligent Selection of Language Model Training Data'. In: *Proceedings of the ACL 2010 Conference Short Papers*. ACLShort '10. Uppsala, Sweden: Association for Computational Linguistics, 2010, pp. 220–224. doi: 10.5555/1858842.1858883.
- [131] M. Kubat, S. Matwin et al. 'Addressing the curse of imbalanced training sets: one-sided selection'. In: *ICML*. Vol. 97. Nashville, USA. 1997, pp. 179–186.
- [132] J. G. Adeva et al. 'Automatic text classification to support systematic reviews in medicine'. In: *Expert Systems with Applications* 41.4, Part 1 (2014), pp. 1498–1508. issn: 0957-4174. doi: 10.1016/j.eswa.2013.08.047.
- [133] F. Pedregosa et al. *sklearn.ensemble.RandomForestClassifier*. 2019. URL: <https://scikit-learn.org/0.20/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.
- [134] W. Koehrsen. *Hyperparameter Tuning the Random Forest in Python*. 2018. URL: <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74> (visited on 03/11/2019).
- [135] B. C. Wallace et al. 'Active Learning for Biomedical Citation Screening'. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '10. Washington, DC, USA: ACM, 2010, pp. 173–182. isbn: 978-1-4503-0055-1. doi: 10.1145/1835804.1835829.
- [136] B. C. Wallace et al. 'Semi-automated screening of biomedical citations for systematic reviews'. In: *BMC Bioinformatics* 11.1 (Jan. 2010), p. 55. issn: 1471-2105. doi: 10.1186/1471-2105-11-55.
- [137] M. Miwa et al. 'Reducing systematic review workload through certainty-based screening'. In: *Journal of Biomedical Informatics* 51 (2014), pp. 242–253. issn: 1532-0464. doi: 10.1016/j.jbi.2014.06.005.
- [138] K. Hashimoto et al. 'Topic detection using paragraph vectors to support active learning in systematic reviews'. In: *Journal of Biomedical Informatics* 62 (2016), pp. 59–65. issn: 1532-0464. doi: 10.1016/j.jbi.2016.06.001.
- [139] A. van Altena et al. 'Training sample selection: impact on screening automation in diagnostic test accuracy reviews'. In: *Research Synthesis Methods* (2021). doi: 10.1002/jrsm.1518.
- [140] J. Thomas, J. McNaught and S. Ananiadou. 'Applications of text mining within systematic reviews'. In: *Research Synthesis Methods* 2.1 (2011), pp. 1–14. doi: 10.1002/jrsm.27.
- [141] S. Jonnalagadda and D. Petitti. 'A new iterative method to reduce workload in the systematic review process'. In: *International journal of computational biology and drug design* 6 (2013), p. 5. doi: 10.1504/IJCBDD.2013.052198. PMID: 23428470.

- [142] A. Sestino and A. De Mauro. 'Leveraging Artificial Intelligence in Business: Implications, Applications and Methods'. In: *Technology Analysis & Strategic Management* (2021), pp. 1–14. doi: 10.1080/09537325.2021.1883583.
- [143] A. De Mauro, M. Greco and M. Grimaldi. 'Understanding Big Data through a systematic literature review: The ITMI model'. In: *International Journal of Information Technology & Decision Making* 18.04 (2019), pp. 1433–1461. doi: 10.1142/S0219622019300040.
- [144] M. Favaretto et al. 'What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade'. In: *PloS one* 15 (Feb. 2020), pp. 1–20. doi: 10.1371/journal.pone.0228987.
- [145] A. Parlina, K. Ramli and H. Murfi. 'Theme mapping and bibliometrics analysis of one decade of big data research in the scopus database'. In: *Information* 11.2 (2020), p. 69. doi: 10.3390/info11020069.
- [146] E. Mohammadi and A. Karami. 'Exploring research trends in big data across disciplines: A text mining analysis'. In: *Journal of Information Science* (2020). doi: 10.1177/0165551520932855.
- [147] A. Gates et al. 'Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools'. In: *Systematic reviews* 8.1 (2019), pp. 1–11. doi: 10.1186/s13643-019-1222-2.
- [148] J. Clark et al. 'A full systematic review was completed in 2 weeks using automation tools: a case study'. In: *Journal of Clinical Epidemiology* 121 (2020), pp. 81–90. issn: 0895-4356. doi: 10.1016/j.jclinepi.2020.01.008.
- [149] R. Borah et al. 'Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry'. In: *BMJ open* 7.2 (2017). doi: 10.1136/bmjopen-2016-012545.
- [150] J. Clark et al. 'The Impact of Systematic Review Automation Tools on Methodological Quality and Time Taken to Complete Systematic Review Tasks: Case Study'. In: *JMIR Medical Education* 7.2 (2021), e24418. doi: 10.2196/24418.
- [151] A. M. Scott et al. 'Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey'. In: *Journal of Clinical Epidemiology* 138 (2021), pp. 80–94. issn: 0895-4356. doi: 10.1016/j.jclinepi.2021.06.030.
- [152] V. R. Walker et al. 'Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr'. In: *Environment International* 159 (2022), p. 107025. issn: 0160-4120. doi: 10.1016/j.envint.2021.107025.
- [153] A. Arno et al. 'The views of health guideline developers on the use of automation in health evidence synthesis'. In: *Systematic Reviews* 10.1 (2021), pp. 1–10. doi: 10.1186/s13643-020-01569-2.
- [154] V. Unnikrishnan et al. 'Love thy Neighbours: A Framework for Error-Driven Discovery of Useful Neighbourhoods for One-Step Forecasts on EMA data'. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. 2021, pp. 295–300. doi: 10.1109/CBMS52027.2021.00080.
- [155] W. Abdelkader et al. 'Machine Learning Approaches to Retrieve High-Quality, Clinically Relevant Evidence From the Biomedical Literature: Systematic Review'. In: *JMIR medical informatics* 9.9 (2021), e30401. doi: 10.2196/30401.

Summary

Big Data is a term that has been around for many years and it is often understood as the use and manipulation of large volumes of data. Over the years more aspects of Big Data have been recognised by researchers and institutions, such as its velocity, variety, and value. Nowadays, many definitions exist and researchers have wildly different understandings of Big Data. We believe that, without an unambiguous definition, communication is hampered, resulting in missed opportunities for both the developers as well as users of Big Data technologies.

A group of researchers that may benefit greatly from applying Big Data technologies to partially automate their work are systematic reviewers. The nature of their work involves reading, dissecting, and selecting considerable numbers of scientific papers. Over the years many tools have been developed to support reviewers. However, it is unclear how often these are used and how well they work.

In this thesis we start with an exploration of a common understanding of the term Big Data. The focus then shifts to systematic reviewers and their use of tools to deal with Big Data. Lastly, we propose a new method that may improve these tools. In this thesis we aim to: uncover a common understanding of Big Data in the (bio)medical research field; aid in improving the adoption of automation tools among systematic reviewers; and, contribute to the effectiveness of automation tools. Our work is divided into three parts, each of which has been summarised below.

Understanding Big Data

While Big Data is a key component of many (bio)medical studies, it has yet to receive a formal definition. Chapter 2 pursued a better understanding of the topics covered by the term Big Data through a data-driven systematic approach using text analysis of scientific (bio)medical literature.

Our study built upon findings of previous qualitative research that analysed fifteen definitions and identified four key Big Data themes. We have revisited these and other definitions of Big Data, and consolidated them into eight additional themes, resulting in a total of twelve themes. We collected manual annotations of Big Data themes in (bio)medical literature and showed a strong presence of the original four themes. We noted that these themes are defined broadly, thereby capturing many of the other eight themes in them. These results indicated that, at that time, the understanding of Big Data was mostly captured in broad terms.

Since our study, others have researched the definition of Big Data. All of them gravitated to more or less the same conclusion: it remains difficult to draw a well-defined boundary around the concepts of Big Data. One study did a series of interviews with researchers. They noted that out of 39 participants only one could pinpoint a formal definition of Big Data. Nevertheless, they found a commonality in the responses of participants. In our opinion the definition of Big Data is less important than attaching the term to whatever a researcher thinks is appropriate. This follows from our belief that, to improve the spread and uptake of Big Data solutions, the solutions should be findable. Even if two researchers were to use the term Big Data in a, often subtly, different way they might be interested in the same thing.

Skeptics argue that Big Data is just a hype term, representing nothing new or at best just an extension of what has been done for decades. This was also the case in our surroundings, which motivated us to assess the value of the term Big Data when used by researchers in their publications, this research is described in Chapter 3. We measured the value of the term Big Data by feeding a large number of (bio)medical research papers into a machine learning method. These research papers belonged to two groups: those that contained the term Big Data and those that did not. We found that the machine learning method had a high performance in distinguishing characteristics of papers in each of the groups. Moreover, analysing the most commonly used words in the Big Data papers, we found that the use of the term Big Data within a publication seemed to indicate a distinct type of research in the biomedical field. We concluded that value can be attributed to the term Big Data when used in a publication. This conclusion strengthened our belief that it is useful to attach the term wherever a researcher thinks it is applicable, thereby making the work findable for others.

Solutions to a data deluge

Systematic reviews are a cornerstone of evidence-informed decision making, they bring together the findings from multiple studies in a structured, reliable, and preferably unbiased way. However, the process is mostly performed manually and very time-consuming. With the rapid expansion of scientific information produced and research questions to be addressed, there is a growing workload on reviewers, making the current practice unsustainable without the aid of automation tools.

In Chapter 4 we investigated why the adoption of automation tools among systematic reviewers seemed to be lagging and we identified potential barriers and facilitators for their adoption. To this end we deployed surveys and found that automation tools were not widely used among the participants. The results provided evidence and confirmed the conclusions and recommendations of previous work, which was based on expert opinion. When tools were used, participants mostly learn about them through their colleagues, peers, or organisation. Tools were often chosen on the basis of user experience, either by own experience or from colleagues or peers. Lastly, licensing, steep learning curve, lack of support, and mismatch to workflow were often reported by participants as relevant barriers.

After our study others have investigated the adoption of automation tools. One study executed a systematic review which heavily relied on automation tools in every step of the process. They found that a review may be significantly sped up by the use of tools. Furthermore, many concluded that the perceived efficiency of a tool was very important for its adoption. In other words, if the user believes that a tool will benefit them they will start using it.

The results above imply that the main blocking factor for adoption is not the actual performance of automation tools, but rather by the users' sentiment towards the tool. Sentiment is influenced by many factors, opinion of peers and the ease of access were most often indicated by systematic reviewers to be important in their choice of tools. This highlights the importance that organisations and best practices in a field can have for the adoption of automation tools for systematic reviews. In the field of systematic reviews guideline developers decide the rules that systematic reviewers should adhere to for scientifically solid work. Therefore, we argue that this should be the first group to be convinced of the reliability of automation tools.

Applying solutions in practice

In Chapter 5 we introduced an approach to improve the performance of systematic review automation tools. A systematic review consists of retrieval, appraisal, and synthesis of evidence. During the appraisal phase, the relevancy of scientific papers found during the

retrieval step is determined. In many reviews the number of studies to appraise is very large. Therefore, in our study we focussed on a subset of automation tools that support the appraisal process of a systematic review by using text mining to predict the relevancy of each study that needs to be appraised. Using the predictions, the reading order could be adjusted so that the reviewer sees the studies that are most likely to be relevant first.

To predict how relevant a study is a prediction method first needed to learn from previous systematic reviews where the relevant studies were appraised by researchers, a process called *training*. The algorithm learned how to make correct predictions from each given sample. A rule of thumb in machine learning is to use all available data, because more examples used during training will usually yield better predictions. However, each systematic review investigated a unique research question. The question arose whether using all available data actually results in the best performing algorithm. After all, when considering which data to use for a review about Alzheimer's disease, another review about Alzheimer might provide a better training set than a review about cancer. Therefore, excluding the review about cancer from the training data might improve the algorithm's performance because the remaining training data is less diluted. Our proposed approach chose which data to use during training based on a metric that quantifies the similarity between reviews based on text features. This approach led to less data for training, but the selection had a high similarity to the data from the review that we were predicting, potentially improving the performance of the automation tool.

In Chapter 5 we introduced our approach. We selected only the most similar data and then trained and tested our prediction algorithm. In contrast to our hypothesis, we found that algorithms trained on more data performed better. However, algorithms trained for reviews that had a research subject similar to other reviews in the dataset got better results. We concluded that our proposed approach had the potential to improve prediction performance in those cases.

To test our conclusion, in Chapter 6 we applied our approach in an active learning setting. In this setting the tool made a few predictions and presented them to the user. The user appraised these predictions and gave feedback to the tool. The tool assessed the feedback and made new predictions. This process was repeated until a stopping criterion was reached (e.g., all relevant items were found). Because the information provided by the user was the most relevant to the review at hand it should weigh heavily in the prediction process. Reducing the number of samples in the initial training data should therefore affect the machine learning performance. In Chapter 6 we found that the performance of an active learning tool trained using less data was significantly higher compared to a tool using all available data. Therefore, we concluded that our approach improved prediction performance when using active learning. Moreover, another benefit of this approach was that the reduction in data also reduced the computational effort of training the machine learning method. This meant that training was faster, leading to less downtime for users and a potential cost reduction for IT infrastructure.

Conclusion

In this thesis we found a definition for the term Big Data and that the term holds value when used in (bio)medical science literature. We then described barriers and facilitators for the adoption of automation tools among systematic reviewers, who face Big Data challenges. Lastly, we introduced a new method that boosts the performance of systematic review automation tools while reducing the computational cost.

Samenvatting

Big Data is een term die al vele jaren bestaat en die vaak begrepen wordt als het gebruik en de manipulatie van grote hoeveelheden data. In de loop der jaren zijn er meer aspecten van Big Data erkend door onderzoekers en instellingen, zoals de snelheid, variëteit en waarde ervan. Tegenwoordig zijn er veel definities en onderzoekers hebben verschillende opvattingen over Big Data. Wij zijn van mening dat, zonder eenduidige definitie, de communicatie belemmerd wordt, met als gevolg dat er kansen blijven liggen voor zowel de ontwikkelaars als gebruikers van Big Data-technologieën.

Een groep onderzoekers die veel baat kan hebben bij het toepassen van Big Data-technologieën om hun werk gedeeltelijk te automatiseren, zijn *systematic reviewers*. De aard van hun werk omvat het lezen, begrijpen en selecteren van behoorlijke aantallen wetenschappelijke artikelen. In de loop der jaren zijn er veel tools ontwikkeld om reviewers te ondersteunen. Het is echter onduidelijk hoe vaak deze worden gebruikt en hoe goed ze werken.

In dit proefschrift beginnen we met een verkenning van een gemeenschappelijke definitie van de term Big Data. Daarna verschuift de focus verschuift naar systematic reviewers en hun gebruik van tools om met Big Data om te gaan. Ten slotte stellen we een nieuwe methode voor die deze tools kan verbeteren. In dit proefschrift willen we: een gemeenschappelijke definitie van Big Data in het (bio)medische onderzoeksveld blootleggen; bijdragen aan het gebruik van automatiseringstools onder systematic reviewers en aan de effectiviteit van deze tools. Ons werk is opgedeeld in drie delen, die hieronder zijn samengevat.

Begrijpen van Big Data

Hoewel Big Data een belangrijk onderdeel is van veel (bio)medische onderzoeken, heeft het nog geen formele definitie. Hoofdstuk 2 streefde naar een beter begrip van de onderwerpen die onder de term Big Data vallen. Dit deden we met een datagedreven systematische aanpak die werkte door tekstanalyse toe te passen op wetenschappelijke (bio)medische literatuur.

Ons onderzoek bouwde voort op bevindingen van eerder kwalitatief onderzoek dat vijftien definities analyseerde en vier belangrijke Big Data-thema's identificeerde. Deze en andere definities van Big Data hebben we onder de loep genomen en samengevoegd tot acht extra thema's, wat in totaal twaalf thema's opleverde. Ook verzamelden we handmatige annotaties van Big Data-thema's in (bio)medische literatuur en toonden een sterke aanwezigheid van de oorspronkelijke vier thema's. We merkten op dat deze thema's breed worden gedefinieerd, waardoor veel van de andere acht thema's erin zijn opgenomen. Deze resultaten gaven aan dat het begrip van Big Data in die tijd vooral in brede bewoordingen was vastgelegd.

Sinds ons onderzoek hebben anderen onderzoek gedaan naar de definitie van Big Data. Allemaal kwamen ze tot min of meer dezelfde conclusie: het blijft moeilijk om een goed gedefinieerde grens te trekken rond de concepten van Big Data. Een studie deed een reeks interviews met onderzoekers. Ze merkten op dat van de 39 deelnemers er maar één een formele definitie van Big Data kon geven. Maar toch vonden ze overeenkomsten in de antwoorden van de deelnemers. Naar onze mening blijkt hier uit dat de definitie van Big

Data minder belangrijk is dan de term toe te passen op alles waar een onderzoeker het van toepassing acht. Dit vloeit voort uit ons geloof dat, om de verspreiding en acceptatie van Big Data-oplossingen te verbeteren, de oplossingen vindbaar moeten zijn. Zelfs als twee onderzoekers de term Big Data op een, vaak subtiel, verschillende manier zouden gebruiken, zouden ze in hetzelfde geïnteresseerd kunnen zijn.

Sceptici beweren dat Big Data slechts een hype-term is, niets nieuws of in het beste geval slechts een uitbreiding van wat al tientallen jaren wordt gedaan. Dit was ook het geval in onze omgeving, wat ons motiveerde om de waarde van de term Big Data aan te tonen. In dit onderzoek, beschreven in Hoofdstuk 3, keken we naar het gebruik van de term Big Data in wetenschappelijke publicaties. We hebben de waarde van de term Big Data gemeten door een groot aantal (bio)medische publicaties aan een machine learning-methode te geven. De publicaties behoorden tot twee groepen: degenen die de term Big Data bevatten en degenen die dat niet deden. We ontdekten dat de machine learning-methode goed presteerde bij het onderscheiden van kenmerken van de publicaties in de groepen. Bovendien ontdekten we bij het analyseren van de meest gebruikte woorden in de Big Data-papers dat het gebruik van de term Big Data in een publicatie leek te wijzen op een ander type onderzoek op biomedisch gebied. We concludeerden dat waarde kan worden toegekend aan de term Big Data bij gebruik in een publicatie. Deze conclusie versterkte onze overtuiging dat het nuttig is om de term toe te passen op alles waar een onderzoeker het van toepassing acht, waardoor het werk voor anderen vindbaar wordt.

Oplossingen voor een stortvloed aan data

Systematic reviews zijn een hoeksteen van door onderzoek gedreven besluitvorming (evidence-informed decision making). Reviewers brengen de bevindingen van meerdere onderzoeken op een gestructureerde, betrouwbare en onpartijdige manier samen. Het proces wordt echter meestal handmatig uitgevoerd en kost veel tijd. Door de snelle groei van wetenschappelijke informatie die wordt geproduceerd en onderzoeksvragen die moeten worden beantwoord, is er een groeiende werkdruk voor reviewers. Hierdoor is de huidige manier van werken onhoudbaar zonder de hulp van automatiseringstools.

In Hoofdstuk 4 onderzochten we waarom het gebruik van automatiseringstools onder systematic reviewers achterbleef en we identificeerden mogelijke barrières en ondersteunende factoren. We hebben enquêtes ingezet en ontdekten dat automatiseringstools niet veel werden gebruikt onder de deelnemers. De resultaten leverden bewijs voor de conclusies en aanbevelingen van eerder werk dat was gebaseerd op de mening van deskundigen. Wanneer tools werden gebruikt, leren deelnemers deze meestal kennen via hun collega's, collega's of de organisatie. Tools werden vaak gekozen op basis van gebruikerservaring, hetzij door eigen ervaring, hetzij van collega's. Ten slotte werden licenties, een steile leercurve, gebrek aan ondersteuning en het niet kunnen inpassen in de gebruikelijke werkwijze vaak door deelnemers als relevante belemmeringen gemeld.

Na ons onderzoek hebben anderen het gebruik van automatiseringstools onderzocht. Eén studie voerde een systematic review uit die in zo veel mogelijk stappen van het proces een automatiseringstool inzette. Ze ontdekten dat een review aanzienlijk kan worden versneld door het gebruik van tools. Bovendien concludeerden veel onderzoeken dat het *idee* dat een tool helpt om een review efficiënter te laten verlopen erg belangrijk was voor de acceptatie ervan. Met andere woorden, als de gebruiker denkt dat een tool hem zal helpen, dan zal hij deze gaan gebruiken.

De bovenstaande resultaten impliceren dat de belangrijkste barrières niet de daadwerkelijke prestatie van een automatiseringstools is, maar eerder het sentiment van de gebruikers ten opzichte van de tool. Sentiment wordt beïnvloed door vele factoren. De mening van de omgeving en een gemakkelijke toegang tot de tool werden door systematic reviewers het vaakst genoemd. Dit benadrukt het belang dat organisaties en best practices in een veld kunnen hebben voor de adoptie van automatiseringstools. Op het gebied van systematic reviews bepalen richtlijnontwikkelaars de regels waaraan systematic reviewers

zich moeten houden. Daarom stellen we dat dit de eerste groep moet zijn die overtuigd is van de betrouwbaarheid van automatiseringstools.

Toepassen van oplossingen in de praktijk

In Hoofdstuk 5 hebben we een aanpak geïntroduceerd om prestaties te verbeteren van automatiseringstools die gebruikt kunnen worden voor systematic review. Een systematic review bestaat uit het zoeken, beoordelen en samenvoegen van bewijs. Tijdens de beoordelingsfase wordt de relevantie van wetenschappelijke artikelen bepaald die tijdens de zoek-stap zijn gevonden. In veel reviews is het aantal te beoordelen artikelen erg groot. Daarom hebben we ons gericht op een groep van automatiseringstools die het beoordelingsproces van een systematic review ondersteunen. Dit doen we door tekstanalyse te gebruiken om de relevantie te voorspellen van elk onderzoek dat moet worden beoordeeld. Met behulp van de voorspellingen kan de leesvolgorde worden aangepast, zodat de reviewer de artikelen die waarschijnlijk relevant zijn als eerste ziet.

Om te voorspellen hoe relevant een onderzoek is leert een algoritme van eerdere systematic reviews waarin de relevante artikelen door onderzoekers zijn aangeduid, een proces wat *trainen* genoemd wordt. Een vuistregel bij machine learning is om alle beschikbare data te gebruiken, omdat meer voorbeelden tijdens de training meestal betere voorspellingen opleveren. Elk systematic review onderzoekt echter een unieke onderzoeksvraag. Hierdoor vroegen we ons af of het gebruik van alle beschikbare data ook daadwerkelijk resulteert in het best presterende algoritme. Immers, bij het kiezen welke data gebruikt moeten worden om een review over de ziekte van Alzheimer te voorspellen, zou een andere review over Alzheimer een betere trainingsset kunnen bieden dan een review over kanker. Daarom zou het uitsluiten van de review over kanker uit de trainingsset de prestaties van het algoritme kunnen verbeteren, omdat de resterende data niet 'verdund' worden. Onze voorgestelde aanpak koos de trainingsset op basis van een statistiek die de gelijkenis tussen reviews kwantificeert op basis van tekstenkenmerken. Deze aanpak leidde tot kleinere trainingssets, waarbij de selectie een hoge gelijkenis had met de gegevens uit de reviews die we voorspelden, wat mogelijk de prestaties van de automatiseringstool verbeterde.

In Hoofdstuk 5 hebben we onze aanpak geïntroduceerd. We hebben alleen de meest vergelijkbare gegevens geselecteerd en vervolgens ons voorspellingsalgoritme getraind en getest. In tegenstelling tot onze hypothese, vonden we dat algoritmen die waren getraind op meer data beter presteerden. Echter, algoritmen die waren getraind voor reviews met een onderzoeksvraag die vergelijkbaar was met andere reviews in de dataset, kregen betere resultaten. We concludeerden dat onze voorgestelde aanpak het potentieel had om de voorspellingsprestaties in die gevallen te verbeteren.

Om onze conclusie te testen, hebben we in Hoofdstuk 6 onze aanpak toegepast met een active learning algoritme. Hierbij deed de tool enkele voorspellingen en presenteerde deze aan de gebruiker. De gebruiker beoordeelde deze voorspellingen en gaf feedback aan de tool. De tool beoordeelde de feedback en deed nieuwe voorspellingen. Dit proces werd herhaald totdat een stopcriterium was bereikt (bijvoorbeeld: alle relevante items werden gevonden). Omdat de door de gebruiker verstrekte informatie het meest relevant was voor de betreffende review, moet deze zwaar wegen in het voorspellingsproces. Het verminderen van het aantal voorbeelden in de initiële trainingsgegevens zou daarom van invloed moeten zijn op de prestaties van het algoritme. In Hoofdstuk 6 vonden we dat de prestatie van een active learning tool die met minder gegevens was getraind, significant hoger was in vergelijking met een tool die alle beschikbare gegevens gebruikte. Daarom concludeerden we dat onze aanpak de prestaties verbeterde bij het gebruik van active learning. Bovendien was een aanvullend voordeel van deze aanpak dat de vermindering van gegevens ook de computerkracht die nodig was voor het trainen van de machine learning-methode verminderde. Dit betekende dat de training sneller klaar was, wat kan

leiden tot een snellere doorlooptijd voor de gebruikers van de tool en dat er mogelijk bespaard kon worden op de IT-infrastructuur.

Conclusie

In dit proefschrift hebben we een definitie gevonden voor de term Big Data en zagen we dat de term waarde heeft bij gebruik in (bio)medische wetenschappelijke literatuur. Vervolgens hebben we barrières en ondersteunende factoren beschreven voor het gebruik van automatiseringstools onder systematic reviewers, die worden geconfronteerd met Big Data-uitdagingen. Ten slotte hebben we een nieuwe methode geïntroduceerd die de prestaties van tools voor automatisering van systematische reviews verbetert en tegelijkertijd de benodigde computerkracht verlaagt.

List of acronyms

UvA University of Amsterdam

IT Information Technology

NIST National Institute of Standards and Technology

TM Topic Modelling

DOI Digital Object Identifier

LDA Latent Dirichlet Allocation

V's Big Data aspects

For example: Volume, Velocity, Variety, Veracity, Value, and Variability.

BD Big Data

NBD non-Big Data

PMC PubMed Central

ROC Receiver Operating Characteristic

AUC Area Under the Curve

FOR False Omission Rate

FDR False Discovery Rate

TAM Technology Acceptance Model 2

DTA Diagnostic Test Accuracy

TF Term Frequency

WSS Work Saved over Sampling

WSS@95 Work Saved over Sampling @ 95%

TF-IDF Term Frequency Inverse Document Frequency

SIMILAR selected data

Name for approach used in data selection studies. This approach used data selected with a similarity metric.

ALL all data

Name for approach used in data selection studies. This approach used all the data available.

RANDOM random data

Name for approach used in data selection studies. This approach used a random selection of the available data.

IB Inverse Burden

BIC Bayesian Information Criterion

AIC Akaike Information Criterion

HTML Hypertext Markup Language

A standardised system for describing layout and styling in (web)documents.

SUS System Usability Scale

ICD-10 International Classification of Diseases, 10th revision

SVM Support Vector Machine

Portfolio

PhD training	Year	ECTS
General Courses		
AMC World of Science	2016	0.7
Didactical Skills	2017	0.6
Scientific Writing	2017	1.5
Project Management	2017	0.6
Specific courses		
Machine Learning for bioinformatics and systems biology, BioSB	2015	1.2
e-Science	2016	1.2
Machine Learning, Stanford via Coursera	2016	2.2
Entrepreneurship	2017	0.6
Presentations		
Oral "Text mining the definition of Big Data", KEBB Seminar (Amsterdam)	2016	0.5
Oral "Towards a better understanding of Big Data in biomedical research", KEBB Seminar (Amsterdam)	2017	0.5
Oral "Usage of Automation Tools in Systematic Reviews", KEBB Seminar (Amsterdam)	2018	0.5
Seminars, workshops and master classes		
Coffee & Data #3 (Amsterdam)	2015	0.2
Anatomische les (Amsterdam)	2015, 2017	0.1
e-Health? Dat zal je leren! (Amsterdam)	2017	0.2
Conferences		
ICTOpen 2016 (Amersfoort)	2016	0.5
4th National eScience Symposium: Science in a Digital World (Amsterdam)	2016	0.25
7th International Digital Health conference (London, UK)	2017	1
Conference and Labs of the Evaluation Forum (CLEF) 2017 (Dublin, Ireland)	2017	0.25
IEEE International Conference on Big Data 2017 (Boston, USA)	2017	1
Other		
Weekly reading club	2016-2018	1
Weekly KEBB seminar	2016-2018	2

Teaching

Tutoring, mentoring

Tutor for the HPC cloud practicum, AMC graduate school	2017 - 2018	0.2
Tutor for the pattern recognition practicum, BioSB	2017	0.1

Lecturing

Guest lecturer for e-Science course, AMC graduate school	2016	0.1
Guest lecturer No-SQL, bachelor Medische informatiekunde	2016 - 2018	0.3

Supervising

1-month master Medical Informatics internship, Noman: Collecting PDF, layout aware text extraction and annotating sections into IMRaD format	2018	0.5
8-month master Medical informatics internship, Shuxin: Active learning for workload reduction of automated screening in systematic reviews	2018	2

List of publications

Publications in this thesis

- 1 Improving active learning performance through training sample selection 2021
A.J. van Altena, A.H. Zwinderman, S.D. Olabarriaga
In Submission.
- 2 Training sample selection: impact on screening automation in diagnostic test accuracy reviews 2021
A.J. van Altena, R. Spijker, M.M.G. Leeflang, S.D. Olabarriaga
Research Synthesis Methods. 2021;12(6):831-841. DOI: 10.1002/jrsm.1518
- 3 Usage of automation tools in systematic reviews 2019
A.J. van Altena, R. Spijker, S.D. Olabarriaga
Research Synthesis Methods. 2019;10(1):72-82. DOI: 10.1002/jrsm.1335
- 4 Usage of the term Big Data in biomedical publications: a text mining approach 2019
A.J. van Altena, P.D. Moerland, A.H. Zwinderman, S.D. Olabarriaga
Big Data and Cognitive Computing. 2019;3(1),13. DOI: 10.3390/bdcc3010013
- 5 Understanding Big Data themes from scientific biomedical literature through topic modeling 2016
A.J. van Altena, P.D. Moerland, A.H. Zwinderman, S.D. Olabarriaga
Journal of Big Data. 2016;3(1):1-21. DOI: 10.1186/s40537-016-0057-0

Other publications

- 6 Effect of parental and ART treatment characteristics on perinatal outcomes 2021
M. Pontesilli, M.H. Hof, A.C.J. Ravelli, **A.J. van Altena**, A.T. Soufan, B.W. Mol, E.H. Kosteljik, E. Slappendel, D. Consten, A.E.P. Cantineau, and others
Human Reproduction 2021;36(6):1640-1665. DOI: 10.1093/humrep/deab008
- 7 Analysis of the term 'big data': Usage in biomedical publications 2017
A.J. van Altena, P.D. Moerland, A.H. Zwinderman, S.D. Olabarriaga
2017 *IEEE International Conference on Big Data* 2017 (p. 1253-1258). DOI: 10.1109/Big-Data.2017.8258051
- 8 Predicting publication inclusion for diagnostic accuracy test reviews using random forests and topic modelling 2017
A.J. van Altena, S.D. Olabarriaga
Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum 2017 (p. 1-9).
- 9 (Bio) medical Publications in the Age of Big Data: Yes, They Are Different 2017
A.J. van Altena, S.D. Olabarriaga
Proceedings of the 2017 International Conference on Digital Health 2017 (p. 221-222). DOI: 10.1145/3079452.3079474

-
- 10 A neuroscience gateway for handling and processing population imaging studies 2015
M. Caan, J. Teeuw, S. Shahand, M.M. Jaghoori, J. Huguet, **A.J. van Altena**, and S.D. Olabarriaga
*Proceedings of the 1st MICCAI 2015 Workshop on Management and Processing of images
for Population Imaging (MICCAI MAPPING 2015)* 2015 (p. 15-22).
- 11 A multi-infrastructure gateway for virtual drug screening 2015
M.M. Jaghoori, **A.J. van Altena**, B. Bleijlevens, S. Ramezani, J.L. Font, S.D. Olabarriaga
Concurrency and Computation: Practice and Experience 2015;27(16):4478-4490. DOI:
10.1002/cpe.3498
- 12 A grid-enabled virtual screening gateway 2014
M.M. Jaghoori, **A.J. Van Altena**, B. Bleijlevens, S.D. Olabarriaga
2014 6th International Workshop on Science Gateways 2014 (p. 24-29). IEEE. DOI:
10.1109/IWSG.2014.11

Dankwoord

Iedereen die me kent zal vast weten dat ik uitstellen tot een kunstvorm verheven heb. Op 24 september 2015 registreerde ik me bij de graduate school zodat ik kon starten met mijn promotietraject. Dit dankwoord schrijf ik op 1 september 2022, bijna zeven jaar later. Het dankwoord is het allerlaatste stukje dat ik schrijf voordat dit boek naar de drukker gaat, daarom vind ik het misschien wel het lastigste om af te maken. Nadat ik de laatste zin van deze pagina's typ is het geheel 'af', dit kan ik in mijn hoofd nog niet helemaal plaatsen en dat is vast de reden dat ik bovenstaande zinnen al minstens achttien keer herschreven heb.

In het voorwoord haal ik een stuk uit het Ilias van Homerus aan om te illustreren dat zelfs de Griekse (half)goden samenwerkten om hun wonderen tot stand te brengen. Het mag een wonder heten dat dit boek eindelijk af is en daar hebben een hoop mensen aan bijgedragen. Hieronder benoem ik een selectie van de mensen die een bijdrage hadden aan dit boek of aan mijn leven tijdens het schrijven ervan. Vaak is het leveren van een schop onder mijn kont de beste vorm van hulp die ik kan gebruiken. Gelukkig zijn er veel mensen geweest die deze service aan mij hebben geleverd. Aan iedereen die zich hierdoor aangesproken voelt: mijn diepste dank, jullie zijn mijn Griekse (half)goden.

Speciale dank aan

Sílvia, *muito obrigado* voor je eeuwige geduld en wijze lessen. Je bent een integer persoon en dit komt sterk terug in jouw onderzoeksstijl. Tevens ben je een goede docent die me heeft weten op te leiden tot een gedegen onderzoeker. De scherpe, maar eerlijke, kritiek die je gaf heb ik altijd gewaardeerd, al liet ik dat op het moment zelf niet altijd blijken. Het plezier dat je hebt in doceren en opleiden is aanstekelijk en ik zal in dezen mijzelf altijd proberen te spiegelen aan jouw voorbeeld. Ik hoop dat jij en Ruud genieten van het leven in Portugal en ik beloof om samen met Maaïke langs te komen om onkruid uit de tuin te trekken.

Koos, het ogenschijnlijke gemak waarmee je complexe onderwerpen doorgrondt en vervolgens kan toepassen is verbluffend. Je vriendelijke aanpak komt zonder onnodige franjes en heeft me, zeker richting het einde van het traject, geholpen om door te zetten. Ook wil ik jou en Sílvia bedanken voor het creëren van een veilig omgeving voor een beginnend onderzoeker waar vrijheid is voor ideeën, fouten gemaakt en geanalyseerd kunnen worden en waar collegialiteit een belangrijke rol speelt.

Mariska, Perry en René, zonder jullie waren de hoofdstukken in dit boek nooit tot stand gekomen, laat staan geaccepteerd voor publicatie. Door te putten uit jullie grote pool van kennis en ervaring zijn mooie hypothesen aan de tand gevoeld in onze onderzoeken. Bedankt voor de hulp bij het uitdenken van hypothesen en methodes, de feedback die jullie gaven op de uitkomsten en de aanwijzingen en herschrijfacties op onze papers.

Alle oud-collega's van de e-Science groep, KEBB en KIK, waaronder: Shayan, Mahdi, Luis, Felipe, Barbera, Aldo, Mélanie, Maryam, Paul, Marcela, Tinka en nog een sloot anderen. Bedankt voor de lunches, jullie collegialiteit en gezelligheid.

Rodney, bedankt voor het maken van het prachtige ontwerp die op de omslag van dit boek staat.

Sander en Thijs, bedankt dat jullie mijn paranimfen willen zijn op 9 december. Jullie zijn beide goede vrienden en al heb ik jullie tijdens verschillende levensfasen ontmoet, toch zijn er redelijk wat parallellen. Zo kunnen we altijd goede (serieuze of minder serieuze) gesprekken voeren en vormen we onze eigen 'kleine slimme groepjes'. Ik hoop dat er nog veel (persoonlijke) hoogtepunten zullen zijn die we samen kunnen vieren.

Rinke, Thijs en Wouter, goede vrienden kan je alleen toevallig tegenkomen maar de vorming van ePisa moet wel voorbestemd zijn geweest. Bedankt voor de slechte humor, hilarische anekdotes, koud bier en chocomel, de vriendschap en gezelligheid. Ik hoop dat onze samenkomsten onder het genot van goed eten nog vaak zullen leiden tot wijze uitspraken.

Ameli, de bedachtzame gesprekken met jou aan de keukentafel op de van Bijnkershoecklaan hebben knopen in mijn gedachten losser weten te maken en mijn ergernissen weten te relativiseren. Als ik iemand zoek om samen Oreo-chocoladerepen mee te eten en slechte series te kijken, weet ik je te vinden.

Femke en Jeroen, bedankt voor de goede etentjes en uiteraard jullie oneindige interesse in de voortgang van mijn boekje.

Papa en mama, bedankt voor een opvoeding waarbij nieuwsgierigheid aangemoedigd werd en waar altijd vragen gesteld konden worden, ook als die vragen soms niet te beantwoorden waren: "mama, hoeveel sterren zijn er eigenlijk?". Als ik vanuit het niks een plan of nieuwe hobby bedacht had, wisten jullie er altijd wel een mouw aan te passen. Jullie hebben me geleerd om plezier te hebben in het uitvoeren van nieuwe ideeën of het uit elkaar halen van spullen om de werking ervan te begrijpen. Ik blijf elke dag weer nieuwe toepassingen vinden voor deze vaardigheden.

Maaïke, in de (vele) jaren dat we elkaar kennen zijn we alleen maar dichterbij elkaar gegroeid en ben ik meer van je gaan houden. We hebben een vrijwel perfecte balans weten te vinden van ondersteunen en vrij laten. Zelfs als we meerdere maanden op elkaars lip in een bus leven. De aankondig dat ik een PhD ging doen ving je op met 10% uitlachen en 90% support. Je kent al mijn sterke punten en accepteert al mijn zwaktes en weet deze ook vaak aan te vullen. Daar heb ik dankbaar gebruik van gemaakt tijdens het schrijfproces van enkele papers. Zonder jouw herschrijfacties was ik nu waarschijnlijk nog steeds op zoek naar de perfecte bewoording. Ik hoop dat we de komende jaren al onze plannen waar kunnen maken, of niet, als we er tegen die tijd toch anders over blijken te denken. We zien wel wat er gebeurt en ik zal altijd van je blijven genieten, in de woorden van Freddie: "you're my best friend".

αύχένα τε στιβαρὸν καὶ στήθεα λαχνήεντα,
δῦ δὲ χιτῶν', ἔλε δὲ σκῆπτρον παχύ, βῆ δὲ θύραζε
χωλεύων: ὑπὸ δ' ἀμφίπολοι ῥώοντο ἄνακτι
chrύσειαι ζωῆσι νεήνισιν εἰοικυῖαι.
τῆς ἐν μὲν νόος ἐστὶ μετὰ φρεσίν, ἐν δὲ καὶ αὐδὴ
καὶ σθένος, ἀθανάτων δὲ θεῶν ἅπο ἔργα ἴσασιν.