Generating High Precision Classification Rules for Screening of Irrelevant Studies in Systematic Review Literature Searches

THE UNIVERSITY OF
SYDNEY

A thesis submitted in fulfilment of the requirements for
the degree of Doctor of Philosophy
School of Information Technologies
University of Sydney

Henry Petersen

2016

Abstract

Systematic reviews aim to produce repeatable, unbiased, and comprehensive answers to clinical questions. Systematic reviews are an essential component of modern evidence based medicine, however due to the risks of omitting relevant research they are highly time consuming to create and are largely conducted manually. This thesis presents a novel framework for partial automation of systematic review literature searches. We exploit the ubiquitous multi-stage screening process by training the classifier using annotations made by reviewers in previous screening stages. Our approach has the benefit of integrating seamlessly with the existing screening process, minimising disruption to users.

Ideally, classification models for systematic reviews should be easily interpretable by users. We propose a novel, rule based algorithm for use with our framework. A new approach for identifying redundant associations when generating rules is also presented. The proposed approach to redundancy seeks to both exclude redundant specialisations of existing rules (those with additional terms in their antecedent), as well as redundant generalisations (those with fewer terms in their antecedent). We demonstrate the ability of the proposed approach to improve the usability of the generated rules. The proposed rule based algorithm is evaluated by simulated application to several existing systematic reviews. Workload savings of up to 10% are demonstrated.

There is an increasing demand for systematic reviews related to a variety of clinical disciplines, such as diagnosis. We examine reviews of diagnosis and contrast them against more traditional systematic reviews of treatment. We demonstrate existing challenges such as target class heterogeneity and high data imbalance are even more pronounced for this class of reviews. The described algorithm accounts for this by seeking to label subsets of non-relevant studies with high precision, avoiding the need to generate a high recall model of the minority class.

Acknowledgements

First of all, I would like to thank my supervisors Josiah Poon, Simon Poon, and Clement Loy for years of advice, guidance, and support.

I am also grateful to Mariska Leeflang for her generous and invaluable assistance.

Lastly, I would like to thank my family: my parents, John and Christine, my sister Leah, my brother Joseph, my great aunt Jenny, and although no longer with us, my grandmother Shirley. I am eternally grateful for your unconditional love, support, and patience. Without you I would not be who I am today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Systematic reviews are a key component of modern evidence based medicine. For a given clinical query, systematic reviews seek to provide a repeatable, unbiased, and comprehensive answer based on an accumulation of all available, relevant evidence. Due to the massive volume of published evidence available (for example, as of 2015 PubMed indexes over 25 million studies), medical practitioners are reliant on systematic reviews to keep abreast of relevant work. However, despite the importance of systematic reviews the process for conducting them is highly time consuming and largely manual.

This thesis presents a novel approach for partial automation of the systematic review process. We demonstrate our approach can provide real workload savings for reviewers by simulated application to the literature searches for real reviews.

A major benefit of our approach is its seamless integration with both existing manual screening process. We exploit the current screening process to obtain training data without and additional labelling phase.

1.1 Motivation

Despite the importance of systematic reviews for modern evidence based medicine, the process by which they are conducted is largely manual and highly time consuming. It is not unusual for a single review to take months, or even years, from inception to publication (a review conducted by Sampson et al. [87] in 2008 found an average time of 61 weeks between the start of citation screening and publication). This can have practical consequences for the currency of the final review. The ability of practitioners to access important knowledge is also impacted.

It should be quite obvious that increased automation of the literature screening process through the application of machine learning has the potential to drastically reduce the burden on already overworked review authors [93]. However, literature screening for systematic reviews have a number of unique challenges when viewed as a classification task. For example, due to the high cost of omitting relevant information there is a need for near perfect recall over relevant studies.

Supervised machine learning requires a human oracle to annotate data, which is then used to train a classification model. The unique nature of each review means that a separate classifier must be trained for each new literature search. While training data could be obtained by randomly selecting a subset of citations for an initial manual screening, it is worth investigating whether or not such an approach could be improved.

Systematic review literature searches are typically conducted as a multi-stage triage process (an introduction to systematic reviews and their literature searches is given in section 2.2). Reviewers will first screen a large volume of citations based on title alone, removing any obviously irrelevant studies. Remaining citations are then screened based on title and abstract, and then again on full text. At each review stage (with the exception of the first), this implies the reviewers are already in possession of a number of studies which have been annotated as not relevant to the topic of the review at hand. To date no research has considered the potential such excluded studies may have to influence an automated screening algorithm.

Such an investigation would be highly worthwhile.

Despite the practical challenges in their creation, the number of systematic reviews being conducted is only increasing. Systematic reviews have traditionally focused on clinical questions related to treatments. In recent times however, there has been an increasing demand for reviews from other areas such as diagnosis and aetiology. For example, in 2007 the Cochrane Collaboration launched the *Cochrane Diagnostic Test Accuracy Working Group* to assist in the creation and dissemination of high quality diagnostic reviews. These areas bring with them their own unique issues and challenges.

As the volume and breadth of reviews increases, new challenges appear that effect the literature search process. For example, the US National Library of Medicines MeSH (Medical Subject Headings) ontology [75] (which is used to index citations databases such as MEDLINE) contains an entry that can be used to indicate a study reports a randomised controlled trial. Such a label is useful when pruning citations for use in a treatment review, as they typically seek to only include RCTs (Randomised Controlled Trials). However, there is no corresponding label in MeSH for diagnostic test accuracy study. Differences between traditional and emerging reviews need to be identified and addressed when seeking to automate the literature screening process.

1.2 Aims

The primary aim of this research is to reduce the workload required in screening abstracts for systematic reviews. We define workload in terms of number of citations screened. An inherent part of classification for systematic review literature screening is the requirement that relevant studies not be missed. We therefore have the additional requirement that recall on relevant citations must be 1 (or at worst equivalent to the recall achieved by a human reviewer).

We also aim to address the question of how best to integrate machine learning into the systematic review literature screening process. Of particular interest is the multi-stage triage process used by human reviewers when screening citations. In order to train a prospective classifier, prior work has sought to generate labels as the screening process to which the classifier will be applied is already underway. By utilising the annotations provided in earlier screening stages, we aim to build the classifier without the need for an initial labelling stage.

The increased prevalence of systematic reviews for clinical fields other than treatment (for example, diagnosis) is also a motivation. We aim to identify any differences between review fields with respect to how it effects the classification task. In particular, we hypothesise that diagnostic reviews require additional reviewer workload compared to reviews of treatment, as well as having comparatively heterogeneous target classes and a lower quality of available meta-data when searching for relevant studies. This analysis would be conducted to see whether any unique challenges or properties exist, and whether they could indicate directions for future work.

1.3 Contributions

- **Empirically demonstrate diagnostic reviews contain increased workload and target class heterogeneity when compared to treatment**

  A comparative analysis of diagnostic test accuracy reviews and treatment reviews was conducted, and it was identified that diagnostic test accuracy reviews as a group contained higher rates of data imbalance and a broader target class. In addition to indicating a particularly challenging subset of reviews for special focus when developing classification algorithms, the practical demonstration of the difference between review fields is useful for review authors and research librarians.

- **Model for citation classification that excludes irrelevant citations with high precision**

A classification algorithm for systematic review literature searches was proposed that focuses on modelling subsets of irrelevant citations with high precision, rather than modelling the entire set of relevant citations with high recall. A major benefit of such an approach is that it avoids having to model the broader target class found in more challenging reviews such as those of diagnosis.

- **Model for automation of systematic review literature searches that uses annotations from prior screening stages to build the classifier**

  In contrast to previous work that trains classifiers with annotations obtained at the stage for which the resultant classifier will be applied, we propose a model where annotations from previous screening stages are used. In addition to avoiding the need for a dedicated labelling stage, this has the benefit of complementing alternate classification approaches. Should one desire, our algorithm could be trained and applied between screening stages before switching to an existing classification system.

- **Initial analysis on real data for recommended parameterisation of classification models**

  The algorithm developed in this thesis requires several parameters: namely a minimum confidence threshold and P-value for valid rules. We perform an initial sensitivity analysis of these parameters for several real datasets, and determine initial recommendations for the P-value threshold. While further work to produce concrete recommendations is required, our analysis provides a starting point upon which further analysis can be built.

- **Alternate definitions for association rule redundancy, with corresponding rule mining algorithm**

  Classical rule redundancy seeks to exclude rules that fail to improve an existing, more general rule. We propose an alternate definition of redundancy that is both more permissive in identifying when an improvement is made over generalisations, and has the ability to remove general rules which are redundant with respect to a more specific rule.

- **Demonstrated ability for alternate redundancy definitions to improve the interpretability of the rule based literature screening algorithm**

  A benefit of rule based classifiers is the ability for users to inspect and understand the generated model. We demonstrate that the proposed approach to rule redundancy is able to increase the descriptive power of individual rules, without significantly impacting classifier performance.

1.4 Outline

The rest of this thesis is organised as follows. Chapter 2 provides an overview of existing work related to this thesis. This includes an overview of systematic reviews and the literature search process, as well as supervised learning and rule mining.

Chapter 3 analyses the literature search process for systematic reviews of diagnosis, and compares them against those for more traditional treatment reviews. Several key differences, including increased target class heterogeneity and data imbalance for diagnostic reviews, are identified and are used to motivate subsequent chapters.

Our algorithm for automated citation screening is then presented in chapter 4. An evaluation by simulated application on real data is presented. We also examine the algorithms sensitivity to changes in its parameterisation.

Chapter 5 examines the rule mining process, in particular the exclusion of redundant rules. Deficiencies in existing definitions are noted. An alternate definition for redundant rules is proposed, and an algorithm is given for mining rules is given. Chapter 6 then demonstrates the improved descriptive power of classifiers generated using the proposed rule mining algorithm for systematic review screening.

Finally, chapter 7 summarises the work reported in this thesis. Contributions and limitations are discussed, and directions for future work are proposed.

Chapter 2

Background

In this chapter we will provide an overview of existing knowledge relevant to this body of research. Like many projects this work touches on a number of different fields, a full treatment of which is far beyond the scope of this document. As such we provide here a treatment only of those areas relevant to this research, with references provided to further information for the interested reader.

A significant part of this work concerns the extraction of association rules from text. To this end, we provide an overview of association rule mining in section 2.1, including rule generation, measures of rule quality, and elimination of redundant rules. This section provides the necessary context for the work on rule generation and redundancy in chapter 5, although it is also useful throughout the remainder of the thesis.

The primary motivation for this work concerns the literature search process for systematic reviews, and methods for its automation. We provide an introduction to systematic reviews, and analyse prior research aimed at automating the corresponding literature search process in section 2.2.

2.1 Association Rule Mining

Association rule mining [3, 1] addresses the task of finding interesting associations between sets of items in transactional data. More formally, assume the existence of a set of N items $\Upsilon = \{\upsilon_1, \upsilon_2, \ldots, \upsilon_N\}$, and a set of transactions T. Each transaction $t \in T$ is a subset of items in Υ (i.e. $t \in \Upsilon$). The association rule mining task aims to identify the set $\Re$ of all interesting rules of the form $X \Rightarrow Y$ where X and Y are disjoint sets of items (i.e. $X \subseteq \Upsilon, Y \subseteq \Upsilon, X \cap Y = \emptyset$). By convention, for such a rule the set of items X is referred to as the antecedent, and the set Y is referred to as the consequent.

An important part of mining association rules involves defining exactly what constitutes an interesting rule. Traditionally association rule mining has used support and confidence to identify rules [3]. We note that alternative interestingness measures is an active area of research [44, 43, 42, 56, 91], and provide a brief introduction to the topic in sub-section 2.1.1.

The terminology used above (items and transactions) when defining the association rule mining task reflects its origins in the domain of market basket analysis; the identification of interesting patterns in customer purchase data [3, 2]. It is however applicable in any domain where the objective is to identify patterns in binary data. In the nearly two and a half decades since it was first proposed association rule mining has found application in a wide range of fields including bioinformatics [68], text mining [7, 45], and medical domains [12, 111]. Adaptations have also been made to account for spatial [111, 60], and temporal data [61].

The number of possible association rules grows exponentially with the number of attributes. For real data, searching over the space of possible rules can quickly become intractable. In order to deal with this, efficient search algorithms must be developed to traverse the search space. We cover rule generation algorithms in 2.1.2.

A central part of the work covered in this thesis concerns the concept of rule redundancy. Redundant rules are those that describe knowledge contained in other rules. Identification and removal of redundant rules is important both to improve the quality of generated rules, as well as to limit the search space for rule generation algorithms. We introduce existing approaches to rule redundancy in section 2.1.3.

We now present a summary of the existing literature relevant to this thesis.

2.1.1 Interestingness Measures

As mentioned in the introduction to this section, association rule mining has traditionally been conducted using a support and confidence framework. Given a rule $X \Rightarrow Y$ and a function m where $m(X)$ is the frequency of X in T, support and confidence can be defined as follows:

$$\mathit{supp}(X \Rightarrow Y) = m(X \cup Y)/|T| \tag{2.1}$$

$$\mathit{conf}(X \Rightarrow Y) = m(X \cup Y)/m(X) \tag{2.2}$$

Put another way, support is defined as the percentage of transactions in T containing all items from both X and Y. Confidence is the percentage of transactions containing the items in X that also contain the items in Y. When mining rules the user selects a minimum acceptable value for support and confidence, then aims to identify all rules exceeding this threshold.

An interesting property of the support function is that it is anti-monotonic. In other words, $X \subset Y \Rightarrow \mathit{supp}(X) \geq \mathit{supp}(Y)$. This property is attractive from a computational standpoint when generating association rules (a point discussed further in section 2.1.2). However, in addition to support and confidence, association rules can and have been generated using a wide range of interestingness measures.

For a given rule $X \Rightarrow Y$, an interestingness measure considers the partial

Measure	Formula
Support [3]	$\frac{m(X)}{M}$
Confidence [3]	$\frac{m(XY)}{m(Y)}$
Interest [16]	$\frac{M \times m(XY)}{m(X)m(Y)}$
Leverage [81]	$\frac{m(XY)}{M} - \frac{m(X)m(Y)}{M^2}$
χ^2	$\frac{M^5 \times Leverage(X \Rightarrow Y)^2}{m(X)m(\neg X)m(Y)m(\neg Y)}$
Fisher's P	$\sum_{i=0}^{min(m(X\neg Y),m(Y\neg X))} \frac{\binom{m(X)}{m(XY)+i}\binom{x(\neg X)}{m(\neg X\neg Y)+i}}{\binom{M}{m(Y)}}$

Table 2.1: Several common interestingness measures for a rule $X \Rightarrow Y$ expressed in terms of partial frequency counts.

frequency counts of the co-occurrences of X and Y in $\mathcal{D}$ to determines the strength of the association between X and Y. Figure 2.1 shows each of the partial counts, as well as their effect on the strength of the association. Different interestingness measures do so in different ways; for example confidence measures the conditional probability of Y given X, while lift measures the ratio of the joint and marginal probabilities (see Table 2.1 for a summary of several well known interestingness measures).

Association rules have also been evaluated using measures of statistical significance. Such tests are useful in that they provide confidence that discovered rules will hold in unseen data. This is often a desirable property. To this end, we evaluate rule quality with statistical significance measures in this work. In line with much of the literature [40, 99, 106], Fisher's Exact Test [30] is used. Fisher's Exact Test defines an exact statistical significance test of the association between two categorical variables, under the assumption of fixed margins.

We note that due to computational issues in computing the binomial coefficients in the formula given in Table 2.1, direct computation of Fishers P can be difficult. In line with similar literature [40], instead of directly computing P values

	Y	$\neg Y$	
X	$m(XY)$	$m(X\neg Y)$	$m(X)$
$\neg X$	$m(\neg XY)$	$m(\neg X\neg Y)$	$m(\neg X)$
	$m(Y)$	$m(\neg Y)$	M

Figure 2.1: Table showing the partial frequency counts used when computing interestingness for a rule $X \Rightarrow Y$. Unshaded cells support the rule, unshaded cells do not.

we instead work with the natural log of Fishers P.

A more detailed treatment of the differences between the various interestingness measures is likely beyond the scope of this work. We note that several comprehensive reviews exist [56, 91], and direct the interested reader to the literature for further information.

2.1.2 Dealing with Combinatorial Complexity

Computational complexity presents a substantial challenge when generating association rules. Given a dataset with N items as well as some minimum threshold for support and confidence, there are $\Sigma_{i=1}^{N-1}\binom{N}{i}2^{N-i}$ possible rules which need to be evaluated. The majority of association rule mining algorithms address this by using an approach based on frequent itemsets. Such algorithms generate rules in a two stage process:

1. Generate all frequent itemsets $\mathcal{F}$ (those sets of items which occur in T with at least some minimum level of support).

2. Compute $\Re$ by evaluating all rules which can be built using $\mathcal{F}$.

Perhaps the best known example of this approach is the a-priori algorithm of Agrawal and Srikant [4]. In order to identify frequent itemsets, the a-priori algorithm exploits the fact that the support function is anti-monotonic. Put more explicitly, any subset of a frequent itemset will itself be frequent [4].

The a-priori algorithm is initialised by finding all length 1 frequent itemsets (individual items whose frequency is above the specified threshold). It then iteratively finds potentially frequent itemsets of length n using unions of frequent itemsets of length $n-1$. Exact supports are then computed for potentially frequent itemsets, with the process repeated until no further itemsets can be identified.

For a given frequent itemset $F \in \mathcal{F}$, all possible rules $X \Rightarrow Y s.t. X, Y \subset F, X \cap Y = \emptyset$ are checked to find those with the minimum required confidence. From the definition for confidence in equation 2.2, computing the confidence requires frequencies for the rule $X \Rightarrow Y$ requires F and X. As $\mathcal{F}$ is a frequent itemset, the monotonicity of the support function implies that X (and Y) are also frequent. Hence by indexing computed supports when finding frequent itemsets, the confidence for all rules can therefore be checked without the need for any additional frequency calculations.

Since the a-priori algorithm was originally proposed, there has been a substantial body of work addressing the itemset generation process (for example, alternate algorithms such as ECLAT [114, 113] and FP-growth [37, 38]). Developments such as frequent closed itemsets (proposed by Pasquier et al. [78]) also allow for a much smaller number of itemsets to generated, improving the efficiency of rule mining without reducing the expressive power of the result set.

In section 2.1.1 we noted that interestingness measures such as Fishers P and χ^2 have also been used. These interestingness measures do not obey the anti-monotonicity assumption employed when mining rules with support. When evaluating rules using such measures, it is possible that even very infrequent rules could be considered interesting. A frequent itemset approach could still be employed, but it then becomes possible that interesting, infrequent rules could be missed.

Work exists on alternate search strategies for identifying association rules in the case where support based pruning is inappropriate. Hämäläinen [40, 41] employs a branch and bound strategy using lower bounds on the natural log of Fishers P to prune the search space in a search for general dependency rules with single attribute consequents. Each node tracks attributes that could form possible rules (based on whether a rule with that attribute as consequent could potentially contain a significance value lower than the required threshold). The search is pruned at nodes where no possible consequent attribute exists. They also employ a novel method for propagating impossible consequent attributes to unvisited nodes.

2.1.3 Rule Redundancy

The number of interesting rules identified as a result of many association rule mining algorithms is a significant concern. In practice, the number of rules produced can be prohibitively large, and present a barrier to their interpretation by users [2, 5, 13, 94, 112, 10, 11].

Much research has focused on the idea of redundant rules [2, 112, 10]. Such work seeks to remove rules that encode knowledge described in equivalent or better fashion by other rules. Such rules are often artifacts of independent attributes. An example of rule redundancy can be seen as follows.

Consider some hypothetical data from which the association *Overweight* $\Rightarrow$ *Diabetes* is mined. Another rule mined might be *Overweight* $\cdot$ *Football Fan* $\Rightarrow$ *Diabetes*. Although the two associations may hold with equivalent strength in the data, the addition of the condition *Football Fan* has no bearing on the strength of the association between weight and diabetes. The second association only holds due to the independence of *Football Fan* and the other two variables. In other words, *Overweight* $\Rightarrow$ *Diabetes* holds regardless of the presence or otherwise of *Football Fan*. The second rule is a redundant specialisation of the first.

Several authors have proposed formal means for defining redundant rules [2, 112, 10]. Given a non-redundant rule $A \Rightarrow B$, Aggarwal [2] sought to exclude as redundant any rule $C \Rightarrow D$ that must have equal or lower support and confidence, independent of any distributional qualities of the data. For example, given the rule $X \Rightarrow YZ$, it can be shown that the rules $XY \Rightarrow Z$ and $XZ \Rightarrow Y$ have equivalent support but must have equal or lower confidence.

In contrast to redundancy based approaches for eliminating spurious associations, Webb [104, 105] use the idea of productive rules. A rule $X \Rightarrow Z$ can be considered redundant if there exists another rule $Y \Rightarrow Z$ with equivalent support where $Y \subset Z$. The productive criterion instead removes rules where no generalisation exists with equivalent confidence. More formally:

$$\mathit{productive}(X \Rightarrow Z) \ \mathit{iff} \ \mathit{confidence}(X \Rightarrow Z) - \max_{Y \subset X} \mathit{confidence}(Y \Rightarrow Z) > 0$$

Productivity is a stronger criterion than redundancy in that any redundant rule will contain a generalisation with equivalent confidence. It is however possible for a rule to have no generalisation with equivalent support, but to have one with equal or better confidence.

The bulk of work on redundancy for association rules assumes that associations are mined using a support and confidence framework. Support of an itemset (or rule) is anti-monotonic with respect to the number of attributes it contains ($\mathit{supp}(X) \geq \mathit{supp}(Y) \ \forall X \subset Y$). This property does not hold for many other interestingness measures, such as statistical tests including Fishers P or χ^2.

A definition of redundancy suitable for use with a general goodness measure (under the assumption of single attribute consequents) was first proposed in 2010 by Hämäläinen [40]. Hämäläinen defines a rule R to be redundant if some more general rule (i.e. a rule whose antecedent is a subset of the antecedent of R) has equal or better utility with respect to some goodness measure. The definition is repeated formally in definition 2.1. More detail on redundancy for generalised goodness measures is provided in section 5.1.

Definition 2.1. ***Classical redundancy*** *Consider two rules $X \Rightarrow Z$ and $XQ \Rightarrow Z$ where X and Q are disjoint sets of items, and Z is a single item of value a. Let $M(\cdot)$ be an increasing measure of rule interestingness. Rule $XQ \Rightarrow Z$ is redundant with respect to rule $X \Rightarrow Z$ if $M(XQ \Rightarrow Z) \leq M(X \Rightarrow Z)$.*

Hämäläinen also proposes an algorithm (Kingfisher) for use finding non-redundant general dependency rules using this definition [40]. Kingfisher is the first algorithm able to efficiently perform such a search. The algorithm uses a branch and bound search over the space of possible rules, and employs several additional pruning heuristics to control the size of the search. The work described in section 5 uses a variant on the Kingfisher algorithm, and more detail on the search process is given in section 5.3. Detail on the pruning heuristics employed by Kingfisher is given in section 5.3.1.

We note however that when comparing rules based on some arbitrary goodness measure, complications can arise due to complex interactions between constituent attributes. That such interactions can give rise to spurious associations has been studied [68, 106], however less attention has been payed to how it might obscure useful relationships. We examine this further in chapter 5, in particular with the concept of specialisation redundancy.

Most research has looked to remove redundant specialisations that add nothing over simpler rules. This is not always appropriate, and authors such as Webb [106] and Liu et al. [62] have looked at models of redundancy that remove spurious generalisations.

Removing spurious generalisations was first looked at in 2001 by Liu et al. [62], in their work on non-actionable rules. For a given rule r_0 and the set of its decedents $R = \{r_1, r_2, \cdots, r_N\}$, they define a r_0 to be *non-actionable* if it is not interesting over the domain where instances matching at least one antecedent in R are removed. Essentially, they claim a rule must cover some unique set of instances (with respect to the set of its specialisations) in which the relationship described still holds. If a rule does not cover such a set of instances, the rule has no utility with respect to the set of its specialisations.

A similar concept to non-actionable rules has been proposed by Webb in his work on self sufficient itemsets [106]. This work builds upon the concept of an *exclusive domain* for a given itemset. Formally, the given an itemset s and its specialisations S, the exclusive domain of s is defined to the domain of s minus the union of the domains of all itemsets in S. After generalising the concepts of redundancy and productivity [104, 105] for use with itemsets, rather than association rules, an itemset is defined to be self sufficient if it is productive and non-redundant both with respect to the entire data and its exclusive domain.

In many respects, self sufficient itemsets can be considered an extension of non-actionable rules for use in an itemset context. However, it is also noteworthy that itemsets must also be productive and non-redundant. This pruning of both specialised and general itemsets is somewhat similar to our work with robust redundancy in chapter 5, although it is performed in an itemset context. As described above, we also examine methods to avoid pruning specialisations where redundancy is likely to be an artifact of interactions between constituent attributes.

2.2 Systematic Reviews

An important concept in modern health care is the idea of evidence based medicine: the requirement that clinical policy and practice should be based on an examination of all available, relevant knowledge [86]. Systematic reviews form a key component of evidence based medicine, and are widely regarded as the highest form of medical evidence [110]. A systematic review is a special type of literature review (and often meta-analysis), which addresses a specific research question with the aim of being comprehensive, unbiased, and repeatable [35, 36].

There is much active research into the efficient generation and dissemination of high quality systematic reviews [26, 34, 46, 53, 55, 70, 90, 96, 98]. Guidelines are regularly published to ensure reviews are conducted and reported according to the highest possible standards [25, 27, 54]. In addition, organisations such as the Cochrane Collaboration have been created to facilitate creation and maintenance of systematic reviews, as well as their dissemination through the widely known Cochrane Library of Reviews [23]. Cochrane also publish their own guidelines for conducting systematic reviews to be published in the Cochrane Library [35][36].

As evidence based practice relies on high quality systematic reviews, there is an implication that a substantial cost exists should review authors omit relevant work. This claim is well established [28]. As a consequence, review authors go to great lengths to ensure reviews are as comprehensive as possible, and to minimise errors. Although a number of resources have been developed to facilitate efficient screening, literature searches for systematic reviews are largely manual and highly time consuming [87].

In the last decade, there has been increasing interest shown by the machine learning and information retrieval communities in improving the efficiency of systematic reviews by addressing the manual nature of their creation [13, 20, 32, 48, 63, 66, 92, 102]. In the remainder of this chapter we provide an introduction the process of conducting systematic reviews, and summarise existing work concerning the application of machine learning for the corresponding literature searches. This section provides context for the motivation behind the work reported in subsequent chapters.

2.2.1 Systematic Review Process

While the exact process for conducting a systematic review varies according to the type of clinical question (i.e. diagnosis, intervention, aetiology), all systematic reviews can be said to follow several major steps [77]. These include question and inclusion criteria formulation, literature search, literature screening, quality assessment, and data synthesis, analysis and interpretation.

A comprehensive treatment of the entire systematic review process is beyond the scope of this thesis, and would likely obscure the pertinent information required to provide context for this and later chapters. Hence a summary of the entire systematic review process will not be presented here. Instead a summary of the first three stages is presented, which are the sections most relevant to this work. For a more complete treatment of the quality assessment, meta-analysis, and reporting stages of systematic reviews the interested reader is directed to literature such as Wright et al. [110], Tsafnat et al. [98], or resources such as the Cochrane handbooks for reviews of interventions [35] and DTA (Diagnostic Test Accuracy) [36].

A summary of the major steps in the systematic review process is given in Figure 2.2.

2.2.1.1 Question and Inclusion Criteria Formulation

Systematic reviews begin with the formulation of a highly specific research questions and associated inclusion criteria. For example, one Cochrane Diagnostic Test Accuracy review [15] obtained from the Diagnostic Test Accuracy Working Group lists its objectives as:

> To compare the diagnostic accuracy of diffusion-weighted MRI (DWI) and CT for acute ischaemic stroke, and to estimate the diagnostic accuracy of MRI for acute haemorrhagic stroke.

Inclusion criteria for Cochrane systematic reviews are formulated according to specific categories that depend on the type of clinical question being answered. For example, Cochrane reviews of Diagnostic Test Accuracy, separate criteria are formulated for the type of study, index and comparator tests, target condition and the desired reference standard [36]. A similar set of criteria (referred to as the PICO criteria—Population, Intervention, Comparison, Outcome) exists for questions related to interventions [35].

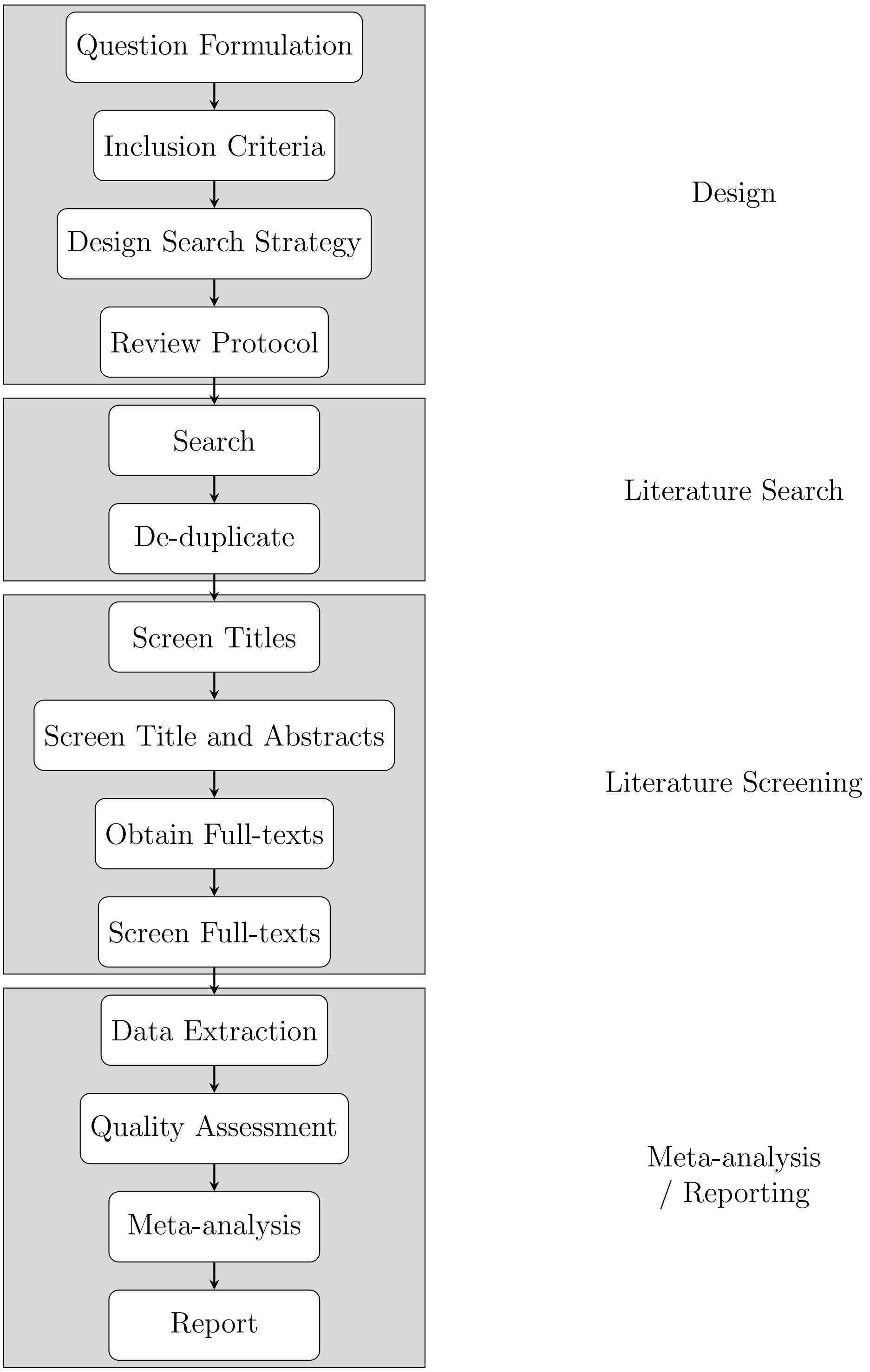


Figure 2.2: Overview of the systematic review process.

2.2.1.2 Literature Search

Once inclusion criteria for the review have been identified, the next step in the review process is to determine which resources are to be searched. Many resources exist indexing available medical literature, of which the most commonly used for Cochrane reviews of diagnostic test accuracy are MEDLINE which is maintained by the US National Library of Medicine, and EMBASE, which is published by Elsevier [29]. MEDLINE is available for free online through PubMed, and also via subscription through providers such as Ovid.

When a citation is indexed in MEDLINE, a medical librarian will manually annotate it with several concept headings from the US National Library of Medicines Medical Subject Headings (MeSH) ontology. The MeSH ontology provides a controlled vocabulary of medical concepts, which a useful for constructing searches in MEDLINE. As of 2015, MeSH contains 27149 headings organised into a hierarchy of 12 levels by degree of specificity [75]. An example of this is given in the MeSH factsheet: “At the most general level of the hierarchical structure are very broad headings such as ’Anatomy’ or ’Mental Disorders’. More specific headings are found at more narrow levels of the twelve-level hierarchy, such as ’Ankle’ and ’Conduct Disorder’”. Citations in EMBASE are annotated with concepts from a similar ontology EMTREE.

Although each of these databases are quite large (as of 2015 MEDLINE indexes over 19 million references from over 5600 journals [74] and EMBASE indexing over 20 million from 8600 journals [29]) it has been noted that there is a relatively low degree of overlap between MEDLINE and EMBASE (depending on topic, the overlap can be anywhere from 10% to 87% [109]). As such it is recommended to search multiple resources when conducting systematic reviews of diagnostic test accuracy. For example, for systematic reviews of treatment Cochrane recommend at a minimum [35]:

- MEDLINE

- EMBASE
- Cochrane register of randomised controlled trials

A similar Cochrane Register of Diagnostic Test Accuracy Studies is being developed for reviews of diagnostic test accuracy, however it has not yet been completed [36].

Literature searches for Cochrane systematic reviews of treatments are conducted by identifying references containing certain relevant MeSH headings and free text terms. Typically, several sets of terms are identified, and references containing at least one term from each are collected. An example of search terms used in the review *Galactomannan detection for invasive aspergillosis in imumunocompromized patients* by Leeflang et al. [57] is presented in Figure 2.3 (further detail on the search for this review can be found in Section 4.5.1. In order to achieve the very high levels of recall required for systematic reviews, Cochrane reviews of interventions will usually identify three sets of MeSH headings relating to the desired index test, the target condition and a methodological filter to limit results to references describing randomised controlled trials.

Literature searches for systematic reviews of diagnostic test accuracy are similar, however the methodological search filter is often omitted [58]. In order to avoid negatively impacting the results of the search, the filters need to have near perfect recall while still maintaining sufficient precision to justify their use. While much research has been done on developing highly sensitive DTA filters [59, 51, 52, 107, 108, 85, 50], the broader community has yet to develop a consensus on their use in DTA reviews (for example the Cochrane handbook for DTA reviews recommends against the ”routine use of methodology search filters”)[36].

In addition to MEDLINE and EMBASE a number of specialist databases exist (e.g. CINAHL, MDEION, DARE, C-EBLM and ARIF) that may be searched as part of conducting a systematic review. Many authors will also utilise resources such as personal communications, hand searching potentially relevant journals and screening grey literature (conference abstracts, unpublished or partial results, aspergillosis[MeSH Terms] OR

Aspergillosis[Text Word] OR

aspergillus[MeSH Terms] OR

Aspergillus[Text Word] OR

aspergill*

AND

"Nucleic Acid Amplification Techniques"[MeSH] OR

Polymerase Chain Reaction[tw] OR

PCR[tw] OR

nucleic acid amplification[tw] OR

immunosorbent assay[tw] OR

immunoassay[tw] OR

ELISA[tw] OR

EIA[tw] OR

"immunoassay"[MeSH Terms]

Figure 2.3: Example of Pubmed search query used in the review *Galactomannan detection for invasive aspergillosis in imumunocompromized patients.* Note the logical conjunction of two separate sets of terms.

theses etc.). In practice the number and type of resources employed is limited according to the scope of the review and resources available.

2.2.1.3 Literature Screening

References returned by the literature search are then manually screened to determine if they meed the inclusion criteria for the review. Literature screening for systematic reviews is a multi-stage process [35, 36]. Depending on the number of citations returned from the initial search, an author may apply a brief sanity check, screening all citations based on title and removing those which are obviously irrelevant to the review [64]. In the next stage two reviewers independently examine the title and abstract for every citation, with any citations possibly meeting the inclusion criteria are selected. The full text is then obtained for all selected citations. The full citations are then screened again by both reviewers. In order to meet the goals of a repeatable and unbiased analysis, the reasons for exclusion in the second stage are often recorded, and published along with the review. An example of a typical literature screening process is shown in Figure 2.4.

At this stage of the review, it is commonplace for authors to screen hundreds or even thousands of citations. Despite this, the number of references included in the final review is often much smaller, possibly one or two orders of magnitude smaller than the total number of references screened.

The rationale behind the use of multiple reviewers is to minimise the risk of error. The need for multiple reviewers, combined with the often very large number of references to be screened make this process labour intensive and extremely time consuming. A review conducted by Sampson et al. [87] in 2008 found an average time of 61 weeks between the start if citation screening and publication. Karimi et al. [50] note that when screening citations, each individual document may require several minutes or more to process. It is not difficult to see that even small reductions in the number of citations viewed could result in a significant reduction in the time and effort required to complete this step.

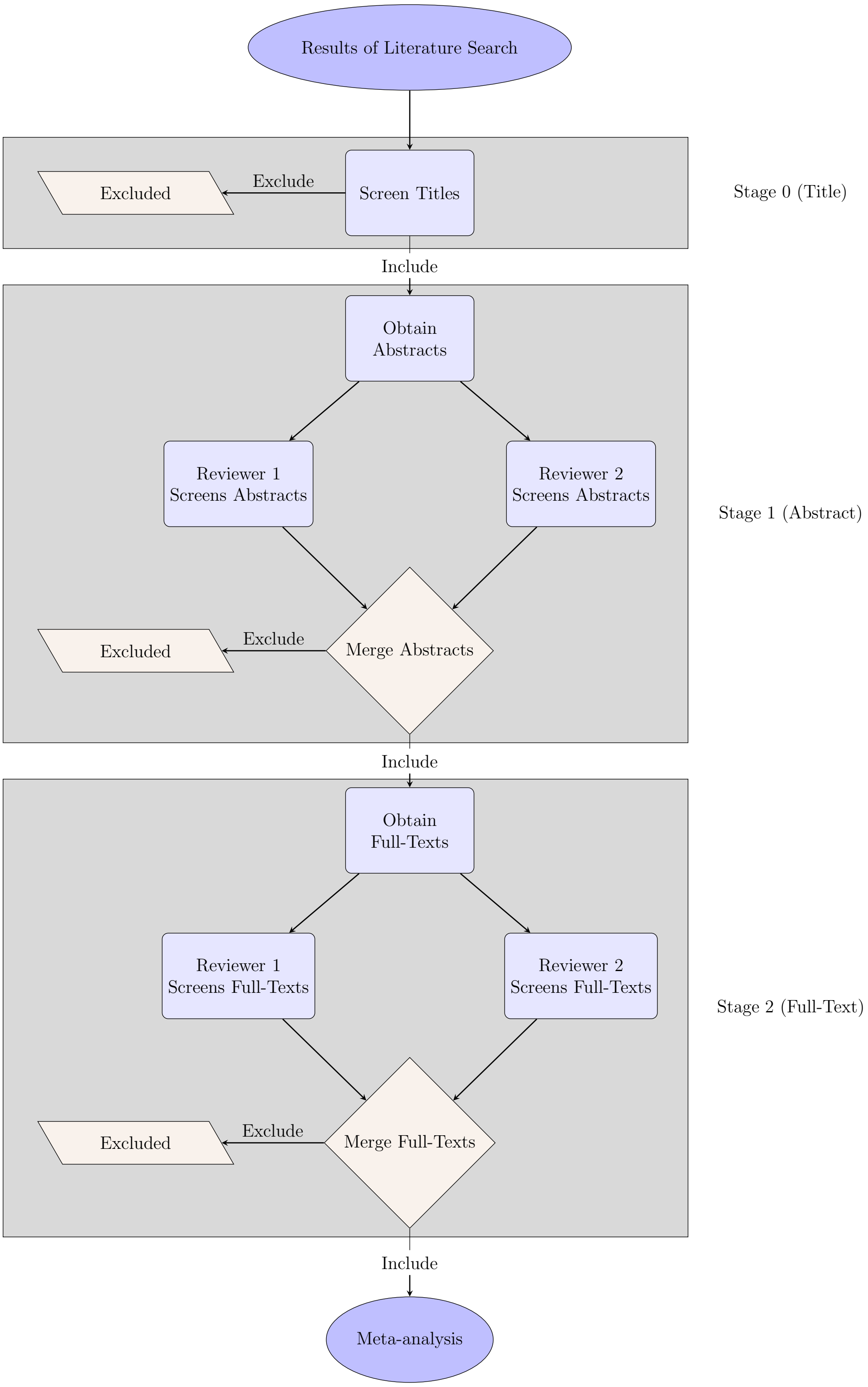


Figure 2.4: Overview of the literature screening process for systematic reviews.

2.2.2 Systematic Reviews as a Classification Task

When considered as a classification task, systematic review literature searches have several defining characteristics. Perhaps most important is the extremely high cost associated with false negatives (citations that are incorrectly excluded) [88]. This means that any classifier must have perfect recall on relevant citations (or at least have recall equal to that of a human reviewer).

In addition to the above restriction on recall, classification for systematic reviews must deal with highly imbalanced training data. It is not unusual for a literature search to consider thousands (or even 10's of thousands) of citations to identify less than 100 relevant studies [67]. Imbalanced training data is an established problem within the machine learning community [69]. The relative lack of relevant citations with which to train any prospective classifier can have the effect of biasing the model toward non-relevant studies, which works to directly counter the requirement for high recall on relevant studies.

In addition, the question of how best to incorporate classification into the systematic review process must be considered. The traditional supervised classification paradigm requires a human oracle to annotate a set of training instances which are then used to build a classifier for use on an unlabelled test set. However such an approach does not easily fit the existing systematic review literature screening process outlined in section 2.2.1.3. Ideally, any application of machine learning techniques to systematic review literature screening should minimise disruption to the existing process. There has been significant research into various possible solutions, including active learning [102], ranking query results [50, 49] and adjusting the process to account for separate training and test phases [32].

We note that in addition to classification of studies as relevant or otherwise for systematic review literature searches, machine learning has been applied to a number of other systematic review problems. For example, identification of high quality citations [8], estimating risk-of-bias [71], and automated assignment of MeSH headings [97]. In this thesis we focus on the topic of classifying citations as relevant or otherwise to a given review, and as such we do not present such work. The interested reader is directed to the relevant literature for such a summary.

The remainder of this section is divided into three areas. Firstly, a summary of existing work seeking to address the requirement for near perfect recall under the conditions of highly imbalanced training data is given. Approaches for including automated classification into the screening process are then considered, and finally a summary of methods for evaluating prospective classifiers is given.

2.2.2.1 High Recall Classification with Data Imbalance

Existing literature has typically used standard machine learning algorithms with slight modifications to allow for tuning to bias toward the minority class. Classifiers such as neural networks [22], support vector machines [102, 17], and Bayesian classifiers [65] have been modified and evaluated. It is worth noting that without modification, the above algorithms all struggle to adequately model the target class, due to the high rates of imbalance and high cost associated with recall when compared against precision [21].

The first work to directly address the systematic review screening problem was that of Cohen et al. [22] with their modified voting perceptron algorithm. A perceptron classifier classifies instances using a linear function of their attributes, and are typically built by modifying individual feature weights for misclassifications in the training data. They extend the algorithm proposed by Freund and Schapire [31] by introducing a learning rate parameter, where errors in training are penalised differently depending on whether they are false negatives or false positives. They found that by employing a 20:1 ratio in learning rate between false positives and false negatives, they could achieve a desired recall of 95%.

Matwin et al. [65] employ a similar approach, but work build upon the naive Bayesian classifiers rather than voting perceptrions. Their factorised complement naive Bayes (fCNB) algorithm is an extension on the complement naive Bayes classifier of Rennie et al. [84]. Complement Naive Bayes classifies using a Bayesian framework, building a separate model of each class and then assigning the class which maximises the posterior probability. The fCNB algorithm of Matwin et al. [65] introduces a weighting parameter, which is used to bias the classifier toward the minority class of relevant citations.

Cohen's voting perceptron and Matwin's fCNB were both evaluated on the same 15 reviews. Interestingly, neither classifier was found to consistently outperform the other, with both algorithms attaining superior performance on at least one of the tested reviews [18, 66]. Both algorithms however struggled at extremely high rates of imbalance.

Although focused on the separate task of work prioritisation, Cohen et al. [17, 19] also proposed an approach based on support vector machines. Support vector machines are a widely used class of classification algorithm which classify instances using a separating hyperplane [24]. They can efficiently handle non-linear classification boundaries via the use of kernel functions to map instances into high dimensional features spaces. Cohen et al. [17, 19] built their model using a linear kernel, and ranked citations by their relevance to the review using the signed margin distance.

Although intended for the separate task of work prioritisation, the performance for review classification is evaluated and compared to the above fCNB and voting perceptron algorithms by Cohen [18]. Similar to the comparison of fCNB and the voting perceptron, each algorithm outperformed the others on at least one tested review.

Wallace et al. [102] also employ a SVM based approach, but in contrast to Cohen et al. [17, 19] who focus on work prioritisation they use an active learning approach to train and apply their classifier. Rather than generating a separate training and test set a-priori, active learning instead relies on an iterative process where the algorithm selects unlabelled instances for annotation by a human oracle. Active learning shows much promise as an approach to systematic review literature classification, and has been used in several recent works [72]. The algorithm typically selects those instances for which labels would provide the most information, thus maximising the efficiency of time spent on annotation by human experts.

The algorithm proposed by Wallace et al. [102, 103] builds upon the SIMPLE strategy for selecting unlabelled instances proposed by Tong et al. [95]. The SIMPLE algorithm selects at each stage those instances closest to the current hyperplane; in other words those about which it is most uncertain. Wallace indicates that such a strategy can be inappropriate for domains such as systematic reviews where there is a high discrepancy in misclassification cost between classes. This is due to the algorithms sensitivity to its initialisation, as it then becomes unlikely that interesting regions far from the initial hyperplane will be identified. Wallace mitigate for this by first employing a random sampling of instances for labelling before employing a more traditional active learning approach to refine the decision boundary.

Within the literature on imbalanced data, a common approach is to resample the training set to generate a more balanced sample. This has been shown to improve performance in modelling minority classes, and has been been explored in the context of systematic reviews [103, 89]. Wallace et al. [103] employ undersampling in their active learning approach described above. Due to the high cost of misclassification for relevant studies, they propose a variant called *aggressive undersampling*. Aggressive undersampling instead selects those majority instances closest to the hyperplane for removal, rather than randomly sampling from the majority class.

2.2.2.2 Incorporating Classification into the Review Process

Traditional supervised machine learning is based on providing an algorithm with a set of labelled training instances, which are used to build a mapping function from unlabelled instances to labels. The systematic review process (in its current form) does not easily support such a paradigm, as no source of labelled training data exists before a review begins. However in order to effectively support the application of machine learning in systematic reviews, such an approach must be developed.

Existing literature can be divided into two schools; those which attempt to modify the annotation process to generate a set of training data, and those which take a less obtrusive approach based on active learning. In the first case, Frunza et al. [32] describe an approach based on having authors manually screen some percentage of all citations, which are then used as training data to build a classifier to be run on the remaining articles. Given a set of unlabelled abstracts under consideration for inclusion in the review, two separate authors annotate some common subset, which is then designated as training data. The trained classification model is then used in place of one of the authors, reducing the overall manual workload. A feature of this approach is that all citations are considered by at least one human reviewer, which helps minimise the potential for error. Such an approach fits well with the traditional supervised machine learning paradigm, and provides a way to include work which has been evaluated using the classical testing set / training set approach (e.g. the work of Cohen et al. [22, 18]).

Wallace et al. [102, 103, 101] describe an active learning approach, where the classifier is built in an iterative process. Their algorithm has been integrated with the publicly available Abstrackr annotation tool [101], an online citation screening tool which has previously been used in systematic reviews [82]. Here the algorithm particularly selects those citations for which manual annotation would provide the greatest improvement, until classifications can be made with sufficient confidence. Although the classification tool is still in beta and not available for public use, existing evaluations have shown much promise [83], although small error rates do exist.

Finally, it is worth noting that work exists addressing the similar task of identifying studies to update existing reviews [22, 20]. The review update task is similar to classification for the initial review, however it fits much better with the traditional classification model in which separate training and test sets are used (i.e. annotations from the original search can be used to train the classifier for the update task). It does suffer a separate challenge however of changing vocabularies, and difficulty modelling ground-breaking research which departs significantly from studies included in the original version of the review.

2.2.2.3 Evaluating Performance for Systematic Reviews

Traditional machine learning algorithms can be evaluated in terms of metrics such as precision, recall, or their mean value (f-score). A feature of classification for systematic reviews is the desire for perfect (or near perfect) recall over the class of relevant citations. As such there is a much greater value placed on recall over relevant citations when compared to precision [103, 83]. Due to this, evaluating using a single measure such as the f-score is inappropriate.

Many researchers take the approach of separately reporting metrics for workload saving and recall over relevant citations [83, 103]. This is the approach taken in our work. Within this framework, many individual metrics are used, such as precision, workload (analogous to the percentage of citations remaining), burden [103] (similar to workload but accounting for multiple reviewers), recall, and raw number of errors [83].

Cohen et al. [22] have also proposed the Work Saved Over Sampling at 95% metric (WSS95) for use evaluating systematic reviews. Under the assumption that an acceptable error rate on the class of relevant citations is 5%, they note that a similar error rate could be achieved by simply randomply sampling 95% of the citations to be screened. They propose that workload savings (measured in terms of the number of citations to be screened) should be evaluated with respect to this baseline. WSS@95 has been used by a number of researchers as an alternative to reporting raw workload savings [22, 65, 47]. WSS is defined as follows (where TN, FN, and N are the true negative, false negative, and raw number of citations respectively):

$$WSS@95 = (TN + FN)/N - 0.05$$

A major issue for evaluating screening algorithms is obtaining data on which to measure performance. Many authors use private data (for example, Wallace et al. [103] and [83]). Alternatively, many authors choose to evaluate using a set of 15 reviews originally used by Cohen et al. [22] which are publicly available for download.

The Cohen datasets comprise the PubMed ids, along with reviewer annotations for citations from 15 separate reviews. Reviewer annotations indicate whether or not a citation survived screening based on abstract, as well as the final triage status of the study. More detailed reasons for exclusion exist for some studies removed during later screening stages (for example, wrong population, or not available in English).

In order to maximise the ability to reproduce results and compare algorithms, the Cohen datasets include only citations which can be found in the TREC 2004 Genomics Track document corpus (a subset of MEDLINE). Citations that did not match a PMID in the document corpus were removed. This is important to note, as were the searches to be rerun and additional citations combined with those already in the dataset, it would be unclear as to where in the screening process they were removed.

Chapter 3

Challenges in Literature Screening for Systematic Reviews

Systematic reviews form a key component of modern evidence based medicine. For a given clinical question, their purpose is to provide a review and often meta-analysis that is as unbiased, comprehensive, and repeatable as possible. Systematic reviews are widely used throughout the medical field to guide policy and practice, and are widely considered as the highest form of medical evidence [110].

In order to meet the stringent requirements of unbiased, repeatable, and comprehensive literature searches, systematic reviews are necessarily very time consuming. Potential consequences for omitting relevant studies are severe [88], and review authors will go to great lengths to avoid such an event. Overwhelmingly, literature searches for systematic reviews reviewer are still conducted manually, with authors often screening thousands (if not more) of studies. Timescales of months or years are not uncommon. More detail on the process of conducting literature searches for systematic reviews is presented in section 2.2.

In recent times, the machine learning community has shown an increased interest in improving the efficiency of literature searches for systematic reviews. This is a difficult problem with numerous idiosyncrasies and challenges. When viewed from the perspective of a machine learning problem, systematic review literature searches have the following two key characteristics:

- Classifiers must identify relevant studies with perfect (or near perfect) recall. This is due to the potential consequences should a reviews conclusions be drawn on incomplete information.
- Highly imbalanced training data. Often thousands of studies will be screened with only a handful of studies included.

Evidently, satisfying these requirements is no simple task. We provide an overview of the existing work on classification for systematic review literature searches in section 2.2.2.

Despite the challenges mentioned above, recent studies have demonstrated that very high performance is attainable [83, 76]. However it must be noted that classification for systematic reviews is by no means a solved task, with issues such as particularly high levels of data imbalance a concern [18, 66, 83].

Traditionally, the majority of systematic reviews have focused on clinical questions related to interventions. However the last decade has seen a substantial increase in demand for reviews related to other types of clinical questions (such are diagnostic test accuracy or prognostic reviews) [79]. The expanding scope of systematic reviews has created a number of challenges. For example, literature search and screening for diagnostic reviews are widely considered more challenging than those of interventions [58].

While the medical community has noted a number of challenges facing authors of DTA reviews [26, 58], there has been no formal analysis on the differences between reviews of diagnosis and interventions when considered as a classification problem. We provide such an analysis in this chapter. We demonstrate that diagnostic reviews form a particularly challenging subset of the systematic review classification task. In particular, we hypothesise and obtain evidence for three specific challenges inherent to DTA reviews when compared against reviews of interventions:

- Diagnostic reviews require more citations to be screened.
- Diagnostic reviews have a broader (more heterogeneous) target class.
- The quality of meta-data for diagnostic studies in common databases is generally lower.

By doing so we aim to demonstrate that classification for diagnostic test accuracy reviews forms a particularly challenging subset of the general systematic review classification task. These findings are then used as a motivation for work in later chapters.

3.1 Specific Challenges in the Systematic Review Process

As discussed in section 2.2 the data imbalance problem for systematic review classification is well established. While efforts have been made to address this issue, there is still much progress to be made [18, 66]. In particular, attention needs to be directed at subsets of the systematic review classification problem with the greatest level of imbalance.

Traditionally, systematic reviews have focused on questions related to medical interventions, however the last few decades has seen increased demand for reviews from other areas (i.e. aetiology, diagnosis, prognosis, etc.). In particular, there has been a substantial increase in demand for reviews of diagnostic test accuracy (DTA) leading to the formation of the Cochrane diagnostic test accuracy working group in 2003.

This section outlines our three hypotheses regarding technical challenges in classification for DTA reviews. These hypotheses relate to differences in the literature search process between systematic reviews of DTA and treatment (for a summary of the various literature screening stages see section 2.2.1.3). Hypothesis

A relates to the screening process as a whole, while Hypotheses B and C relate to stage 2 and stage 1 screening respectively.

We also describe one or more expected manifestations for each hypothesis. By testing for the existence of these manifestations in real data, we then obtain evidence to support the hypothesised challenges.

3.1.0.1 Hypothesis A: Workload

A major practical issue when conducting systematic reviews is the workload generated by the volume of citations needing to be screened. Most IR research for systematic reviews has focused specifically on how to deal with the very high rates of class-imbalance caused by this volume of data. Substantial progress has been made, however it can by no means be considered a solved problem.

We hypothesise that the number of citations to be screened at each stage of the literature search process is higher for DTA reviews than for those of the treatment. This increases the already large class-imbalance between the number of included and excluded studies, thereby again increasing the difficulty of what was already very challenging. Assuming this to be true, one could then expect the following manifestations:

- The mean number of search results to be screened will be higher for DTA reviews than for those of treatment.
- The mean number of full-text articles to be screened will be higher for DTA reviews than for those of treatment.
- The number of included studies as a percentage of the number of full-text articles screened would be lower for DTA reviews than for treatment.

3.1.0.2 Hypothesis B: Target Class Heterogeneity

The relative heterogeneity of what exactly constitutes a DTA study can be problematic when screening literature for DTA reviews. Quoting from Whiting et al.

[107], diagnostic test accuracy studies “are heterogeneous, exploring a range of diagnostic techniques and strategies, and are likely to have been conducted using a variety of methods”. In addition, there are examples (such as some cohort studies) where one could derive sensitivity and specificity despite the authors not having explicitly calculated them. The ideal DTA filter should be highly sensitive and would include studies such as these.

We hypothesise that due to this increased difficulty, the percentage of irrelevant citations that cannot be identified on title and abstract alone will be larger for DTA reviews than for treatment. Assuming this to be true, we can expect the following manifestations:

- The mean number of full-text articles to be screened will be higher for DTA reviews than for those of treatment.
- The number of included studies as a percentage of the number of full-text articles screened would be lower for DTA reviews than for treatment.

Intuitively, if a given study type is more challenging to identify than another, it can be expected that an author would need to expend greater effort on discerning similar studies. This increased effort could take the form of additional time to screen individual citations, or screening more citations in greater detail (i.e. examining the full-text article). Due to the high cost of false negative classifications, it is reasonable to assume that any ambiguity in the initial screening stage would be resolved by obtaining the full-text article rather than putting more effort on the title and abstract. As such, assuming DTA studies to be inherently more challenging to identify than randomized controlled trials, we would expect to observe more full-text articles being screened when conducting DTA reviews.

3.1.0.3 Hypothesis C: Suitability of Metadata

Appropriate use of high quality metadata (i.e. MeSH terms for MEDLINE) in literature searches is crucial to generate a manageable number of citations while still remaining confident that no relevant ones would be omitted. It is common to identify thousands of citations at this stage. It follows that as the quality of the available metadata decreases, the total number of citations one would need to screen to maintain this confidence would increase.

It has been noted within the literature that the metadata in many medical databases are more suited to describing concepts related to treatment as opposed to diagnosis [36]. For example, while high quality MeSH terms exist for study types such as randomized controlled trials, the same cannot be said for studies of diagnostic test accuracy. From Whiting et al. [107]: "Although MEDLINE includes a number of medical subject headings (MeSH) that capture outcome measures used in test accuracy studies (e.g. sensitivity and specificity), these terms are not specific to test accuracy studies and are inconsistently applied by indexers".

We hypothesise that the quality of metadata is typically lower for DTA reviews than for treatment. Therefore we can expect the following manifestations in literature searches for systematic reviews:

- The mean number of search results to be screened will be higher for DTA reviews than for those of treatment.
- The number of full-text articles retrieved as a percentage of the total search results would be lower for DTA reviews than for treatment.

3.2 Evaluation

We have identified five expected manifestations of the stated hypotheses on the literature searches for DTA reviews. For clarity, each manifestation and its associated hypothesis are shown in Table 3.1. In order to test these claims, summaries of the literature search and screening stages were extracted from a sample of DTA and treatment reviews. Data collected included the number of references retrieved by the original search (SR), the number of references for which full-text papers

Manifestation	Description	Hypothesis A: Increased Workload	Hypothesis B: Increased Target Class Heterogeneity	Hypothesis C: Decreased Suitability of Metadata
FT	The mean number of full-text articles screened would be higher for DTA reviews than for treatment	Yes	-	-
SR	The mean number of search results would be higher for DTA reviews than for treatment	Yes	-	Yes
INC/FT	The number of included studies as a percentage of the number of full-text articles screened would be lower for DTA reviews than for treatment	-	Yes	-
INC/SR	The number of included studies as a percentage of the total number of search results would be lower for DTA reviews than for treatment	Yes	-	-
FT/SR	The number of full-text articles retrieved as a percentage of the total search results would be lower for DTA reviews than for treatment	-	-	Yes

Table 3.1: List of expected manifestations (differences between DTA and treatment reviews) for all hypotheses.

were screened (FT), the number of references included in the final meta-analysis (INC), and the paired ratios between each of the collected statistics.

Systematic reviews can be conducted and reported according to varying standards of rigour. This could be problematic for the purposes of our evaluation, as ideally the variation between two samples should be restricted to one review type (i.e. DTA or treatment). For systematic reviews published by the Cochrane collaboration, authors are required to follow strict guidelines outlined in the Cochrane handbooks for treatment and DTA reviews [36, 35]. Reviews published by Cochrane are widely regarded as meeting very high procedural and reporting standards, and their published guidelines for reviews of DTA and treatment contain a number of shared protocols. As we wish to restrict differences between the samples to whether the reviews are of treatment or DTA, the analysis reported in this chapter is performed exclusively on a subset of the Cochrane database.

As of the search date (2013/7/12), Cochrane had published 13 complete systematic reviews of DTA (one from each of the acute respiratory infections [ARI], airways, back, bone, joint, and muscle trauma [BJMT], eyes and vision, gynecological cancer, pregnancy, renal, and stroke Cochrane review groups [CRG], two from the infectious diseases CRG, and three from the Back CRG). A copy of each DTA review was obtained. For each DTA review, 15 non-withdrawn treatment reviews were selected at random from those published by the corresponding CRG. The number of treatment reviews was chosen to provide a sufficiently large sample for statistical analysis while requiring a reasonable time for collection and data extraction. Stratifying the data in this way was intended to account for any variation in search procedures across CRGs, as well as the availability of data within each field generally. A summary of the number of selected treatment reviews for each CRG is presented in Table 3.3. A list of each selected diagnostic and treatment review is included in the Appendix A. The desired statistics were then manually scraped from the values reported in the literature search summary from each review.

Cochrane review groups	DTA reviews	Treatment reviews
Acute respiratory infections	1	15
Airways	1	15
Back	3	45
Bone, joint, and muscle trauma	1	15
Eyes and vision	1	15
Gynecological cancer	1	15
Infectious diseases	2	30
Pregnancy	1	15
Renal	1	15
Stroke	1	15
Total	13	195

Table 3.2: Summary of the total number of DTA and treatment reviews randomly selected for inclusion in our analysis, ordered by CRG.

It is important to recall that depending on the specific conditions of each review (DTA or treatment) changes in the search process may be made to find the desired balance between search sensitivity and reviewer workload. Using the values reported by the reviewers (as opposed to manually re-running searches, possibly with the inclusion of more or less sensitive filters) had the added benefit of taking into account the review authors conclusions for the specific domain of each review.

Not all reviews reported the number of citations obtained at each stage of the literature search (e.g. some would report only the number of included and full-text articles). Where values were missing or unclear, we made an attempt to contact the review authors by email. If no data could be obtained, a blank value was recorded and the review would be omitted from analyses involving the missing statistical data. For computational reasons, extracted values equal to 0 were also omitted. A summary of the number of extracted values for all data types is given in Table 3.3. For example, of the 195 randomly selected treatment reviews, the number of full-text articles examined could not be obtained from 62 reviews, hence the number of collected data points for the number of full-text articles in treatment reviews is 133 (as reported in row 2 of Table 3.3).

Based on prior experience, we expected that the number of reported studies for the literature searches would be heavily skewed. This expectation is supported by comparing the mean and median values for each of the statistics from the collected treatment reviews (see Table 3.4); for 5 out of the 6 statistics the mean is approximately twice the value of the median. For example, the number of reported search results collected includes a number of values describing unusually large literature searches. Such values significantly affect the skewedness of the collected data, substantially increasing the mean without affecting the median.

In order to compensate for the level of skewness, all reported statistical comparisons are performed using an unequal variance t test on ranked data (i.e. as an approximation to a non-parametric test); each individual data point is replaced

	DTA	Treatment
DATA_{INC}	13/13	186/195
DATA_{FT}	12/13	133/195
DATA_{SR}	13/13	101/195
$\text{DATA}_{\text{INC / FT}}$	12/13	126/195
$\text{DATA}_{\text{INC / SR}}$	13/13	95/195
$\text{DATA}_{\text{FT / SR}}$	12/13	92/195

Table 3.3: Summary of the sample sizes (number of reviews reporting nonzero values) for evaluating each of the expected manifestations.

	Mean	Median	Mean / Median
DATA_{INC}	19.56	11.0	1.78
DATA_{FT}	71.89	33.00	2.18
DATA_{SR}	1799.04	900.00	2.00
$\text{DATA}_{\text{INC / FT}}$	0.394	0.357	1.11
$\text{DATA}_{\text{INC / SR}}$	0.033	0.013	2.47
$\text{DATA}_{\text{FT / SR}}$	0.099	0.046	2.13

Table 3.4: Ratio between mean and median for collected treatment reviews.

	Mean_{DTA}	$\text{Mean}_{\text{Treat}}$	$\text{Mean}_{\text{DTA}}/\ \text{Mean}_{\text{Treat}}$
DATA_{FT}	191.92 (n=13,s=233.51)	71.89 (n=133,s=154.76)	2.67
DATA_{SR}	5144.23 (n=13,s=4109.78)	1799.04 (n=101,s=2530.11)	2.86
$\text{DATA}_{\text{INC / FT}}$	0.191 (n=13,s=0.11)	0.394 (n=126,s=0.24)	0.49
$\text{DATA}_{\text{INC / SR}}$	0.021 (n=13,s=0.036)	0.033 (n=95,s=0.049)	0.63
$\text{DATA}_{\text{FT / SR}}$	0.087 (n=13,s=0.124)	0.100 (n=92,s=0.156)	0.87

Table 3.5: Summary of mean values for collected statistics.

	Mean_{DTA}	$\text{Median}_{\text{DTA}}$	$\text{Mean}_{\text{Treat}}$	$\text{Median}_{\text{Treat}}$
DATA_{FT}	110.67 (n=12,s=27.64)	113.0	68.51 (n=133,s=41.16)	67.0
DATA_{SR}	85.54 (n=13,s=27.84)	94.0	52.76 (n=101,s=31.62)	52.0
$\text{DATA}_{\text{INC / FT}}$	35.67 (n=12,s=24.69)	29.0	71.63 (n=126,s=39.60)	73.5
$\text{DATA}_{\text{INC / SR}}$	40.54 (n=13,s=31.12)	35.0	55.27 (n=95,s=30.76)	56.0
$\text{DATA}_{\text{FT / SR}}$	47.5 (n=12,s=30.18)	45.5	52.02 (n=92,s=29.97)	53.5

Table 3.6: Summary of ranked data for collected statistics.

by its index in the sorted set of data. If multiple data points shared a common value the ranked values were averaged. Summaries of the unranked and ranked data are presented in Table 3.5 and Table 3.6.

To further illustrate the ranking process, the mean number of search results obtained (as reported in Table 3.5) was 5144.23 for DTA reviews and 1799.04 for treatment reviews. When the 13 DTA and 101 treatment data points were combined and sorted however, the mean position for DTA reviews was 85.54 and that for the treatment reviews was 52.76 (as reported in Table 3.6).

	Hypothesis A: Workload	Hypothesis B: Target class heterogeneity	Hypothesis C: Suitability of Metadata
Total articles screened	Increase $5144.2_{DTA} > 1799.0_{TR}$ (P=.002)	-	Increase $5144.2_{DTA} > 1799.0_{TR}$ (P=.002)
Full-text articles obtained	Increase $191.9_{DTA} > 71.9_{TR}$ (P<.001)	-	Decreased as a % of total articles screened $0.087_{DTA} < 0.100_{TR}$ (P=.65)
Included Articles	Decrease as a % of total articles screened $0.021_{DTA} < 0.033_{TR}$ (P=.14)	Decreased as a % of full-text articles obtained $0.191_{DTA} < 0.394_{TR}$ (P<.001)	-

Table 3.7: Summary linking each hypothesis, expected manifestation, and literature screening stage.

3.2.1 Results

The results section is divided into one section for each of the three proposed hypotheses. Summaries of each hypothesis, along with the expected and observed manifestations are presented in Table 3.7.

3.2.1.1 Hypothesis A: Workload

Comparing the mean absolute number of the search results obtained we observe a 186% increase for reviews of DTA when compared to reviews of interventions (5144.2 vs 1799.0). There was strong evidence that this difference was statistically significant (unequal variance t test on ranked data, P=.002). Similarly for the mean number of full-text articles obtained we can observe an increase of 167% (191.9 vs 71.9). Again, there was very strong evidence that this difference was statistically significant (unequal variance t test on ranked data, P<.001).

We note not only the statistically significant difference in means, but also the substantial difference in effect size. The magnitude of the difference supports the claim that identification of relevant papers is noticeably more complicated for DTA reviews than for those of treatment, and also that there is an increase in difficulty both for authors and any prospective IR system.

Considering the number of included studies as a proportion of the total search results, a decrease of approximately 35% is observed for DTA reviews when compared to reviews of treatment (0.021 vs 0.033). However, despite the large magnitude of the difference there is insufficient evidence to claim statistically significance (unequal variance t test on ranked data, P=.14). However, caution is urged in concluding that no difference exists (see section 3.2.2).

3.2.1.2 Hypothesis B: Target Class Heterogeneity

Comparing the number of included studies as a percentage of full-text articles examined, an increase of approximately 106% is observed for DTA reviews when compared to those for treatment (0.191 vs 0.394). Very strong evidence was obtained that this difference was significant (unequal variance t test on ranked data, P<.001).

Again, we note the substantial difference in the observed effect size here. Its magnitude indicates the increased practical difficulty of screening a potentially relevant article for inclusion in a DTA review.

3.2.1.3 Hypothesis C: Suitability of Metadata

As stated in the results section for Hypothesis A, strong evidence was obtained to support an increase in the mean absolute number of search results obtained when comparing reviews of DTA and treatment (unequal variance t test on ranked data, P=.002). When looking at the number of full-text articles retrieved as a percentage of total search results, one can observe a decrease of approximately 13% for DTA reviews when compared to treatment reviews (0.087 vs 0.100). However, there is insufficient evidence to identify a statistically significant difference (unequal variance t test on ranked data, P=.65). As with the observed mean number of included studies as a percentage of search results, the caution is urged in concluding that no difference exists, and discuss possible reasons in section 3.2.2.

3.2.2 Analysis

As observed from the reported P values in Table 3.7, there is very strong evidence that the number of articles at each stage of the screening process is higher for DTA reviews than for those of treatment, in support of hypothesis A (and hypothesis C in the case of an increased number of raw search results). This demonstrates a significant increase in the required workload for systematic reviews of diagnostic test accuracy. In addition, very strong evidence is obtained in support of hypothesis B. However, the p-values obtained for both the number of included and full-text articles retrieved as a percentage of the total search results were insufficient to ascertain a statistically significant difference between the means for DTA and treatment reviews.

As reported in Table 3.5 and 3.6, the standard deviation for all results is quite large. In addition, our analysis is limited in that only 13 completed Cochrane DTA reviews existed as on the search date. This small sample size, combined with the large standard deviations results in relatively low power. There is a possibility that the negative results reported for the included and full-text articles as a percentage of total search results were type II errors. This possibility is enhanced by the relatively large magnitude of the differences in sample means (see Table 3.5). Of course, it is impossible to say for certain until more data is available.

The authors note that while the analysis does not support the claim of suboptimal metadata for DTA reviews, such a claim is not new and is supported by previously published literature. In addition to the lack of a definitive MeSH term for DTA studies, the Cochrane Handbook for reviews of DTA studies [36] notes that many index and reference tests employed during DTA studies have no corresponding MeSH term. From the handbook: a “database of names used to describe index tests and reference standards is being built”. However it is not complete as yet and due to the size of databases like MEDLINE and EMBASE, it is unlikely to be able to be applied retrospectively.

The reported results (summarized in Table 3.7) combined with the substantial difference in observed effect sizes lead us to conclude that the analysis supports the claim that DTA reviews present additional IR challenges. The magnitude of the difference in effect sizes is of particular importance as it implies a practical difference in the level of effort required for DTA and treatment reviews. We note the limitations of the study due to the small sample size of available DTA reviews. Further analysis needs to be done when more data is available.

It is interesting to note that the expected manifestations of hypothesis B (increased target class heterogeneity) could be said to drive the expected increase in workload during stage 2 screening described in hypothesis A. Similarly, hypothesis C (sub-optimal meta-data) could be said to drive the increased workload in stage 1. This provides an interesting guide to any future work on the application of classifications to diagnostic reviews; by addressing these challenges the comparative difficulty of DTA reviews can be reduced.

We would also like to note that the hypotheses discussed in this chapter could have additional manifestations throughout the review in addition to those in the literature search and screening stages. For example, the increased range of study designs and analysis methodologies (hypothesis B) could lead to increased difficulty in performing or interpreting any subsequent meta-analysis. As the focus of this work is the literature search/screening stages of DTA reviews (and due to the inability to observe such manifestations in our data) we do not consider such manifestations in our work, however such a study in future may be interesting.

3.3 Summary

We demonstrate an increase in practical difficulty when screening literature for DTA reviews as compared to treatment. In addition, some potential causes for this additional difficulty are presented. Three main conclusions can be drawn from this study with respect to diagnostic review literature searches:

1. The overall reviewer workload during literature screening is higher for DTA reviews than for treatment. This is evidenced by the larger number of citations obtained at each stage of the literature screening process.

2. The target class of studies included in DTA reviews is broader than the corresponding class for reviews of treatment. This is evidenced by the lower number of included studies as a percentage of full-text articles screened.

3. We obtain partial statistical evidence to support claims of the relative unsuitability of available meta-data for DTA reviews. We observe a significantly greater number of studies retrieved for screening, however do not observe a statistically significant difference when considered as a percentage of included studies.

Of particular note is the broader target class for studies of diagnostic test accuracy. As classification for systematic review literature screening requires perfect recall, it requires the classifier to generate a model encompassing the entire target class. However, due to the imbalanced data problem any classifier for systematic reviews will naturally have a tendency to bias toward the non-target class. As such, methods must be developed that are very robust to a heterogeneous target class despite imbalanced training data.

Despite specific examples of deficiencies in available meta-data for diagnostic studies (e.g. the lack of a MeSH term for diagnostic test accuracy studies [107]), we failed to find sufficient evidence of the corresponding manifestations in the study data. While we hypothesise that this is likely due to poor statistical power caused by the small available sample size, we obviously cannot draw any conclusions without supporting evidence. Such evidence would imply a greater level of class imbalance when classifying studies for diagnostic reviews when compared to reviews of interventions. A future analysis with higher statistical power would be of great interest.

Finally, we would like to highlight the utility of the results reported in this chapter for fields outside the domain of automated classification. As noted at the beginning of this chapter, while it is generally accepted that the literature search and screening stages for diagnostic reviews are more complex than for those of treatment, to the best of our knowledge this is the first formal study of the effects on overall workload. Such data could be of interest when authors are planning reviews, for example to aid in determining appropriate resource allocation to a project.

Chapter 4

Excluding Irrelevant Studies using Annotation from Earlier Screening Stages

Building classifiers for systematic reviews literature screening requires building extremely high recall classifiers with highly imbalanced training data. Despite this challenge, existing approaches have shown some promise in correctly identifying relevant studies [32, 83, 103, 100]. However authors such as Cohen [18] and Matwin et al. [66] have noted that improvements still need to be made for reviews with particularly few relevant studies. A recent study in 2015 by Rathbone et al. [83] evaluating Abstrackr [101, 100] also noted that while performance was generally very good, errors would still be made with certain reviews.

Imbalanced training data is an acknowledged problem when building classifiers for systematic review literature screening. In chapter 3, we noted that certain sub-fields of systematic reviews exist for which the data imbalance problem is even more pronounced. For one such class (diagnostic test accuracy reviews) we empirically validated this assertion. Among the causes for this include increased heterogeneity of the target class of relevant studies (with respect to the more traditional reviews of interventions), and relatively poor meta-data for studies in popular databases.

A requirement for any classifier used to screen studies for systematic reviews is that it achieve perfect (or near perfect) recall on the target class. In the former case, this amounts to building a highly accurate model of the entire class of relevant studies. Intuitively, we can see that the difficulty of this task increases with the heterogeneity of the target class.

This task is further complicated as the number of relevant studies decreases as a percentage of studies screened (due to the imbalanced training data problem). This decrease is a direct consequence of the relatively poor quality meta-data. Indeed, the aforementioned study by Rathbone et al. [83] attributes several studies misclassified by Abstrackr to high data imbalance. Clearly opportunities for further improvements exist. As the popularity of diagnostic (or indeed other challenging fields such as aetiology) reviews increases, automated screening systems will need to be improved to cope with the additional challenges.

This chapter proposes a novel approach for partial automation of systematic review literature searches. Rather than focus on the more challenging task of identifying relevant citations with perfect recall, we instead invert the task and attempt to identify subsections of irrelevant articles with perfect precision. Our approach utilises annotations made during an initial title only based screening. Although not universally applied, many authors choose to perform the additional screening stage with evidence suggesting it can improve the overall efficiency of the search [64]. We propose to leverage annotations made during the initial title based screening to remove additional articles prior to screening based on title and abstract. This not only integrates seamlessly with existing screening practices, but is complementary to existing work on automated screening such as that of Wallace et al. [101, 100]. As the classifier is trained using annotations made in previous screening stages, it could be applied prior to the application of an alternate system. To the best of our knowledge, this is the first time classifiers for systematic reviews have been trained using annotations from prior screening stages.

The rest of this chapter is structured as follows. Section 4.1 provides a review of the systematic review literature search process and requirements for classification algorithms. The effects of a heterogeneous target class on classification for systematic reviews is then discussed in section 4.2, and a recap of the multi-stage triage process used for literature screening is provided in section 4.3. These sections also provide proposals for modifying the literature screening process. These proposals are combined and a novel algorithm for semi-automated literature screening is presented and discussed in detail in section 4.4. Finally an evaluation of the proposed approach on real data is given in section 4.5, and conclusions are drawn in section 5.5.

4.1 Requirements for Automated Screening

In chapter 2 we discussed systematic reviews and the major issues concerning the automation of their literature searches. Two major requirements were raised that need to be considered by any prospective algorithms: the need to guarantee suitable recall over the class of relevant citations, and how the classifier should be trained and integrated into the literature screening process.

In the former case, any classification algorithm is required to achieve extremely high recall on relevant citations. Although potentially mitigated by the process with which assigned labels are used in excluding citations, there is a high cost associated with false negatives. This is made even more challenging due to issues such as imbalanced training data and broadly defined target classes (see chapter 3). Human reviewers are willing to spend substantial effort in manually screening citations. If a system misses relevant work then reviewers are unlikely to use it regardless of any expected workload savings [83].

In practice however, we note that even the best human reviewers make mistakes. Typically, this is mitigated by having multiple reviewers screen each individual citation. In the event that a prospective classifier were to be used in place of a single reviewer, then the goal instead becomes to replicate the recall of a human annotator.

Evaluating the performance of a human reviewer to generate a target recall is challenging. Reviewer error rates are likely to vary across clinical domain, as well as by reviewer experience and even individual reviews. We are aware of no comprehensive study to date which addresses this issue. As such, previous literature [21, 65] has often used a proxy value of 95%.

As mentioned above, the method with which the classifier will be included into the screening process can aid in ensuring recall is maximised. Although using a classifier as an exclusive annotator to screen a subset of citations has been considered [83], most researchers have considered classification as a potential replacement for a single reviewer [80, 33].

In addition to reviewer overhead, the method by which training data is obtained and employed will modify the screening process for reviewers. It is desirable that as little overhead as possible be required of human reviewers when using any prospective system. Additional annotation time, or modifications to established procedures can act as a barrier to practical use. While users would obviously be motivated to use new systems if sufficient workload savings can be achieved, minimisation of disruption to existing processes is still desirable.

Similar to the above goal of minimising disruption to existing processes is the desire for white-box, deterministic classifiers. A white-box classifier is one where the learned process for assigning labels to instances can be viewed and interpreted by a human user. A deterministic classifier is one where application of a given algorithm to identical sets of data will always produce the same result.

That human reviewers are willing to spend literally months to years screening citations to ensure all relevant work is included is indicative of the importance of accurate literature searches. As above, while reviewers can always be persuaded by sufficiently comprehensive evaluation, reviewers are reluctant to give up manual control. Again, white-box, deterministic classifiers are desirable.

We can now state a list of criteria against which to measure classification algorithms for systematic reviews, and which we use as goals in the following work. The list contains both major and minor criteria; major criteria are those which must be met, with minor being those which are desirable but not essential. Major Requirements:

- Perfect (or near perfect) recall over relevant studies must be achieved.

Minor Requirements:

- Classifiers should ideally minimise disruption to the existing process.
- Classifiers should ideally be white-box.
- Classifiers should ideally be deterministic.

4.2 Target-Class Heterogeneity

In chapter 3 we established the existence of sub-classes within the systematic review literature screening problem for which it is even more difficult than normal to model the target class of relevant citations. Specifically, diagnostic test accuracy reviews were found to include a broader range of studies when compared to more traditional reviews of treatment. Existing challenges such as the data imbalance problem already complicate accurately modelling relevant studies for systematic reviews. The increasing prevalence of diagnostic reviews necessitates development of methods for dealing with these additional challenges.

In this section, we propose to address concerns relating to heterogeneous target classes by exploiting the relationship between precision and recall for two class classification problems. Essentially, we recast the problem of modelling relevant

		Assigned	
		X	Y
Actual	X	a	b
	Y	c	d

Figure 4.1: Confusion matrix for two class classification task with labels X and Y.

citations with high recall to instead model irrelevant citations with high precision. We summarise this relationship in section 4.2.1. We then describe how we apply it to the problem of systematic review literature screening in section 4.2.2.

4.2.1 Precision vs. Recall for Two Class Classifiers

For two class classification problems, we consider the task of maximising precision on one class under the constraint of perfect recall. A duality exists between this task and maximising recall on the second class under the constraint of perfect precision. This can be seen using the following example.

Figure 4.1 shows a confusion matrix for a two class classifier, where instances are assigned as a label either X or Y. In this figure, the number of true positives for class X is denoted a, true negative as d, false positives c, and false negatives b. The precision and recall for each class can be computed as follows:

$$precision_X = \frac{a}{a+c} \tag{4.1}$$

$$recall_Y = \frac{d}{d+c} \tag{4.2}$$

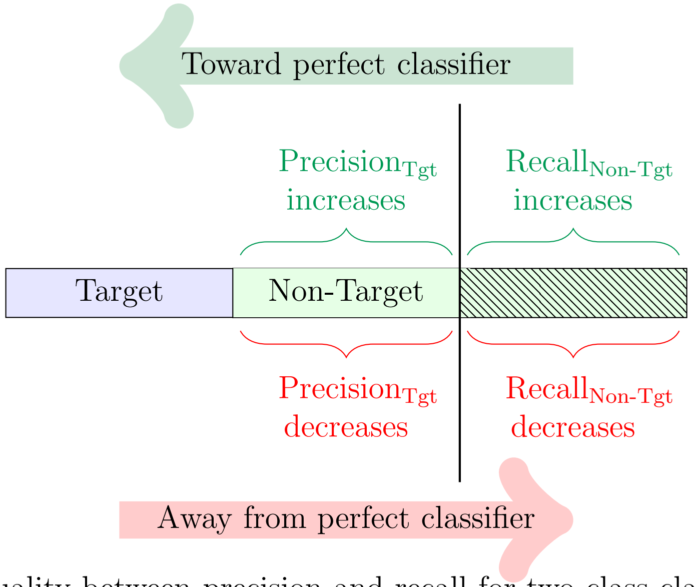

Toward perfect classifier
Precision$_{\text{Tgt}}$ increases
Recall$_{\text{Non-Tgt}}$ increases
Target
Non-Target
Precision$_{\text{Tgt}}$ decreases
Recall$_{\text{Non-Tgt}}$ decreases
Away from perfect classifier

Figure 4.2: Duality between precision and recall for two class classification problems.

$$recall_X = \frac{a}{a+b} \tag{4.3}$$

$$precision_Y = \frac{d}{d+b} \tag{4.4}$$

We can see that any errors in the precision for class X or recall for class Y arise solely from the number of instances in cell c. The requirement that precision for class X and recall for class Y be equal one is satisfied if and only if the value in cell c is 0. A similar observation can be made with cell b regarding precision for class Y and recall for class X.

Maximising precision for one class and recall for the other therefore requires minimising the same type of error. It follows that a classifier may have perfect recall on one class if and only if it has perfect precision on the other. Similarly, with every misclassification that decreases recall on the first class, precision on the second will also decrease.

We further demonstrate this relationship by considering Figure 4.2. Assume the existence of some hypothetical data with two underlying classes, represented by the horizontal rectangle. The left section of the rectangle (coloured blue) represents those instances whose ground truth label is the target class. The right section (coloured green) represents those instances belonging to the alternate, non-target class.

Our goal is to produce a classifier which models the target class with near-perfect recall, and maximises precision under this constraint. We can represent our classifier using a vertical line through the set of non-target instances. All instances to the right of the line are correctly classified as belonging to the non-target class. Instances to the left are classified as belonging to the target class.

The perfect classifier in Figure 4.2 is one where the vertical line is positioned directly over the boundary between target and non-target instances. As the line moves to the right, our precision over target instances decreases, as does our recall over non-target instances. If we bring the line back to the left, we increase both the precision over target instances, and recall over non-target instances. Moving the line to the left of the ideal boundary violates the requirement for perfect recall over target instances, and precision over non-target instances.

We note that for many problems the cost of violating the recall constraint on the target class far outweighs the cost of any decrease in precision. In such a case, we contend that it is attractive to model the non-target class with high precision.

Under a requirement for perfect recall on the target class, we can see that modelling the target class with high precision is equivalent to the task of modelling the non-target class with high recall. Errors in recall will be produced as we attempt to model the correct boundary between target and non-target instances. An alternate approach however is to attempt to generate high precision models of the non-target class by building classifiers for subsets.

Consider the shaded region of the non-target instances in Figure 4.2. This region is a subset of the non-target class. By modelling such a region, our errors are likely to fall around the boundary between this subset and other, non-target instances. False negatives do not impact the recall over the actual target class. Similarly, false positives are likely to still belong to the non-target class as a whole. Hence they are also unlikely to violate the condition of perfect recall over the target class.

4.2.2 Inverting the Classification Task

The relationship between precision and recall described above allows us to avoid modelling the entire domain of relevant citations. While modelling all non-relevant citations is an identical (and equivalently difficult) problem, we can focus instead on subsets of non-relevant studies. By identifying coherent groups of non-relevant studies and developing high precision models, we can then effectively and confidently prune citations without the need (or with reduced need) for human intervention.

This approach has some similarity to the idea of clinical queries filters [52, 59, 85], or classifying citations based on quality [8]. A key difference is that such filters allow a user to identify re-usable classes of studies (for example randomised controlled trials, diagnostic studies etc.) which are often used in screening for multiple reviews. Our aim is to develop a methodology for training domain specific filters for specific reviews. In addition, methodological filters are often used to limit search results based on requirements for relevant studies, being applied to keep only those studies meeting some pre-set requirement. As discussed above, our goal is to model non-relevant, rather than relevant citations.

The question of how to select subsets of irrelevant studies for classification must now be considered. One possibility would be to apply unsupervised learning to cluster a set of non-relevant citations, grouping them into conceptually separate topics for classification. One complication with this approach is that it requires human intervention to label a sufficiently large set of citations before the algorithm can be trained and applied. An alternative approach that has no such drawback is outlined in section 4.3.

4.3 The Multi-stage Literature Screening Process

Literature screening for systematic reviews is typically conducted as a multistage triage process (see section 2.2.1.3 for a more detailed treatment of the literature screening process). After a highly sensitive search over multiple databases is conducted, multiple reviewers will first screen all citations on title, then title and abstract, and finally full-text, removing those which can confidently be considered as outside the scope of the review at each stage.

The reasons for a study being excluded during screening can vary according to the stage at which they are removed. A study might be removed based on title due addressing the wrong population or condition (e.g. *allergic aspergillosis* as opposed to *invasive aspergillosis*). Reasons for removing studies at later stages of the screening process may include more detailed questions relating to study design or data reporting.

It is also possible that a study could be removed at a later stage of the review for reasons commonly applied at an earlier stage. A study may survive screening based on title due to ambiguity concerning the exact condition under investigation, but be readily excluded once the abstract is examined. For example, Methley et al. note that many titles omit important descriptive keywords [70].

The identification of studies which fail to be removed at prior screening stages is one that can be exploited in combination with our previously stated goal of modelling non-relevant studies with high precision. The remainder of this section outlines how annotations from prior literature screening stages can be leveraged to train a classifier for this purpose.

At the start of a given screening stage (with the exception of the first), we are in possession of the following data:

1. A set of citations labelled as not relevant to the review

2. A set of as yet unlabelled citations that will be removed for the same reason

as a study in the previous stage

3. A set of citations that will be removed for other reasons

4. A set of citations which will not be removed

The citations that have been labelled as not relevant were so labelled with less information than is now available. For example, at the beginning of the abstract screening stage, all removed citations must have been removed based on title alone. As discussed above, it is possible that citations in the second set will exist that are very similar to some of the already excluded citations. Rather than require human annotators to identify these studies manually, we would like to employ supervised learning techniques to do so automatically.

Although the labels for excluded citations were applied using title alone, the classification algorithm need not be bound by such a restriction. We propose that citations in the second set can be identified by training a classifier based on the labels applied in the previous stage, and representing citations using features derived based on title and abstract.

4.4 Classification Rules

This section outlines the algorithm used to generate the rule based classifier used to exclude irrelevant studies. The algorithm is broken down into three major steps, each of which is described in detail below. First, titles and abstracts are obtained for all citations, preprocessing is applied and features are extracted. Secondly, separate training and evaluation sets are generated. Finally the rule generation algorithm is applied to the training data to build the classifier. A graphical summary of this process is presented in Figure 4.3.

At several points in the rule generation process we note multiple possible choices that exist. These alternate courses of action are summarised in section

4.4.4. A sensitivity analysis examining the performance of the various approaches is performed in section 4.5.3, which is used to recommend a single approach.

4.4.1 Selecting the Training Data

In section 4.2 we discussed systematic review classification, and outlined the utility of inverting the problem and identifying irrelevant citations with high precision. Section 4.3 noted the failure of existing methods to adequately consider annotations from prior literature screening stages. We now show that these two observations can be combined to produce a novel approach toward semi-automation of the systematic review literature screening process.

As discussed in section 4.3, citations removed in a given screening stage can be divided into two groups. The first covers those citations that are removed for similar or identical reasons to citations removed in prior stages, but were not previously removed due to lack of evidence in title. The second is those citations that will be removed for reasons not previously used to prune citations. In the latter case, we have no way to identify such citations without input from a human oracle. In the first case however, we note that due to the annotations made in the previous stage we are already in possession of a number of citations which have been labelled as not being relevant to the topic of the review.

We propose to use this labelled data to train a classifier to exclude additional irrelevant citations. This process will work by training the classifier on the full set of citations screened in stage 0, with the target class being those citations that were excluded. However instead of extracting features exclusively from the titles (similar to the reviewers task in assigning the initial annotations), we extract features from both the title and abstract. Essentially, we aim to identify studies excluded for the same reasons as citations in stage 0 where the relevant features are present in the abstract. A graphical representation of this approach is outlined in Figure 4.4.

When training the classifier, it is important that the generated rules are not

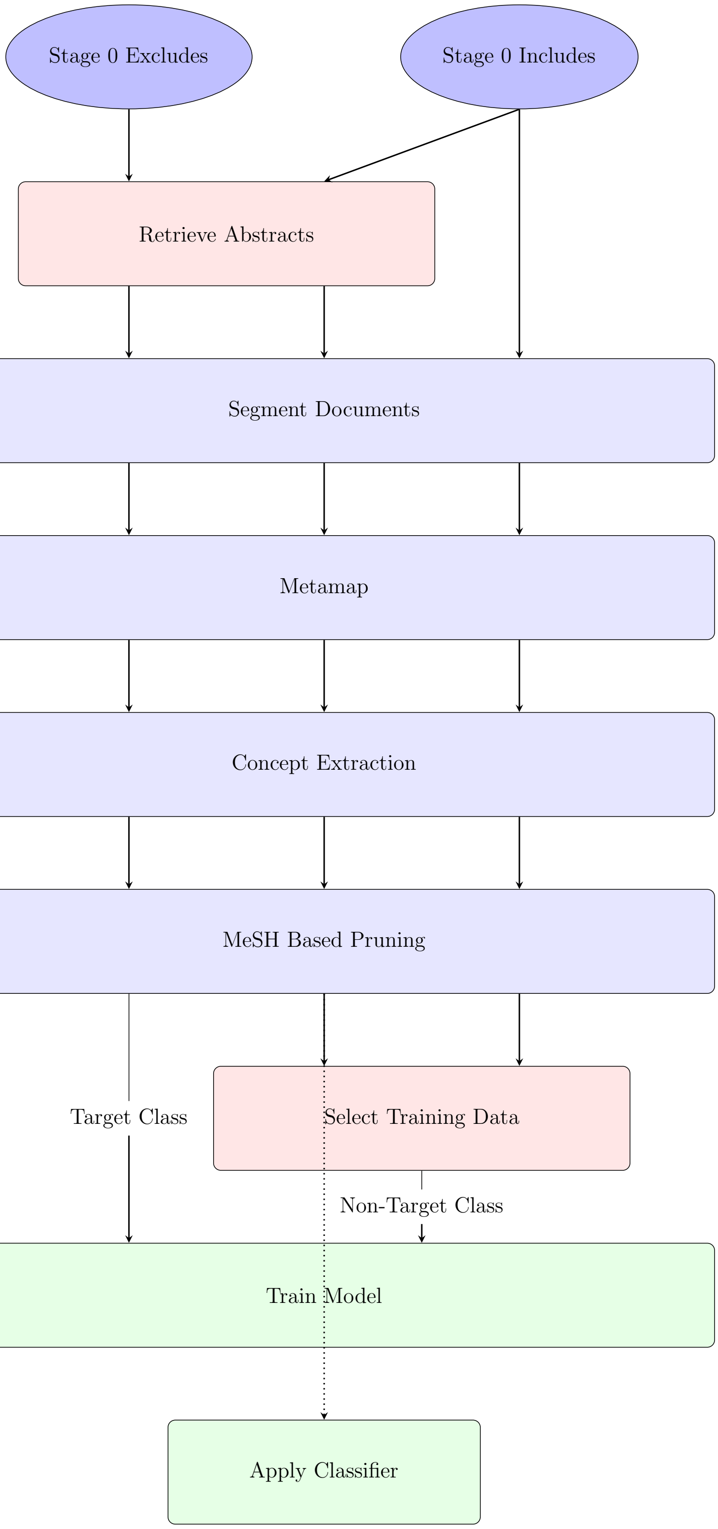


Figure 4.3: Summary of the rule generation process.

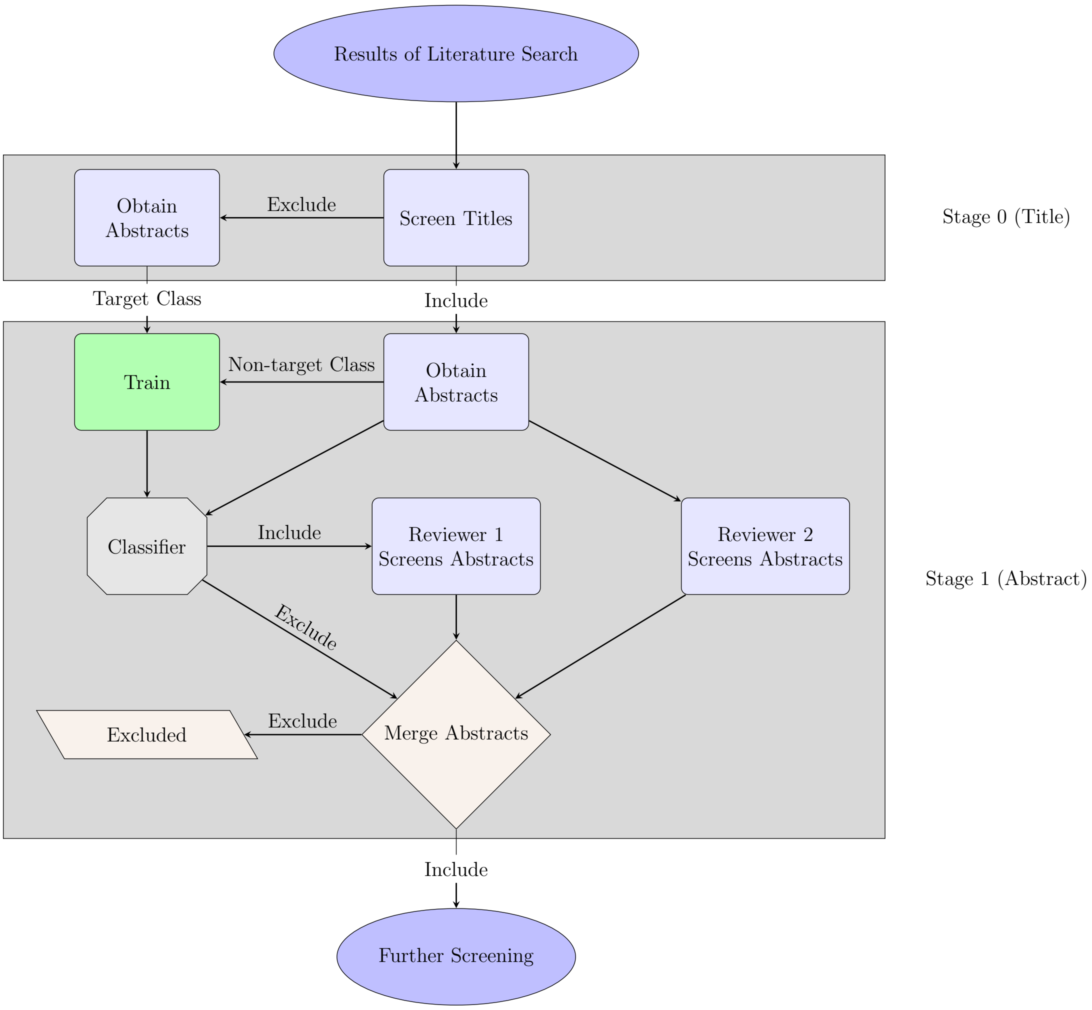


Figure 4.4: Modified screening process for stage 1 utilising annotations from stage 0.

so specific that they do not match any citations not removed based exclusively on title. Similarly it is crucial that they are sufficiently specific that they produce an acceptable error rate. The modified classification pipeline shown in Figure 4.4 requires two sets of citations to be input. The first is the training set, which consists of two subsets with citations labelled as either *irrelevant* or *potentially relevant*. The second is the test set, which is the citations that are to be screened by the classifier.

Two possibilities exist to guarantee that appropriate precision is achieved in identifying non-relevant citations. The first involves tuning precision as a parameter of the rule mining algorithm. If this approach is taken, then care must be taken both to select a sufficiently strict setting to avoid erroneously excluding relevant studies, and to select a sufficiently loose setting so as to produce a non-zero work saving.

Alternatively, we can train rules on a different set of data, adding more at the test stage to increase the probability of a match. Rule mining could be performed based on the title and abstract from excluded citations, but only the titles from citations which have not yet been excluded. By using this reduced training data and developing sufficiently precise rules, we would aim to generate a highly accurate model of non-relevant citations which could then be matched based on the abstracts of as yet unlabelled studies.

We note that removing data when training the classifier has the potential to negatively impact performance. Without sufficient counter examples, the model for the target class of non-relevant citations could be overly general. Whether or not a sufficiently precise model can be trained requires evaluation, which we perform in section 4.5.3.

4.4.2 Feature Extraction

We now present an overview of the preprocessing steps used to generate descriptive features from citations. As such, a rule might be created indicating a citation should be excluded if it contains the words *allergic* and *aspergillosis*. Consider the sample abstract from the Aspergillosis data shown in Figure 4.5 (which a human author considered relevant to the review). Such a rule would exclude this abstract as it contains both the terms *allergic* and *aspergillosis*. However all occurrences of *aspergillosis* co-occur with the term *invasive* (indicating relevance), and have no direct link to the single occurrence of the term *allergic*.

To address this, we consider whether or not citations should be segmented into individual sentences before features extraction. If we choose to work with citations

Mycoserological tests in the diagnosis of invasive ***aspergillosis***. [Polish] Introduction: Long lasting exposure to Aspergillus antigens may result in ***allergic*** diseases and, in immunocompromised persons, in deep infections. The lack of distinctive symptoms and signs of invasive ***aspergillosis*** as well as doubtful value of culture make mycoserological tests essential in the diagnostics of systemic ***aspergillosis***. Objective: The purpose of this article was: I) a retrospective evaluation of the results of mycoserological tests performed in patients suspected of systemic mycosis at our Department in the years 2002-2006. II) a comparison of various serological methods applied in the laboratory diagnostics of invasive ***aspergillosis***. Material and methods: A total of 1086 serum samples from patients suspected of generalized mycosis were tested using: indirect haemagglutination test (IHA), double diffusion (DD), and latex agglutination (LA). The diagnostic usefulness of those tests was compared with reported data concerning enzyme immunoassay (EIA). Results: Antibodies against Aspergillus were detected in 226 positive tests results (20.8%) in IHA (titre >=1:160) and in 264 positive ones (24.3%) in DD. Circulating fungal antigen galactomannan was detected considerably less frequently than antibodies - positive LA tests were obtained only in 50 out of 1086 (4.60%) serum samples. Conclusions: It appears that enzyme immunoassay (EIA) is the most useful mycoserological method in the early diagnosis of invasive ***aspergillosis*** because of its high sensitivity in detecting both fungal antigen and specific antibody against Aspergillus in serum. Now that it is not available at our institution, the best solution is simultaneous use of IHA, DD and LA. Copyright 2007 Cornetis

Figure 4.5: Sample abstract from the Aspergillosis data. Occurrences of the words *allergic* and *aspergillosis* are highlighted.

as entire documents, then a single set of features is extracted for a given title and abstract. When building the classifier this set of features is then presented to the training algorithm as a single instance. In the case where we segment citations into sentences, we extract a separate set of features for each sentence in a citation. Each set of features is then annotated with the label of the citation to which it belongs, and is presented to the training algorithm as a separate instance. When applying the complete classifier, a citation is said to match a rule if that rule matches at least one of its constituent sentences.

To tokenise sentences we use the sent_tokenizer method for NLTKs default tokeniser (nltk.tokenize.sent_tokenize). Our data contained several titles which contained multiple sentences. In this case titles were not tokenised (so all titles were treated as a single sentence).

Finally, we note that that segmenting citations into sentences will not necessarily guarantee an improvement in performance, and the effect of such a choice should be evaluated. The effect of segmenting citations into sentences or processing them as entire documents is examined as part of the sensitivity analysis in section 4.5.3.

4.4.2.1 Concept Extraction and Pruning

Instances (documents or sentences) are then transformed into feature vectors used to train the classifier. This was accomplished by mapping instances to relevant concepts from the UMLS ontology [14]. UMLS is a curated vocabulary of medical terms maintained by the US National Library of Medicine (NLM), containing over 2 million terms for 900,000 concepts.

To map instances to concepts we used the publicly available (under license) MetaMap tool [9]. Metamap has been widely used within biomedical classification literature [17, 39, 97]. It provides a highly configurable tool to parse biomedical text and map concepts into the UMLS ontology. Figure 4.6 shows a sample of fielded metamap output.

```
00000000—MMI—22.35—diagnosis aspect—C1704338—[qlco]—[”diagnosis”-tx-1-”diagnosis”-noun-0]—TX—67:9—x.x.x.x.x.x.x
00000000—MMI—16.18—Hypersensitivity skin testing—C0037296—[diap]—[”SKIN TESTS”-tx-1-”skin tests”-noun-0]—TX—49:10—E01.370.225.812.871;E05.200.812.871;E05.478.594.890
00000000—MMI—9.75—Lung Diseases, Fungal—C0024116—[dsyn]—[”pulmonary mycoses”-tx-1-”pulmonary mycoses”-noun-0]—TX—80:17—C01.703.534;C08.381.472;C08.730.435
00000000—MMI—3.53—immunological status—C0599818—[lbpr]—[”immunological status”-tx-1-”status immunologic”-noun-0]—TX—12:63̂3:11—
00000000—MMI—3.43—Current (present time)—C0521116—[tmco]—[”CURRENT”-tx-1-”current”-adj-0]—TX—4:7—
00000000—MMI—3.43—Electrical Current—C1705970—[npop]—[”Current”-tx-1-”current”-adj-0]—TX—4:7—
00000000—MMI—3.43—Serologic—C0205473—[ftcn]—[”Serologic”-tx-1-”serologic”-adj-0]—TX—22:9—
00000000—MMI—3.42—Diagnosis—C0011900—[fndg]—[”DIAGNOSIS”-tx-1-”diagnosis”-noun-0]—TX—67:9—E01
00000000—MMI—3.42—Diagnosis Study—C1704656—[resa]—[”DIAGNOSIS”-tx-1-”diagnosis”-noun-0]—TX—67:9—
```

Figure 4.6: Sample fielded MetaMap indexing output for the sentence ”The current status of serologic, immunologic and skin tests in the diagnosis of pulmonary mycoses.”

	Aspergillosis	Alzheimers
MeSH	3661	22035
UMLS	7158	12107

Table 4.1: Number of features generated using entire UMLS ontology vs. MeSH

UMLS contains concepts from a wide range of vocabularies, some of which are less useful for classification. Using the full UMLS ontology to generate features also complicates the rule generation process, as association rule mining generally scales poorly as the number of features is increased. In order to limit the number of features to a manageable size, we only use those features which are present in MeSH (the US NLM's vocabulary for indexing biomedical literature in MEDLINE). For two tested datasets (see section 4.5.1) this was found to reduce the number of generated feature by roughly two thirds. Table 4.1 shows the number of features computed using MeSH headings vs. the entire UMLS vocabulary.

For a single MeSH concept, it is possible that multiple possible terms could be used to indicate its relevance to a piece of text. For example, the high level MeSH code *Diagnosis* (E01) contains the following additional entry terms:

- Antemortem Diagnosis
- Diagnoses and Examinations
- Examinations and Diagnoses
- Postmortem Diagnosis

Although organised together as a single conceptual entity in MeSH, each of the above entry terms constitutes a separate concept in the UMLS metathesaurous. As we limit our analysis to use those concepts which appear in MeSH, we can account for the polysemous nature of UMLS concepts by merging entry terms where possible. However, this has the potential to cause errors where a single MeSH heading contains two entry terms that describe concepts we with to consider separately (for example, it may not be appropriate to merge postmortem and antemortem diagnosis). An evaluation of the effect of merging entry terms is reported in section 4.5.3.

4.4.3 Training the Model

Once features have been extracted and an appropriate training set created, we train the classifier. We have elected to use a rule based approach in our work. This is to meet the desire for a white-box approach as outlined in section 4.1. We note here that our approach is not specific to any single algorithm however, and the rule based approach employed in this work could be substituted for an alternative algorithm to model non-relevant citations if desired.

We identify rules using statistical significance based on Fisher's Exact Test as a measure of rule quality. Instances are treated as a bag-of-words over all extracted features, and association rules are mined between combinations of descriptive features and a label indicating whether a citation was excluded based on title.

We generate classification rules by applying Hämäläinen's Kingfisher [41] algorithm with a fixed consequent. The Kingfisher algorithm uses a branch and bound approach over the search space of possible rules. It computes lower bounds on the P-value for rules at a given node, and employs an efficient method of propagating knowledge about when subtrees can be pruned to efficiently generate rules. A more detailed treatment of rule mining and the Kingfisher algorithm is given in chapter 5. A modified version of the algorithm is also presented in section 5.3.

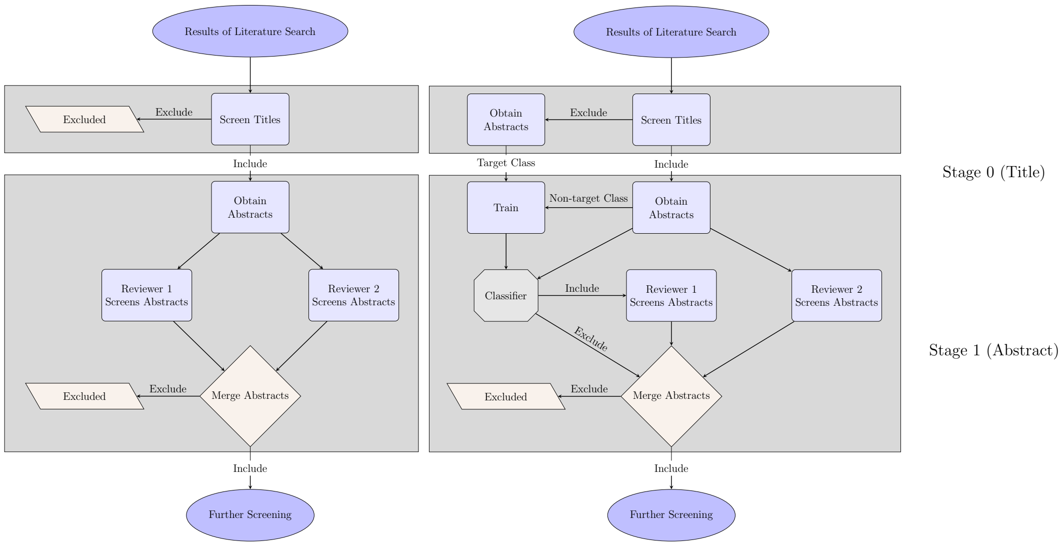
In addition to the P-value threshold parameter, we also modify the Kingfisher algorithm to utilise a confidence threshold for valid rules. The purpose is to allow users to tune to rules to guarantee a suitably precise model of excluded citations is developed. For a given association rule $X \Rightarrow Z$, confidence is defined to be the frequency of combined feature set XZ divided by the frequency of X. For classification rules, confidence is equivalent to precision. This gives two algorithmic parameters that require tuning: minimum confidence and the goodness threshold for rules. Choices for these parameters are discussed further in section 4.5.4.

4.4.4 Summary

This section outlined the preprocessing and rule extraction process used to generate and apply classification rules to identify irrelevant citations. The approach is novel both for its focus on excluding irrelevant citations with high precision, and for its use of reviewer annotations from prior screening stages to inform decisions later on in the screening process.

Figure 4.7 compares the title and abstract stages of a traditional literature screening process for systematic reviews against one including the proposed algorithm. It is interesting to note two things, firstly the data and annotations input to the training algorithm are all generated prior to any abstract based screening by either reviewer. This is useful in that there appears to be no change in the process from the perspective of the human reviewers. The only difference is that reviewer 1 receives a different, smaller set of citations than reviewer 2 for screening based on abstract.

Secondly, the application of our classifier prior to abstract based screening by human reviewers suggests that our approach would integrate well with existing work by authors such as Frunza et al. [32] or [101]. These algorithms rely on authors annotating a set of abstracts prior to the application of a classifier. Our algorithm could be used as a preprocessing step to reduce the overall number of abstracts for screening, after which annotations could be sought as appropriate



(a) Traditional (b) Modified

Figure 4.7: Comparison of the traditional and modified screening processes for stage 0 and 1 screening.

and alternative algorithms applied.

Excluding parameters of the rule generation algorithm, there are three points in the process where multiple courses of action are available; the level of segmentation used, choice of features, and the choice of training data. A brief summary of each is as follows:

Segmentation Level Classification rules for irrelevant citations can either be extracted at a document or sentence level. If rules are extracted at a document level, a rule will be considered to match a citation if all terms in the rule antecedent appear somewhere in the citation. If rules are extracted at a sentence level, terms from the antecedent must occur together in a single sentence. Training data is created by segmenting each citation, and labelling each sentence with the reviewer annotation of the citation to which it belongs. After rules have been generated, a citation is then excluded if at least one of its sentences match a rule.

Feature Selection Fielded MMI output returns a number of UMLS concepts related to the input data. For each concept we are given a preferred name, along with a set of tree codes for concepts which are also found in MeSH (any concepts which are not also present in MeSH are pruned). A single MeSH heading may map to multiple UMLS concepts. Features can be generated either by using the UMLS concept preferred names, or the tree codes associated with the relevant MeSH concept. We evaluate document representation using tree codes versus concept names in our evaluation.

Training Data The rational behind the classification approach presented in this chapter involves modelling irrelevant citations using abstract data which was not available to reviewers in the first screening stage. The abstract data for irrelevant citations should therefore be involved when training the rules, however it is unclear how to use abstract data for citations which have not yet been excluded. Including abstract data for all studies when training rules risks generating rules that are too specific and will not match any additional studies. Excluding abstract data increases the chance that more citations will be matched, but may decrease the quality of the generated rules.

As part of the evaluation given in section 4.5, a sensitivity analysis is performed examining the effect of each of the above choices. The result of this is a recommendation for each, which is then used to evaluate both the parameters of the rule extraction algorithm and the potential workload savings of the approach described in this chapter.

4.5 Evaluation

In this section we present an evaluation of the performance of the algorithm described earlier in this chapter. We do so by simulating its application on two real systematic reviews. Details on each review are given in section 4.5.1, and evaluation metrics are described in section 4.5.2.

Review Stage	Aspergillosis		Alzheimers	
	# Citations at start	# Citations Removed	# Citations at start	Citations Removed
Stage 0 (Titles only)	4377	3412 (79.0%)	2097	556 (26.5%)
Stage 1 (Abstracts)	965	818 (96.6%)	1541	-
Stage 2 (Full-text)	147	60 (98.0%)	-	1393 (92.9%)
Data Extraction	87	40 (99.1%)	148	0 (92.9%)

Table 4.2: Summary of data used for the evaluation in chapter 4.

The algorithm described in section 4.4 has several parameters, including choice of features, training methodology, and training data. The trade-offs involved with these parameters are examined in section 4.5.3. Parameters of the rule generation algorithm (goodness threshold and minimum rule confidence) are examined in section 4.5.4. The final workload savings are then covered in section 4.5.5.

4.5.1 Data Collection

The systematic review data used in this chapter was drawn from the literature searches for two Cochrane reviews: *Galactomannan detection for invasive aspergillosis in imumunocompromized patients* by Leeflang et al. [57], and an unpublished review on the accuracy of diagnostic biomarkers for Alzheimers disease. A list of citations obtained, along with corresponding annotations, during an updated literature search for the above review were kindly provided by the review authors. Citations were annotated with the final review decision for each stage (although references were screened by multiple reviewers, individual reviewer annotations are not relevant to the methodology described in this chapter).

We refer to the two data sets as *Aspergillosis* and *Alzheimers* respectively. The number of citations retrieved and their movement through the screening process is summarised in Table 4.2.

The literature search process for the Aspergillosis data covered screening of citations retrieved from three databases (Medline, Embase, and Web of Knowledge). After duplicates were removed from the combined search results, titles were screened by a single author to remove obviously irrelevant citations. Two authors then screened the remaining abstracts and removed any citations which they agreed were not relevant to the final review. This process was then repeated with full texts for the remaining citations. This resulted in 87 citations which were considered relevant for inclusion in the review.

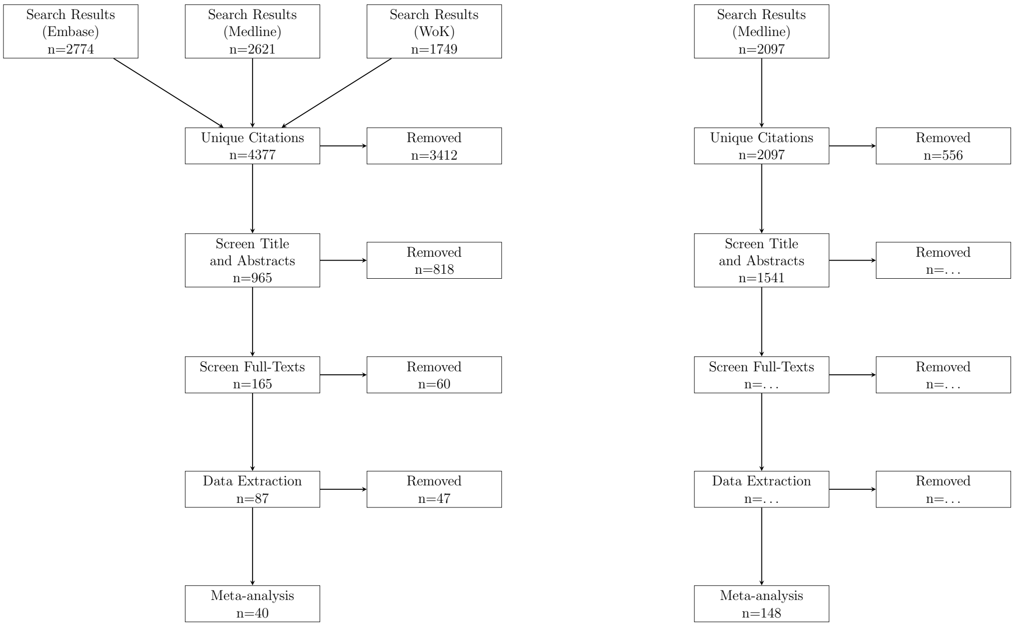
A further 47 citations were removed during the data extraction process. Reasons for these further exclusions included things like studies reporting insufficient data, duplicating results from other studies, or analyses that were based on in-house tests inappropriate for inclusion in the review. For the purposes of simulating the screening process for the analysis reported in this chapter, we consider such studies to have been included in the review. This is due to the fact that we aim to identify citations concerning non-relevant topics. The topics covered in the above 47 citations were considered relevant; the decision to exclude was made based upon information to which our classifier would not have access.

The literature search process for the Alzheimers data screened citations retrieved from both Medline and Embase. However, as we were only able to obtain annotations for the Medline portion of the studies, we do not consider citations found exclusively in Embase. In all, the Medline search returned 2097 studies, with 1541 abstracts retrieved. That 3/4 of citations remained after title based screening differs substantially from the Aspergillosis data, where the majority of citations retrieved were removed at the first stage. Abstracts were screened by two reviewers, and 148 of the 1541 abstracts retrieved were eventually judged relevant to the review.

A graphical summary (similar to a PRISMA flowchart [73]) for both datasets is also presented in Figure 4.8.

4.5.2 Metrics

We now outline the evaluation metrics used in this chapter, and how they relate to the annotated citations described in section 4.5.1. We start by describing the gold



(a) Aspergillosis (b) Alzheimers

Figure 4.8: Citation flow diagram for Aspergillosis and Alzheimers literature search. Only title based screening and final inclusion annotations were available for the Alzheimers review data, hence no values are reported for other stages.

standard against which the performance of the proposed approach is measured. As we are using classifiers where the target class is those citations which are not relevant to a given review, we also explicitly define the terms true positive, false positive, false negative, and true negative. Note that as we treat non-relevant citations as the target class (as opposed to relevant citations), the meaning of the terms positive and negative in this context are inverted with respect to much of the existing literature.

Gold Standard We define the gold standard annotations according to the combined decisions of the human reviewers. For the aspergillosis data described in section 4.5.1, this means we have 4377 total annotations with 87 includes and 4290 excludes. We measure the performance of our approach against its ability to replicate these annotations.

True Positive (TP) A true positive classification occurs when a citation is excluded by both the classifier and the gold standard annotations.

False Positive (FP) A false positive classification occurs when a citation is excluded by the classifier but included by the gold standard annotations.

False Negative (FN) A false negative classification occurs when a citation is included by the classifier but excluded by the gold standard annotations.

True Negative (TN) A true negative classification occurs when a citation is included by both the classifier and the gold standard annotations.

As discussed earlier, the aim of classification for systematic reviews is to exclude as many irrelevant citations as possible without removing relevant studies (so maximising the number of true positives conditioned on having nearly zero false positives). Therefore, based on the above definitions we employ three separate evaluation metrics in our analysis; reviewer workload, recall, and precision. In the discussion throughout this thesis, recall is used to refer to the recall over relevant citations, and precision refers to precision over non-relevant citations. We use these definitions unless otherwise specified. We define these metrics as follows:

Reviewer Workload We define the reviewer workload as the percentage of citations requiring screening by a human reviewer. Referring to the proposed workflow outlined in Figure 4.7, reviewer workload is defined as the number of abstracts included by the classifier divided by the total number of abstracts obtained $\frac{TN+FN}{N}$. This metric directly reflects the practical benefits in terms of time saved for human reviewers.

Recall We define the recall to be the percentage of actually relevant citations retained by the classifier. According to the definitions given above, this is defined as the number of true negatives divided by the combined number of true negatives and false positives ($\frac{FP}{TP+FN}$). Note that the sensitivity is analogous to the recall for relevant citations. We use the term sensitivity in this case to avoid confusion between recall for relevant citations, and recall for the classifiers target class (non-relevant citations).

Precision Precision is defined to be the percentage of citations excluded by the classifier that are genuinely not relevant to the review ($\frac{TP}{TP+FP}$). Similar to the logic used for sensitivity, we use the term precision when talking about the classifiers ability to model its target class (non-relevant citations).

Finally, we consider the question of what exactly constitutes appropriate recall for a prospective classifier. The goal for any classification algorithm should be to exclude with recall equivalent to an expert human reviewer. Although individual reviewers will seek to not exclude any relevant studies, in practice this is unlikely to be achieved. Even among expert human reviewers, disagreements over the relevance of individual citations will occur.

To our knowledge there has been no comprehensive study to date examining error rates for human reviewers. As such, we use a recall level of 95% as our target for acceptable performance. This threshold is in line with previous, comparable literature [22].

4.5.3 Sensitivity Analysis

In section 4.4 we outlined the algorithm used to process citations and generate rules to identify non-relevant studies. We noted three points in the algorithm at which multiple design choices could to be made; namely the data used to train the classifier, the features extracted from the data, and the level of segmentation used before feature extraction from citations. We now evaluate the effect that each design choice has one the resultant classifier.

For each section, results figures report Log P threshold for valid rules is shown on the x-axis, with a separate figure created for different minimum confidence thresholds (0.5, 0.9, 0.95, and 0.98 respectively). The major y-axis (corresponding to the solid lines) reports the remaining reviewer workload, while the minor y-axis reports the recall over relevant citations. Each value is reported along with the 95% confidence interval, with values for UMLS concepts reported in red and MeSH headings reported in blue. Circled data points indicate a statistically significant improvement (P=.05) over the comparable data point.

4.5.3.1 Feature Selection

We begin by examining the performance of the two different features types proposed for document representation (as outlined in section 4.4.2.1). Figures 4.9, 4.10, 4.11, and 4.12, 4.13, 4.14, 4.15, and 4.16 compare the reviewer workload and sensitivity obtained using UMLS concepts vs. MeSH tree codes for a range of parameterisations of the rule extraction process. Each individual figure displays the performance for all four combinations of the other algorithmic choices examined in this section (document vs. sentence level segmentation, and reduced vs. full training data).

From the values reported, it can been seen that there is very little difference in performance between the two sets of features. For the Aspergillosis data, there was no statistically significant difference (P=.05) for any points with a minimum confidence level of 0.9 or 0.95. For the values with minimum confidence of 0.98, MeSH headings performed better (although the actual difference in performance was very small) in several cases and never worse than the UMLS concepts. When training with a minimum confidence of 0.5, recall over relevant studies was improved for concept names with respect to tree codes in one case, with all other points reporting no statistically significant difference.

For the Alzheimers data, all tested minimum confidence thresholds resulted in at least several datapoints for which concept names outperformed tree codes. Points where tree codes outperformed concept names do exist, however are relatively rare (13/128 instances where concept names were better, and 3/128 where tree codes were better).

Although the magnitude of the difference was not particularly great, concept names appeared to slightly outperform tree codes over the tested data and parameters. The effect of replacing concept names with tree codes is that different entry

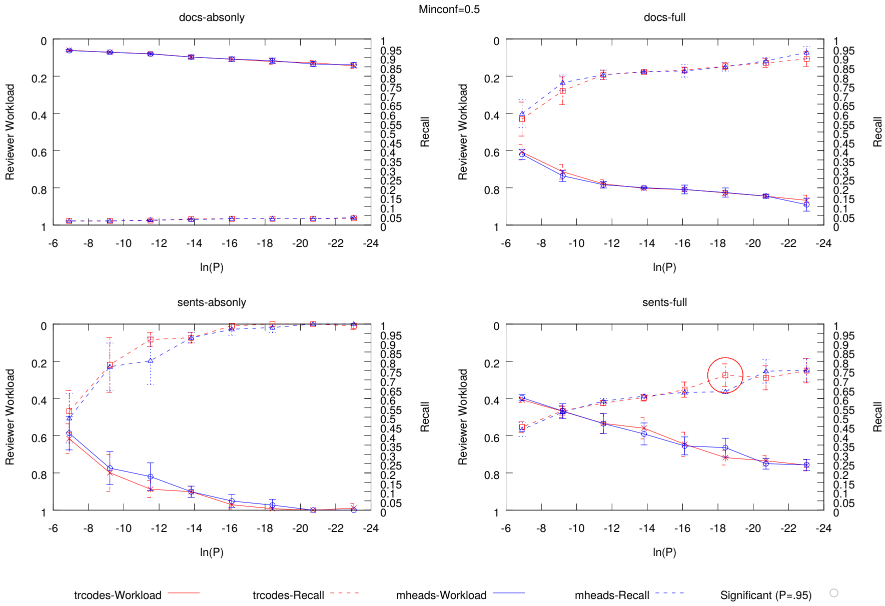


Figure 4.9: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for tree code features vs. MeSH headings with a minimum confidence of 0.5 on Aspergillosis data. Performance is reported for all combinations of segmentation level (document vs. sentence) and training data (full vs. irrelevant abstracts only).

terms for the same concept will be merged into a single feature. That the performance of concept names as features was better (but still comparable) to that of tree codes suggests the occasional benefits of accounting for polysemous features are not sufficient to support merging MeSH entry terms in general. While MeSH contains a very large number of terms covering a wide range topics, the context in which various entry terms can be considered to describe the same concept is likely to differ between reviews. An example of this can be seen in the Alzheimers data. The MeSH term *Diagnosis* (tree code E01) contains several entry terms, among them *Antemortem Diagnosis* and *Postmortem Diagnosis*. While in some contexts the two entry terms could indicate a common concept, a review examining diagnostic methods for a given disease is likely to want to distinguish between the two.

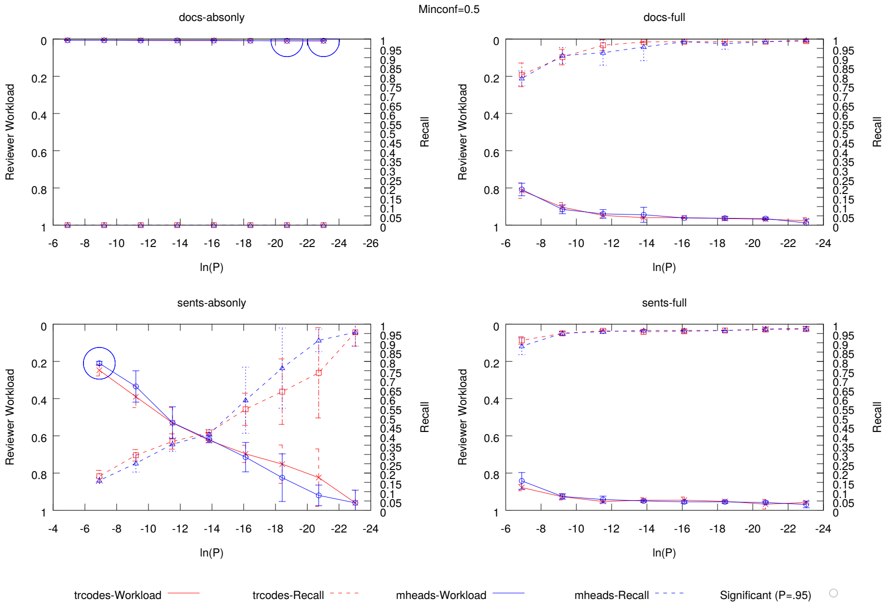


Figure 4.10: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for tree code features vs. MeSH headings with a minimum confidence of 0.5 on Alzheimers data. Performance is reported for all combinations of segmentation level (document vs. sentence) and training data (full vs. irrelevant abstracts only).

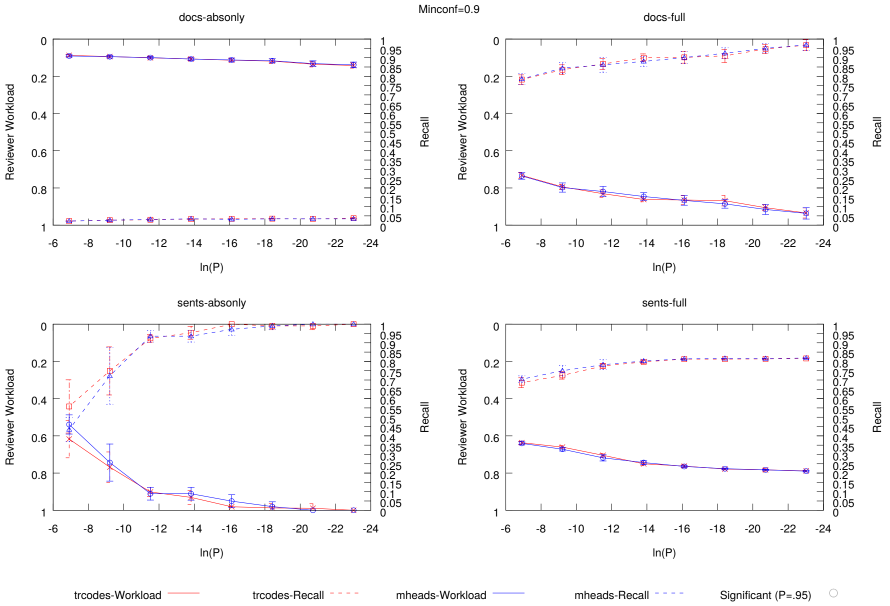


Figure 4.11: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for tree code features vs. MeSH headings with a minimum confidence of 0.9 on Aspergillosis data. Performance is reported for all combinations of segmentation level (document vs. sentence) and training data (full vs. irrelevant abstracts only).

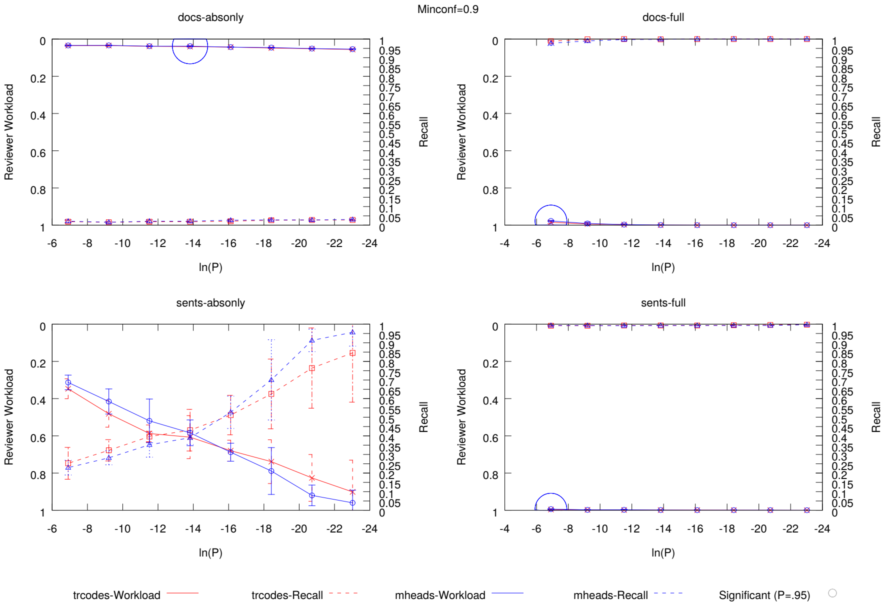


Figure 4.12: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for tree code features vs. MeSH headings with a minimum confidence of 0.9 on Alzheimers data. Performance is reported for all combinations of segmentation level (document vs. sentence) and training data (full vs. irrelevant abstracts only).

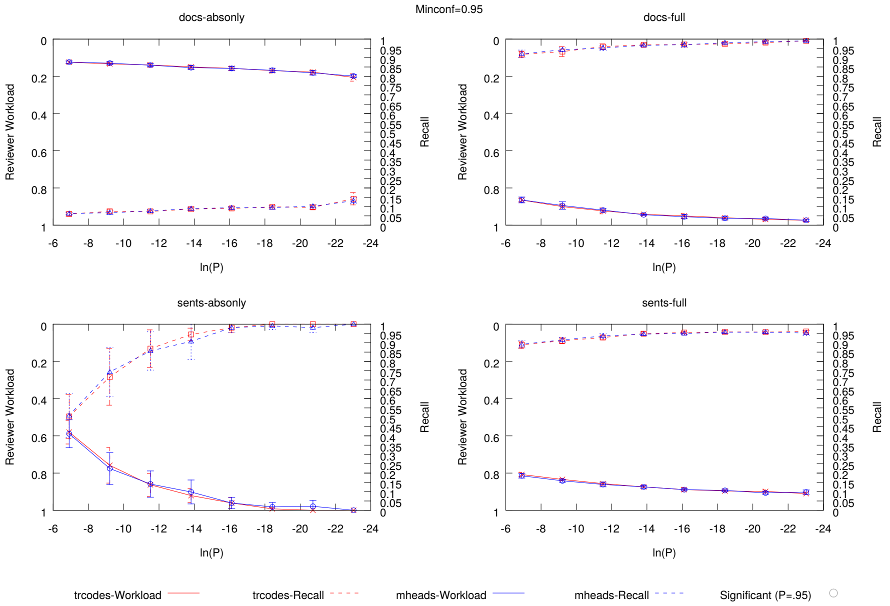


Figure 4.13: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for tree code features vs. MeSH headings with a minimum confidence of 0.95 on Aspergillosis data. Performance is reported for all combinations of segmentation level (document vs. sentence) and training data (full vs. irrelevant abstracts only).

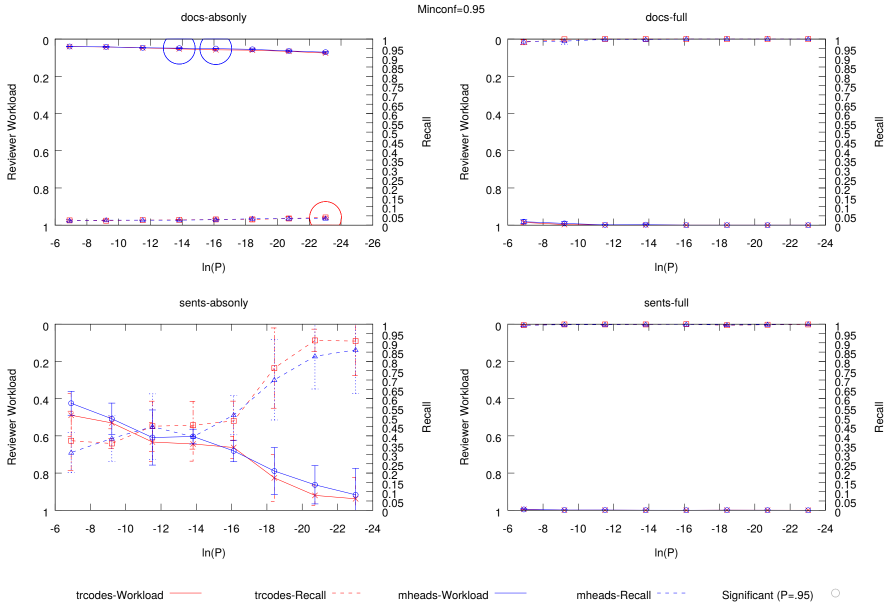


Figure 4.14: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for tree code features vs. MeSH headings with a minimum confidence of 0.95 on Alzheimers data. Performance is reported for all combinations of segmentation level (document vs. sentence) and training data (full vs. irrelevant abstracts only).

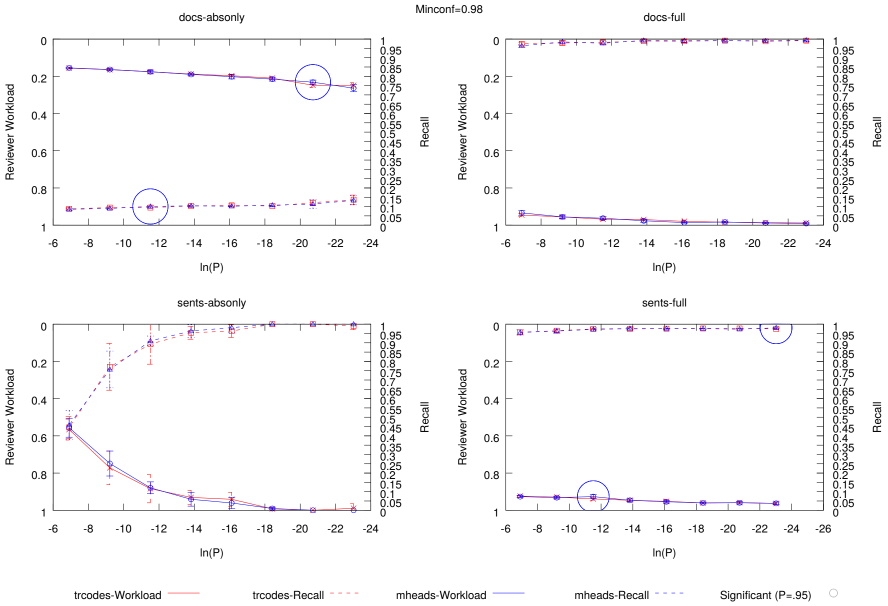


Figure 4.15: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for tree code features vs. MeSH headings with a minimum confidence of 0.98 on Aspergillosis data. Performance is reported for all combinations of segmentation level (document vs. sentence) and training data (full vs. irrelevant abstracts only).

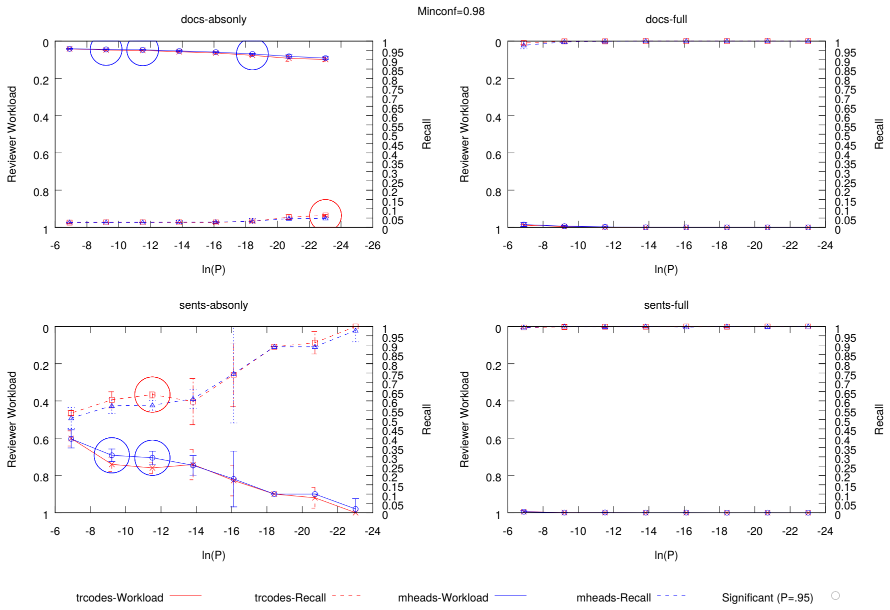


Figure 4.16: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for tree code features vs. MeSH headings with a minimum confidence of 0.98 on Alzheimers data. Performance is reported for all combinations of segmentation level (document vs. sentence) and training data (full vs. irrelevant abstracts only).

4.5.3.2 Training Data

Figures 4.17, 4.18, 4.19, 4.20, 4.21, 4.22, 4.23, and 4.24 compare the recall and remaining workload for rules built using all available data when training against rules built using titles from all citations, and abstracts only from citations that were removed based on title alone. For the analysis reported in this section we refer to the first method as *full*, and the latter as *absonly*.

Observing the results when using absonly with document level segmentation (the first row of subfigures), we can see that absonly performs poorly. In all cases, recall with absonly and document level segmentation is unacceptably low, often below 10%. At the same time we can see the reviewer workload is also extremely low, indicating that combining the absonly training method with document level segmentation produces rules that match a very large number of citations.

That absonly training performs poorly with document level segmentation is not surprising. Citations are excluded if they match at least one rule. If rules are too general they are likely to match a large number of relevant citation in addition to non-relevant ones, substantially lowering recall. Rules are generated from training data by identifying features that co-occur regularly in the target instances. Avoiding overly general rules therefore requires training data with a sufficient number of counter examples where features that do not indicate non-relevance also co-occur in the non-target instances.

When using absonly training, we only consider the title part of the non-target instances (ignoring the text contained in the abstract). This will reduce the chance of counter examples for terms co-occurring in the target instances. By also using sentence level segmentation we require that terms co-occur within individual sentences, rather than anywhere within a citations. This has the effect of lowering the number of feature sets examined, reducing the possibility of a spurious combination of features being tested.

Performance with absonly training data is much better when used with sentence level segmentation. While recall is generally quite low (around 50%) with lower goodness (ln(P)) thresholds, recall appears to improve sharply as the strictness of the threshold is increased. Also of particular interest is the lack of variation in performance of absonly training with sentence level segmentation as the minimum confidence threshold is changed. This is in contrast to training with full data, which appears to require an appropriately strict confidence threshold in order to achieve sufficient recall.

It is interesting to note that training with full data generally varies to a much greater extent as the minimum confidence threshold is varied. For example, looking at the performance on the Aspergillosis data with a low minimum confidence threshold (such as in Figure 4.17) we can see that recall with full training data fails to reach 95% recall. As the minimum confidence is increased to 0.9 and 0.95 in figures 4.19 and 4.21 respectively, recall increases sharply, and is much more consistently above 0.95.

Finally, we note that although both absonly and full training data approach full workload as minimum confidence and goodness threshold increase, full training data tends to approach at a slower rate. For example, comparing the performance with sentence level segmentation (the second row of sub-figures) with a minimum confidence of 0.95 on the Aspergillosis data (figures 4.21) we can see workload approaching 1 for absonly training at roughly the same rate recall approaches 1. In contrast, non-zero workload savings are achieved with full training data for all tested confidence thresholds.

Similar observations can be made with the Alzheimers data. Looking at the performance with a minimum confidence of 0.5 (Figure 4.18), we can see that as the goodness threshold increases, recall with absonly training is always further from 1 than reviewer workload. Training with full data however still manages workload savings of greater than 0.05% even when recall is above 0.95%.

It is important to note that the minimum confidence thresholds required to obtain acceptable recall for the two data sets differ substantially. For example the Alzheimers data performed poorly with all tested minimum confidence thresholds

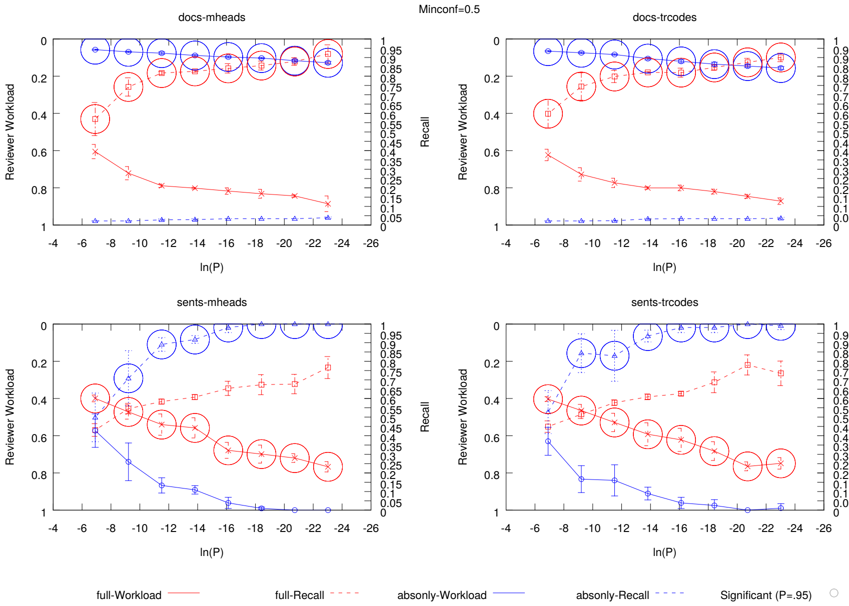


Figure 4.17: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for training with full data vs. irrelevant abstracts only with a minimum confidence of 0.5 on Aspergillosis data. Performance is reported for all combinations of segmentation level (document vs. sentence) and features (tree codes vs. MeSH headings).

other than 0.5, with rules either failing to generate appropriate recall or generating very small workload savings.

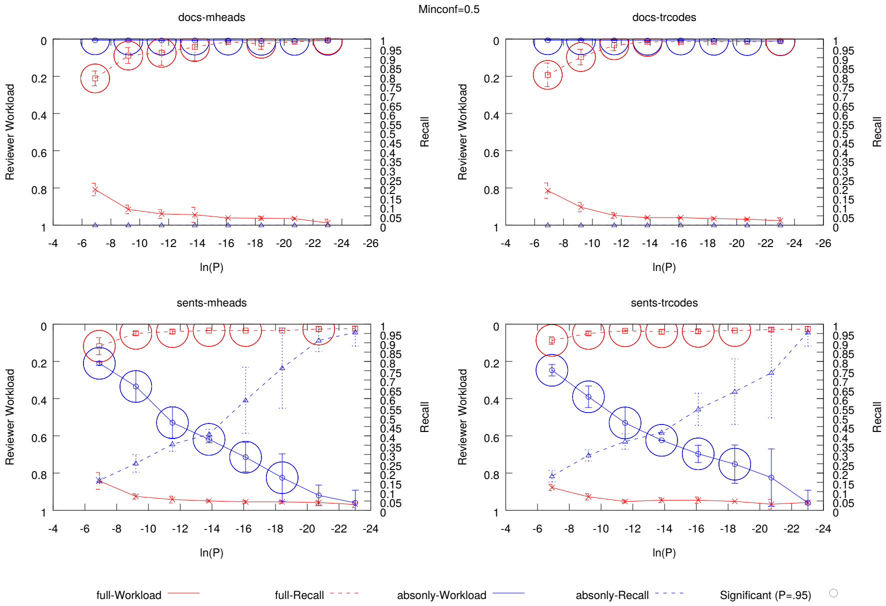


Figure 4.18: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for training with full data vs. irrelevant abstracts only with a minimum confidence of 0.5 on Alzheimers data. Performance is reported for all combinations of segmentation level (document vs. sentence) and features (tree codes vs. MeSH headings).

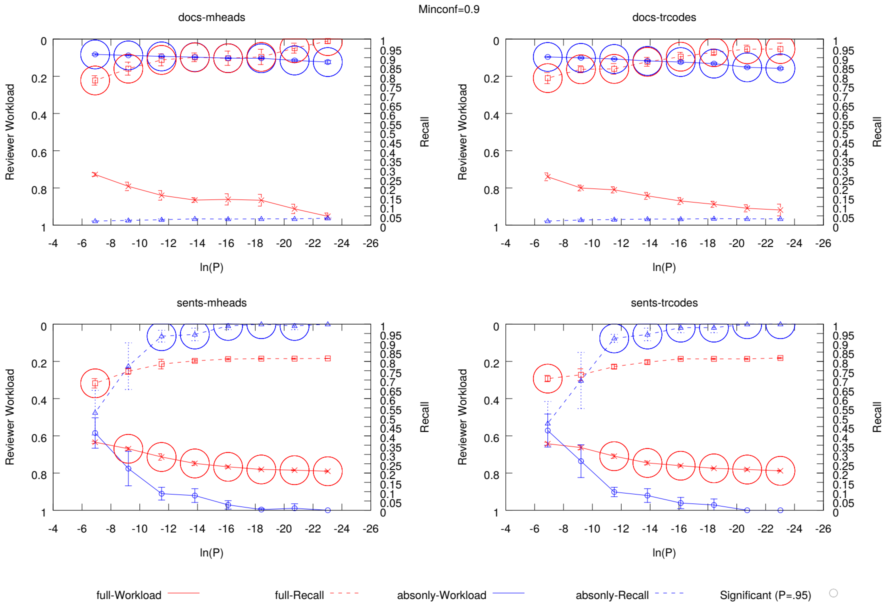


Figure 4.19: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for training with full data vs. irrelevant abstracts only with a minimum confidence of 0.9 on Aspergillosis data. Performance is reported for all combinations of segmentation level (document vs. sentence) and features (tree codes vs. MeSH headings).

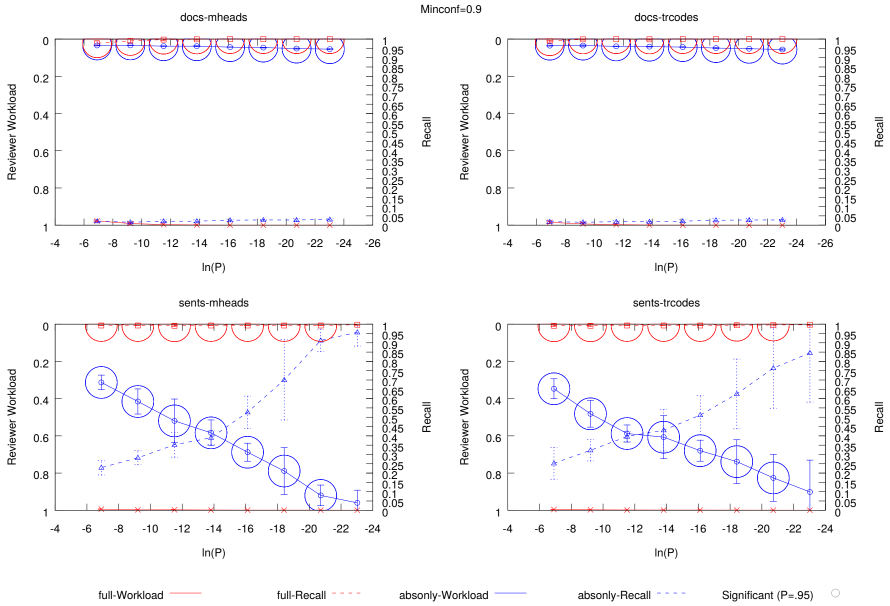


Figure 4.20: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for training with full data vs. irrelevant abstracts only with a minimum confidence of 0.9 on Alzheimers data. Performance is reported for all combinations of segmentation level (document vs. sentence) and features (tree codes vs. MeSH headings).

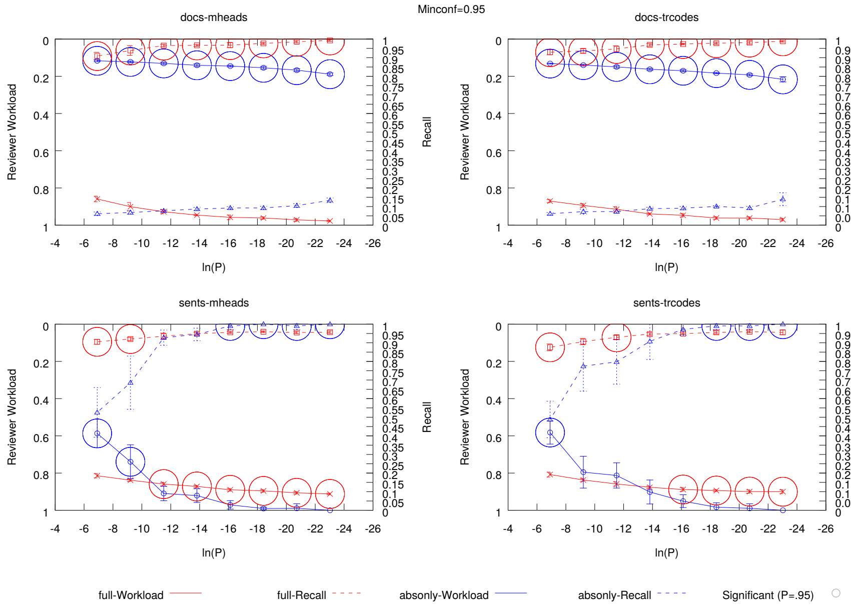


Figure 4.21: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for training with full data vs. irrelevant abstracts only with a minimum confidence of 0.95 on Aspergillosis data. Performance is reported for all combinations of segmentation level (document vs. sentence) and features (tree codes vs. MeSH headings).

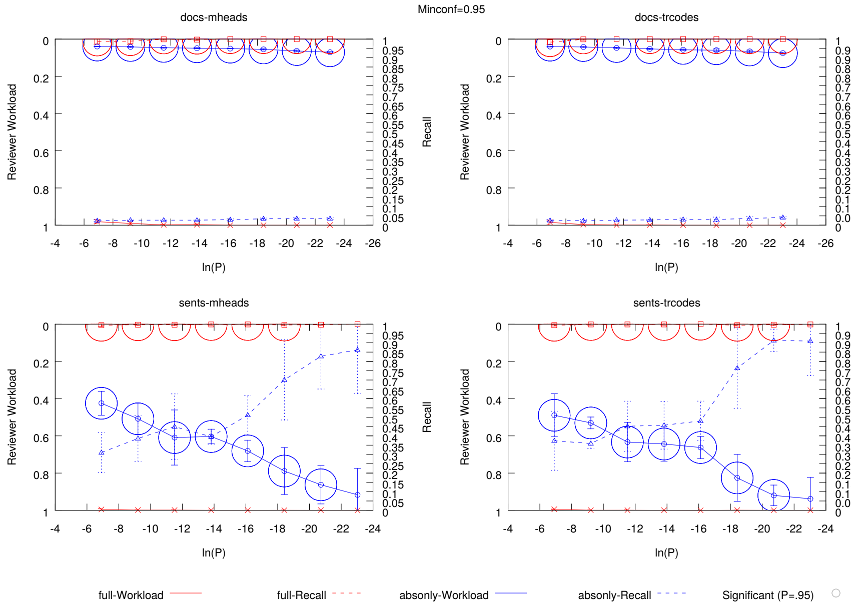


Figure 4.22: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for training with full data vs. irrelevant abstracts only with a minimum confidence of 0.95 on Alzheimers data. Performance is reported for all combinations of segmentation level (document vs. sentence) and features (tree codes vs. MeSH headings).

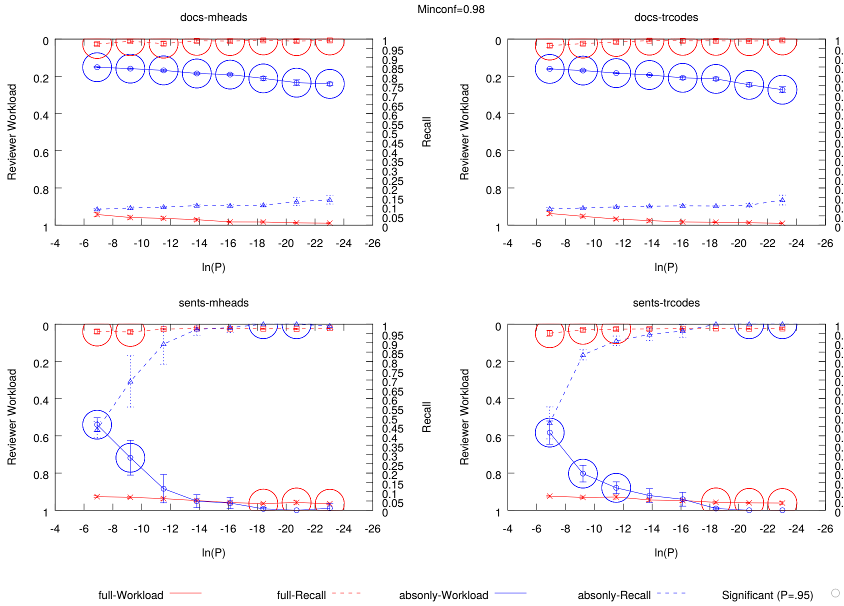


Figure 4.23: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for training with full data vs. irrelevant abstracts only with a minimum confidence of 0.98 on Aspergillosis data. Performance is reported for all combinations of segmentation level (document vs. sentence) and features (tree codes vs. MeSH headings).

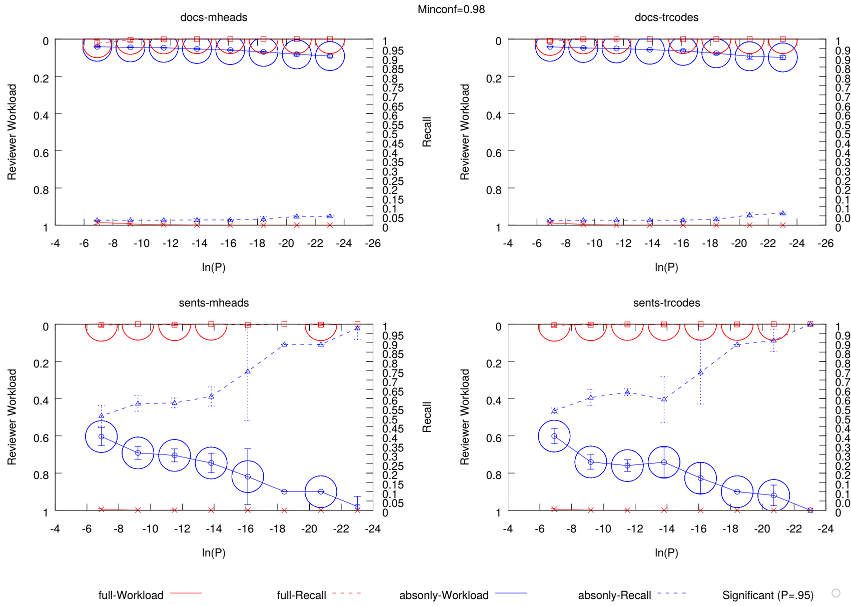


Figure 4.24: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for training with full data vs. irrelevant abstracts only with a minimum confidence of 0.98 on Alzheimers data. Performance is reported for all combinations of segmentation level (document vs. sentence) and features (tree codes vs. MeSH headings).

4.5.3.3 Segmentation Level

Figures 4.25, 4.26, 4.27, and 4.28, 4.29, 4.30, 4.31, and 4.32 compare the reviewer workload and recall obtained using sentence and document level segmentation. While the recall produced using document level segmentation depends heavily on whether or not full training data is being used, sentence level segmentation is much more consistent (this was discussed in section 4.5.3.2).

It is interesting to note the difference in performance between the two datasets. For the Alzheimers data, recall for sentence level segmentation is consistently as good or better than document level segmentation. For training with full data, this is more apparent with lower minimum confidence values (0.5 and 0.9). Recall for both methods approaches 1 with higher thresholds, as does remaining reviewer workload. Training with full data and document level segmentation performs poorly for reasons discussed in section 4.5.3.2).

However for the Aspergillosis data, performance is more dependent upon minimum confidence threshold. At lower confidence thresholds (and with full training data), document level segmentation appears to generate better recall than sentence level segmentation. Better workload savings are generated with sentence level segmentation, although in all cases both document and sentence level segmentation fail to generate recall above 0.95. With higher confidence thresholds both document and sentence level segmentation generate acceptable recall, however reviewer workload with sentence level segmentation appears to be slightly lower.

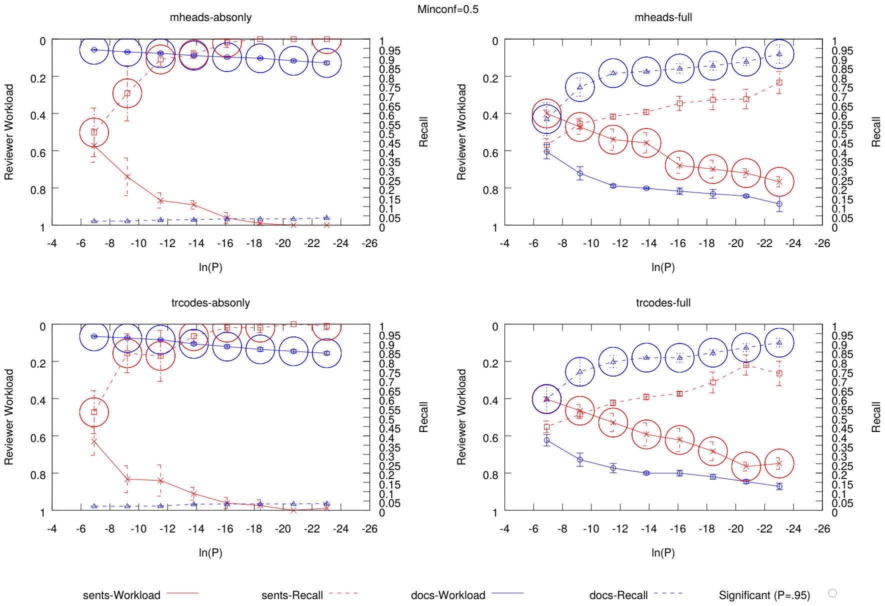


Figure 4.25: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for document vs. sentence level segmentation with a minimum confidence of 0.5 on Aspergillosis data. Performance is reported for all combinations of training data (full data vs. irrelevant abstracts only) and features (tree codes vs. MeSH headings).

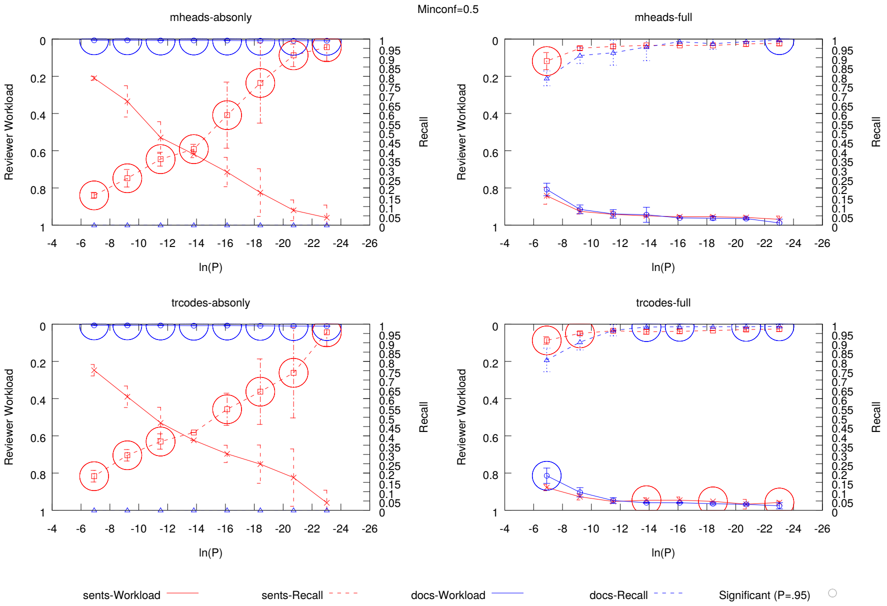


Figure 4.26: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for document vs. sentence level segmentation with a minimum confidence of 0.5 on Alzheimers data. Performance is reported for all combinations of training data (full data vs. irrelevant abstracts only) and features (tree codes vs. MeSH headings).

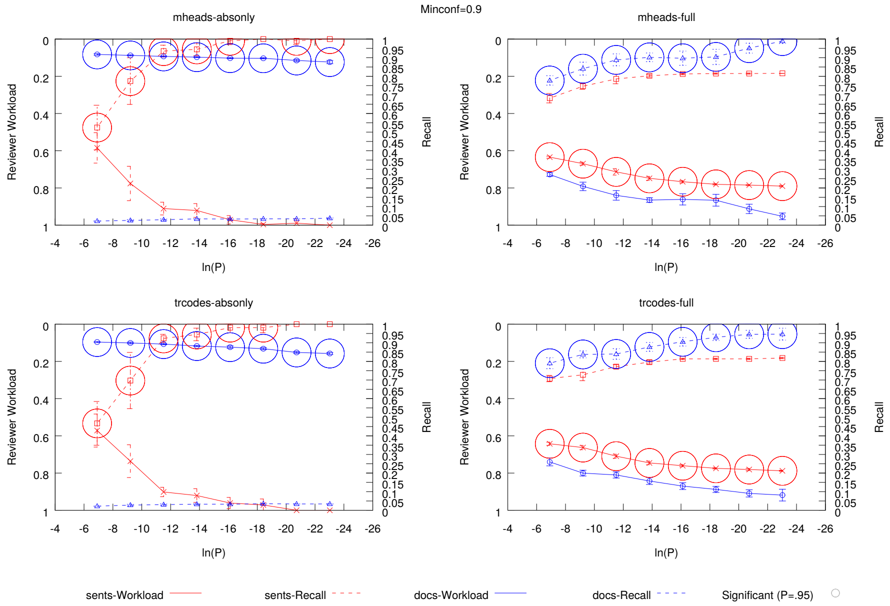


Figure 4.27: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for document vs. sentence level segmentation with a minimum confidence of 0.9 on Aspergillosis data. Performance is reported for all combinations of training data (full data vs. irrelevant abstracts only) and features (tree codes vs. MeSH headings).

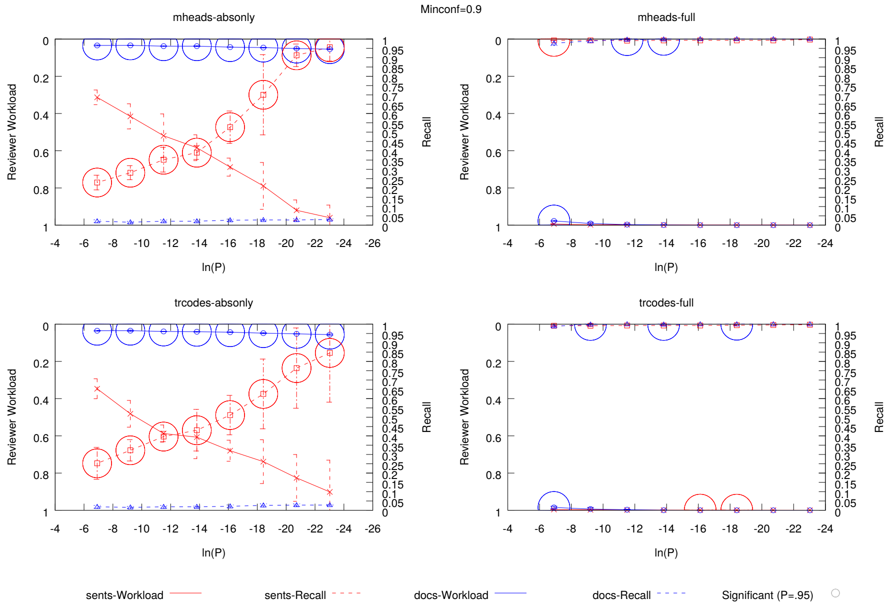


Figure 4.28: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for document vs. sentence level segmentation with a minimum confidence of 0.9 on Alzheimers data. Performance is reported for all combinations of training data (full data vs. irrelevant abstracts only) and features (tree codes vs. MeSH headings).

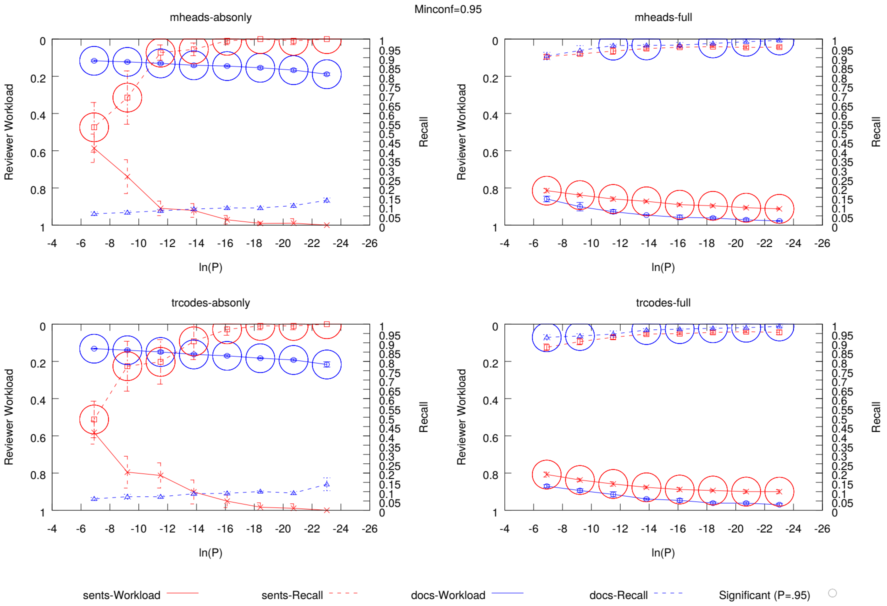


Figure 4.29: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for document vs. sentence level segmentation with a minimum confidence of 0.95 on Aspergillosis data. Performance is reported for all combinations of training data (full data vs. irrelevant abstracts only) and features (tree codes vs. MeSH headings).

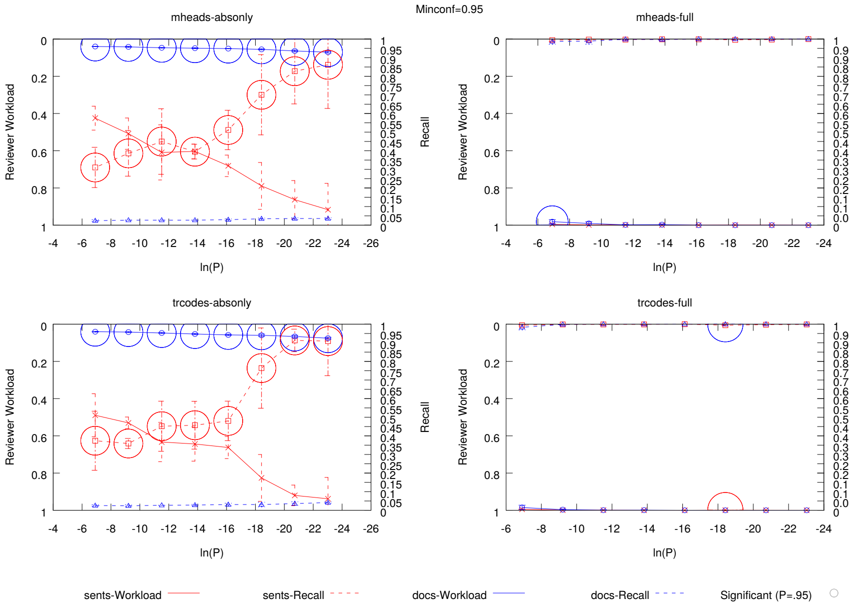


Figure 4.30: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for document vs. sentence level segmentation with a minimum confidence of 0.95 on Alzheimers data. Performance is reported for all combinations of training data (full data vs. irrelevant abstracts only) and features (tree codes vs. MeSH headings).

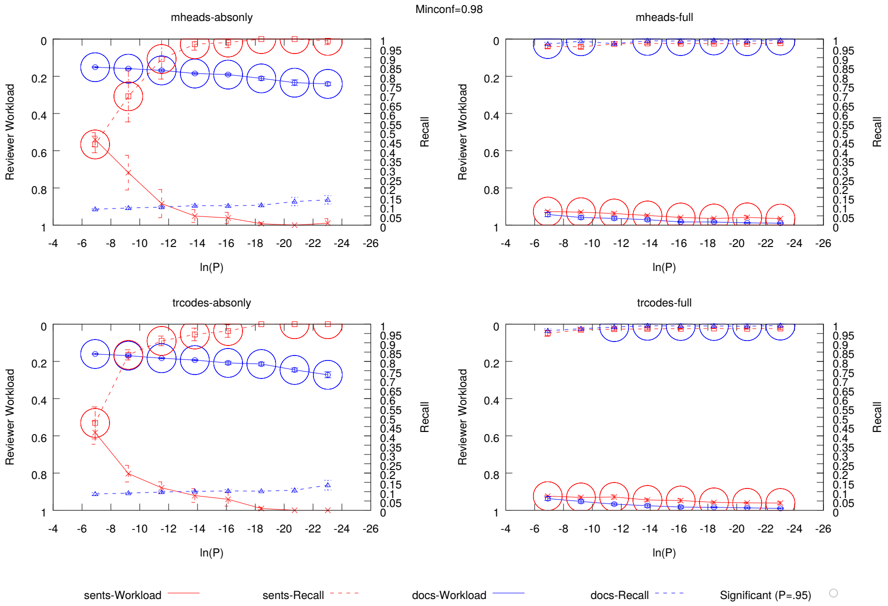


Figure 4.31: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for document vs. sentence level segmentation with a minimum confidence of 0.98 on Aspergillosis data. Performance is reported for all combinations of training data (full data vs. irrelevant abstracts only) and features (tree codes vs. MeSH headings).

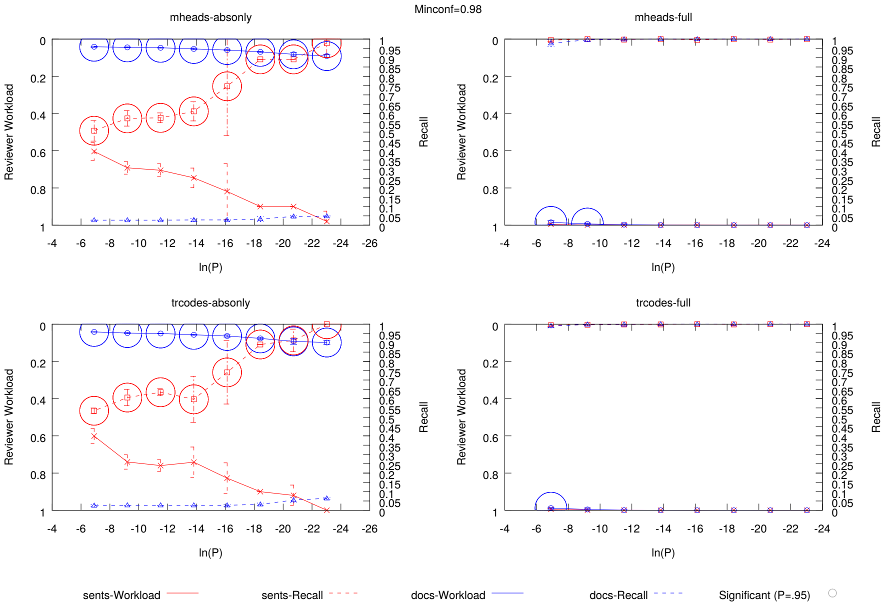


Figure 4.32: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for document vs. sentence level segmentation with a minimum confidence of 0.98 on Alzheimers data. Performance is reported for all combinations of training data (full data vs. irrelevant abstracts only) and features (tree codes vs. MeSH headings).

4.5.3.4 Summary

In the previous sections we analysed the effects of excluding abstracts for the non-target class during training, UMLS concepts vs. tree codes as features, and whether or not to segment citations at a document or sentence level.

Firstly, we were able to see that while excluding abstracts from non-target citations did occasionally result in good performance (with other appropriate parameter settings), it tended to produce either unacceptable recall or very low workload savings. Training using full data was in general more consistent than excluding non-target abstracts.

In the case of UMLS concepts vs. tree codes, it became apparent that for the tested data there was little to no benefit from using tree codes as opposed to UMLS concepts.

Sentence level segmentation also tended to outperform document level segmentation. In the case where non-target abstracts were excluded from training this was due to an inability to effectively model non-target instances. In the case where full training data was used, sentence level segmentation tended to produce a more consistent workload saving once a sufficient level of recall was produced.

4.5.4 Rule Generation Parameterisation

Figures 4.33 and 4.34 compare recall and remaining reviewer workload against goodness threshold for a range of minimum confidence levels. Figure 4.33 reports results on the Aspergillosis data, while Figure 4.34 reports results for the Alzheimers data. Rules were generated using full training data with UMLS concepts as features and sentence level segmentation (see section 4.4.4). Results were generated as the average over 5 runs using a random 70% sample of the data to generate rules. Rule performance was measured on the full set of citations for which abstracts were obtained.

For all tested confidence levels, the rate at which recall improved with goodness

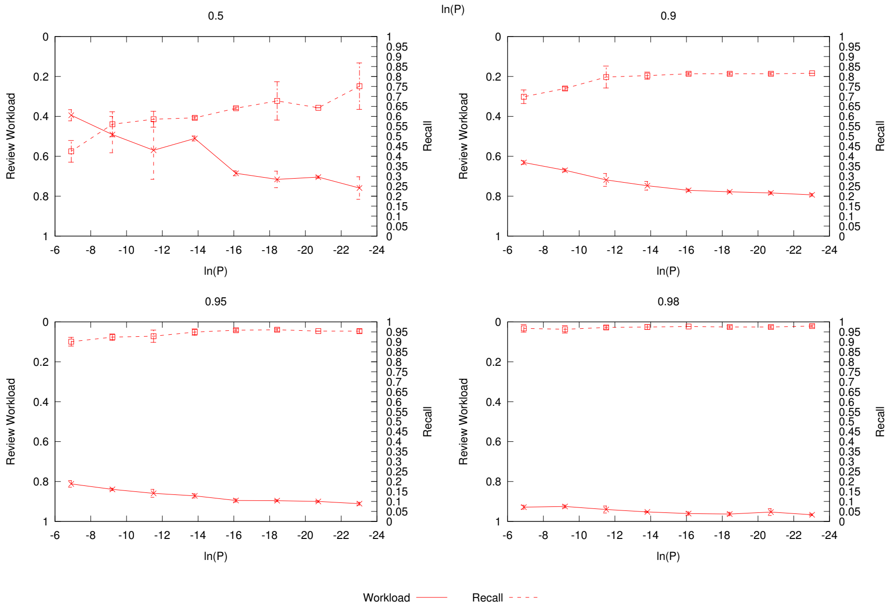


Figure 4.33: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for choice of ln(P) threshold on Aspergillosis data.

threshold drops sharply at the third data point. This corresponds to a goodness threshold P of 0.00001, or ln(P) of -11.513. Recall with the next strictest goodness threshold (P = 0.000001, ln(P) = -13.816) did not appear to differ, although in several cases (e.g. Aspergillosis data with minimum confidence thresholds of 0.9 and 0.95) reviewer workload did slightly rise.

It is interesting to note that for the selected algorithmic choices outlined in section 4.5.3, rule performance appears to peak at roughly the same goodness threshold. For this reason, in our following experiments we use ln(P) = -11.513 as our goodness threshold. It would be interesting to examine whether this threshold generalised for other reviews. It is likely that the size of the tested data is inversely proportional to the required threshold (both the Aspergillosis and Alzheimers data contain in the order of several thousand citations). A more detailed analysis with additional data would be interesting, but is left for future work.

Although both datasets achieve good performance with similar goodness thresholds, the minimum confidence required is much more data dependent. Figures 4.35

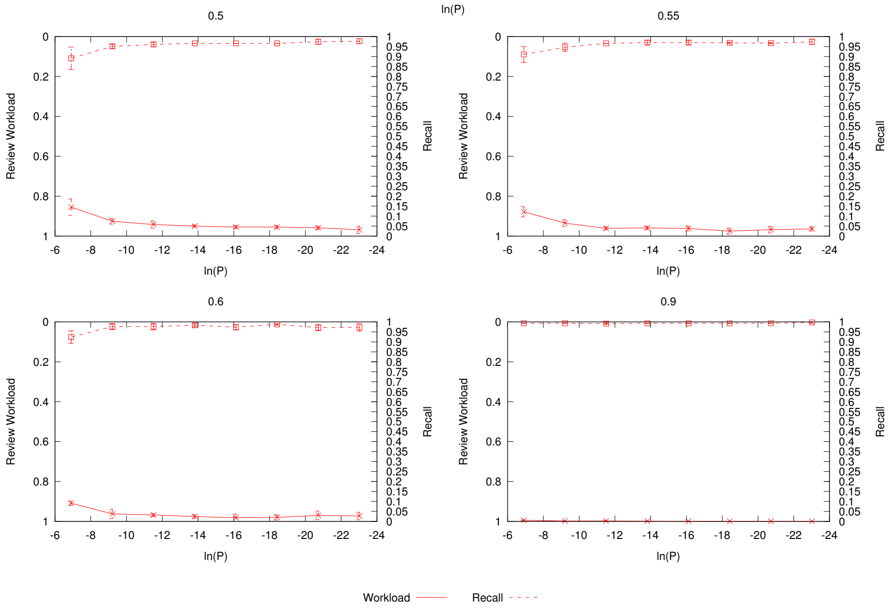


Figure 4.34: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for choice of ln(P) threshold on Alzheimers data.

and 4.36 show recall and reviewer workload versus minimum confidence threshold for the Aspergillosis and Alzheimers data respectively. Results are reported with goodness thresholds of -11.513 and -9.210.

For the Alzheimers data in Figure 4.36 we can see that with a goodness threshold of -11.513, all measured confidence thresholds give an acceptable level of recall. Remaining reviewer workload drops slightly as minimum confidence is increased, and achieves its best performance around 0.5.

If the goodness threshold is lowered slightly to -9.210, we observe a similarly strong level of recall for all tested confidence levels. Of note is that a slight drop in recall is observed with a minimum confidence of 0.5 when compared to 0.51, although this is likely a statistical anomaly (note the large confidence interval). This is pleasing to note, as it indicates that performance with low minimum confidence levels is robust even for weaker goodness thresholds.

However minimum confidence threshold appears to have a much greater impact on rule performance with the Aspergillosis data. From the results reported in

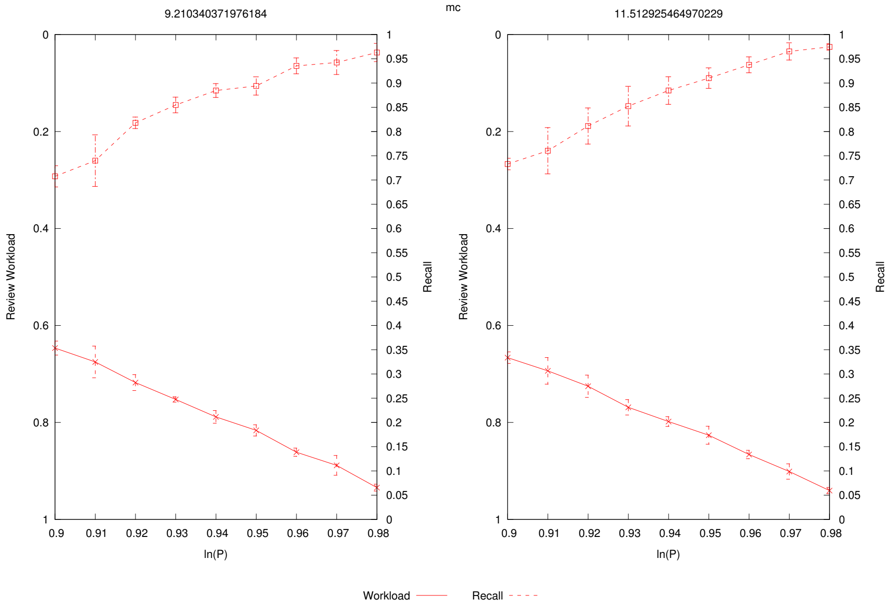


Figure 4.35: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for choice of minimum confidence threshold (mc) on Aspergillosis data.

Figure 4.35 we can see that acceptable recall is achieved, but not without a very strict minimum confidence threshold of 0.97 or higher. Performance is similar as the strictness of the goodness threshold is lowered, although the mean recall with a minimum confidence of 0.97 drops slightly below 0.95. While this particular measurement does have a larger than normal confidence interval with respect to other runs, it indicates the sensitivity of the Aspergillosis data to the selected minimum confidence level.

It is not immediately obvious why the Aspergillosis data should have such a greater recall to the selected minimum confidence level. Unfortunately, with only two data sets it is hard to speculate as to the cause. Two things are apparent: firstly, further analysis with additional data would be of interest. Secondly, in the meantime it is worthwhile examining the effectiveness of rule generation algorithms to identify one that is as robust as possible as minimum confidence for rules is varied.

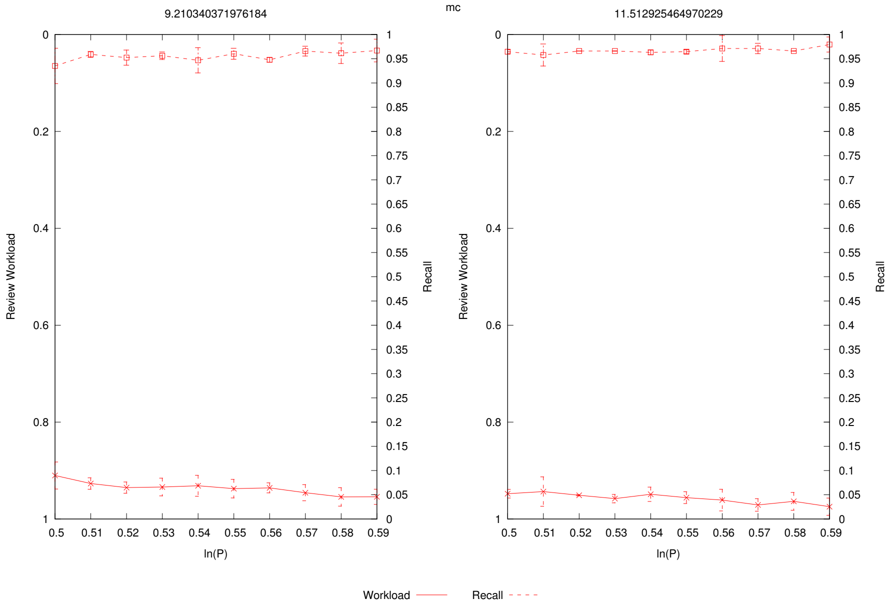


Figure 4.36: Sensitivity analysis (reviewer workload and recall vs. goodness threshold) for choice of minimum confidence threshold (mc) on Alzheimers data.

4.5.5 Workload Savings

In sections 4.5.3 and 4.5.4 we analysed parameter choices for both the rule generation and classification algorithms. Rule generation choices are outlined in section 4.5.3.4. For the classifier, a goodness threshold of ln(P)=-11.513 was found to perform appropriately for each tested data set. Although minimum confidence was much more dependent on choice of data, values of 0.51 and 0.97 were chosen for the Alzheimers and Aspergillosis data respectively.

Table 4.3 shows the recall and workload savings generated using the full data sets for both the Aspergillosis and Alzheimers data. Results with the selected minimum confidence levels are highlighted. With the selected parameters, workload savings of 10.9% for the Aspergillosis data and 5.6% for the Alzheimers data are obtained. In both cases a recall over relevant citations of greater than 95% is achieved.

These savings are significant, and have the potential to be of practical benefit to reviewers. Of particular note is that the approach outlined in this chapter

Alzheimers			Aspergillosis		
Min. Conf.	Workload Saving	Recall	Min. Conf.	Workload Saving	Recall
0.50	5.6%	0.966	0.90	35.4%	0.701
0.51	5.6%	0.966	0.91	35.1%	0.701
0.52	5.6%	0.966	0.92	28.9%	0.804
0.53	5.6%	0.966	0.93	23.7%	0.862
0.54	5.6%	0.966	0.94	20.6%	0.873
0.55	4.3%	0.966	0.95	19.4%	0.896
0.56	4.3%	0.966	0.96	13.0%	0.942
0.57	4.3%	0.966	0.97	10.9%	0.954
0.58	4.3%	0.966	0.98	5.6%	0.954

Table 4.3: Final workload savings for Aspergillosis and Alzheimers data. Highlighted cells correspond to pre-selected parameter settings.

can be used to complement existing research on review automation (for example, Abstrackr [101, 100]). Our approach could be applied before additional screening with other algorithms, further reducing the burden on human reviewers.

As discussed in section 4.5.4, it is interesting to note how dependent the results obtained with two datasets are on the choice of minimum confidence parameter. Workload saving for the Alzheimers data varies very little as the minimum confidence in increased, and recall does not change. In contrast, if the minimum confidence is lowered for the Aspergillosis data then a notable decrease in both recall and workload savings is observed. Further analysis with more data is warranted to understand the cause of this behaviour, and research into algorithmic changes that improve robustness to changes in minimum confidence are warranted.

4.6 Summary

In this chapter we proposed a method for semi-automated screening of citations for systematic reviews. Our approach is novel in several ways:

- Our method utilises reviewer annotations made based on title alone to inform decisions made based on title and abstract. No previous work exists in which annotations at one stage of a review are used to screen during another.
- Instead of seeking to identify relevant studies with high recall, we seek to exclude non-relevant studies with high precision. This approach helps to addresses the increasingly heterogeneous target classes found in modern systematic reviews.

We demonstrated the effectiveness of our approach on two existing systematic reviews, obtaining workload savings of between 5 and 10%. In addition to the above, an additional benefit of our approach is its compatibility with existing work on automating the literature screening process (for example, Abstrackr [100, 101]). As training data can be generated automatically from annotation made at prior screening stages, it can be applied before continuing the screening process with other algorithms.

We also presented a sensitivity analysis examining the effect of different parameter choices on the performance of the algorithm. Although good recommendations appear to have been found for most, it was interesting to note the significant variance in performance with minimum confidence thresholds for different data. We note that further analysis with additional review would be interesting, both to discover the cause behind the discrepancy and to mitigate its effect for practical use of our algorithm.

Chapter 5

Identifying Redundant Association Rules with Increased Accuracy

Association rule mining (ARM) is one of the fundamental tasks in data mining. The goal of ARM is to identify interesting relationships between groups of attributes for some data. More formally, let $\mathcal{A} = a_1, a_2, \ldots, a_N$ be a set of M attributes. We then define $\mathcal{N}$ data $\mathcal{D} = d_1, d_2, \ldots, d_N$, where each individual datum or instance is a subset of $\mathcal{A}$ (i.e. $d_i \subseteq \mathcal{A} \forall d_i \in \mathcal{D}$). The association rule mining task seeks to find all *interesting* rules of the form $X \Rightarrow Y$, where X and Y are disjoint subsets of $\mathcal{A}$. By convention, the sets X and Y are referred to as the antecedent and consequent.

Within the literature, there are several sub-problems that have been studied. Approaches have been developed to look for rules such as those with fixed [99], or single attribute [41] consequents, or negative associations [6, 41] (e.g. rules of the form $X \Rightarrow \neg Z$). In this chapter we focus on the problem of positive rules from binary data with single attribute consequents.

It is a well known problem that association rule mining algorithms can return a very large number of rules. The number generated can be so large as to obscure their interpretation and provide a barrier to practical use of the results [112]. Generated only the most interesting rules is therefore an important task. This can be broken into two tasks; identification of truly interesting rules, and the removal of those which are simply redundant artifacts of other rules. Existing approaches for both are covered in sections 2.1.1 and 2.1.3 respectively.

This chapter addresses the issue of rule redundancy, specifically the fact that existing approaches to rule redundancy are based on incomplete information. When comparing two rules $X \Rightarrow Z$ and $XQ \Rightarrow Z$, information on data containing only part of the antecedent is ignored. This is particularly problematic when dealing with noisy or incomplete data. We present an alternate approach to redundancy that makes use of this information in section 5.2.

Section 5.4 describes our evaluation and experimental results. Finally, conclusions are drawn in section 5.5.

5.1 Classical Redundancy

A key problem when generating association rules is measuring how interesting a rule is. Traditionally, this has been done using *support* and *confidence* (which are analogous the sample probability of the rule, and the conditional probability of Y given X respectively). Rules are then considered interesting if they meet some minimum thresholds *minsup* and *minconf*.

Many alternate approaches for measuring interestingness have been proposed [16, 81]. Several good reviews on interestingness measures exist [91, 56], so for brevities sake we do not give a complete coverage here and direct the interested reader to the literature.

Following the seminal work by Agrawal et al. [4] association rules are typically generated using a two stage process based on frequent itemsets. Frequent itemsets are those sets of attributes which have support equal to at least *minsup*. The search typically proceeds by first identifying all frequent itemsets, then evaluating the interestingness for all rules that can be generated from them.

The search for frequent itemsets is aided by the well known anti-monotonicity property of the support function. Algorithms are able to ignore all children of non-frequent itemsets, as it is known that all their descendants will have equal or lower frequencies. For this reason, searching for rules using a support and confidence framework is attractive. However using such an approach suffers from a number of drawbacks, including that infrequent interesting associations will be missed and no real guarantee that the rules will hold in future data.

A number of researchers have generated rules using statistical significance measures such as χ^2 or Fisher's P [40, 41, 99, 104]. This both removes the problem of infrequent interesting associations, and requires only a single, well understood threshold. The search is complicated however by the fact that statistical significance is not monotonic, exponentially increasing the size of the search space.

Prior work has often employed heuristics such as maximum rule lengths, fixed consequents, frequency thresholds, or other heuristics in order to control the size of the search space [40]. To our knowledge, only one algorithm exists which is able to identify all significant, non-redundant rules. Hämäläinen's Kingfisher algorithm, first proposed in 2010, uses a tight bound on Fisher's P to restrict the search space. We extend this approach in our work (see section 5.2 for further details).

It is well established that the number of rules identified during mining can often be so large as to hamper their interpretation [2]. In order to control for this, the concept of *redundancy* is often used. When mining association rules, we consider a rule to be *redundant* if it adds no additional value to existing rules. The identification and removal of redundant rules is an important task. Pruning based on redundancy aims to remove these confounding rules, and return only those representing interesting patterns in the data.

An example of rule redundancy can be seen through the addition of independent attributes. Consider a hypothetical study of supermarket transactions which identifies that people who buy a soft drink will also buy chips. Further analysis may also identify that people who buy soft drink on Tuesday will buy chips. However the addition of the requirement that it be Tuesday does not improve the quality of the association. It is likely that the rule simply exists because the association between people buying soft drink and chips holds regardless of whether it is Tuesday or not. The association between people buying soft drink on Tuesday and buying chips is a redundant rule.

In much of the discussion throughout the remainder of this chapter, we use the concepts of rule *generalisations* and *specialisations*. We provide here a definition for both terms. Let $X \Rightarrow Z$ be a rule. Rule $Y \Rightarrow Z$ is a *generalisation* of $X \Rightarrow Z$ if Y is a proper subset of X. Similarly, rule $Y \Rightarrow Z$ is a *specialisation* of $X \Rightarrow Z$ if Y is a proper superset of X.

A rule is considered *redundant* if the relationship it describes is adequately described by one or more of its generalisations. Existing literature typically takes the view that a more specialised rule adds no value if its interestingness is less than or equal to that of it's generalisation. This concept was recast for a general goodness measure by Hämäläinen [41], which we repeat in definition 5.1.

Definition 5.1. ***Classical redundancy*** *Consider two rules $X \Rightarrow Z$ and $XQ \Rightarrow Z$ where X and Q are disjoint sets of items, and Z is a single item of value a. Let $M(\cdot)$ be an increasing measure of rule interestingness. Rule $XQ \Rightarrow Z$ is redundant with respect to rule $X \Rightarrow Z$ if $M(XQ \Rightarrow Z) \leq M(X \Rightarrow Z)$.*

5.2 Robust Redundancy

In section 5.1 we discussed the idea of rule redundancy. A rule is considered redundant when it adds nothing to a simpler rule. When using the classical definition of redundancy, we define adding nothing by not increasing its interestingness value (see definition 5.1). However such a comparison is made using incomplete information.

Depending on the interestingness measure $M(\cdot)$ in use, $M(X \Rightarrow Z)$ is computed using the frequencies XZ, $\neg XZ$, $X\neg Z$, and $\neg X\neg Z$. We note that comparing rules $M(X \Rightarrow Z)$ and $M(XQ \Rightarrow Z)$ using their interestingness measures does not consider transactions including only part of the rule antecedent. Namely, it does not consider the frequencies $X\neg QZ$, $\neg XQZ$, $X\neg Q\neg Z$, and $\neg XQ\neg Z$.

When performing association rule mining on real data, one must deal with several issues. Complex relationships between variables and noisy data can act to confound the analysis. This problem can be further complicated by a lack of control over the data collection process. Such noise could artificially raise or lower the measured interestingness value of a rule, which could lead to interesting rules being erroneously excluded.

We propose to employ such information in an attempt to avoid excluding interesting rules. In addition, we also propose an approach for identifying situations where seemingly interesting rules are simply artifacts of groups of their specialisations. We refer to these approaches as specialisation and generalisation redundancy respectively, which are outlined in detail throughout the remainder of this section. Section 5.3 then outlines the approach we take to efficiently compute the required partial frequencies, and presents the algorithm used to generate rules.

5.2.1 Specialisation Redundancy

We propose an alternate approach to redundancy in definition 5.2. This approach augments the classical approach given in definition 5.1 by not eliminating a rule $XQ \Rightarrow Z$ if the partial frequencies can be used to demonstrate that adding the attributes in Q add value. This is accomplished by computing the strength of the association between X and Z conditioned on Q, and comparing it against the strength of the marginal association. If the conditional association between X and Z improves over the strength of the previous association, then we have obtained evidence that the addition of the variables in Q adds value to the existing rule.

$P(X)$	0.3
$P(Y)$	0.3
$P(Z\|X, Y)$	0.8
$P(Z\|X, \neg Y)$	0.4
$P(Z\|\neg X, Y)$	0.4
$P(Z\|\neg X, \neg Y)$	0.6

Table 5.1: Marginal and conditional probabilities for several combinations of variables used in the motivating example for robust specialisation redundancy.

Definition 5.2. ***Robust Specialisation Redundancy*** *Consider two rules $X \Rightarrow Z$ and $XQ \Rightarrow Z$ where X and Q are disjoint sets of attributes, and Z is a single attribute. Let $M(\cdot)$ be an increasing measure of rule interestingness. Rule $XQ \Rightarrow Z$ is specialisation redundant with respect to rule $X \Rightarrow Z$ if $M(XQ \Rightarrow Z) \leq M(X \Rightarrow Z)$, and $M(X \Rightarrow Z|Q) \leq M(X \Rightarrow Z)$.*

Computing the conditional association requires the frequencies for $\neg XQZ$ and $\neg XQ\neg Z$ (in addition to the already used frequencies XQZ and $XQ\neg Z$). We deliberately do not consider to association between X and Z conditioned on $\neg Q$, as the rule we are seeking to obtain evidence for is $XQ \Rightarrow Z$, which contains Q.

5.2.1.1 Example

Consider hypothetical data where each individual datum samples 3 binary variables (referred to as X, Y, and Z). Assume that there are 1000 data points, and assume the probabilities expressed in Table 5.1. From these probabilities, it can be observed that a strong dependency exists between Z and the itemset XY. We now examine the quality of the rules $X \Rightarrow Z$, $Y \Rightarrow Z$, and $XY \Rightarrow Z$ as we vary the joint probability of variables X and Y.

Figure 5.1 plots the quality of these rules against the conditional probability of X given Y. We focus on the situation where the conditional probability is less than the marginal (i.e. the probability of X is lower than the marginal given the presence of Y). For larger conditional probabilities we can observe that the quality of the rule $XY \Rightarrow Z$ is superior to that of rule $X \Rightarrow Z$. However as the overlap

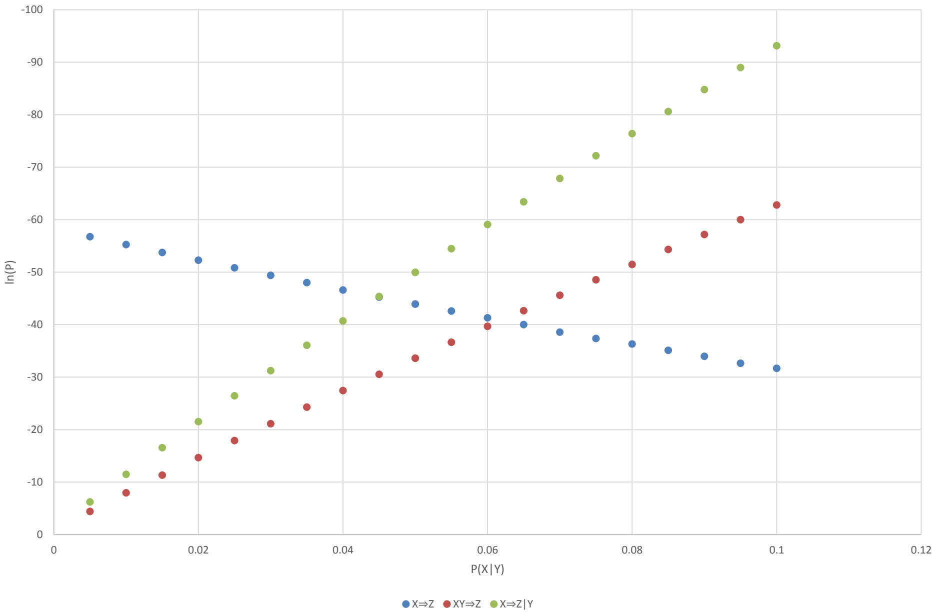


Figure 5.1: ln(P-values) for rules $X \Rightarrow Z$, $Y \Rightarrow Z$, and $XY \Rightarrow Z$ versus conditional probability of X and Y.

between data containing X and data containing Y decreases, the quality of the more general rule $X \Rightarrow Z$ surpasses its specialisation. As a consequence the rule $XY \Rightarrow Z$ is removed as redundant, obscuring the true underlying structure of the data.

Holding the marginal probabilities constant, the number of data containing both X and Y decreases along with the conditional probability of X given Y. In order to support the rule $XY \Rightarrow Z$, we require data containing both (or neither) XY and Z. Hence, as the number of data with XY and Z decreases, we decrease the amount of evidence available to evaluate it. That the general rule $X \Rightarrow Z$ surpasses the true rule $XY \Rightarrow Z$ in quality as the conditional probability decreases is a reflection of this fact.

When comparing rules $X \Rightarrow Z$ and $XY \Rightarrow Z$ using the proposed robust redundancy approach (i.e. including the conditional dependencies), we are able to make more effective use of available data to evaluate the rules. In the current example, rule $XY \Rightarrow Z$ is retained as non-redundant for conditional probabilities greater than ∼0.045. This is in contrast to classical redundancy, where the threshold for retaining $XY \Rightarrow Z$ is ∼0.062. Although in both cases the conditional probability of X given Y eventually reaches a point where insufficient evidence for the specialised rule exists, the range of values for which robust redundancy can still retain $XY \Rightarrow Z$ is increased.

5.2.2 Generalisation Redundancy

It is possible for general rules to exist that only appear interesting due to the presence of interesting specialisations. Definition 5.3 outlines a concept we call *Robust Generalisation Redundancy.* We note that in contrast to specialisation redundancy where a rule is made redundant with respect to another rule, generalisation redundancy requires that a rule is redundant with respect to the entire set of other rules. A rule $X \Rightarrow Z$ is generalisation redundant if for all non-redundant specialisations $XQ \Rightarrow Z$, the rule $X \Rightarrow Z|\neg Q$ is uninteresting (has a goodness

$P(X)$	0.5
$P(Y)$	0.5
$P(Z\|X, Y)$	0.5
$P(Z\|X, \neg Y)$	0.1
$P(Z\|\neg X, Y)$	0.1
$P(Z\|\neg X, \neg Y)$	0.1

Table 5.2: Marginal and conditional probabilities for several combinations of variables used in the motivating example for robust generalisation redundancy.

value less than the required threshold for interesting rules).

If a rule $X\neg Q \Rightarrow Z$ is interesting, we obtain evidence that the generalised rule is interesting even in the absence of the terms in Q. If, after identifying all other interesting rules, we cannot find evidence that $X \Rightarrow Z$ is interesting in the absence of the additional terms in its specialisations, we consider it redundant. Computing the conditional association on $\neg Q$ uses the frequencies $X\neg QZ$ and $X\neg Q\neg Z$. Therefore, by applying both specialisation and generalisation redundancy we consider all frequencies in the sample data.

Definition 5.3. ***Robust Generalisation Redundancy*** *Consider a rule $X \Rightarrow Z$ and the complete set of its non-redundant specialisations $\mathcal{R}$. Let $M(\cdot)$ be an increasing measure of rule interestingness, and α be the corresponding goodness threshold. Rule $X \Rightarrow Z$ is generalisation redundant with respect to $\mathcal{R}$ if $M(X \Rightarrow Z|\neg Q) \leq \alpha$ for all rules $XQ \Rightarrow Z$ in $\mathcal{R}$.*

5.2.2.1 Example

Consider hypothetical data recording prescriptions containing combinations drugs along with a binary patient outcome. Assume that there are two such drugs (X and Y), which work in combination to produce a positive outcome. Neither drugs will produce a positive outcome on its own (in which case, a baseline probability of 0.1 is used). The exact probabilities used in this example can be found in Table 5.2.

It is fairly intuitive to see that when the conditional probability of X given

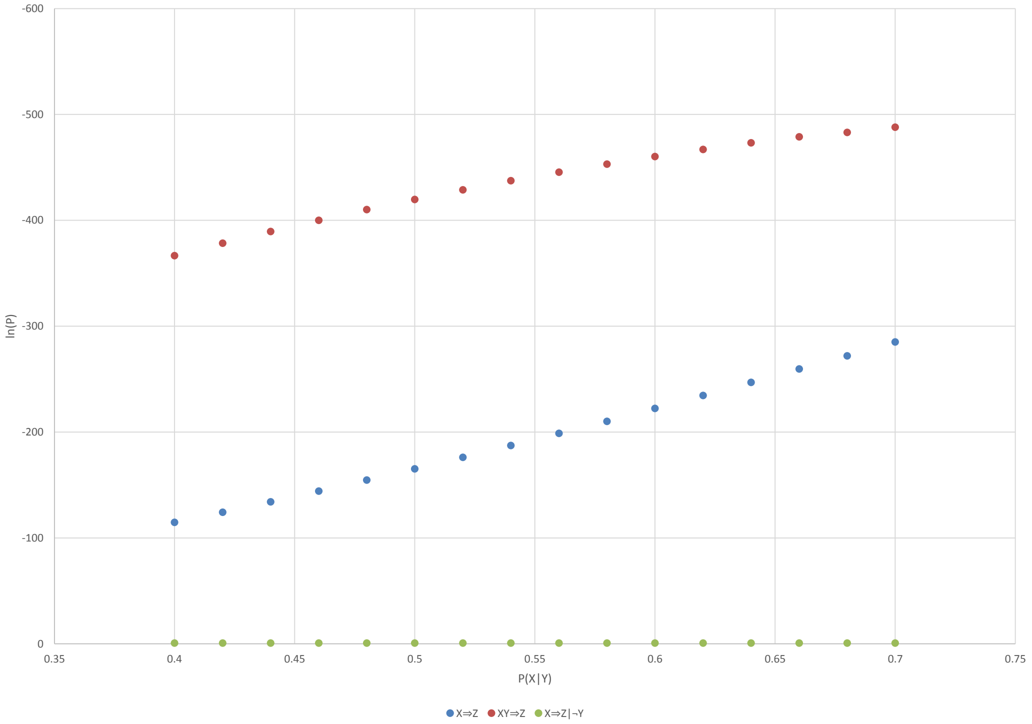


Figure 5.2: ln(P-values) for rules $X \Rightarrow Z$, $Y \Rightarrow Z$, and $XY \Rightarrow Z$ versus conditional probability of X and Y.

Y is 1, the measured quality of the rules $X \Rightarrow Z$, $Y \Rightarrow Z$, and $XY \Rightarrow Z$ will be identical and maximal. As the conditional probability decreases, so will the quality of each of these rules, with the quality of the general rules decreasing at the greatest rate. However despite the underlying structure of the data indicating that neither X or Y alone support a positive outcome, the strength of these associations is likely to remain quite high.

When comparing the rules $X \Rightarrow Z$ and $XY \Rightarrow Z$, by examining the strength of the rule $X \Rightarrow Z|\neg Y$ (i.e. conditioned on the absence of the additional terms Y) we can see that there is no evidence to support the rule $X \Rightarrow Z$ without also including the features Y. As $XY \Rightarrow Z$ is the only identified specialisation of $X \Rightarrow Z$, and there is no evidence to indicate $X \Rightarrow Z$ is valid without the additional features, we therefore consider it redundant.

Finally, we acknowledge that such an approach could potentially over-fit and remove valid general rules. We address this concern in the following section on redundancy chaining.

5.2.3 Redundancy Chaining

Classical redundancy as defined in definition 5.1 is transitive. If a rule $XQY \Rightarrow Z$ is redundant with respect to a generalisation $XQ \Rightarrow Z$, and $XQ \Rightarrow Z$ is also redundant with respect to $X \Rightarrow Z$, then $XQY \Rightarrow Z$ will be redundant with respect to $X \Rightarrow Z$. This result is straightforward to prove.

Unfortunately the same relation does not hold for the proposed robust redundancy approaches. A proof by example is given for specialisation redundancy in in lemma 5.1.

Lemma 5.1. ***Robust specialisation redundancy is not transitive.***

Consider three rules $A \Rightarrow D$, $AB \Rightarrow D$, and $ABC \Rightarrow D$ generated from the data in Table 5.3 using the log of Fishers P.

As the interestingness of the rules $AB \Rightarrow D$ and $A \Rightarrow D|B$ is worse than that of the rule $A \Rightarrow D$, $AB \Rightarrow D$ is redundant w.r.t. $A \Rightarrow D$.

Attr.	fr	Rule	ln(P)
ABCD	10	$A \Rightarrow D$	-19.33
ABD	10	$AB \Rightarrow D$	-8.10
ACD	10	$ABC \Rightarrow D$	-3.78
AD	10	$A \Rightarrow D\|B$	-18.75
BC	30	$A \Rightarrow D\|BC$	-20.56
BD	10	$AB \Rightarrow D\|C$	-7.83
CD	10		
D	10		

Table 5.3: Sample data and rules for lemma 5.1.

As the interestingness of the rules $ABC \Rightarrow D$ and $AB \Rightarrow D|C$ is worse than that of the rule $AB \Rightarrow D$, $ABC \Rightarrow D$ is redundant w.r.t. $AB \Rightarrow D$.

As the interestingness of the rule $A \Rightarrow D|BC$ is better than that of the rule $A \Rightarrow D$, the rule $ABC \Rightarrow D$ is non-redundant w.r.t. $A \Rightarrow D$. ■

The lack of a transitive definition for specialisation redundancy creates an interesting possibility. Assume the existence of a rule r that is specialisation redundant with respect to one or more generalisations $r_0, \dots r_i$. Let those generalisations $r_0, \dots r_i$ themselves be redundant with respect to rules $r_{i+1}, \dots r_n$. Despite being a redundant specialisation of other rules, it could be argued that r should be kept as it is non-redundant with respect to all non-redundant generalisations.

We take the view that in such a situation, the rule r should be considered non-redundant. This more permissive approach to rule inclusion is preferable as it minimises the chance of interesting rules being removed. We also feel it is more intuitive, as it avoids the possibility that a rule is omitted from the computed rule set despite being interesting with respect to all other returned rules.

Lemma 5.2. ***Using redundant rules when evaluating generalisation redundancy allows for additional rules to be included.***

Consider three rules $A \Rightarrow D$, $AB \Rightarrow D$, and $ABC \Rightarrow D$ generated from the data in Table 5.4 using the confidence measure with a threshold of 0.6.

According to definition 5.3 and the confidence scores for the above rules, $AB \Rightarrow D|\neg C$ is uninteresting so $AB \Rightarrow D$ is redundant w.r.t. $ABC \Rightarrow D$.

Attr.	fr	Rule	Conf
ABCD	60	$A \Rightarrow D$	0.90
AB	20	$AB \Rightarrow D$	0.75
ACD	10	$ABC \Rightarrow D$	1.00
AD	10	$A \Rightarrow D \vert \neg B$	1.00
		$A \Rightarrow D \vert \neg(BC)$	0.50
		$AB \Rightarrow D \vert \neg C$	0.00

Table 5.4: Sample data and rules for lemma 5.2.

Assume we evaluate generalisation redundancy WITHOUT redundant rules. Then uninteresting rule $A \Rightarrow D|\neg(BC)$ implies $A \Rightarrow D$ is redundant and will be pruned.

Alternately, assume we evaluate generalisation redundancy WITH redundant rules. As $A \Rightarrow D|\neg B$ is interesting $A \Rightarrow D$ is non-redundant and will not be pruned.

We also prove that whether or not redundant attributes are counted effects generalisation redundancy in lemma 5.2. For similar reasons to specialisation redundancy, we elect not to allow redundant rules to influence the redundancy of another rule. In contrast to specialisation redundancy however, this policy will lead to the exclusion of additional rules (as generalisation redundancy requires a rule is uninteresting with respect to ALL its specialisations, comparing against additional rules only raises the chance of inclusion).

Not allowing redundant rules to provide evidence for keeping otherwise redundant generalisations has the potential to produce the following interesting situation. Assume the existence of a rule $r_1 : Y \Rightarrow A$ where $conf(Y \Rightarrow A) = 1$ and $supp(Y) = supp(A)$. Then for all rules of the form r_1 $X \Rightarrow A$ where $Y = XQ$ (i.e. generalisations of $Y \Rightarrow A$), the frequency of the set $X\neg QA$ will be 0, implying that the rule $X\neg Q \Rightarrow A$ will be uninteresting. By the definition of generalisation redundancy, r_1 is the only possible non-redundant rule with consequent A.

While it may in fact be desirable to keep such a rule, care must be taken to avoid confounding caused by the addition of frequent, independent attributes.

We demonstrate how such confounding might occur by providing an extension of the above example. Consider the rule $YZ \Rightarrow A$ for some variable Z where $supp(ZA) = 1$. It is simple to see that $conf(YZ \Rightarrow A) = 1$, $supp(YZ) = supp(A)$, and $freq(Y\neg ZA) = 0$. As such, the rule $Y \Rightarrow A$ will be considered redundant.

By Occam's razor, we prefer a more general rule over one of its specialisations unless we can obtain evidence to suggest otherwise. If we apply only generalisation redundancy, we can violate this principle as no evidence is ever considered to support $YZ \Rightarrow A$ over $Y \Rightarrow A$. In the worst case, for a given consequent only one, highly specific rule will be selected with all others being made redundant. We therefore suggest that specialisation (or classical) redundancy should usually be used before generalisation redundancy. This ensures that evidence exists that each rule improves upon its generalisations. We note however that in some cases (such as those where we prefer to generate more specific rules), generalisation redundancy may be applied first.

5.2.3.1 Example

In section 5.2.3, it was proposed that redundant specialisations should be removed before generalisations. A good example of the potential cost of pruning generalisations before specialisations can be observed in the well known Mushroom data from the UCI Machine Learning Repository (see section 5.4.1 for more detail).

Table 5.9 shows the number average number of rules generated for the Mushroom data with a range of thresholds. For clarity the relevant parts of that table are reproduced in Table 5.5. An additional column has been included showing the number of rules generated when pruning only redundant generalisations.

From the results in Table 5.5, we can see that using only generalisation pruning for two of the three tested data produces a significantly lower number of rules than with any other approach. In fact, for all tested thresholds only a handful of rules remain. This result can be explained by examining the unpruned rules and observing two things. Firstly, on the vast majority of rules produced contain

Dataset	α	No Prune	Classic	Robust Specialisations	Robust Generalisations	Robust (Both)
Mushroom	-1250	61767.80 ± 158.39	409.70 ± 5.32	568.70 ± 7.42	19.80 ± 0.37	227.40 ± 2.95
	-1375	37501.10 ± 4080.51	308.10 ± 7.70	342.40 ± 10.14	19.90 ± 0.81	166.00 ± 6.38
	-1500	22634.50 ± 92.39	229.70 ± 5.67	239.40 ± 5.97	15.20 ± 0.54	125.50 ± 1.67
	-1625	22049.80 ± 39.14	191.30 ± 2.54	196.30 ± 2.85	11.50 ± 0.42	114.80 ± 1.77
	-1750	19980.00 ± 2498.40	140.70 ± 5.51	141.70 ± 5.51	7.20 ± 0.25	93.30 ± 6.17
	-1875	7507.80 ± 78.31	88.60 ± 3.75	89.60 ± 3.75	4.80 ± 0.25	56.70 ± 1.08
	-2000	6430.80 ± 522.10	38.90 ± 5.39	39.90 ± 5.39	2.30 ± 0.40	34.70 ± 4.15

Table 5.5: Number of rules generated with different redundancy approaches (including pruning only generalisations) for the Mushroom data.

the same consequent (feature #48). Secondly, the unpruned set of rules always contains the following rule with feature #48 as the consequent.

$$1 \cdot 24 \cdot 34 \cdot 36 \cdot 38 \cdot 53 \cdot 58 \cdot 85 \cdot 86 \cdot 90 \cdot 94 \cdot 102 \cdot 110 \Rightarrow 48$$

The mushroom data contains no instances with only the antecedent or consequent of the above rule (i.e. the set of instances containing the antecedent is identical to the set containing the consequent). For a given generalisation $X \Rightarrow A$ of this rule, this implies that the frequencies of the sets $X\neg QA$, $\neg X\neg QA$, and $X\neg Q\neg A$ will be 0, guaranteeing $X \Rightarrow A|\neg Q$ will be uninteresting. However all other discovered rules (with antecedent #48) are generalisations of this rule. As we do not allow redundant specialisations to be used in pruning, this rule renders every single generalisation redundant.

The presence of these highly specific rules with no violations presents a problem for generalisation pruning. The justification we use when removing generalisations is that they are uninteresting in the absence of the additional attributes used in their specialisations. Although for a given data it may in fact be desirable to keep such a specialised rule, it is crucial to first test whether or not the addition of the extra features improves over its generalisations.

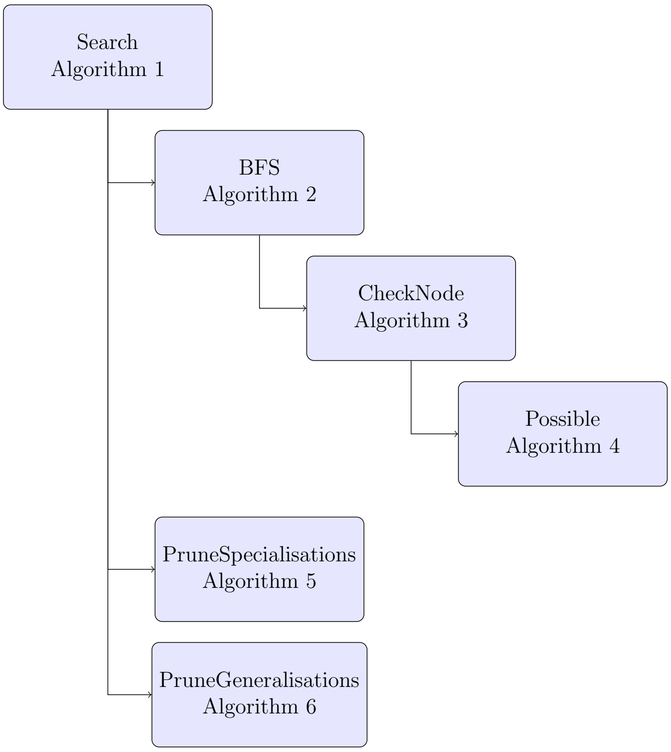


Figure 5.3: Overview of dependencies between algorithms used in the rule generation process.

5.3 Rule Generation Algorithm

The algorithm we use is a variant on Hämäläinen's Kingfisher algorithm [40, 41]. Pseudocode for our algorithm is given in Algorithms 1 to 6. Algorithms 1 to 4 are similar to Hämäläinen's Kingfisher [40] excepting that no minimality based pruning is performed. Additionally, bounds in Algorithm 4 are computed in line with the definition of robust redundancy given in definition 5.2. A diagram showing the dependencies between algorithms is also given in Figure 5.3. This section describes the process used, and the differences between our approach and Kingfisher.

We search for all non-redundant rules using the natural log of the Fisher's P measure (a decreasing measure). Let $\mathcal{D}$ be a dataset with $\mathcal{N}$ items over $\mathcal{A}$ attributes. We use a three stage process:

1. All potentially non-redundant rules with some minimum log P-value are identified.

Algorithm 1: Search($\mathcal{D}$, $\mathcal{A}$, M, α)
Search algorithm for non-robust redundant association rules.

```
input  : A set of data $\mathcal{D}$ over attributes $\mathcal{A}$, an increasing interestingness
         measure $M(\cdot)$, and a corresponding threshold $\alpha$
output: A set of rules $\mathcal{R}$

// Step 1:  Find potentially interesting rules
determine minf
r $\leftarrow$ Level1nodes($\mathcal{D}$,$\mathcal{A}$,$M$,$\alpha$,minf)
l $\leftarrow 2$
nls $\leftarrow$ |r|
while nls $\geq$ l do
    nls $\leftarrow 0$
    for $i \leftarrow 1$ to $|\mathcal{A}|$ do
        nls $\leftarrow$ nls + Bfs(r.children[i],l,0)
    end
    l $\leftarrow$ l + 1
end

// Steps 2 and 3:  Prune redundant rules
$\mathcal{R} \leftarrow$ PruneSpecialisations($\mathcal{R}$)
$\mathcal{R} \leftarrow$ PruneGeneralisations($\mathcal{R}$, $\alpha$)

return $\mathcal{R}$
```

Algorithm 2: BFS(n, l, t)

BFS search for potentially non-redundant rules.

```
input  : Root node n, target search level l, current level t
output: The number of level l nodes remaining in the tree

nls ← 0
if t = l − 2 then
    for $i \leftarrow 1$ to $|\mathsf{n}.children| - 1$ do
        Y ← n.children[i].set
        for $j \leftarrow i + 1$ to $|\mathsf{n}.children|$ do
            Z ← n.children[j].set
            $\mathsf{X} \leftarrow \mathsf{Y} \cup \mathsf{Z}$
            Create new node child = Node(X)
            nls ← nls + 1
            n.children[i].insert(child)
            if Checknode(child) = false then
                delete child
                nls ← nls − 1
                for $\forall$ nodes v = Node($\mathsf{Y}_m$) where $\mathsf{X} = \mathsf{Y}_m\mathsf{A}_m$ do
                    v.possible[m] = false
                end
            end
            delete n.children[$|n.children|$]
        end
    end
else
    for $i \leftarrow 1$ to $|\mathsf{n}.children|$ do
        nls ← nls + Bfs(n.children[i],l,t + 1)
    end
end
if $|\mathsf{n}.children| = 0$ then
    delete node n
end
return nls
```

Algorithm 3: Checknode(v_X)

Generate rules from a node and check if it can have valid descendants.

input : Node v_X to check

output: Boolean value indicating whether or not children of the v_X can produce interesting, non-redundant rules.

```
for $\forall$ Y $\subset$ X where $|Y| = |X| - 1$ do
    $Par_Y \leftarrow$ searchTree(Y)
    if $Par_Y$ not found then
        return false
    end
    for $i \leftarrow 1$ to $|\mathcal{A}|$ do
        $v_X.possible[i] \leftarrow v_X.possible[i]$ & $Par_Y.possible[i]$
        $v_X.pbest[i] \leftarrow$ min($v_X.pbest[i]$,$Par_Y.pbest[i]$)
    end
    if $v_X.possible = \emptyset$ then
        return false
    end
end
setfreq(X) = calcFreq(X)
for $\forall$ $A_i \in \mathcal{A}$ do
    $v_X.possible[i] \leftarrow v_X.possible[i]$ & possible(-)
    if (($A_i \in X$) and ($v_X.possible[i]$)) then
        val $\leftarrow M(X \backslash \{A_i\} \Rightarrow A_i)$
        if val $\leq \alpha$ then
            add rule $X \backslash \{A_i\} \Rightarrow A_i$ to $\mathcal{R}$
            $v_X.pbest[i] \leftarrow$ min(val,$v_X.pbest[i]$)
        end
    end
end
if $v_X.possible = \emptyset$ then
    return false
end
return true
```

Algorithm 4: Possible(v_X,X,A)

Check if rules generated from a node or its descendants with a given consequent can be non-redundant

```
input  : Node $v_X$ being checked, set X associated with that node, and
         consequent attribute $A_j$
output: True if rules with consequent A generated using $v_X$ or its
         descendants can be interesting and non-redundant

if $|X| < minf$ then
    return false
end
if $A \notin X$ then
    if $|X| > |A|$ then
        bnd = LB1($|A|, |\mathcal{D}|$)
    else
        bnd = LB2($|X|, |A|, |\mathcal{D}|$)
    end
else
    bnd $\leftarrow$ LB3($|X|, |X \setminus \{A\}|, |A|, |\mathcal{D}|$)
    bndonq $\leftarrow$ LB3($|X|, |X|, |X|, |\mathcal{D}| - |X \neg A| - |A \neg X|$)
end
if bnd $> \alpha$ then
    return false
end
if $A \in X$ then
    if bnd $\geq v_X.pbest[j]$ and bndonq $\geq v_X.pbest[j]$ then
        return false
    end
end
return true
```

Algorithm 5: PruneSpecialisations($\mathcal{R}$)
Prune redundant specialisations in $\mathcal{R}$

```
input  : Set of rules $\mathcal{R}$
output: Set of non (specialisation) redundant rules $\mathcal{R}$
for $\forall$ $\mathsf{A} \in \mathcal{A}$ do
    $\mathcal{R}_\mathsf{A} = \{\mathsf{R} \in \mathcal{R}$ s.t. $\mathsf{R} = \mathsf{X} \Rightarrow \mathsf{A}\}$
    Sort $\mathcal{R}_\mathsf{A}$ in increasing order on length of the antecedent
    for $i \leftarrow 1 \rightarrow |\mathcal{R}_\mathsf{A}| - 1$ do
        Let $\mathcal{R}_\mathsf{A}[i] = \mathsf{X} \Rightarrow \mathsf{A}$
        for $j \leftarrow i + 1 \rightarrow |\mathcal{R}_\mathsf{A}|$ do
            Let $\mathcal{R}_\mathsf{A}[j] = \mathsf{Y} \Rightarrow \mathsf{A}$
            $\mathsf{Q} = \mathsf{Y} \backslash \mathsf{X}$
            $\mathsf{CondM} \leftarrow \mathtt{M(}\mathsf{setFreq}[XQA], \mathsf{setFreq}[XQ], \mathsf{setFreq}[QA], \mathsf{setFreq}[Q]\mathtt{)}$
            if $\mathsf{X} \subset \mathsf{Y}$ then
                if $\mathtt{M(R_2)} \leq \mathtt{M(R_1)}$ and $\mathsf{CondM} \leq \mathtt{M(R_1)}$ then
                    delete $\mathsf{R}_2$
                end
            end
        end
    end
end
return $\mathcal{R}$
```

Algorithm 6: PruneGeneralisations($\mathcal{R}$, α)
Prune redundant generalisations in $\mathcal{R}$

```
input  : Set of rules $\mathcal{R}$ and an interestingness threshold $\alpha$
output: Set of non (generalisation) redundant rules $\mathcal{R}$
for $\forall$ R $\in \mathcal{R}$ do
    Keep(R) $= false$
    HasSpec(R) $= false$
end
for $\forall$ A $\in \mathcal{A}$ do
    $\mathcal{R}_A = \{R \in \mathcal{R}$ s.t. $R = X \Rightarrow A\}$
    Sort $\mathcal{R}_A$ in decreasing order on length of the antecedent
    for $i \leftarrow 1 \rightarrow |\mathcal{R}_A| - 1$ do
        if Keep($\mathcal{R}_A[i]$) $= false$ and HasSpec($\mathcal{R}(A)[i]$)$=true$ then
            continue
        end
        Let $\mathcal{R}_A[i] = X \Rightarrow A$
        for $j \leftarrow i + 1 \rightarrow |\mathcal{R}_A|$ do
            Let $\mathcal{R}_A[j] = Y \Rightarrow A$
            $Q = Y \backslash X$
            CondM $\leftarrow$
             M(setFreq$[XA\neg Q]$, setFreq$[X\neg Q]$, setFreq$[A\neg Q]$, setFreq$[\neg Q]$)
            if $X \subset Y$ then
                HasSpec(X) $= true$
                if CondM $\leq \alpha$ then
                    Keep(X) $= true$
                end
            end
        end
    end
end
for $\forall$ R $\in \mathcal{R}$ do
    if Keep(R) $= false$ then
        delete R
    end
end
return $\mathcal{R}$
```

2. Rules identified in stage 1 are examined and specialisation redundant rules are pruned.

3. Remaining rules are examined and generalisation redundant rules are pruned.

The search in stage 1 consists of a bfs over itemsets. For a level k node X corresponding to attributes $\{x_1, x_2, \ldots, x_k\}$, the P-values of the k rules $X \setminus \{x_i\} \Rightarrow x_i$ are computed, with those meeting the desired minimum threshold kept. As each node is considered, the frequency of the set X is calculated, with P-values for rules being computed using frequencies computed for the parents. The number of iterations over the dataset is therefore limited to the number of nodes considered.

In order to control the size of the search space, each node maintains a length $|\mathcal{A}|$ bit vector of possible consequents (attributes A where the rule $XQ \Rightarrow A$ is possible). These vectors are initialised as the bitwise and of the vectors for a nodes parents. As the node is processed, lower bounds on the log Fisher's P value are then computed for all rules of the form $XQ \setminus \{A\} \Rightarrow A$ for all A in $\mathcal{A}$. If the bounds for all attributes exceed the relevance threshold (the vector of possible consequents is 0), the node is removed and no further descendants are generated. The bounds used were first reported by Hämäläinen [40], and are reproduced here in Table 5.6.

Each node X also contains a vector with the best previous P-value for rules with consequent $x_i \in X$. Similar to the possible bit vector, these vectors are merged from parents when the node X is created. As the Kingfisher algorithm employs classical redundancy, if the bound on P-values for rules with a given consequent exceeds the corresponding value in this vector that consequent can also be considered impossible. In order for an attribute to be considered an impossible consequent with the proposed redundancy in definition 5.2, an additional test must be applied. We need to test that the bound on the rule $XQ \Rightarrow A|Q$ is also worse than the previous best value.

The Fishers P-value for rule $XQ \Rightarrow A|Q$ takes its smallest value when the number of instances containing sets $QA\neg X$ and $QX\neg A$ are 0 and QXA and

$\mathit{bnd1}(\lvert A\rvert, \mathcal{N}) =$	$\dfrac{f(A)!f(\neg A)!}{\mathcal{N}!}$
$\mathit{bnd2}(\lvert X\rvert, \lvert A\rvert, \mathcal{N}) =$	$\dfrac{f(\neg X)!f(A)!}{\mathcal{N}!(f(A)-f(X))!}$
$\mathit{bnd3}(\lvert XA\rvert, \lvert X\rvert, \lvert A\rvert, \mathcal{N}) =$	$\dfrac{f(A)!f(\neg A)!(\mathcal{N}-f(XA))!}{\mathcal{N}!f(\neg A)!f(A\neg X)!}$

Table 5.6: Lower bounds for Fishers P as computed by Hämäläinen [40]. The function $f(\cdot)$ returns the frequency of its argument in $\mathcal{D}$

$Q\neg X\neg A$ are as large as possible. This occurs when $freq(QXA) = freq(XA)$, $freq(Q\neg X\neg A) = freq(\neg X\neg A)$. We therefore compute the bound for $XQ \Rightarrow A|Q$ using bnd3 from Table 5.6 with parameters $f(XA) = freq(XA)$, $f(X) = freq(XA)$, $f(A) = freq(XA)$, and $\mathcal{N} = freq(XA) + freq(\neg X\neg A)$.

The Kingfisher algorithm employs two additional pruning steps to control the size of the search space. They are covered in section 5.3.1.

The time and space complexity of the Kingfisher algorithm are exponential with regard to the number of attributes [40]. The algorithm employed in stage 1 of our search differs from Kingfisher only in the pruning strategies employed, and as such maintains the same worst case time and space complexity. Although the less aggressive pruning employed in our work does result in increased complexity, we note that in practice performance is still reasonable. Section 5.4.4 presents an empirical evaluation of the speed and memory requirements of our approach on several datasets.

The running time for the searches in stage 2 and 3 are quadratic in the number of rules tested (in general this is dwarfed by the initial search in stage 1). When comparing two rules $X \Rightarrow A$ and $XQ \Rightarrow A$, specialisation redundancy requires the computation of $M(X \Rightarrow A|Q)$, and generalisation redundancy requires $M(X \Rightarrow A|\neg Q)$. For Fisher's P, this requires us to obtain the frequencies for Q, $\neg Q$, AQ, and $A\neg Q$.

If the number of rules is sufficiently small, it may be preferable to obtain these frequencies by iterating over $\mathcal{D}$ as required. We note however that for a rule $X \Rightarrow A$ to be generated in our search, the node $X = \{A\}$ (and all its generalisations) must be considered. Our implementation therefore creates a map from sets to their frequency as each node is constructed. Obviously, this adds additional space requirements which may be avoided by computing frequencies as required.

5.3.1 Pruning the Search Space

As mentioned previously, the Kingfisher algorithm employs two additional pruning steps to control the size of the search space. The first, referred to as the *lapis philosophorum* principle, deals with the case where all rules of the form $XQ \Rightarrow A$ become impossible at a given node $X\{A\}$. In such a case, A is also an impossible consequence for children of the parent node X, and it's possible consequents vector can be updated. This greatly improves the efficiency of the search, and is also applied in our approach.

The latter pruning step is pruning based on *minimality.* A rule $X \Rightarrow A$ is considered minimal iff $P(A|X) = 1$. It can be proven (Hämäläinen [41]) that for a given minimal rule $X \Rightarrow A$ any rule of the form $XQ \Rightarrow A$ or $XQA \Rightarrow B$ will be either classically redundant or not significant. The Kingfisher algorithm therefore employs pruning based on minimality to restrict the search space.

Unfortunately, pruning based on minimality cannot be employed when searching for rules with robust redundancy. We now prove that with robust redundancy it is possible for a specialisation of a minimal rule to be both significant and non-redundant.

Lemma 5.3. ***Given data $\mathcal{D}$, an increasing statistical goodness measure $M(\cdot)$, and a rule $X \Rightarrow A$ such that $P(A|X) = 1$, there may exist a rule $XQ \Rightarrow A$ such that $M(X \Rightarrow A) < M(X \Rightarrow A|Q)$.***

That $X \Rightarrow A$ is minimal implies that the frequency of the set $X\neg A$ is 0. The frequencies of the sets XA, $\neg XA$, and $\neg X\neg A$ are unknown.

Let Q be a set of attributes whose corresponding rows in $\mathcal{D}$ exactly match the sets XA and $\neg X\neg A$. $M(X \Rightarrow A)$ increases with each occurrence of XA and $\neg X\neg A$, and decreases with each occurrence of $\neg XA$. It is easy to observe that $freq(XA) = freq(XQA)$, $freq(\neg X\neg A) = freq(\neg XQ\neg A)$, and $freq(\neg XA) \geq freq(\neg XQA)$. Assuming $\mathcal{D}$ contains at least one occurrence of $\neg XA$, $M(X \Rightarrow A|Q)$ will therefore be greater than $M(X \Rightarrow A)$. ■

Lemma 5.4. ***Given data $\mathcal{D}$, an increasing statistical goodness measure $M(\cdot)$, and a rule $X \Rightarrow A$ such that $P(A|X) = 1$, there may exist a rule $XQA \Rightarrow B$ such that $M(XA \Rightarrow B) < M(XA \Rightarrow B|Q)$.***

That $X \Rightarrow A$ is minimal implies that the frequency of the set $X\neg A$ is 0. This implies that $freq(XA) \geq freq(XQA)$. The frequencies of the sets XA, $\neg XA$, and $\neg X\neg A$ are unknown.

Let Q be a set of attributes whose corresponding rows in $\mathcal{D}$ exactly match the sets CB and $\neg C\neg B$ where $C = XA$. It is easy to observe that $freq(C) = freq(CQ)$, $freq(\neg C\neg B) = freq(\neg CQ\neg B)$, and $freq(\neg C) \geq freq(\neg CQ)$. Assuming $\mathcal{D}$ contains at least one occurrence of $\neg CB$, $M(C \Rightarrow B|Q)$ will therefore be greater than $M(C \Rightarrow B)$ (or $M(XA \Rightarrow B|Q) > M(XA \Rightarrow B)$). ■

As such, we do not employ pruning based on minimality when searching for rules with robust redundancy.

5.4 Evaluation

We evaluate the performance of the proposed robust pruning with respect to three characteristics. These are the total number of rules generated, the overall quality of the rules, and the efficiency of the rule generation process.

Experiments were run on a PC running Ubuntu Linux, with an Intel I7-4500 processor and 8gb RAM. The rule generation algorithm was implemented in C++.

Performance is also reported for rules generated with the classical definition of redundancy (definition 5.1). We generate these rules using the Kingfisher algorithm [41]. Baseline performance of rules generated with no redundancy based pruning is also reported.

We examine the results from three perspectives; the size of the generated rule set, quality of the generated rules, and the efficiency of the rule generation process. We first present an overview of the data used in our evaluation.

5.4.1 Data

We perform our evaluation on multiple data sets which we describe below. Most test collections are standard datasets. We also report results for several additional domains such as text.

- **Mushroom** Descriptions of mushrooms originally collected from the 1981 Audobon Society Field Guide to North American Mushrooms. This data is widely used, and is freely available from the UCI Machine Learning Repository [1].

- **T10I4D100K** An artificial dataset representing market basket data. Originally produced using the now unavailable generator from the IBM Almaden Quest research group, this data was obtained from the Frequent Itemset Mining Dataset Repository [2].

- **T40I10D100K** An artificial dataset representing market basket data. Originally produced using the now unavailable generator from the IBM Almaden Quest research group, this data was obtained from the Frequent Itemset Mining Dataset Repository [2].

- **Diabetes** Collection of real world data reporting traditional Chinese medical herbal prescriptions for diabetes. Includes both the herbs prescribed and a

[1] https://archive.ics.uci.edu/ml/datasets/Mushroom

[2] http://fimi.ua.ac.be/data/

Name	# Instances	# Attributes	Agv. Instances Length	Avg. Attribute Freq.
Aspergillosis	4377	101	15.93 ± 0.26	680.51 ± 66.05
Mushroom	8124	119	23.00 ± 0.00	1624.80 ± 358.73
Diabetes	1915	204	10.26 ± 0.11	105.21 ± 30.09
Fertility	766	215	15.73 ± 0.32	59.62 ± 14.21
Insomnia	460	112	13.48 ± 0.25	55.38 ± 11.10
T10I4D100K	100000	870	10.10 ± 0.02	1161.18 ± 74.73
T40I10D100K	100000	942	39.61 ± 0.05	4204.36 ± 249.57

Table 5.7: Summary of datasets used in the evaluation.

binary classification of the patient outcome as ’good’ or ’bad’.

- **Fertility** Collection of real world data reporting traditional Chinese medical herbal prescriptions for fertility. Includes both the herbs prescribed and a binary classification of the patient outcome as ’good’ or ’bad’.

- **Insomnia** Collection of real world data reporting traditional Chinese medical herbal prescriptions for insomnia. Includes both the herbs prescribed and a binary classification of the patient outcome as ’good’ or ’bad’.

- **Aspergillosis** Text documents (titles and abstracts) for articles considered for inclusion in a systematic review on Aspergillosis [58]. Each document is converted to a binary vector indicating the presence or absence of each of 100 words, as well as a binary variable indicating whether the title and abstract was potentially relevant to the review. The words selected were those with the greatest discriminative power when identifying articles relevant to the review.

Descriptive statistics for each of the data are provided in Table 5.7.

All values reported were obtained as the average of 10 independent experiments. For each experiment, the data were randomly divided into a 50/50 test/training split. Rules were generated using the training data, and evaluated on the hold out test set. Results are reported with their 95% confidence interval. When reporting statistically significant differences, all results are tested with a P-value of 0.05.

Interestingness for rules is measured using the natural log of P values computed using Fishers Exact test. This is a decreasing, strictly negative measure (lower values indicate stronger dependencies). The thresholds for interesting rules were chosen to strike a balance between permissiveness and execution time, and differ between data and experiments. Thresholds are reported along with results.

5.4.2 Size of the Rule Set

Tables 5.8 and 5.9 show the number of rules generated for each dataset, pruning approach, and threshold. For all tested data and thresholds, we can observe that each pruning approach is able to identify and eliminate a substantial percentage of rules when compared to generating rules without pruning. We also observe that for almost all tested data, the number of rules generated with the proposed robust redundancy is significantly lower than the number generated by the classical approach (P=.05). For the three exceptions (Insomnia data with thresholds -30 and -35, and Mushroom with threshold -2000), no significant difference in means is observed.

The cases where no significant difference in the number of non-redundant rules is observed occur when using the strictest tested thresholds. In addition, the difference between the mean number of rules generated appears to increase as the interestingness threshold is relaxed. The number of rules appears to converge as the bound on interesting rules is tightened, and diverge as it is relaxed. This supports the conclusion that our proposed approach is able to produce a practical number of rules from a larger number of potentially interesting associations. This quality is desirable as it allows the use of relaxed interestingness thresholds, lowering the risk of missing potentially useful associations.

Dataset	α	No Prune	Classic	Robust Specialisations	Robust (Both)
Aspergillosis	-50	29548.60 ± 4461.94	389.40 ± 31.39	391.60 ± 31.48	268.30 ± 21.10
	-75	4207.80 ± 352.07	119.20 ± 6.36	120.10 ± 6.41	88.90 ± 2.82
	-100	1073.60 ± 124.63	55.90 ± 5.90	56.20 ± 5.87	41.70 ± 3.45
	-125	383.10 ± 33.97	26.90 ± 2.55	27.00 ± 2.57	22.20 ± 1.77
	-150	163.90 ± 16.29	19.50 ± 1.55	19.50 ± 1.55	16.50 ± 1.28
	-175	89.90 ± 6.74	14.60 ± 1.08	14.70 ± 1.04	12.90 ± 1.02
	-200	49.40 ± 5.33	10.00 ± 1.04	10.20 ± 1.10	9.20 ± 0.77
Diabetes	-15	34543.20 ± 12832.45	1327.80 ± 99.71	1676.30 ± 235.23	823.60 ± 83.15
	-20	11816.40 ± 2574.97	613.00 ± 52.78	699.90 ± 73.42	394.40 ± 38.01
	-25	4152.80 ± 227.20	343.40 ± 18.07	365.00 ± 21.57	224.90 ± 10.38
	-30	2731.10 ± 229.94	244.50 ± 10.73	248.60 ± 11.75	162.40 ± 9.20
	-35	1656.20 ± 102.51	180.80 ± 7.58	183.40 ± 7.73	126.30 ± 5.47
	-40	1198.80 ± 86.62	143.60 ± 7.21	145.10 ± 7.38	102.50 ± 4.95
	-45	782.30 ± 78.51	107.10 ± 11.24	107.20 ± 11.29	78.80 ± 7.13
Fertility	-15	362405.60 ± 86789.45	595.30 ± 45.95	618.30 ± 48.06	352.80 ± 25.40
	-20	141740.50 ± 41787.17	278.00 ± 15.74	283.80 ± 15.80	176.80 ± 12.03
	-25	42389.60 ± 11074.22	176.10 ± 5.92	178.20 ± 6.07	111.20 ± 5.71
	-30	20686.90 ± 4609.10	132.50 ± 14.70	133.20 ± 15.04	85.00 ± 9.83
	-35	12736.90 ± 2347.48	113.60 ± 6.68	114.00 ± 7.06	72.80 ± 4.27
	-40	7713.60 ± 1830.05	90.50 ± 7.11	90.80 ± 7.11	62.40 ± 6.12
	-45	4718.80 ± 687.50	74.50 ± 10.29	74.60 ± 10.33	48.70 ± 7.68
Insomnia	-15	12812.70 ± 2399.55	624.00 ± 48.35	747.80 ± 64.69	402.10 ± 33.42
	-20	3180.50 ± 998.68	243.60 ± 37.06	269.00 ± 40.89	161.20 ± 24.88
	-25	876.70 ± 362.33	104.50 ± 15.75	112.50 ± 14.98	72.30 ± 12.27
	-30	271.10 ± 36.97	43.60 ± 6.41	46.80 ± 6.78	32.00 ± 4.42
	-35	136.40 ± 23.64	24.30 ± 5.93	25.40 ± 5.86	17.40 ± 3.59
	-40	59.40 ± 8.13	10.60 ± 1.25	12.40 ± 2.19	8.20 ± 1.20
	-45	31.60 ± 7.02	5.20 ± 1.70	6.10 ± 1.93	4.60 ± 1.47

Table 5.8: Average number of rules generated across 10 runs with different redundancy approaches and goodness thresholds for text and herbal prescription data.

Dataset	α	No Prune	Classic	Robust Specialisations	Robust (Both)
Mushroom	-1250	61767.80 ± 158.39	409.70 ± 5.32	568.70 ± 7.42	227.40 ± 2.95
	-1375	37501.10 ± 4080.51	308.10 ± 7.70	342.40 ± 10.14	166.00 ± 6.38
	-1500	22634.50 ± 92.39	229.70 ± 5.67	239.40 ± 5.97	125.50 ± 1.67
	-1625	22049.80 ± 39.14	191.30 ± 2.54	196.30 ± 2.85	114.80 ± 1.77
	-1750	19980.00 ± 2498.40	140.70 ± 5.51	141.70 ± 5.51	93.30 ± 6.17
	-1875	7507.80 ± 78.31	88.60 ± 3.75	89.60 ± 3.75	56.70 ± 1.08
	-2000	6430.80 ± 522.10	38.90 ± 5.39	39.90 ± 5.39	34.70 ± 4.15
T10I4D100K	-500	17287.60 ± 191.08	6114.80 ± 53.65	6114.80 ± 53.65	4302.40 ± 43.14
	-750	3484.70 ± 73.63	1568.00 ± 28.03	1568.00 ± 28.03	1217.30 ± 21.15
	-1000	750.40 ± 31.17	411.90 ± 12.19	411.90 ± 12.19	353.50 ± 9.51
	-1250	169.70 ± 3.67	99.80 ± 2.51	99.80 ± 2.51	85.30 ± 2.30
	-1500	76.70 ± 4.39	41.70 ± 2.73	41.70 ± 2.73	36.50 ± 2.47
	-1750	28.50 ± 3.82	16.90 ± 1.55	16.90 ± 1.55	15.60 ± 1.31
	-2000	2.90 ± 1.53	2.90 ± 1.53	2.90 ± 1.53	2.90 ± 1.53
T40I10D100K	-2000	297056.60 ± 16747.78	5477.00 ± 211.44	5477.00 ± 211.44	3675.70 ± 133.93
	-2125	227409.30 ± 30167.75	4165.60 ± 181.86	4165.60 ± 181.86	2874.70 ± 116.69
	-2250	80480.40 ± 31503.32	3001.90 ± 159.61	3001.90 ± 159.61	2195.80 ± 93.08
	-2375	32533.60 ± 24486.95	1746.40 ± 372.79	1746.40 ± 372.79	1323.10 ± 264.55
	-2500	5693.70 ± 611.02	660.20 ± 78.37	660.20 ± 78.37	528.70 ± 69.71
	-2625	1933.20 ± 631.18	341.90 ± 58.24	341.90 ± 58.24	282.20 ± 47.46
	-2750	615.10 ± 272.55	193.50 ± 57.87	193.50 ± 57.87	172.50 ± 51.47

Table 5.9: Average number of rules generated across 10 runs with different redundancy approaches and goodness thresholds for traditional data.

We now examine the performance of exclusively removing redundant specialisations. Given a rule $X \Rightarrow Z$, the proposed approach for pruning specialisations given in definition 5.2 uses the conditional association $X \Rightarrow Z|Q$ to provide an additional chance to obtain evidence for keeping rule $XQ \Rightarrow Z$ (with respect to the classical approach defined in section 5.1). As such, all specialisations that survive pruning based on classical redundancy will also survive our proposed robust pruning. Using the robust pruning approach to remove only specialisations will always return at least as many rules as the classical approach. This is supported by the results presented in Tables 5.8 and 5.9, where the mean number of rules generated with with the proposed approach is always greater, or not significantly different from the number of rules generated with the classical approach.

5.4.3 Rule Quality

Finally, we look at the performance of the generated rules. As evidence has been given that we should not prune generalisations without first pruning specialisations, no results are reported for pruning generalisations exclusively. Tables 5.10 and 5.9 shows the average log P-values for each of the tested redundancy methods and thresholds. Despite the smaller size of the generated rule set, it can be seen that in all cases the performance of robust pruning is equivalent or slightly better than for rules generated with classically based pruning.

5.4.4 Efficiency

The expanded search needed to identify rules using robust redundancy increases the amount of time and space required. The main factor that effects both computational time and memory requirements is the number of nodes generated during the search. This can be seen by observing the similarity of the trends for the number of nodes generated (Figure 5.4) against time (5.5) and memory (5.6). As the robust specialisation redundancy approach proposed in section 5.2.1 is more permissive than classical redundancy, the pruning employed during the search must

Dataset	α	No Prune	Classic	Robust Specialisations	Robust (Both)
Aspergillosis	-50	-57.44 ± 3.85	-73.19 ± 2.67	-73.25 ± 2.64	-77.29 ± 2.86
	-75	-91.80 ± 3.04	-115.51 ± 2.75	-115.36 ± 2.73	-121.00 ± 2.09
	-100	-125.78 ± 5.52	-155.54 ± 9.36	-155.37 ± 9.22	-165.06 ± 7.93
	-125	-157.39 ± 5.65	-206.80 ± 10.15	-206.60 ± 10.25	-215.07 ± 9.12
	-150	-197.29 ± 8.45	-240.32 ± 7.31	-240.32 ± 7.31	-249.13 ± 7.18
	-175	-224.52 ± 6.51	-265.57 ± 7.95	-265.10 ± 7.75	-273.72 ± 8.71
	-200	-259.87 ± 11.65	-297.70 ± 13.49	-296.33 ± 14.14	-304.60 ± 11.22
Diabetes	-15	-12.17 ± 1.72	-16.78 ± 0.90	-15.30 ± 1.28	-17.67 ± 1.39
	-20	-18.97 ± 2.38	-26.30 ± 1.74	-24.48 ± 1.91	-27.65 ± 2.27
	-25	-30.79 ± 1.34	-36.92 ± 1.58	-35.67 ± 1.53	-39.42 ± 1.54
	-30	-35.59 ± 2.38	-43.34 ± 2.05	-43.11 ± 2.05	-47.00 ± 2.27
	-35	-44.02 ± 1.78	-50.70 ± 1.22	-50.44 ± 1.18	-54.14 ± 1.27
	-40	-48.68 ± 2.62	-55.41 ± 2.06	-55.15 ± 2.07	-58.88 ± 1.94
	-45	-56.41 ± 3.62	-62.17 ± 4.00	-62.14 ± 4.01	-66.32 ± 4.32
Fertility	-15	-14.63 ± 1.86	-20.65 ± 1.35	-20.18 ± 1.29	-21.54 ± 1.23
	-20	-19.21 ± 3.37	-33.41 ± 3.24	-33.00 ± 3.06	-34.13 ± 3.22
	-25	-29.80 ± 3.75	-43.14 ± 3.31	-42.84 ± 3.30	-42.72 ± 3.14
	-30	-34.94 ± 3.51	-54.64 ± 3.28	-54.49 ± 3.30	-55.67 ± 2.92
	-35	-38.45 ± 4.19	-56.66 ± 4.09	-56.55 ± 4.10	-57.15 ± 4.08
	-40	-44.74 ± 5.33	-62.70 ± 6.00	-62.55 ± 5.92	-63.21 ± 5.56
	-45	-50.15 ± 3.88	-66.96 ± 5.25	-66.92 ± 5.28	-66.96 ± 5.36
Insomnia	-15	-10.67 ± 1.41	-13.27 ± 0.74	-12.83 ± 0.80	-13.05 ± 0.72
	-20	-16.50 ± 2.50	-18.77 ± 1.76	-18.68 ± 1.75	-18.61 ± 1.78
	-25	-24.19 ± 3.16	-24.32 ± 1.88	-24.58 ± 1.97	-24.39 ± 2.06
	-30	-31.51 ± 1.70	-30.45 ± 1.52	-30.45 ± 1.64	-30.64 ± 1.62
	-35	-35.02 ± 3.08	-33.93 ± 2.99	-34.37 ± 3.04	-35.13 ± 3.09
	-40	-44.25 ± 4.01	-41.28 ± 2.06	-40.81 ± 2.88	-42.70 ± 2.77
	-45	-57.66 ± 5.25	-55.98 ± 6.25	-55.34 ± 6.45	-58.16 ± 7.72

Table 5.10: Average rule performance across 10 runs using average log P-values with different redundancy approaches and goodness thresholds for text and herbal prescription data.

Dataset	α	No Prune	Classic	Robust Specialisations	Robust (Both)
Mushroom	-1250	-1542.40 ± 8.59	-1604.50 ± 9.79	-1535.21 ± 8.13	-1622.80 ± 9.53
	-1375	-1674.28 ± 39.12	-1700.65 ± 14.68	-1676.63 ± 14.71	-1733.22 ± 20.36
	-1500	-1853.86 ± 17.29	-1793.91 ± 19.56	-1787.94 ± 19.16	-1842.73 ± 17.71
	-1625	-1860.36 ± 11.36	-1843.53 ± 10.95	-1840.13 ± 11.21	-1869.33 ± 10.59
	-1750	-1886.27 ± 24.84	-1903.90 ± 8.54	-1904.92 ± 8.53	-1915.04 ± 9.57
	-1875	-2058.56 ± 15.00	-1996.28 ± 15.97	-1997.05 ± 15.92	-2016.88 ± 13.73
	-2000	-2064.61 ± 13.26	-2040.31 ± 15.01	-2040.44 ± 14.65	-2044.43 ± 13.54
T10I4D100K	-500	-654.97 ± 3.75	-677.78 ± 2.91	-677.78 ± 2.91	-691.13 ± 3.47
	-750	-901.33 ± 7.01	-923.40 ± 5.67	-923.40 ± 5.67	-934.32 ± 6.11
	-1000	-1155.35 ± 11.23	-1171.87 ± 9.15	-1171.87 ± 9.15	-1174.81 ± 8.79
	-1250	-1496.84 ± 15.98	-1489.76 ± 14.07	-1489.76 ± 14.07	-1497.53 ± 14.53
	-1500	-1679.61 ± 30.62	-1690.91 ± 30.69	-1690.91 ± 30.69	-1702.46 ± 30.36
	-1750	-1839.38 ± 44.39	-1865.57 ± 41.17	-1865.57 ± 41.17	-1869.08 ± 40.72
T40I10D100K	-2000	-2205.18 ± 32.30	-2282.00 ± 17.71	-2282.00 ± 17.71	-2302.10 ± 18.75
	-2125	-2201.06 ± 42.97	-2339.99 ± 33.51	-2339.99 ± 33.51	-2360.88 ± 33.30
	-2250	-2305.75 ± 57.95	-2408.56 ± 28.13	-2408.56 ± 28.13	-2423.45 ± 26.10
	-2375	-2410.26 ± 100.15	-2470.00 ± 58.98	-2470.00 ± 58.98	-2481.85 ± 56.55
	-2500	-2585.09 ± 42.48	-2621.87 ± 51.48	-2621.87 ± 51.48	-2631.06 ± 52.50
	-2625	-2691.38 ± 36.23	-2748.60 ± 26.65	-2748.60 ± 26.65	-2756.40 ± 25.11
	-2750	-2686.73 ± 40.39	-2706.56 ± 33.16	-2706.56 ± 33.16	-2710.19 ± 33.49

Table 5.11: Average rule performance across 10 runs using average log P-values with different redundancy approaches and goodness thresholds for traditional data.

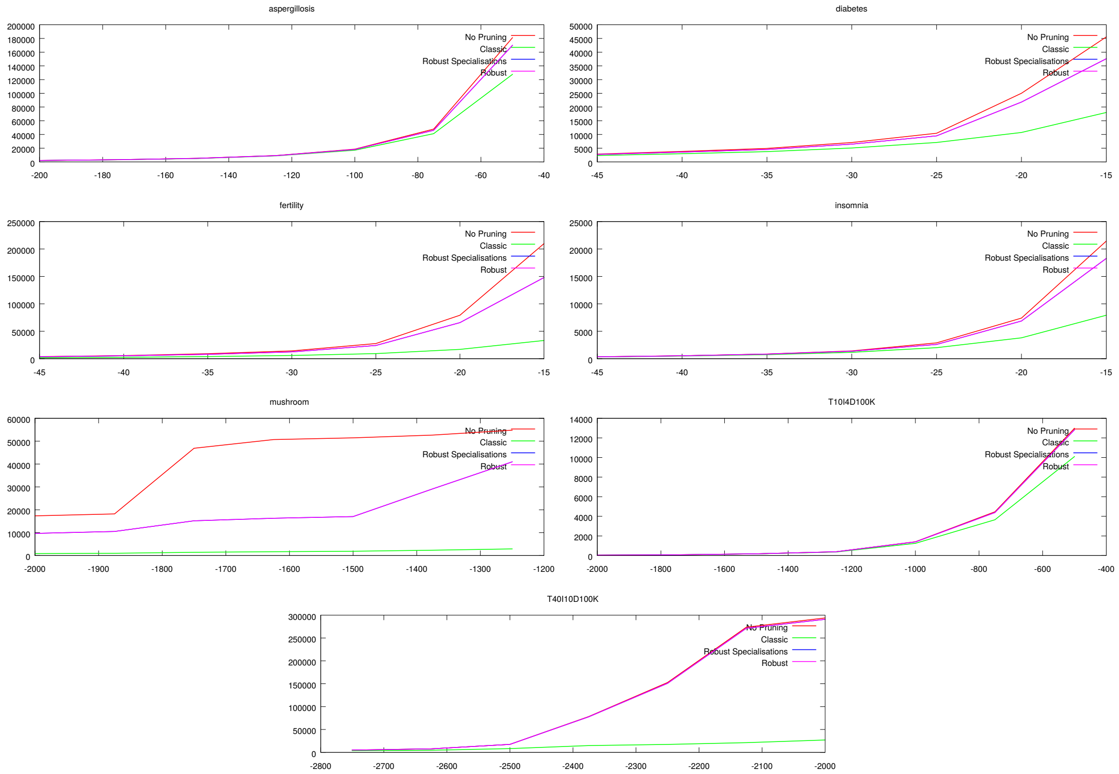


Figure 5.4: Number of nodes generated vs. goodness threshold during the search for rules for each data and redundancy approach.

be less aggressive. We now consider the performance of the search space pruning with robust redundancy.

Figure 5.4 shows the number of nodes generated when searching for each data. It can be observed that the difference in number of nodes generated when searching with robust and classical redundancy varies substantially. Some data, such as the Aspergillosis and T10I4D100K sets differ very little, while the greatest difference is observed for the Mushroom and T40I10D100K data.

As discussed in section 5.3.1 two main approaches are used when using classical redundancy to prune the search space; lapis philosophorum, and minimality. Of those only the lapis philosophorum principle is used when searching with robust redundancy (in fact it is used for all searches regardless of redundancy approach). Therefore two factors can contribute to the increased size of the search space; the lack of minimality based pruning, and the computation of the second bound on

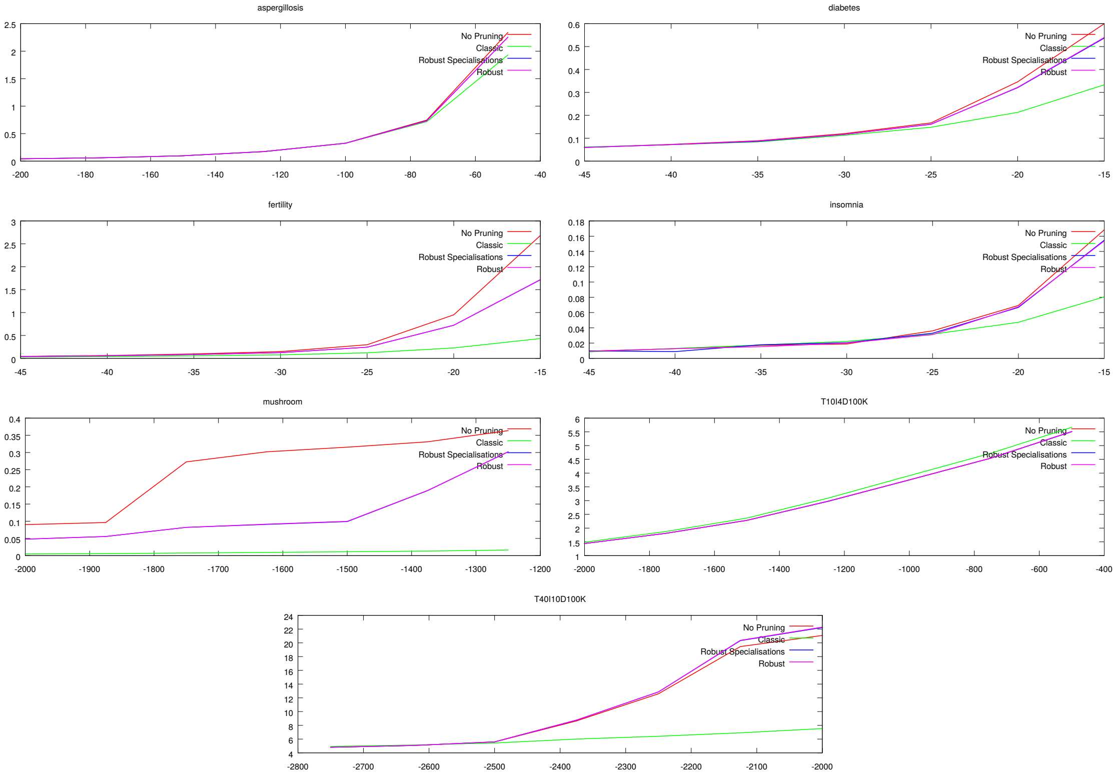


Figure 5.5: Average search time vs. goodness threshold when searching for rules for each data and redundancy approach.

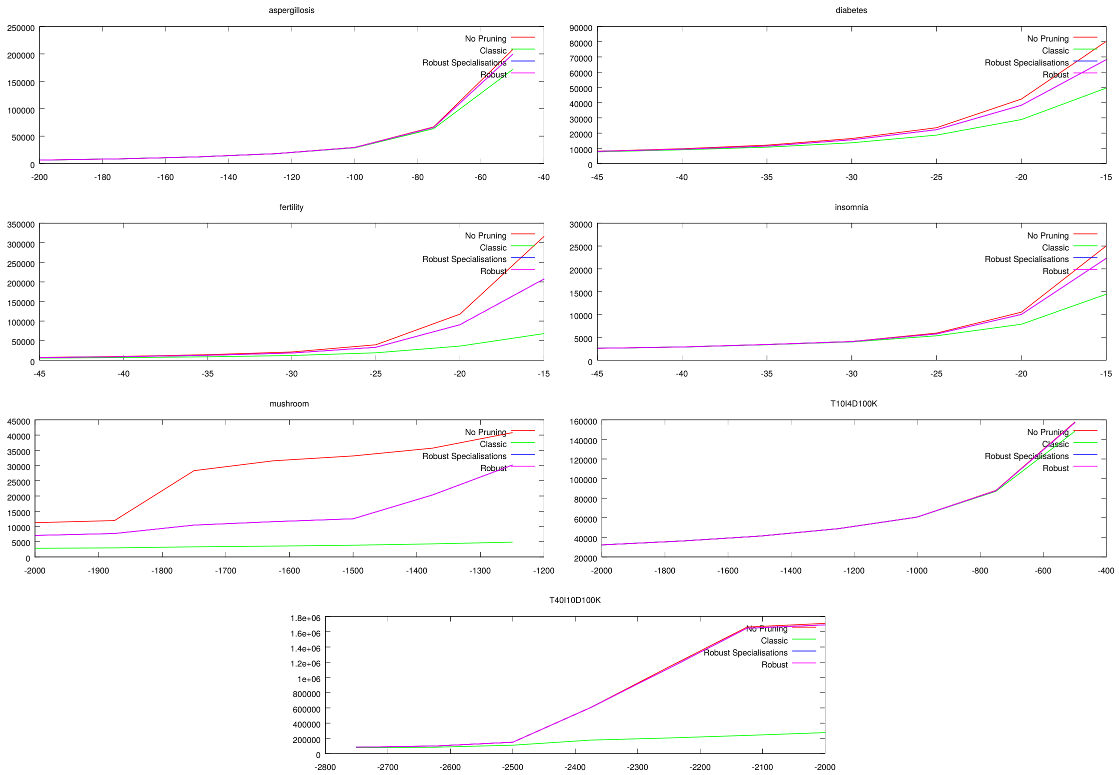


Figure 5.6: Memory usage (kb) vs. goodness threshold when searching for rules for each data and redundancy approach.

$M(X \Rightarrow A|Q)$ in algorithm 4.

In addition to comparing the number of nodes generated when searching with classical versus robust redundancy, it is interesting to examine the number of nodes generated when no redundancy based pruning is used. In general, there appears to be a substantial difference between the number of nodes generated without pruning when compared against using robust redundancy. This implies that the bounds computed in algorithm 4 have a reasonable effect on the size of the search space.

There are three exceptions to the above observation; namely the Aspergillosis, T10I4D100K, and T40I10D100K data. In the case of the Aspergillosis and T10I4D100K data we note there is also little difference between the number of nodes generated using robust and classical redundancy. The implication here is that the majority of the pruning is being done by lapis philosophorum.

However, for the T40I10D100K data there is a substantial difference between the performance with robust and classical redundancy.

Figure 5.5 reports the time required for all searches (including pruning in algorithms 5 and 6). In all cases it the required search time was quite manageable. Even in the case of the T40I10D100K data, all searches completed in less than 30 seconds (all other data completed much quicker). In practice a lack of memory appears to become an issue long before time required for the search.

5.5 Summary

In this chapter we have proposed a novel approach to identifying and removing redundant rules, which we refer to as robust redundancy. Previous approaches compared rules using only the interestingness computed with their respective contingency tables. Such a comparison fails to take into account information included in instances containing only part of the antecedent. Robust redundancy is able to use this information to discover interesting specialisations that would be erroneously removed with a classical approach.

We have also proposed to remove rules which are redundant artifacts of their non-redundant specialisations. Unlike previous approaches [62, 106] that evaluate generalisations based on their exclusive domain with respect to the set of all specialisations, we base our method on comparisons to individual rules. We also present the first work using both specialisation and generalisation redundancy in a rule based context (as opposed to the work of Webb [106] with itemsets). When combined with the removal of redundant specialisations, we demonstrate for multiple data that we are able to produce smaller overall rule sets which hold as well or better in future data.

Chapter 6

Evaluating Classifiers Using Proposed Redundancy Methods

In chapter 4 we introduced a novel, rule based approach for screening citations in systematic reviews. We simulated its application to the literature searches for two real systematic reviews, and demonstrated that it had the potential to provide significant workload savings without unreasonably reducing recall on relevant citations. As part of our analysis we looked at recommendations for various parameter choices when applying the algorithm to real data. It was noted that while promising performance could be achieved with both tested reviews, the algorithm was quite sensitive to the minimum confidence parameter for valid classification rules.

A major requirement for any classification algorithm in systematic review literature review searches is that the recall over relevant citations must be as high as possible. If classification rules are built for non-relevant citations, maximising recall on relevant citations is equivalent to maximising the confidence of generated rules. By requiring that all rules have at least some given minimum confidence level, we attempt to prevent the rule mining algorithm from generating overly general rules, maximising recall.

In section 4.1 we listed additional requirements for systematic review classifiers including determinism, white-box classifiers, and a training methodology which minimises disruption for reviewers. The purpose behind these requirements is to increase user confidence by maximising their ability to inspect and interpret the resulting classifier. The desire for white-box classifiers was a motivating factor in our decision to use rule based classifiers in our work. However, an acknowledged problem within the association rule mining literature is that the number of rules generated when mining can be impractically large [2, 94, 112, 10, 11]. The number of rules generated can often be a barrier to interpretation by users.

One method used to eliminate unwanted rules is based on the concept of redundancy. For classification rules $X \Rightarrow Z$, the classical notion of rule redundancy seeks to identify when an otherwise valid rule $XY \Rightarrow Z$ adds no information to the rule $X \Rightarrow Z$. Rule $XY \Rightarrow Z$ would be considered redundant with respect to rule $X \Rightarrow Z$. Identifying redundant rules can be useful for facilitating understanding and interpretation of the rule set, as well as controlling the size of the search space.

Chapter 5 extended the classical approach to rule redundancy. Additional information was utilised to prevent the incorrect exclusion of more specialised rules $XY \Rightarrow Z$. We demonstrated that instances containing only part of a rule antecedent (e.g. $X\neg Y$ or $\neg XY$) could be used to obtain evidence that rule improved over one of its generalisations $X \Rightarrow Z$. It was also shown how to identify cases where it was preferable to keep more specialised rules of the form $XY \Rightarrow Z$ and prune the general rules $X \Rightarrow Z$.

We show in this chapter that by incorporating the robust specialisation and generalisation redundancy approaches outlined in chapter 5 we can improve the interpretability of our classifier. Compared to classical redundancy, our approach is able to identify more specific rules which had previously been discarded as redundant. We are also able to remove shorter rules which are spurious generalisations of more specific associations. Little to no increase in the size of the overall rule set is observed, and similar classification performance is achieved.

Recall that a goal of this work is to produce models of non-relevant citations that are both interpretable and highly precise. A tendency to produce more specific rules as opposed to more general associations benefits both of these objectives. The descriptive power of a rule increases with the number of terms it contains, allowing the user to better understand the concepts behind each association. Also, by reducing the number of general associations we aim to produce rules that describe smaller, more targeted concepts. Using smaller concepts increases the users ability to evaluate whether or not they are suitable for their intended purpose.

We generally prefer to apply specialisation redundancy prior to the application of generalisation redundancy. However, as discussed in section 5.2.3 is may be preferable to simply use generalisation redundancy if the goal is to generate highly specific rules. As highly precise models of non-relevant citations are the goal of this work, we also evaluate the performance of generalisation redundancy on its own.

The rest of this chapter is structured as follows: Firstly, the data used in this chapter is briefly described in section 6.1. We then examine the performance of the redundancy approaches outlined in chapter 5 when applied to real review data in section 6.2. The sensitivity of rule length and overall set size is evaluated in section 6.3, with the sensitivity for workload and recall evaluated in section 6.4. Finally, conclusions are drawn in section 6.5.

6.1 Data

Experiments in this chapter were performed on the same systematic review data used in chapter 4. This data was drawn from the literature searches for two Cochrane reviews:

- *Galactomannan detection for invasive aspergillosis in imumunocompromized patients* by Leeflang et al. [57]

- An unpublished review on the accuracy of diagnostic biomarkers for Alzheimer's disease.

A summary of the number of citations in each review is provided in Table 4.2. For more information the reader is directed to section 4.5.1.

6.2 Systematic Review Classification with Different Redundancy Approaches

Tables 6.1 and 6.2 show the workload (percentage of citations remaining), recall, average rule length (number of terms in the antecedent), and number of rules generated for the three redundancy types on the Aspergillosis data. Tables 6.4 and 6.5 report the same results for the Alzheimer's data. Rules were generated by training on the full set of citations using annotations made based on title alone as labels. Performance was measured on citations for which abstracts were obtained with final inclusion annotations used as labels.

Before continuing our analysis, we note the difference in rule set size for the Aspergillosis and Alzheimer's data. From Table 6.5 we can see that the Alzheimer's data tends to produce rule sets with only a handful of rules (between 19 and 26 depending on the type of redundancy used). In contrast, the number of rules produced with the Alzheimer's data (Table 6.2) is much higher. Depending on the type of redundancy used, we generate between 111 and 205 rules.

It is interesting that all three redundancy approaches produced quite different rule sets (as can be seen from looking at the average length of the rule antecedents given in Table 6.2 and Table 6.5). In particular, exclusively applying generalisation redundancy produced rules with substantially larger antecedents than either robust or classical redundancy. The difference in rule length was reduced between robust and classical redundancy (particularly on the Aspergillosis data), however a larger difference was observed with the Alzheimer's data. Due to the much smaller number of rules generated for the Alzheimer's data, it is possible that this is due to the addition of longer rules having a greater individual impact of the mean.

	Classic		Robust		Generalisation Only	
Min. Conf.	Workload	Recall	Workload	Recall	Workload	Recall
0.90	0.646	0.701	0.646	0.701	0.650	0.701
0.91	0.649	0.701	0.649	0.701	0.654	0.701
0.92	0.711	0.804	0.711	0.804	0.718	0.804
0.93	0.763	0.862	0.763	0.862	0.770	0.862
0.94	0.794	0.873	0.794	0.873	0.800	0.885
0.95	0.806	0.896	0.806	0.896	0.812	0.896
0.96	0.870	0.942	0.870	0.942	0.876	0.942
0.97	0.891	0.954	0.891	0.954	0.897	0.954
0.98	0.944	0.954	0.945	0.954	0.948	0.954

Table 6.1: Comparison of workload and recall for different redundancy types with Aspergillosis data. Rules trained using full data. Highlighted cells correspond to pre-selected parameter settings.

	Classic		Robust		Generalisation Only	
Min. Conf.	Ave. Rule Len.	# Rules	Ave. Rule Len.	# Rules	Ave. Rule Len.	# Rules
0.90	1.02 ± 0.02	179	1.03 ± 0.02	180	1.38 ± 0.08	205
0.91	1.01 ± 0.02	177	1.03 ± 0.02	178	1.38 ± 0.08	203
0.92	1.03 ± 0.02	174	1.04 ± 0.02	175	1.39 ± 0.08	197
0.93	1.04 ± 0.03	169	1.05 ± 0.03	170	1.40 ± 0.08	189
0.94	1.07 ± 0.04	164	1.07 ± 0.04	165	1.39 ± 0.08	181
0.95	1.04 ± 0.03	156	1.05 ± 0.03	157	1.39 ± 0.08	176
0.96	1.02 ± 0.02	138	1.03 ± 0.03	139	1.39 ± 0.09	157
0.97	1.04 ± 0.03	130	1.04 ± 0.03	130	1.39 ± 0.09	145
0.98	1.09 ± 0.05	112	1.09 ± 0.06	111	1.38 ± 0.10	119

Table 6.2: Comparison of number and length of rules for different redundancy types with Aspergillosis data. Rules trained using full data. Highlighted cells correspond to pre-selected parameter settings.

We start by observing the effect of substituting robust specialisation redundancy for classical redundancy. Table 6.3 and Table 6.6 compare the number of rules generated using classical and robust specialisation redundancy. The left two columns indicate the number of rules generated by exclusively applying these redundancy approaches. The right two columns give the number of rules after subsequent pruning of generalisations.

We can see that for 8 of the 9 minimum confidence thresholds with the Aspergillosis data, specialisation redundancy allows us to generate additional rules when compared to classical redundancy. Additional rules were generated in all cases with the Alzheimer's data.

That specialisation redundancy is able to prevent the exclusion of interesting, more specialised rules that would previously have been considered redundant is promising. We note however that exclusive application of specialisation redundancy (as opposed to classical redundancy) is not useful for classification purposes. This is due to the fact that additional rules produced by specialisation redundancy

	Only Specialisations		Also Prune Generalisations	
Min. Conf.	Classic	Robust	Classic	Robust
0.90	179	181	179	180
0.91	177	179	177	178
0.92	174	176	174	175
0.93	169	171	169	170
0.94	164	165	164	165
0.95	156	157	156	157
0.96	138	139	138	139
0.97	130	130	130	130
0.98	112	113	112	111

Table 6.3: Comparison of number of rules generated by classic and robust specialisation redundancy training with full Aspergillosis data. First pair of columns prunes only specialisations. Second pair of columns subsequently prunes generalisations.

will always match a subset of another, more general rule. While the generation of the additional specific rules has utility if generalisation redundancy is to be applied (or if we are producing rules exclusively for exploratory purposes), they have no use when the goal is to produce a rule set for classification. Therefore we must perform an additional pruning step based on generalisation redundancy.

Looking at the number of rules reported in the second column of Table 6.3 we can make several observations. Firstly, the number of rules identified when using classical redundancy is the same regardless of whether or not generalisation redundancy is applied (we note this will not always be the case, rather it is a consequence of the training data). Secondly, the application of generalisation redundancy in addition to robust specialisation redundancy is able to prune spurious general rules for several tested thresholds.

Although the change in rule sets was not particularly large for the Aspergillosis data, much larger changes were observed with the Alzheimer's data in Table 6.6. We can see that initial pruning with classical redundancy followed by generalisation redundancy is able to prune one general rule for several tested minimum confidence thresholds (and no additional rules for other tested thresholds). While initial pruning with robust specialisation redundancy doesn't change the final number of rules, comparing the average rule size for classical and robust redundancy in Table 6.5 we can see that those for robust redundancy are substantially larger. This suggests most of the general rules found by classical redundancy are being replaced by the additional rules generated with robust specialisation redundancy. This supports the ability of robust redundancy to create longer, more descriptive rule sets.

We also investigate the exclusive application of generalisation redundancy with no prior pruning (other than removing rules which do not meet the minimum threshold for P-value). For the Aspergillosis data, not pruning specialisations allows the algorithm to consider many additional longer rules. This can be seen by the difference in average rule length (the second to last column in Table 6.2) when compared against either classical or full robust redundancy. Despite this, we note the total size of the rule set (the last column in Table 6.2) is only marginally increased, with the increase in the number of generated rules ranging from 6.3% with a minimum confidence of 0.98 to 14% at 0.91.

Similar results are observed for the Alzheimer's data in Table 6.5, although with this data generalisation redundancy contains an interesting property. In contrast to all other tested redundancy types, the number of rules decreases as the minimum confidence for rules is raised (from 21 to 19).

When pruning specialised rules, using a minimum confidence threshold makes it possible that a previously discovered non-redundant rule may not be identified. As this rule is not discovered, the mining algorithm may instead generate more specific rules that it would otherwise make redundant. Increasing the minimum confidence threshold allows the mining algorithm to generate more specific rules at the expense of general rules. Due to the relatively small size of the rules set, the number of such rules appears to be larger than the number removed. This leads to the decreases observed for classic and full robust redundancy.

For generalisation redundancy, as no specialisation based pruning is applied then the algorithm starts from a point where all rules with suitable P-value are present. Rules will only be pruned if they are redundant with respect to all their discovered specialisations. As the minimum confidence threshold is increased, it is possible that more specialised rules will be removed. In the case that such a rule rendered one of its generalisations non-redundant, this could lead to the removal of one or more of these general rules.

Despite the difference in observed rule sets, we also note from Tables 6.1 and 6.4 that very little performance difference exists when testing with the full data. In the case of the Aspergillosis data, no difference in recall was observed between classical and robust redundancy for any measured minimum confidence level. A very small drop in workload (equivalent to a single study) was observed with the strictest confidence threshold tested (0.98). Applying only generalisation redundancy with

	Classic		Robust		Generalisation Only	
Min. Conf.	Workload	Recall	Workload	Recall	Workload	Recall
0.50	0.054	0.97	0.043	0.97	0.054	0.97
0.51	0.054	0.97	0.043	0.97	0.054	0.97
0.52	0.946	0.97	0.967	0.97	0.946	0.97
0.53	0.946	0.97	0.967	0.97	0.946	0.97
0.54	0.946	0.97	0.967	0.97	0.946	0.97
0.55	0.967	0.97	0.967	0.97	0.967	0.97
0.56	0.967	0.97	0.967	0.97	0.967	0.97
0.57	0.967	0.97	0.967	0.97	0.967	0.97
0.58	0.967	0.97	0.967	0.97	0.967	0.97

Table 6.4: Comparison of workload and recall for different redundancy types with Alzheimer's data. Rules trained using full data. Highlighted cells correspond to pre-selected parameter settings.

no initial specialisation based pruning produced a small spike in recall with a minimum confidence level of 0.93, however reviewer workload was slightly higher for all tested values.

For the Alzheimer's data, no difference in recall was observed for any tested minimum confidence level or robustness approach. Mining rules with robust redundancy did produce a small increase in workload for minimum confidence levels of 0.54 or less.

We are interested to know how sensitive the above observations are to variations in the training data or algorithmic parameters. Essentially, are the observations significant or simply an artifact of the tested training data. Such an analysis is of interest as it may also indicate the potential of each approach on future data.

	Classic		Robust		Generalisation Only	
Min. Conf.	Ave. Rule Len.	# Rules	Ave. Rule Len.	# Rules	Ave. Rule Len.	# Rules
0.50	1.55 ± 0.27	22	2.19 ± 0.45	21	2.52 ± 0.53	21
0.51	1.55 ± 0.27	22	2.19 ± 0.45	21	2.52 ± 0.53	21
0.52	1.55 ± 0.27	22	2.19 ± 0.45	21	2.52 ± 0.53	21
0.53	1.55 ± 0.27	22	2.19 ± 0.45	21	2.52 ± 0.53	21
0.54	1.55 ± 0.27	22	2.19 ± 0.45	21	2.52 ± 0.53	21
0.55	1.65 ± 0.24	26	2.35 ± 0.38	26	2.63 ± 0.57	19
0.56	1.65 ± 0.24	26	2.35 ± 0.38	26	2.63 ± 0.57	19
0.57	1.65 ± 0.24	26	2.35 ± 0.38	26	2.63 ± 0.57	19
0.58	1.65 ± 0.24	26	2.35 ± 0.38	26	2.63 ± 0.57	19

Table 6.5: Comparison of number and length of rules for different redundancy types with Alzheimer's data. Rules trained using full data. Highlighted cells correspond to pre-selected parameter settings.

We analyse the sensitivity of the rule length and set size to variations in training data in section 6.3. An analysis of performance with respect to variations in minimum confidence is given in section 6.4.

	Only Specialisations		Also Prune Generalisations	
Min. Conf.	Classic	Robust	Classic	Robust
0.50	22	36	21	21
0.51	22	36	21	21
0.52	22	36	21	21
0.53	22	36	21	21
0.54	22	36	21	21
0.55	26	44	26	26
0.56	26	44	26	26
0.57	26	44	26	26
0.58	26	44	26	26

Table 6.6: Comparison of number of rules generated by classic and robust specialisation redundancy training with full Alzheimer's data. First pair of columns prunes only specialisations. Second pair of columns subsequently prunes generalisations.

6.3 Sensitivity of Rule Structure to Training Data

This section evaluates the sensitivity of the rule length and set size to variations in training data. This is accomplished by measuring results as the average for rules generated using 30 independent samples with 70% of the training data. Values for rule set size and average rule length for the Aspergillosis and Alzheimer's data are reported in Tables 6.7 and 6.8 respectively. Values highlighted in red are larger than the corresponding values for classical redundancy with statistical significance using an unequal variance t-test (P=0.05). Smaller values are highlighted in blue.

The first observation we make is that the large difference in rule length observed between classical redundancy and exclusive application of generalisation redundancy held for all tested data and minimum confidence thresholds. In addition, the difference in rule set size for the Aspergillosis data was repeated for 8 of the 9 tested thresholds. The difference in rule set size was slightly smaller than when training with the entire data, however the decrease was proportional to the percentage decrease in number of training instances.

The rule set size between generalisation and classical redundancy was much closer for the Alzheimer's data. Statistically significant differences were only observed with the three lowest minimum confidence values (0.5, 0.51, and 0.52). As noted in section 6.2, increasing the minimum confidence threshold for the Alzheimer's data tended to lower the number of rules with generalisation redundancy, and raise the number of rules with classical or robust redundancy. That larger rule sets would be observed with low minimum confidence values, with rule set size converging as the minimum confidence increased is not particularly surprising.

More interesting was the difference in rule length and rule set size for robust redundancy. Despite specialisation redundancy only producing a small number of additional rules, the mean rule length was greater than for classical redundancy with 6 of the 9 tested confidence thresholds with the Aspergillosis data. All tested confidence thresholds with the Alzheimer's data showed significantly longer rules.

	Classic		Robust		Generalisation Only	
Min. Conf.	Ave. Rule Len.	# Rules	Ave. Rule Len.	# Rules	Ave. Rule Len.	# Rules
0.90	1.00 ± 0.00	132.4 ± 1.28	1.01 ± 0.00	133.87 ± 1.36	1.30 ± 0.02	142.07 ± 1.59
0.91	1.01 ± 0.00	130.87 ± 1.26	1.02 ± 0.00	132.47 ± 1.34	1.31 ± 0.02	137.33 ± 1.59
0.92	1.02 ± 0.00	128.53 ± 1.12	1.03 ± 0.01	129.07 ± 1.50	1.32 ± 0.02	134.67 ± 1.45
0.93	1.02 ± 0.00	123.83 ± 2.87	1.03 ± 0.01	126.8 ± 1.28	1.32 ± 0.02	131.67 ± 1.56
0.94	1.03 ± 0.01	119.83 ± 1.23	1.04 ± 0.01	119.13 ± 1.53	1.33 ± 0.02	127.8 ± 1.54
0.95	1.03 ± 0.01	113.8 ± 1.44	1.04 ± 0.01	113.77 ± 1.50	1.33 ± 0.02	120.76 ± 1.25
0.96	1.02 ± 0.01	103.9 ± 1.52	1.03 ± 0.01	102.43 ± 1.81	1.32 ± 0.02	109.67 ± 1.34
0.97	1.03 ± 0.01	91.37 ± 1.32	1.04 ± 0.01	92.57 ± 1.19	1.31 ± 0.02	97.3 ± 1.58
0.98	1.09 ± 0.01	79.93 ± 1.61	1.11 ± 0.01	79.53 ± 1.80	1.32 ± 0.02	79.43 ± 1.60

Table 6.7: Comparison of number and length of rules for different redundancy types with Aspergillosis data. Reporting averages over 30 runs training with 70% of full data. Highlighted cells correspond to pre-selected parameter settings.

Comparing the rule set size, we can see that robust redundancy only produced a larger rule set in one case: using a minimum confidence threshold of 0.93 with the Aspergillosis data. In two cases with the Alzheimer's data (minimum confidence thresholds of 0.51 and 0.53), we even observe a statistically significant decrease in rule set size.

The above observations using the sampled data are pleasing, in that they demonstrate the ability of both generalisation and complete robust redundancy to increase the descriptive power of the generated rules without unduly increasing set size. In fact, in the case of robust redundancy we observed a decrease in set size more often than we observed an increase. Generalisation redundancy was more consistent in increasing the size of the rule set, however also produced larger increases in the size of the rules.

	Classic		Robust		Generalisation Only	
Min. Conf.	Ave. Rule Len.	# Rules	Ave. Rule Len.	# Rules	Ave. Rule Len.	# Rules
0.50	1.57 ± 0.08	9.67 ± 0.96	2.22 ± 0.12	8.97 ± 0.76	2.40 ± 0.12	10.67 ± 0.66
0.51	1.52 ± 0.07	9.7 ± 0.89	2.16 ± 0.13	8.17 ± 0.73	2.37 ± 0.11	11.13 ± 0.74
0.52	1.57 ± 0.07	9.27 ± 0.84	2.14 ± 0.12	8.37 ± 0.77	2.45 ± 0.12	11.03 ± 0.71
0.53	1.54 ± 0.07	9.8 ± 0.80	2.06 ± 0.12	8.57 ± 0.62	2.47 ± 0.12	10.1 ± 0.76
0.54	1.49 ± 0.07	8.93 ± 0.83	2.23 ± 0.12	9.17 ± 0.82	2.48 ± 0.12	9.97 ± 0.89
0.55	1.59 ± 0.07	9.63 ± 1.10	2.20 ± 0.11	9.43 ± 0.75	2.47 ± 0.12	9.7 ± 0.62
0.56	1.60 ± 0.06	9.43 ± 0.77	2.32 ± 0.11	10.17 ± 0.99	2.54 ± 0.13	9.84 ± 0.69
0.57	1.67 ± 0.07	8.9 ± 0.83	2.24 ± 0.11	9.87 ± 0.98	2.53 ± 0.12	9.9 ± 0.85
0.58	1.68 ± 0.06	9.33 ± 0.92	2.22 ± 0.11	10.0 ± 1.01	2.55 ± 0.13	9.23 ± 0.81

Table 6.8: Comparison of number and length of rules for different redundancy types with Alzheimer's data. Reporting averages over 30 runs training with 70% of full data. Highlighted cells correspond to pre-selected parameter settings.

6.4 Sensitivity of Performance to Variations in Minimum Confidence

The two charts in the top row of Figures 6.1 and 6.2 show the performance when pruning only generalisations (with no initial pruning of specialisations) on the Aspergillosis and Alzheimer's data respectively. Results are reported as the average of 30 runs with rules trained over a random sample of 70% of the training data. Rule sets were evaluated using the entire set of citations for which abstracts were obtained.

As the minimum confidence threshold approaches 1, the recall and workload converge. In chapter 4 we established that 0.97 was a good threshold for the Aspergillosis data with a goodness threshold ln(P) of -11.513 (see section 4.5.4). We are interested to find out the relation between performance and the addition of general rules. This is done by the relaxation of the minimum confidence threshold. We also wish to see how these relations differ by redundancy approach.

Looking at the performance with the Aspergillosis data reported in Figure

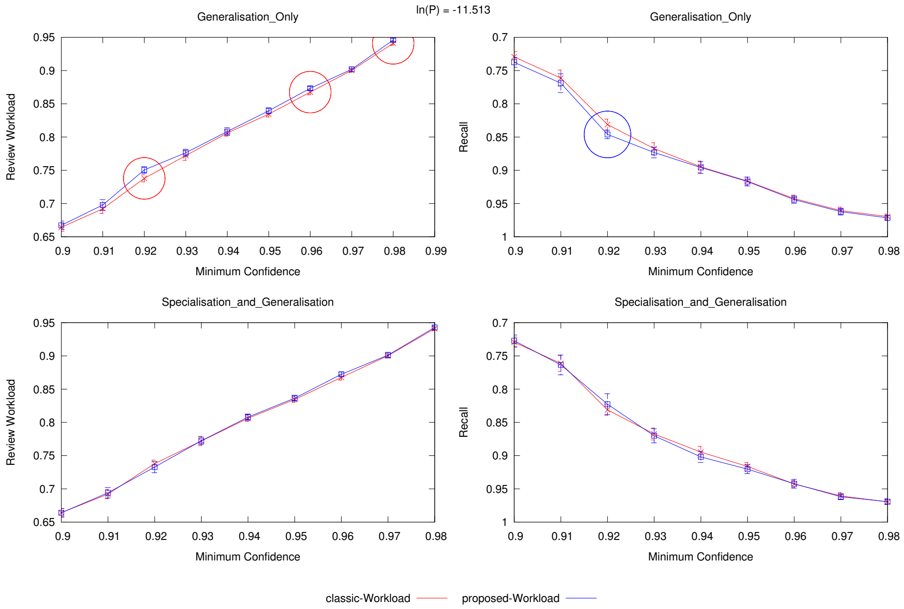


Figure 6.1: Classical vs. Generalisation and Robust redundancy for choice of minimum confidence threshold on Aspergillosis data.

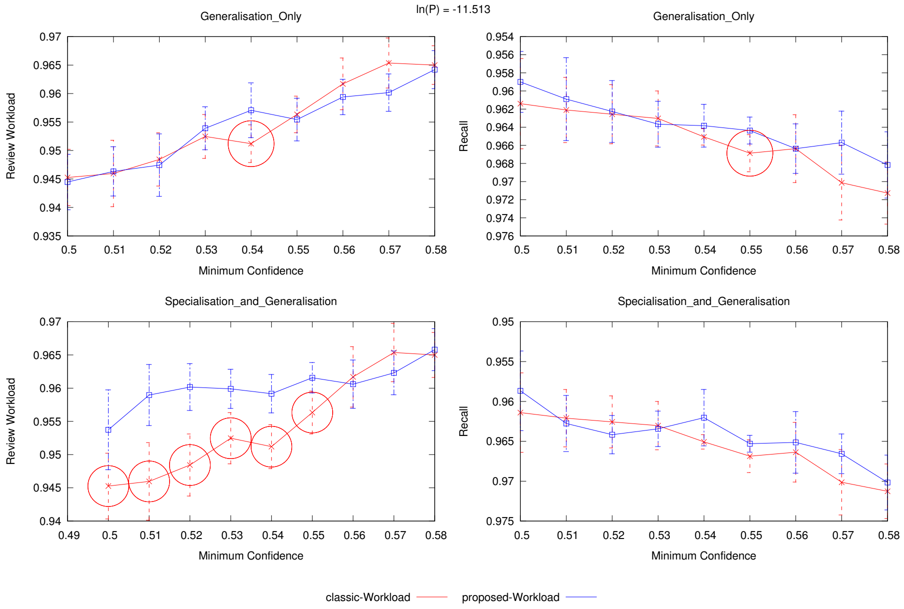


Figure 6.2: Classical vs. Generalisation and Robust redundancy for choice of minimum confidence threshold on Alzheimer's data.

6.1, we can see that as the minimum confidence decreases away from 0.97 the recall with generalisation redundancy decreases at a slower rate than with classical redundancy. The difference in performance is most substantial at minimum confidence level of 0.92, before the recall for both approaches converges. We established in previous sections that generalisation redundancy produces rules with longer antecedents. The increasingly specific rules generate fewer false positives when identifying non-relevant studies.

It is noteworthy that while recall for generalisation redundancy improves over classical redundancy, the opposite effect is observed for reviewer workload (i.e. generalisation redundancy produces a higher reviewer workload). This is unsurprising, as the tendency of classical redundancy to prefer more general rules with lower confidence will produce rules that have a higher chance of matching new citations. It is important to remember that the goal of classification for systematic reviews is to minimise workload under a requirement of sufficient recall. It is preferable to reduce the rate at which recall decreases, even at the expense of workload savings.

In contrast to the Aspergillosis data, there is little difference observed in recall between generalisation and classical redundancy for the Alzheimer's review data reported in Figure 6.2. For almost all tested minimum confidence thresholds with a ln(P) value of -11.513, recall was statistically equivalent. In only one case (generalisation redundancy only with a minimum confidence threshold of 0.55) was any difference observed.

The Alzheimer's review uses an extremely weak minimum confidence threshold of 0.51. As a result, it can be expected that pretty much all very general rules will be identified during rule mining. That generalisation redundancy fails to significantly improve performance indicates that for the Alzheimer's data general rules appear to perform quite well on their own.

It is important to note the size of the confidence intervals for performance on the Alzheimer's data, which are much larger than for the equivalent tests on the

Aspergillosis data. Recall from the start of section 6.2 that the average length of rules for the Alzheimer's data is quite short (often only 10-20 rules). This implies that small changes in the rule set will produce a much larger impact in performance.

The second row of subfigures in Figures 6.1 and 6.2 report the performance of robust redundancy on the Aspergillosis and Alzheimer's data respectively. Robust redundancy produced no significant difference in performance to classical redundancy for the Aspergillosis data, either in terms of recall or workload. Recall was statistically identical for the Alzheimer's data, although there was a noted increase in workload for minimum confidence levels less than 0.55.

We can see by comparing the first two columns in Table 6.6 that robust specialisation redundancy (without pruning generalisations) produces a significantly larger number of rules than with classical redundancy. Looking at the fourth column (as well as the rule set size values in Table 6.8), we can see that many of the general rules are made redundant by generalisation. The resulting set size is quite similar to that for both classical redundancy, and exclusive application of generalisation redundancy. This suggests that full robust redundancy prunes many of the rules found with the classical approach, which are in turn pruned when exclusively applying generalisation redundancy without any initial specialisation based pruning.

The above observation, combined with the increase in workload for full robust redundancy is interesting to note. Data that produce small rule sets are likely quite sensitive to the addition or removal of a few rules. In the case of the Alzheimer's data, highly specific rules are required to produce appropriate performance.

We hypothesise that for data with very small rule sets, such as the Alzheimer's data, it may be worthwhile to skip specialisation based pruning and exclusively apply generalisation redundancy. However due to the fact we have only two datasets, we note that further evaluation before drawing such a conclusion is required.

6.5 Summary

In chapter 4 we proposed a novel classification rule based algorithm for semi-automation of literature screening for systematic reviews. This chapter evaluated the ability of the redundancy approaches covered in chapter 5 to improve the interpretability and performance of that algorithm. We demonstrated that robust redundancy had the ability to increase the descriptive power of individual rules without increasing the size of the overall rule set.

We also evaluated pruning redundant generalisations without the prior step to prune redundant specialisations. Even more substantial increases in descriptive power were observed for this approach, however increases in the number of overall rules of up to 14% were observed for one of the tested datasets.

There appeared to be little performance trade off when increasing descriptive power with the proposed redundancy approaches. Considering performance in terms of recall and workload, minimal difference was observed between redundancy approaches. This is positive in that it indicates increased descriptive power is available without substantial trade off in performance.

One exception to the above claim is the reviewer workload for the Alzheimer's data when pruning rules with robust specialisations and generalisations. For several minimum confidence thresholds, the Alzheimer's data saw a substantial drop off in workload with robust redundancy. We note that no such observation was made with generalisation redundancy.

The Alzheimer's data produced very small rule sets, which increases the potential impact of removing individual rules on the classifiers overall coverage. It is possible that had more highly precise rules been considered, the number used to replace each general rule could have been increased and the coverage of the overall rule set improved.

We hypothesise that for very small rule sets, it may be preferable to employ generalisation redundancy without initial specialisation based pruning. However we note that further evaluation is required to validate this claim.

Chapter 7

Conclusion

Systematic reviews are a crucial component of modern evidence based medicine. However despite their importance for clinical policy and practice, they are extremely difficult to conduct. Literature searches are largely conducted manually, often over a timescale of months or even years [87].

The past decade has seen the machine learning community increasingly try to improve the automation and efficiency of the systematic review process. Much progress has been made, with systems such as Abstrackr [100, 101] showing significant promise. Yet room for improvement exists, for example with reviews containing particularly imbalanced data. Despite this progress much work remains; both in minimising reviewer workload rates without excluding relevant research, as well as complementary issues such as how best to train and integrate classification models into the systematic review process.

This thesis set out to investigate ways in which reviewer workload could be reduced during the literature screening process for systematic reviews. We proposed a novel approach for integrating classification into the systematic review process, and demonstrated its potential to produce real workload savings by simulated application on real data. To maximise user confidence in our classifier we adopt white-box, rule based classifiers. We propose an alternate framework in which to identify and remove redundant rules, and demonstrate its ability to improve the descriptive power of the rules produced. Prior to the work reported in this thesis, there has been little research into how the review process might be exploited to facilitate automation. We provide such an analysis. We also consider issues relating to the automation of diagnostic test accuracy reviews, and how they differ from more traditional reviews such as those of treatment.

7.1 Contributions

This thesis provides several contributions to existing knowledge, which we summarise in the following section.

- **Empirically demonstrate diagnostic reviews contain increased workload and target class heterogeneity when compared to treatment**

  Chapter 3 presents a comparative analysis of diagnostic test accuracy and treatment reviews. While systematic reviews have traditionally focused on questions relating to treatment, reviews from other fields are becoming increasingly common. Diagnostic reviews are found to contain both a higher rate of data imbalance (ratio of relevant to non-relevant studies), as well as a broader target class.

  The identification of diagnostic test accuracy reviews as a particularly challenging subset of the review screening problem, along with several possible reasons, can assist researchers in improving the performance of classification for systematic reviews as a whole. In addition to suggesting diagnostic reviews as a source of challenging test data for future work, it motivates research designing classifiers which are robust to a broadly defined target class.

- **Model for citation classification that excludes irrelevant citations with high precision**

We build upon this work in chapter 4 by proposing a classification model where subsets of non-relevant citations are labelled with high precision. This allows us to reduce reviewer workload, while avoiding the need to completely model the entire target class of relevant citations. Providing a guarantee for recall on some target class by by inverting the problem and excluding subsets of the non-target class with high precision is a novel contribution of this work.

- **Model for automation of systematic review literature searches that uses annotations from prior screening stages to build the classifier**

  We also address the problem of how to obtain training data for automated screening during systematic reviews by analysing the existing screening process for systematic reviews in chapter 4. Systematic reviews typically screen citations in a multi-stage triage process. Citations are first screened based on title alone, then title and abstract, and finally on full text, with citations that can confidently be excluded from the review removed after each stage. For a classifier which is intended to be applied at a given stage (for example, title and abstract), all previous research has relied upon labels applied at this stage to build an appropriate training set. In contrast, we propose that annotations made at prior screening stages (for example, title alone) have utility in building a classifier to model non-relevant citations.

  Training a classifier using annotations from prior screening stages has several benefits. Firstly, these labels can be obtained with no additional modification to the screening process; existing approaches require some modification to the methodology with which citations are selected for annotation. Secondly, building a classifier based upon annotations made in prior screening stages allows for the application of the classifier prior the starting the next stage. This suggests that not only does the proposed model utilise additional information which has previously been discarded, but could be applied in sequence with existing approaches to produce even greater workload savings.

- **Initial analysis on real data for performance evaluation and recommended parameterisation of classification models**

  Section 4.5 contains an evaluation of the proposed classifier by simulated application to two real reviews. A sensitivity analysis of the algorithm to its parameters, feature selection, and training methodology is performed. Initial recommendations are made with respect to consistently performing choices. This process can be followed with future reviews to increase confidence in our conclusions or to obtain evidence to suggest their modification.

- **Alternate definitions for association rule redundancy, with corresponding rule mining algorithm**

  Chapter 5 examines the process of association rule mining, in particular the issue of rule redundancy. We elected to build our classifier using a white-box, rule based approach. When mining association rules, redundancy is important both to restrict the result set to a manageable size, and to prevent the generation of spurious rules which are simply artifacts of other, more interesting associations. Traditionally, rules are considered as redundant when a more general rule can be found that is considered as good or better with respect to some goodness measure. We extend this definition in two ways.

  Firstly, we propose an approach called specialisation redundancy. Specialisation redundancy is similar to the traditional definition described above, however is more permissive in that it requires additional tests to verify that a rule indeed fails to improve over a generalisation of itself. Generalisation redundancy is also proposed. In contrast to specialisation redundancy, generalisation redundancy seeks to remove general rules which fail to add anything when compared against the entire set of their non-redundant specialisations.

- **Demonstrated ability for alternate redundancy definitions to improve the interpretability of the rule based literature screening algorithm**

  In chapter 6 we reevaluate our rule based algorithm on the two real reviews and examine the performance of incorporating robust redundancy. The combination of robust and specialisation redundancy demonstrates the ability to increase the length of generated rules without increasing the overall rule set size. Generalisation redundancy is shown to also increase the length of rules, although in some cases may produce a increase the size of the generated rule set.

  Sections 6.3 and 6.4 examine the performance of the two approaches, and hypothesise as to when each approach should be preferred. Initial analysis suggests that exclusive application of generalisation redundancy should be applied when the rule set is small. Specialisation and generalisation redundancy should be applied together in other cases. We note that due to lack of data, further analysis is required.

7.2 Limitations and Future Work

A number of limitations of the work reported in this thesis must be discussed, both to better understand how the work extends existing knowledge and to provide a guide for future research. One particular limitation concerns the evaluation of our proposed classification model in chapters 4 and 6. While the potential for our algorithm to reduce reviewer workload and the utility of training with annotations from prior stages is established, our analysis is limited in that data for only two reviews was available. Analysis with additional data would be important, both to increase confidence in the conclusions drawn in this work, as well as to better understand the performance of the algorithm on different reviews.

We also note the sensitivity of the proposed algorithm to the minimum confidence threshold employed while mining classification rules. Ideally, we would like principled default setting for all parameters, or at least some proposed guidelines for their selection. Developing such methods with any reasonable certainty has not been possible given the relatively low number of studies with which we have to test our approach, and further analysis is certainly warranted.

The comparative analysis of diagnostic and treatment reviews reported in chapter 3 could also be extended. As of the search date, only 13 diagnostic test accuracy reviews had been published in the Cochrane library. A repetition of the study, both including reviews published after the initial search date, as well as diagnostic reviews published by sources other than Cochrane, would be interesting. This would allow both increased confidence in the conclusions drawn, as well as their generalisation to non-Cochrane reviews.

The methodology proposed in this thesis utilises annotations made based on titles to remove additional non-relevant studies. Although developed for systematic reviews, it is possible that the approach could be applied more broadly to other document screening tasks which also follow a multi-stage triage process. Further work exploring the potential for such generalisation would be of interest. In particular it would be interesting to determine what, if any, adjustments to the process modifications outlined in Chapter 4 are required.

Finally, we note that while the utility of annotations from prior stages can be useful in training classification models for systematic reviews, we have barely scraped the surface in terms of how they might be applied. For example, active learning algorithms such as those tested with Abstrackr could utilise such annotations to guide the model initialisation prior to seeking additional labels from a human oracle. Training with these labels demonstrates a promising new line of enquiry for automation of systematic review literature screening, and a further analysis in future would be highly interesting.

Publications Arising From This Work

- H. Petersen, J. Poon, S. K. Poon, C. Loy, and M. Leeflang. Partially automated literature screening for systematic reviews by modelling non-relevant articles. In *The Third Australasian Workshop on Artificial Intelligence in Health (AIH)*, pages 43–44, 2013

- H. Petersen, J. Poon, S. K. Poon, and C. Loy. Increased workload for systematic review literature searches of diagnostic tests compared with treatments: Challenges and opportunities. *JMIR medical informatics*, 2(1), 2014

Bibliography

[1] C. C. Aggarwal and P. S. Yu. A new framework for itemset generation. In *PODS 98, Symposium on Principles of Database Systems*, pages 18–24, 1998.

[2] C. C. Aggarwal and P. S. Yu. A new approach to online generation of association rules. *Knowledge and Data Engineering, IEEE Transactions on*, 13(4):527–540, 2001.

[3] R. Agrawal, T. Imieli, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM International Conference on Management of Data (SIGMOD)*, pages 207–216. ACM, 1993.

[4] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proceedings of the Twentieth International Conference on Very Large Data Bases (VLDB)*, volume 1215, pages 487–499, 1994.

[5] A. An, S. Khan, and X. Huang. Hierarchical grouping of association rules and its application to a real-world domain. *International journal of systems science*, 37(13):867–878, 2006.

[6] L. Antonie, J. Li, and O. Zaiane. Negative association rules. In *Frequent Pattern Mining*, pages 135–145. Springer, 2014.

[7] M. L. Antonie and O. R. Zaiane. Text document categorization by term association. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM)*, pages 19–26, 2002.

[8] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, and C. F. Aliferis. Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association*, 12(2):207–216, 2005.

[9] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, pages 17–21, 2001.

[10] M. Ashrafi, D. Taniar, and K. Smith. A new approach of eliminating redundant association rules. *Database and Expert Systems Applications*, 3180:465–474, 2004.

[11] M. Ashrafi, D. Taniar, and K. Smith. Redundant association rules reduction techniques. In *Proceedings of the Eighteenth Australian Joint Conference on Artificial Intelligence (AI)*, pages 254–263. Springer Berlin Heidelberg, 2005.

[12] H. Ayadi, M. Torjmen, M. Daoud, Maher Ben J., and J. X. Huang. Correlating medical-dependent query features with image retrieval models using association rules. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 299–308. ACM, 2013.

[13] A. Babashzadeh, M. Daoud, and J. Huang. Using semantic-based association rule mining for improving clinical text retrieval. In *International Conference on Health Information Science*, pages 186–197. Springer, 2013.

[14] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[15] M. Brazzelli, P. A. Sandercock, F. M. Chappell, M. G. Celani, E. Righetti, N. Arestis, J. M. Wardlaw, and J. J. Deeks. Magnetic resonance imaging versus computed tomography for detection of acute vascular lesions in patients presenting with stroke symptoms. *The Cochrane Library*, 4, 2009.

[16] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM International Conference on Management of Data (SIGMOD)*, pages 255–264. ACM, 1997.

[17] A. M. Cohen. Optimizing feature representation for automated systematic review work prioritization. In *Proceedings of the AMIA Annual Symposium*, volume 2008, page 121. American Medical Informatics Association, 2008.

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

[19] A. M. Cohen, K. Ambert, and M. McDonagh. Cross-topic learning for work prioritization in systematic review creation and update. *Journal of the American Medical Informatics Association*, 16(5):690–704, 2009.

[20] A. M. Cohen, K. Ambert, and M. McDonagh. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Med Inform Decis Mak*, 12:33, 2012.

[21] A. M. Cohen, R. T. Bhupatiraju, and W. R. Hersh. Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In *Proceedings of the Text REtrieval Conference (TREC)*, 2004.

[22] A. M. Cohen, W. R. Hersh, K. Peterson, and P. Y. Yen. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc*, 13(2):206–19, 2006.

[23] The Cochrane Collaboration. The cochrane library, 2015. http://www.cochranelibrary.com.

[24] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[25] J. Darrah, R. Hickman, M. O'donnell, L. Vogtle, and L. Wiart. Aacpdm methodology to develop systematic reviews of treatment interventions (revision 1.2). *Milwaukee, WI, USA: American Academy for Cerebral Palsy and Developmental Medicine*, 2008.

[26] W. L Devillé, F. Buntinx, L. M Bouter, V. M Montori, H. C. W. De Vet, D. A. W. M. Van der Windt, and P. D. Bezemer. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC medical research methodology*, 2(1):9, 2002.

[27] J. Eden, L. Levit, A. Berg, S. Morton, et al. *Finding what works in health care: standards for systematic reviews.* National Academies Press, 2011.

[28] M. Egger, P. Juni, C. Bartlett, F. Holenstein, and J. Sterne. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? empirical study. *Health Technol Assess*, 7(1):1–76, 2003.

[29] Elsevier. Embase, 2015. http://www.elsevier.com/online-tools/embase.

[30] R. A. Fisher. *Statistical methods for research workers.* Oliver and Boyd, Edinburgh, 1934.

[31] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296, 1999.

[32] O. Frunza, D. Inkpen, and S. Matwin. Building systematic reviews using automatic text classification techniques. In *Proceedings of the Twentythird International Conference on Computational Linguistics: Posters*, pages 303–311, 2010.

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

[34] P. P. Glasziou and S. L. Sanders. Investigating causes of heterogeneity in systematic reviews. *Statistics in medicine*, 21(11):1503–1511, 2002.

[35] Higgins JPT Green S, editor. *Cochrane Handbook for Systematic Reviews of Interventions.* The Cochrane Collaboration, 2011. http://community.cochrane.org/handbook.

[36] Cochrane Diagnostic Test Accuracy Working Group. *Cochrane handbook for systematic reviews of diagnostic test accuracy.* The Cochrane Collaboration, 2015. http://srdta.cochrane.org/handbook-dta-reviews.

[37] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM International Conference on Management of Data (SIGMOD)*, pages 1–12. ACM, 2000.

[38] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, 8(1):53–87, January 2004.

[39] D. A. Hanauer, M. Saeed, K. Zheng, Q. Mei, K. Shedden, A. R. Aronson, and N. Ramakrishnan. Applying metamap to medline for identifying novel associations in a large clinical dataset: a feasibility analysis. *Journal of the American Medical Informatics Association*, 21(5):925–937, 2014.

[40] W. Hämäläinen. Efficient discovery of the top-k optimal dependency rules with fisher's exact test of significance. In *Proceedings of the Tenth IEEE International Conference on Data Mining (ICDM)*, pages 196–205, Dec 2010.

[41] W. Hämäläinen. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. *Knowledge and Information Systems*, 32(2):383–414, 2012.

[42] X. Huang. Comparison of interestingness measures for web usage mining:

An empirical study. *International Journal of Information Technology & Decision Making*, 6(01):15–41, 2007.

[43] X. Huang and A. An. Discovery of interesting association rules from livelink web log data. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 763–766. IEEE, 2002.

[44] X. Huang, A. An, and N. Cercone. Comparison of interestingness functions for learning web usage patterns. 2002.

[45] X. Huang, F. Peng, A. An, D. Schuurmans, and N. Cercone. Session boundary detection for association rule learning using n-gram language models. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 237–251. Springer, 2003.

[46] C. K. Irwin, K. Yoon, C. Wang, S. J. Hoff, J. J. Zimmerman, T. Denagamage, and A. M. O'Connor. Using the systematic review methodology to evaluate factors that influence the persistence of influenza virus in environmental matrices. *Applied and environmental microbiology*, 77(3):1049–1060, 2011.

[47] X. Ji and P. Yen. Using medline elemental similarity to assist in the article screening process for systematic reviews. *JMIR Medical Informatics*, 3(3):e28, 2015.

[48] S. R. Jonnalagadda, P. Goyal, and M. D. Huffman. Automating data extraction in systematic reviews: a systematic review. *Systematic reviews*, 4(1):1, 2015.

[49] S. Karimi, S. Pohl, F. Scholer, L. Cavedon, and J. Zobel. Boolean versus ranked querying for biomedical systematic reviews. *BMC medical informatics and decision making*, 10(1):58, 2010.

[50] S. Karimi, J. Zobel, S. Pohl, and F. Scholer. The challenge of high recall

in biomedical systematic search. In *Proceedings of the Third International Workshop on Data and Text Mining in Bioinformatics*, pages 89–92, 2009.

[51] M. Kastner, R. B. Haynes, and N. L. Wilczynski. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. *J Clin Epidemiol*, 65(1):116–7; author reply 117–8, 2012.

[52] M. Kastner, N. L. Wilczynski, A. K. McKibbon, A. X. Garg, and R. B. Haynes. Diagnostic test systematic reviews: bibliographic search filters ("clinical queries") for diagnostic accuracy studies perform well. *J Clin Epidemiol*, 62(9):974–81, 2009.

[53] K. Khan, R. Kunz, J. Kleijnen, and G. Antes. *Systematic reviews to support evidence-based medicine*. Crc Press, 2011.

[54] K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes. Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, 96(3):118–121, 2003.

[55] R. Khorsan and C. Crawford. External validity and model validity: a conceptual approach for systematic review methodology. *Evidence-Based Complementary and Alternative Medicine*, 2014, 2014.

[56] S. Lallich, O. Teytaud, and E. Prudhomme. Association rule interestingness: Measure and statistical validation. *Quality Measures in Data Mining*, 43:251–275, 2007.

[57] M. M. Leeflang, Y. J. Debets-Ossenkopp, C. E. Visser, R. JPM Scholten, L. Hooft, H. A. Bijlmer, J. B. Reitsma, P. MM Bossuyt, and C. M. Vandenbroucke-Grauls. Galactomannan detection for invasive aspergillosis in immunocompromized patients. *The Cochrane Library*, 2008.

[58] M. M. Leeflang, J. J. Deeks, C. Gatsonis, P. M. Bossuyt, and Group

Cochrane Diagnostic Test Accuracy Working. Systematic reviews of diagnostic test accuracy. *Ann Intern Med*, 149(12):889–97, 2008.

[59] M. M. Leeflang, R. J. Scholten, A. W. Rutjes, J. B. Reitsma, and P. M. Bossuyt. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *J Clin Epidemiol*, 59(3):234–40, 2006.

[60] J. Li, O. Zaiane, and A. Osornio-Vargas. Discovering statistically significant co-location rules in datasets with extended spatial objects. *Data Warehousing and Knowledge Discovery*, 8646:124–135, 2014.

[61] Y. Li, P. Ning, X. S. Wang, and S. Jajodia. Discovering calendar-based temporal association rules. *Data & Knowledge Engineering*, 44(2):193–218, 2003.

[62] B. Liu, W. Hsu, and Y. Ma. Identifying non-actionable association rules. In *Proceedings of the Seventh ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 329–334, 2001.

[63] Y. Ma. *Text classification on imbalanced data: Application to Systematic Reviews Automation*. PhD thesis, University of Ottawa (Canada), 2007.

[64] F. J. Mateen, J. Oh, A. I. Tergas, N. H. Bhayani, and B. B. Kamdar. Titles versus titles and abstracts for initial screening of articles for systematic reviews. *Clinical epidemiology*, 5:89–95, 2013.

[65] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and P. O'Blenis. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, 17(4):446–453, 2010.

[66] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and P. O'Blenis. Performance of svm and bayesian classifiers on the systematic review classification

task. *Journal of the American Medical Informatics Association*, 18(1):104–105, 2011.

[67] J. McGowan and M. Sampson. Systematic reviews need systematic searchers. *Journal of the Medical Library Association*, 93(1):74, 2005.

[68] M. McGrane and S. K. Poon. Interaction as an interestingness measure. In *2010 IEEE International Conference on Data Mining Workshops*, pages 726–731, 2010.

[69] G. Menardi and N. Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122, 2014.

[70] A. M Methley, S. Campbell, C. Chew-Graham, R. McNally, and S. Cheraghi-Sohi. Pico, picos and spider: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC health services research*, 14(1):579, 2014.

[71] L. A. C. Millard, P. A. Flach, and J. P. T. Higgins. Machine learning to assist risk-of-bias assessments in systematic reviews. *International journal of epidemiology*, 45(1):266–277, 2016.

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

[73] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269, 2009.

[74] US National Library of Medicine. Medline fact sheet, 2015. http://www.nlm.nih.gov/pubs/factsheets/medline.html.

[75] US National Library of Medicine. Mesh fact sheet, 2015. http://www.nlm.nih.gov/pubs/factsheets/mesh.html.

[76] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):5, 2015.

[77] M. Pai, M. McCulloch, W. Enanoria, and J. M. Colford. Systematic reviews of diagnostic test evaluations: what's behind the scenes? *Evidence Based Medicine*, 9(4):101–103, 2004.

[78] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *International Conference on Database Theory (ICDT)*, pages 398–416. Springer, 1999.

[79] H. Petersen, J. Poon, S. K. Poon, and C. Loy. Increased workload for systematic review literature searches of diagnostic tests compared with treatments: Challenges and opportunities. *JMIR medical informatics*, 2(1), 2014.

[80] H. Petersen, J. Poon, S. K. Poon, C. Loy, and M. Leeflang. Partially automated literature screening for systematic reviews by modelling non-relevant articles. In *The Third Australasian Workshop on Artificial Intelligence in Health (AIH)*, pages 43–44, 2013.

[81] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, pages 229–238, 1991.

[82] D. D. Rahn, C. Carberry, T. V Sanses, M. M Mamik, R. M. Ward, K. V. Meriwether, C. K. Olivera, H. Abed, E. M. Balk, M. Murphy, et al. Vaginal estrogen for genitourinary syndrome of menopause: a systematic review. *Obstetrics & Gynecology*, 124(6):1147–1156, 2014.

[83] J. Rathbone, T. Hoffmann, and P. Glasziou. Faster title and abstract screening? evaluating abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic reviews*, 4(1):1–7, 2015.

[84] J. D. Rennie, L. Shih, J. Teevan, D. R. Karger, et al. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, volume 3, pages 616–623, 2003.

[85] G. Ritchie, J. Glanville, and C. Lefebvre. Do published search filters to identify diagnostic test accuracy studies perform adequately? *Health Info Libr J*, 24(3):188–92, 2007.

[86] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: what it is and what it isn't. *BMJ: British Medical Journal*, 312(7023):71, 1996.

[87] M. Sampson, K. G. Shojania, C. Garritty, T. Horsley, M. Ocampo, and D. Moher. Systematic reviews can be produced and published faster. *J Clin Epidemiol*, 61(6):531–536, 2008.

[88] J. Savulescu and M. Spriggs. The hexamethonium asthma study and the death of a normal volunteer in research. *Journal of medical ethics*, 28(1):3–4, 2002.

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

[90] V. Smith, D. Devane, C. M. Begley, and M. Clarke. Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC Medical Research Methodology*, 11(1):1–6, 2011.

[91] P. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Inf. Syst.*, 29(4):293–313, 2004.

[92] J. Thomas. Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation. *OA Evidence-Based Medicine*, 1(2):1–6, 2013.

[93] J. Thomas, J. McNaught, and S. Ananiadou. Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1):1–14, 2011.

[94] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hätönen, and H. Mannila. Pruning and grouping discovered association rules. In *ECML'95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, pages 47–52, 1995.

[95] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[96] S. J. Tregear, L. E. Gavin, and J. R. Williams. Systematic review evidence methodology: providing quality family planning services. *American journal of preventive medicine*, 49(2):S23–S30, 2015.

[97] D. Trieschnigg, P. Pezik, V. Lee, F. De Jong, W. Kraaij, and D. Rebholz-Schuhmann. Mesh up: effective mesh text classification for improved document retrieval. *Bioinformatics*, 25(11):1412–1418, 2009.

[98] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani, and E. Coiera. Systematic review automation technologies. *Syst Rev*, 3(1):74, 2014.

[99] F. Verhein and S. Chawla. Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets. In *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM)*, pages 679–684, 2007.

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

[101] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, and T. A. Trikalinos. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedings of the Second ACM International Health Informatics Symposium (SIGHIT)*, pages 819–824, 2012.

[102] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Active learning for biomedical citation screening. In *Proceedings of the Sixteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 173–182, 2010.

[103] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, and C. H. Schmid. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11:55, 2010.

[104] G. I. Webb. Discovering significant rules. In *Proceedings of the Twelfth ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 434–443. ACM, 2006.

[105] G. I. Webb. Discovering significant patterns. *Machine Learning*, 68(1):1–33, 2007.

[106] G. I. Webb. Self-sufficient itemsets: An approach to screening potentially interesting associations between items. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(1):3, 2010.

[107] P. Whiting, M. Westwood, R. Beynon, M. Burke, J. A. Sterne, and J. Glanville. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. *J Clin Epidemiol*, 64(6):602–7, 2011.

[108] N. L. Wilczynski, K. A. McKibbon, S. D. Walter, A. X. Garg, and R. B. Haynes. Medline clinical queries are robust when searching in recent publishing years. *J Am Med Inform Assoc*, 20(2):363–8, 2013.

[109] S. S. L. Wong, N. L. Wilczynski, and R. B. Haynes. Comparison of top-performing search strategies for detecting clinically sound treatment studies and systematic reviews in medline and embase. *J Med Libr Assoc*, 94(4):451–5, 2006.

[110] R. W. Wright, R. A. Brand, W. Dunn, and K. P. Spindler. How to write a systematic review. *Clin Orthop Relat Res*, 455:23–9, 2007.

[111] O. R. Zaiane, M. Antonie, and A. Coman. Mammography classification by an association rule-based classifier. In *Third International ACM SIGKDD Workshop on Multimedia Data Mining (MDM/KDD'2002) in conjunction with Eighth ACM SIGKDD*, pages 62–69, 2002.

[112] M. J. Zaki. Generating non-redundant association rules. In *Proceedings of the Sixth International Conference on Knowledge discovery and Data Mining (KDD)*, pages 34–43, 2000.

[113] M. J. Zaki. Scalable algorithms for association mining. *Knowledge and Data Engineering, IEEE Transactions on*, 12(3):372–390, 2000.

[114] M. J. Zaki, S. Parthasarathy, M. Ogihara, W. Li, et al. New algorithms for fast discovery of association rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD)*, volume 97, pages 283–286, 1997.

Appendix A

Selected Reviews for DTA vs. Treatment Review Experiment

Type	CRG	Title
diagnosis	airways	Galactomannan detection for invasive aspergillosis in immunocompromized patients
diagnosis	ari	Clinical symptoms and signs for the diagnosis of Mycoplasma pneumoniae in children and adolescents with community-acquired pneumonia
diagnosis	back	Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain
diagnosis	back	Red flags to screen for malignancy in patients with low-back pain
diagnosis	back	Red flags to screen for vertebral fracture in patients presenting with low-back pain

Type	CRG	Title
diagnosis	eyes	Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy
diagnosis	gynaeca	Human papillomavirus testing versus repeat cytology for triage of minor cytological cervical lesions
diagnosis	infectn	Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries
diagnosis	infectn	Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults
diagnosis	muskinj	Physical tests for shoulder impingements and local lesions of bursa, tendon or labrum that may accompany impingement
diagnosis	preg	Second trimester serum tests for Down's Syndrome screening
diagnosis	renal	Cardiac testing for coronary artery disease in potential kidney transplant recipients
diagnosis	stroke	Magnetic resonance imaging versus computed tomography for detection of acute vascular lesions in patients presenting with stroke symptoms
treatment	airways	Anti-histamines for prolonged non-specific cough in children
treatment	airways	Beclomethasone versus placebo for chronic asthma
treatment	airways	Caffeine for asthma

Type	CRG	Title
treatment	airways	Cardioselective beta-blockers for chronic obstructive pulmonary disease
treatment	airways	Combination fluticasone and salmeterol versus fixed dose combination budesonide and formoterol for chronic asthma in adults and children
treatment	airways	Combination inhaled steroid and long-acting beta2-agonist in addition to tiotropium versus tiotropium or combination alone for chronic obstructive pulmonary disease
treatment	airways	Continuous versus intermittent beta-agonists for acute asthma
treatment	airways	Gastro-oesophageal reflux treatment for prolonged non-specific cough in children and adults
treatment	airways	Gold as an oral corticosteroid sparing agent in stable asthma
treatment	airways	Inhaled cromones for prolonged non-specific cough in children
treatment	airways	Inspiratory muscle training for asthma
treatment	airways	Intravenous beta2-agonists for acute asthma in the emergency department
treatment	airways	Physical training for interstitial lung disease
treatment	airways	Singing for children and adults with bronchiectasis
treatment	airways	Troleandomycin as an oral corticosteroid sparing agent in stable asthma

Type	CRG	Title
treatment	ari	Acellular vaccines for preventing whooping cough in children
treatment	ari	Acupuncture for mumps in children
treatment	ari	Acyclovir for treating varicella in otherwise healthy children and adolescents
treatment	ari	Advising patients to increase fluid intake for treating acute respiratory infections
treatment	ari	Antibiotics for preventing complications in children with measles
treatment	ari	Chinese medicinal herbs for sore throat
treatment	ari	Chinese medicinal herbs for the common cold
treatment	ari	Corticosteroids for parasitic eosinophilic meningitis
treatment	ari	Intranasal ipratropium bromide for the common cold
treatment	ari	Macrolides for diffuse panbronchiolitis
treatment	ari	Once or twice daily versus three times daily amoxicillin with or without clavulanate for the treatment of acute otitis media
treatment	ari	Pre-admission antibiotics for suspected cases of meningococcal disease
treatment	ari	Remediating buildings damaged by dampness and mould for preventing or reducing respiratory tract symptoms, infections and asthma
treatment	ari	Vaccines for post-exposure prophylaxis against varicella (chickenpox) in children and adults

Type	CRG	Title
treatment	ari	Vitamin C for preventing and treating the common cold
treatment	back	Acupuncture and dry-needling for low back pain
treatment	back	Acupuncture for neck disorders
treatment	back	Advice to rest in bed versus advice to stay active for acute low-back pain and sciatica
treatment	back	Antidepressants for non-specific low back pain
treatment	back	Arthroplasty versus fusion in single-level cervical degenerative disc disease
treatment	back	Back schools for non-specific low-back pain.
treatment	back	Behavioural treatment for chronic low-back pain
treatment	back	Botulinum toxin for subacute/chronic neck pain
treatment	back	Botulinum toxin injections for low-back pain and sciatica
treatment	back	Braces for idiopathic scoliosis in adolescents
treatment	back	Chinese herbal medicine for chronic neck pain due to cervical degenerative disc disease
treatment	back	Combined chiropractic interventions for low-back pain
treatment	back	Electrotherapy for neck pain
treatment	back	Exercise therapy for treatment of non-specific low back pain
treatment	back	Exercises for adolescent idiopathic scoliosis
treatment	back	Exercises for mechanical neck disorders

Type	CRG	Title
treatment	back	Exercises for prevention of recurrences of low-back pain
treatment	back	Herbal medicine for low back pain
treatment	back	Individual patient education for low back pain
treatment	back	Injection therapy for subacute and chronic low-back pain
treatment	back	Low level laser therapy for nonspecific low-back pain
treatment	back	Lumbar supports for prevention and treatment of low back pain
treatment	back	Manipulation or Mobilisation for Neck Pain
treatment	back	Manual material handling advice and assistive devices for preventing and treating back pain in workers
treatment	back	Massage for low-back pain
treatment	back	Massage for mechanical neck disorders
treatment	back	Mechanical traction for neck pain with or without radiculopathy
treatment	back	Medicinal and injection therapies for mechanical neck disorders
treatment	back	Multidisciplinary biopsychosocial rehabilitation for neck and shoulder pain among working age adults
treatment	back	Muscle relaxants for non-specific low-back pain
treatment	back	Neuroreflexotherapy for non-specific low-back pain

Type	CRG	Title
treatment	back	Non-steroidal anti-inflammatory drugs for low back pain
treatment	back	Patient education for neck pain
treatment	back	Physical conditioning programs for improving work outcomes in workers with back pain
treatment	back	Prolotherapy injections for chronic low-back pain
treatment	back	Radiofrequency denervation for neck and back pain
treatment	back	Rehabilitation after lumbar disc surgery
treatment	back	Single or double-level anterior interbody fusion techniques for cervical degenerative disc disease
treatment	back	Spinal manipulative therapy for acute low-back pain
treatment	back	Superficial heat or cold for low back pain
treatment	back	Surgery for cervical radiculopathy or myelopathy
treatment	back	Surgical interventions for lumbar disc prolapse
treatment	back	Total disc replacement for chronic back pain in the presence of disc degeneration
treatment	back	Transcutaneous electrical nerve stimulation (TENS) versus placebo for chronic low-back pain
treatment	back	Workplace interventions for neck pain in workers

Type	CRG	Title
treatment	eyes	Amniotic membrane transplantation for acute ocular burns
treatment	eyes	Anti-vascular endothelial growth factor for macular edema secondary to central retinal vein occlusion
treatment	eyes	Day care versus in-patient surgery for age-related cataract
treatment	eyes	Environmental and behavioural interventions for reducing physical activity limitation in community-dwelling visually impaired older people
treatment	eyes	Environmental sanitary interventions for preventing active trachoma
treatment	eyes	Ginkgo biloba extract for age-related macular degeneration
treatment	eyes	Interventions for chronic blepharitis
treatment	eyes	Interventions for late trabeculectomy bleb leak
treatment	eyes	Interventions for preventing posterior capsule opacification
treatment	eyes	Interventions for stimulus deprivation amblyopia
treatment	eyes	Interventions for trachoma trichiasis
treatment	eyes	Photodynamic therapy for neovascular age-related macular degeneration
treatment	eyes	Surgery for nonarteritic anterior ischemic optic neuropathy

Type	CRG	Title
treatment	eyes	Surgical implantation of steroids with antiangiogenic characteristics for treating neovascular age-related macular degeneration
treatment	eyes	Vitrectomy with internal limiting membrane (ILM) peeling versus vitrectomy with no peeling for idiopathic full-thickness macular hole (FTMH)
treatment	gynaeca	Adjuvant (post-surgery) chemotherapy for early stage epithelial ovarian cancer
treatment	gynaeca	Chemotherapy for high-grade glioma
treatment	gynaeca	Concomitant chemotherapy and radiation therapy for cancer of the uterine cervix
treatment	gynaeca	Granulocyte transfusions for preventing infections in patients with neutropenia or neutrophil dysfunction
treatment	gynaeca	High dose rate versus low dose rate intracavity brachytherapy for locally advanced uterine cervix cancer
treatment	gynaeca	Hormonal therapy in advanced or recurrent endometrial cancer
treatment	gynaeca	Hyperbaric oxygenation for tumour sensitisation to radiotherapy
treatment	gynaeca	Low molecular weight heparin versus unfractionated heparin for perioperative thromboprophylaxis in patients with cancer
treatment	gynaeca	Music interventions for improving psychological and physical outcomes in cancer patients

Type	CRG	Title
treatment	gynaeca	Nystatin prophylaxis and treatment in severely immunodepressed patients
treatment	gynaeca	Pharmacological treatment of depression in patients with a primary brain tumour
treatment	gynaeca	Retroperitoneal drainage versus no drainage after pelvic lymphadenectomy for the prevention of lymphocyst formation in patients with gynaecological malignancies
treatment	gynaeca	Surgery for cervical intraepithelial neoplasia
treatment	gynaeca	Surgical cytoreduction for recurrent epithelial ovarian cancer
treatment	gynaeca	Topotecan for ovarian cancer
treatment	infectn	Antiamoebic drugs for treating amoebic colitis
treatment	infectn	Antibiotics for treating scrub typhus
treatment	infectn	Antipyretic measures for treating fever in malaria
treatment	infectn	Artemisinin-based combination therapy for treating uncomplicated malaria
treatment	infectn	Artesunate versus quinine for treating severe malaria
treatment	infectn	Azithromycin for treating uncomplicated malaria
treatment	infectn	Chloroquine or amodiaquine combined with sulfadoxine-pyrimethamine for treating uncomplicated malaria
treatment	infectn	Drugs for preventing malaria in pregnant women

Type	CRG	Title
treatment	infectn	Drugs for preventing malaria in travellers
treatment	infectn	Drugs for treating uncomplicated malaria in pregnant women
treatment	infectn	Electronic mosquito repellents for preventing mosquito bites and malaria infection
treatment	infectn	Indoor residual spraying for preventing malaria
treatment	infectn	Insecticide-treated bed nets and curtains for preventing malaria
treatment	infectn	Insecticide-treated nets for preventing malaria in pregnancy
treatment	infectn	Interventions for preventing reactions to snake antivenom
treatment	infectn	Interventions for treating scabies
treatment	infectn	Interventions to improve disposal of human excreta for preventing diarrhoea
treatment	infectn	Intramuscular arteether for treating severe malaria
treatment	infectn	Iron-chelating agents for treating malaria
treatment	infectn	Low level laser therapy for treating tuberculosis
treatment	infectn	Oral vaccines for preventing cholera
treatment	infectn	Probiotics for treating persistent diarrhoea in children
treatment	infectn	Regimens of less than six months for treating tuberculosis
treatment	infectn	Rotavirus vaccine for preventing diarrhoea
treatment	infectn	Steroids for treating cerebral malaria

Type	CRG	Title
treatment	infectn	Sulfadoxine-pyrimethamine plus artesunate versus sulfadoxine-pyrimethamine plus amodiaquine for treating uncomplicated malaria
treatment	infectn	Vaccines for preventing malaria (blood-stage)
treatment	infectn	Vaccines for preventing malaria (pre-erythrocytic)
treatment	infectn	Vaccines for preventing smallpox
treatment	infectn	Vaccines for preventing tick-borne encephalitis
treatment	muskinj	Anaesthesia for hip fracture surgery in adults
treatment	muskinj	Conservative interventions for treating diaphyseal fractures of the forearm bones in children
treatment	muskinj	Conservative versus operative treatment for hip fractures in adults
treatment	muskinj	Dynamic compression plating versus locked intramedullary nailing for humeral shaft fractures in adults
treatment	muskinj	External fixation versus conservative treatment for distal radial fractures in adults
treatment	muskinj	Gamma and other cephalocondylic intramedullary nails versus extramedullary implants for extracapsular hip fractures in adults
treatment	muskinj	Hip protectors for preventing hip fractures in older people

Type	CRG	Title
treatment	muskinj	Hyperbaric oxygen therapy for delayed onset muscle soreness and closed soft tissue injury
treatment	muskinj	Hyperbaric oxygen therapy for promoting fracture healing and treating fracture non-union
treatment	muskinj	Interventions for treating acute elbow dislocations in adults
treatment	muskinj	Interventions for treating mallet finger injuries
treatment	muskinj	Platelet rich therapies for long bone healing in adults
treatment	muskinj	Surgical versus conservative interventions for anterior cruciate ligament ruptures in adults
treatment	muskinj	Surgical versus conservative interventions for treating ankle fractures in adults
treatment	muskinj	Ultrasound and shockwave therapy for acute fractures in adults
treatment	preg	Antibiotics for gonorrhoea in pregnancy
treatment	preg	Anti-D administration in pregnancy for preventing Rhesus alloimmunisation
treatment	preg	Bed rest in hospital for suspected impaired fetal growth
treatment	preg	Duration of treatment for asymptomatic bacteriuria during pregnancy
treatment	preg	Interventions for promoting smoking cessation during pregnancy
treatment	preg	Intracervical prostaglandins for induction of labour

Type	CRG	Title
treatment	preg	Intravenous oxytocin alone for cervical ripening and induction of labour
treatment	preg	Maternal oxygen administration for suspected impaired fetal growth
treatment	preg	Oxytocin versus no treatment or delayed treatment for slow progress in the first stage of spontaneous labour
treatment	preg	Package of care for active management in labour for reducing caesarean section rates in low-risk women
treatment	preg	Risk scoring systems for predicting preterm birth with the aim of reducing associated adverse outcomes
treatment	preg	Third trimester antiviral prophylaxis for preventing maternal genital herpes simplex virus (HSV) recurrences and neonatal infection
treatment	preg	Treatments for iron-deficiency anaemia in pregnancy
treatment	preg	Vibroacoustic stimulation for fetal assessment in labour in the presence of a nonreassuring fetal heart rate trace
treatment	preg	Vitamin C supplementation in pregnancy
treatment	renal	Aldosterone antagonists for preventing the progression of chronic kidney disease

Type	CRG	Title
treatment	renal	Angiotensin converting enzyme inhibitors and angiotensin II receptor antagonists for preventing the progression of diabetic kidney disease
treatment	renal	Antibiotic duration for treating uncomplicated, symptomatic lower urinary tract infections in elderly women
treatment	renal	Antimicrobial agents for treating uncomplicated urinary tract infection in women
treatment	renal	Antiviral medications for preventing cytomegalovirus disease in solid organ transplant recipients
treatment	renal	Biocompatible hemodialysis membranes for acute renal failure
treatment	renal	Corticosteroid therapy for nephrotic syndrome in children
treatment	renal	Double bag or Y-set versus standard transfer systems for continuous ambulatory peritoneal dialysis in end-stage renal disease
treatment	renal	Fluids and diuretics for acute ureteric colic
treatment	renal	Immunosuppressive agents for treating IgA nephropathy
treatment	renal	Immunosuppressive treatment for idiopathic membranous nephropathy in adults with nephrotic syndrome
treatment	renal	Non-immunosuppressive treatment for IgA nephropathy

Type	CRG	Title
treatment	renal	Pharmacological interventions for preventing complications in idiopathic hypercalciuria
treatment	renal	Tidal versus other forms of peritoneal dialysis for acute kidney injury
treatment	renal	Ultrasound use for the placement of haemodialysis catheters
treatment	stroke	Antibiotic therapy for preventing infections in patients with acute stroke
treatment	stroke	Calcium antagonists for acute ischemic stroke
treatment	stroke	Cilostazol versus aspirin for secondary prevention of vascular events after stroke of arterial origin
treatment	stroke	Corticosteroids for aneurysmal subarachnoid haemorrhage and primary intracerebral haemorrhage
treatment	stroke	Fibrinogen depleting agents for acute ischaemic stroke
treatment	stroke	Force platform feedback for standing balance training after stroke
treatment	stroke	Low-molecular-weight heparins or heparinoids versus standard unfractionated heparin for acute ischaemic stroke
treatment	stroke	Music therapy for acquired brain injury
treatment	stroke	Oral anticoagulants versus antiplatelet therapy for preventing stroke in patients with non-valvular atrial fibrillation and no history of stroke or transient ischemic attacks

Type	CRG	Title
treatment	stroke	Puerarin for acute ischaemic stroke
treatment	stroke	Sonothrombolysis for acute ischaemic stroke
treatment	stroke	Therapy-based rehabilitation services for stroke patients at home
treatment	stroke	Thrombolysis for acute ischaemic stroke
treatment	stroke	Triflusal for preventing serious vascular events in people at high risk
treatment	stroke	Water-based exercises for improving activities of daily living after stroke

Table A.1: List of collected review used for DTA vs. Treatment experiments in Chapter 3