Towards data-intensive epidemiology

Explorations in systematic reviews and causal inference

Louise Amanda Claire Millard

A thesis submitted to the University of Bristol in accordance with the requirements for the degree of Doctor of Philosophy in the Faculty of Medicine and Dentistry.

August 2015

Abstract

The field of epidemiology is now experiencing a data deluge, demanding appropriate methods to efficiently analyse large amounts of data. In this thesis we present advances towards data-intensive epidemiology, introducing novel methods and applications of data mining in this field. We focus on two distinct applications.

Our first application is the task of risk of bias assessments of systematic reviews. At present these are a highly manual process, where reviewers identify relevant parts of research articles for a set of methodological elements that affect the risk of bias, in order to make a risk of bias judgement for each of these elements. We use text mining to identify relevant sentences within the text of included articles, to rank articles by risk of bias, and to reduce the number of risk of bias assessments the reviewers need to perform by hand.

The application of text mining to risk of bias assessments also led to the following methodological contributions. We introduce the concept of a rate-constrained ranking task, of which ranking articles for rapid reviews is an example. We derive a novel metric, the rate-weighted area under the ROC curve (rAUC), to evaluate ranking models for rate-constrained ranking tasks. Furthermore, we derive a method to generate confidence bounds around ROC curves, that is particularly appropriate for these types of tasks.

Our second application is the task of choosing hypotheses to test in epidemiological analyses. Currently researchers use prior knowledge about the composition of causal pathways, and their own research interests and preconceptions, to decide which hypotheses to test. Where no strong priors exist it may be preferable to use a systematic approach to identify those to follow up. We present a novel screening step that uses Mendelian randomisation to systematically search a large number of hypotheses for potentially causal relationships that should be investigated further. As an exemplar we search for the causal effects of body mass index (BMI) and find many associations with outcomes that are supported in the literature.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

Acknowledgements

I would like to start by thanking my supervisors – Peter Flach, Julian Higgins, George Davey Smith and Nic Timpson, for their support, time and enthusiasm throughout my PhD.

I would also like to thank co-authors of published work described in this thesis – Meelis Kull, Neil Davies and Kate Tilling. Thanks to Jonathan Sterne and Nello Cristianini, for their valuable comments during my annual reviews.

Thanks to Ida Pu and Robert Zimmer, for their enthusiasm and encouragement during my undergraduate degree, that helped to develop my lasting interest in Computer Science. Thanks to my colleagues and friends in the School of Social and Community Medicine who made every day in the office a supportive and happy place to be, with particular mention to Teri, Kim, Anna, and Nabila.

Finally, a special thanks to Sam, for his continued support and encouragement.

Contents

List of Figures

List of Tables

List of Abbreviations

ALSPAC Avon Longitudinal Study of Parents and Children. **AUC** Area under the ROC curve.

BEDROC Boltzmann-enhanced discrimination of ROC.

BMI Body mass index.

CI Confidence interval.

CROC Concentrated ROC.

DANVA Diagnostic Analysis of Nonverbal Accuracy.

EWAS Environment-wide association study.

FPR False positive rate.

GWAS Genome-wide association study.

HOME Home Observation for Measurement of the Environment.

IV Instrumental variable.

NDCG Normalised discounted cumulative gain.

pAUC Partial AUC.

pheWAS Phenome-wide association study.

rate Predicted positive rate.

rAUC Rate-weighted AUC.

RCT Randomised controlled trial.

RIE Robust initial enhancement.

ROC Receiver operating characteristic.

SD Standard deviation.

SGD Stochastic gradient descent.

SLR Sum of the log ranks.

SNP Single nucleotide polymorphism.

SVM Support vector machine.

TPR True positive rate.

WISC Wechsler Intelligence Scale for Children.

Chapter 1

Introduction

Epidemiological analyses have been highly effective at identifying causal associations between exposures and health outcomes to improve public health. For instance, the early example that smoking causes lung cancer led to a decrease in cigarette smoking in the United Kingdom followed by a reduction of lung cancer prevalence [1]. Traditionally, epidemiological analyses involve deciding which hypothesis to test, and then analysing the selected hypothesis in a collection of individuals to test the association directly. Nowadays, the number of variables available for analysis and hence the number of potential hypotheses is large, and this presents a major challenge for epidemiologists in need of data-intensive methods that can efficiently process the data to establish meaningful relationships between variables.

The increase in data is due to the availability of new sources of data, and new techniques to generate data. Large cohorts exist with data such as phenotypic variables, genetic variation, gene expression, DNA methylation, of which the Avon Longitudinal Study of Parents and Children (ALSPAC) [2] used in Part II of this thesis, is a comprehensive example. There is also a vast array of data available online, examples of which include biological databases such as KEGG [3] and social media websites such as twitter. Furthermore, research articles are an important source of knowledge, and can themselves be analysed to make new inferences. For instance, systematic reviews use research articles to analyse the evidence from multiple studies together, to improve the answer to a research question [4].

The dramatic increase in data availability provides a need for appropriate techniques

to analyse these large datasets. One such approach is data mining, where a potentially large number of hypotheses are generated and explored to identify meaningful associations. The vision explored in this thesis is one of data-intensive epidemiology, where epidemiological analyses use scalable data mining methods to assist or automate the inference of causal associations within large epidemiological datasets. In this thesis we present two applications of data mining to help cope with the data deluge encountered by epidemiologists.

The first application we present assists the process of systematic reviews. Systematic reviews have been highly successful at combining the evidence from multiple studies to answer a research question more comprehensively than is possible from each individual study. At present systematic reviews are performed by hand, but with the ever growing number of published research studies this is not sustainable [5]. Hence systematic reviews need modernising to cope with the amount of data.

Much of the systematic review process involves making inferences from information found in research articles, such as the initial search for relevant research articles, screening the research articles, extracting the information needed to perform the review, and performing a risk of bias assessment [4]. The application of data mining to textual data, known as text mining, can potentially be used to automate these types of tasks. While text mining for systematic reviews is currently an active area of research [6–20], to date there has been little research investigating the automation of risk of bias assessments [21–23]. Hence, this is a focus of our work.

Risk of bias assessments seek to determine if the result of a clinical trial is likely to be biased, which may happen if the study methods are not adequate. For instance, where a participant reports an outcome themselves, they may exaggerate the effect if they know they have received the active drug rather than a placebo. The effect estimate may then be more extreme than the true effect due to the intervention alone. At present reviewers need to manually identify relevant parts of research articles for a set of methodological elements that affect the risk of bias, in order to make a risk of bias judgement for each of these elements. We demonstrate that text mining can be used to assist systematic reviews by identifying text relevant to risk of bias within research articles, and predicting the risk of bias values of clinical trials that the research articles describe.

The second application we present assists hypothesis-driven analyses, where epi-

demiologists choose a hypothesis to test and then perform this analysis in a particular cohort of individuals. The large number of variables in a cohort dataset means that there is an extremely large number of potential hypotheses, as the number of potential exposure-outcome pairs grows exponentially with the number of variables. Where there is no strong prior knowledge of which hypothesis should be tested, it may be preferable to use a systematic approach to search all hypotheses to identify those to follow up. Existing methods that use this approach include genome-wide association studies (GWAS), that search for loci on the genome that are associated with phenotypic traits. GWAS have been highly successful at generating replicable associations. Prior to GWAS, the results of candidate gene studies – the hypothesis-driven alternative to GWAS – were largely non-replicable [24, 25].

In this thesis we present a novel approach to search a potentially large hypothesis space to identify associations that may be causal, using an instrumental variable approach called Mendelian randomisation. We propose a general pipeline to be used as a first step to find potentially interesting associations to then follow up with further analyses. We demonstrate this approach by searching for the causal effects of body mass index (BMI).

1.1 Thesis overview

This thesis consists of two parts. Part I presents our work on assisting systematic reviews through automation of risk of bias assessments. Part II presents our work on assisting hypothesis selection for hypothesis-driven analyses.

Part I begins in Chapter 2, with background on systematic reviews and risk of bias. We introduce our three objectives, of: 1) identifying relevant sentences within research articles, 2) ranking articles by risk of bias and 3) reducing the number of assessments the reviewers need to perform by hand. We give an overview of related work. We present the methods we used to generate our dataset and a summary of this data. We provide an introduction to machine learning and ROC analysis methods, that are used in the subsequent chapters.

Chapters 3 and 4 present contributions to the ROC analysis literature, the motivation of which is provided by rapid reviews. In Chapter 3 we introduce the concepts of rate-oriented and rate-constrained ranking tasks, and show how ranking articles for rapid

reviews can be formulated as a rate-constrained ranking task. We present a new metric, the rate-weighted AUC (rAUC), to evaluate the performance of ranking models for rate-constrained ranking tasks using a novel evaluation of ROC curves. In Chapter 4 we present a new approach to generate confidence bounds for ROC curves, as a series of confidence intervals along the ROC curve. This approach is particularly appropriate for rate-oriented ranking tasks. In Chapter 5 we present our main results for assisting systematic reviews and risk of bias assessments. We use supervised machine learning to train a set of models in order to achieve our objectives described in Chapter 2.

Part II begins in Chapter 6, with an overview of current approaches that search for hypotheses in individual level data, a discussion of the issues of causality and confounding in observational epidemiology, and an overview of the Mendelian randomisation approach, that can help infer causality between two traits. In Chapter 7 we introduce a novel method to search a potentially large number of hypotheses for causal associations using Mendelian randomisation. We perform a proof-of-principle analysis that searches for the causal effects of body mass index, and present the results of this analysis in Chapter 8.

Chapter 9 concludes with a summary of the work described in this thesis and a discussion of future work.

1.2 Main contributions

Our main contributions are as follows.

Part I: Assisting systematic reviews

- We introduce a novel metric, the rate-weighted AUC, to evaluate ranking performance for rate-constrained ranking tasks.
- We propose a formulation of ranking articles for rapid reviews, in terms of a rate-constrained ranking task.
- We introduce an effective approach to generate rate-oriented point-wise confidence bounds for ROC curves.

- We demonstrate that text mining can be used to assist systematic reviews through automation of risk of bias assessments. Specifically we have shown that it is possible to:
  - Rank sentences in research articles by relevance for each risk of bias property individually.
  - Rank articles by risk of bias so that a reviewer can assess articles from low to high predicted risk of bias.
  - Reduce the number of articles that need to be reviewed by hand, by classifying a subset of articles automatically for each risk of bias property individually.

Part II: Assisting hypothesis selection

- We introduce a novel approach to search for causal associations using Mendelian randomisation.
- We demonstrate this approach using body mass index as an exemplar.

The work presented in this thesis has been published in the following peer reviewed research articles. Consistent with the convention in these papers, I refer to 'we' rather than 'I' in this thesis. I am first author of these papers, and have led and performed the work they describe.

> L. A. C. Millard, N. Timpson, K. Tilling, P. A. Flach, and G. Davey Smith, MR-PheWAS: hypothesis-prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization, *Scientific reports*, vol. 5, 2015. [26]
>
> *I contributed to the design of analyses, performed the analyses, wrote the initial version of the manuscript, revised the manuscript.*

> L. A. C. Millard, P. A. Flach, and J. P. T. Higgins, Rate-constrained ranking and the rate-weighted AUC, *in Machine Learning and Knowledge Discovery in Databases*, pp. 386403, Springer, 2014. [27]

I conceived and developed the initial concept, derived the solution, performed the analyses, wrote the initial version of the manuscript, revised the manuscript.

L. A. C. Millard, M. Kull, and P. A. Flach, Rate-oriented point-wise confidence bounds for ROC curves, *in Machine Learning and Knowledge Discovery in Databases*, pp. 404421, Springer, 2014. [28]

I implemented the solutions, contributed to the design of the experiments, performed the experiments, wrote the initial version of the manuscript, revised the manuscript. MK proposed the analytical solution (Theorem 4.1), the parametric extension to a whole ROC table (described in Section 4.5.3), and the bootstrap method (described in Section 4.5.4).

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*, 2015. doi:10.1093/ije/dyv306. [29]

I conceived and developed the initial concept, developed the objectives, designed the experiments, implemented the experiments, wrote the initial version of the manuscript, revised the manuscript, developed the prototype.

In summary, the work presented in this thesis makes strides towards our vision of data-intensive epidemiology. This includes both methodological developments and novel applications of existing data mining methods to epidemiological research tasks. We focus on two specific applications – assisting systematic reviews and assisting hypothesis generation.

Part I

Assisting with the conduct of systematic reviews

Chapter 2

Background and methods

In this chapter we begin by providing background on systematic reviews and risk of bias assessments. We introduce a set of three objectives to assist systematic reviews and risk of bias assessments, and give an overview of related work. We then describe the methods we use to generate our dataset, and provide a summary of this data. Lastly, we provide key background on machine learning methods and ROC analysis, that are used throughout the remainder of this thesis.

2.1 Systematic reviews and risk of bias

Observational analyses test the association between two traits but often suffer from confounding, where an association between these traits is affected by their associations with other variables, known as confounding factors. This may mean an association is found that is due to confounding rather than a causal effect of one trait on another. For example, when testing whether drinking alcohol affects the risk of developing lung cancer, an association may be seen because people who smoke more on average drink more, and smoking is known to increase the risk of developing lung cancer. Smoking confounds the association between alcohol and lung cancer risk. Furthermore, irrespective of confounding, an association between an exposure and outcome may occur due to reverse causation, where the outcome affects the exposure rather than vice-versa, such that the causal direction cannot be determined [30].

The randomised controlled trial (RCT) study design avoids confounding because

participants are randomly assigned to study groups, and this determines the intervention they receive. This means that (in principle) any relationship between the exposure and outcome should not be due to confounding. For instance, an RCT may test the effect of a drug on a disease, where one group is given an active drug and another is given a placebo. Because this assignment is random it will not be associated with confounding factors. Furthermore, because participants are randomly assigned to study groups we know the causal direction is from the study group assignment to the outcome. We note however that it is not always possible to perform a RCT as it may not be ethical to do so (when the interventions are harmful).

RCTs are regarded as the best study design to evaluate the effect of an intervention. Several observational associations have been contradicted by subsequent RCTs [31]. The results of several RCTs can be combined to try to increase the precision of the estimate, in a systematic review.

2.1.1 Systematic reviews

Systematic reviews combine evidence from multiple studies to answer a research question more comprehensively than is possible from an individual study. An important collection of systematic reviews are those published by Cochrane. These are produced using a clearly defined procedure, detailed in the Cochrane Handbook [4]. Systematic reviews are important to determine the extent to which findings reported in individual studies are generalisable [32]. This is because studies within a review usually differ with regards to the population characteristics or study methods, such as the age of participants in the study. These differences may alter the effect of an intervention. For example, an intervention may be more effective in younger compared to older people.

The steps of a systematic review on the effects of an intervention are shown in Figure 2.1. The main activities are shown on the left in blue, and sub-activities are shown next to each of these in grey. The first step of a systematic review is to specify the research question. For systematic reviews of the effect of an intervention this should include the population, intervention, comparison, and outcome attributes, referred to as PICO. Eligibility criteria are then specified using the research question and further details such as the study design. The next step of the review is a search of online databases for articles describing studies that should be included. The search is performed and a

set of titles and abstracts of research articles are returned. This set typically contains a large number of articles that are not appropriate for the review, either because they are irrelevant or do not meet the inclusion criteria. Hence the next step is to screen these articles to remove those that are unsuitable.

Screening is performed in two stages. The first stage involves assessing the title and abstract of each research article to determine whether it can be excluded from the review. Whilst this can remove many articles, there are still some articles remaining after this step that should also be excluded, because the information related to the inclusion criteria is not available in the title or abstract of the article. The second screening step assesses the full text of the remaining articles to determine whether these should be included in the review. After screening, information that is needed for the review is extracted from each research article. This is then followed by a risk of bias assessment, which we discuss in the next section of this chapter. The final step of a review is to synthesise the evidence, where inferences are made using the information extracted from the research articles [4].

The evidence synthesis step seeks to summarize the findings of the studies included in the review. This may include a meta-analysis, which quantitatively combines the results from multiple studies with the aim of providing a more precise estimate. Meta-analyses can only be performed when the studies in a review are sufficiently homogeneous, such that it makes sense to combine the results into a single estimate. One striking example that demonstrates the value of meta-analyses is that of the effect of streptokinase on myocardial infarction [33]. This meta-analysis, published in 1992, included 33 trials and shows that streptokinase reduces the odds of myocardial infarction. However, this association could have been known as early as 1973, had a meta-analysis been performed to combine the results from the existing studies at that time.

Rapid reviews

A rapid review is a specific type of systematic review that needs to be performed under strict time and resource constraints. Rapid reviews should follow the broad principles of systematic reviews, where a medical research question is asked, such as the effect of a drug on a disease, and the evidence from all relevant research articles is compiled to give a better estimate of the drug's effect than each individual study provides. However,

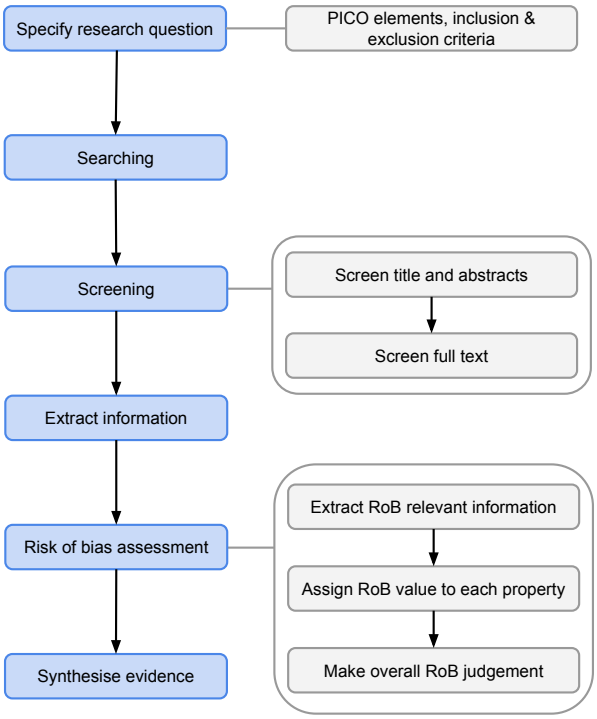


Figure 2.1: Illustration of the systematic review process applied to a question on the effects of an intervention. Main activities are shown on the left in blue, sub-activities are shown in grey next to each main activity. PICO: population, intervention, control and outcome; RoB: risk of bias.

rapid reviews typically take between 3 weeks and 6 months [34], whereas a systematic review can take between 1 and 3 years. In order to perform the review in such a short amount of time, the reviewer must streamline the process.

Currently there is no single 'best practice' for performing a rapid review. As shown in Figure 2.2, typically the search for relevant articles may be iterative, unlike standard systematic reviews. This is because often the number of retrieved articles is too large to be assessed within the allocated time, and so the reviewer may iteratively refine the search query until the number of articles is deemed manageable. The search criteria may also be refined to restrict to publications that can be easily accessed, such as those in a particular language, or only a subset of the available publication databases may be searched [34]. During the extraction of data from articles the reviewers may choose to only collate the information available in the articles and not to contact authors of

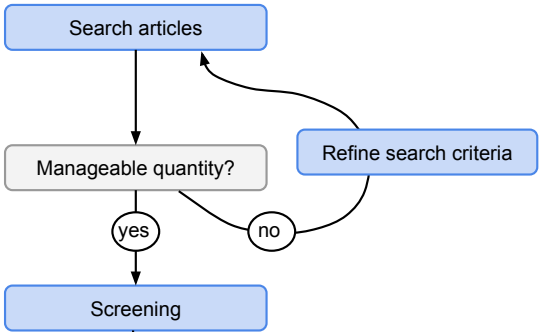


Figure 2.2: Typical iterative search process for rapid reviews.

publications to get further information when this is not stated in the article [34].

There has been little research looking at the effect of these approaches on the results of reviews. Egger et al. [35] investigated the effect of the depth of the literature search on the results of systematic reviews, and found that trials that were difficult to locate (by being unpublished, not in English or not indexed in Medline) were smaller and of lower quality, suggesting that including these studies may actually increase the bias of studies in a review. They suggest that thorough assessments of study quality should take priority over thorough literature searches when time constraints exists, so that the bias due to methodological quality is low for the studies included in the review.

2.1.2 Risk of bias

Bias is any systematic deviation from the truth in the results or inferences of a study [4]. When combining the evidence from several studies in a systematic review, it is important to consider whether each study it includes may be biased. Results of clinical trials may be biased if the study methods are not adequate.

As shown in Figure 2.1, a risk of bias assessment should be performed before synthesising the evidence, because this assessment affects how studies are included in the review. Studies with poor methodological quality may be given less weight in a review or excluded completely. Assessing risk of bias in clinical trials typically involves extraction of information sufficient to assign a judgement on the adequacy of each methodological property affecting risk of bias [36]. A single risk of bias assignment is then made for each study in the review using the individual risk of bias assignments of each methodological property.

Before we discuss the methodological properties that affect the risk of bias we first describe the main types of bias, within which the methodological properties can be grouped.

Types of bias

The five main types of bias in clinical trials are: selection, performance, detection, attrition and reporting bias [4]. Each of these are problematic because, as illustrated in Figure 2.3, they may add a path between the exposure and outcome such that the association between them may not be solely because of a causal effect of the exposure on the outcome.

Selection bias refers to the selection of participants to the study groups in a trial. Participants should be randomly assigned to study groups to ensure the intervention is not associated with confounding factors. For instance, if all participants assigned to a particular group have more severe symptoms then differences in the outcome across groups may be due to this, rather than differences in the intervention.

Performance bias refers to differences in the treatment given to participants, beyond the specific intervention of the trial. For instance, a treatment provider may supplement the treatment of participants receiving a placebo, due to their desire to provide some kind of active treatment, but this may affect the results of the study. Detection bias refers to any differences in the assessment or reporting of the outcome across study groups. This is particularly important for subjective outcomes because, for instance, if a participant knows they are receiving the active treatment rather than a placebo they may exaggerate their response when reporting a subjective outcome.

Attrition bias is caused when participants leave a trial after they have registered. This may cause bias in the study result if exit from a study is associated with the group the participants were assigned to. For example, an intervention may cause adverse effects on participants, causing many in this group to leave the trial. To avoid this bias, where possible (ie. where the outcome can still be ascertained from former participants) an intention to treat analysis is performed where the results are analysed with those who left the trial still in the analysis.

Reporting bias refers to the selective reporting of study results. For instance, researchers may change which test is the main result of a trial depending on the signif-

icance of each finding. The outcomes reported in publications have been shown to be more likely to be significant compared with unreported outcomes [37].

Each type of bias is caused by poor methodological quality, which we now describe.

Methodological sources of bias

The methodological elements that can cause bias are referred to as domains [4]. We also refer to domains as risk of bias properties. In this thesis we focus on three key domains: 1) the method used to generate the random sequence to assign participants to groups, 2) the method use to allocate participants to groups, and 3) whether the participants and study personnel are blinded to the study group assignments of the participants, which we refer to as sequence generation, allocation concealment and blinding respectively. As we describe below, these may cause selection, performance and detection bias. Other domains include selective reporting (causing reporting bias), and incomplete outcome data (causing attrition bias). These may require more complex methods to predict, such as requiring comparisons with initial trial protocols to establish if selective reporting has occurred, and hence we leave these for future work.

Inadequate sequence generation, allocation concealment and blinding have been shown to be associated with more extreme effect estimates [38, 39]. Recently this association has been shown to be driven by the effect of bias for subjective outcomes [40,41].

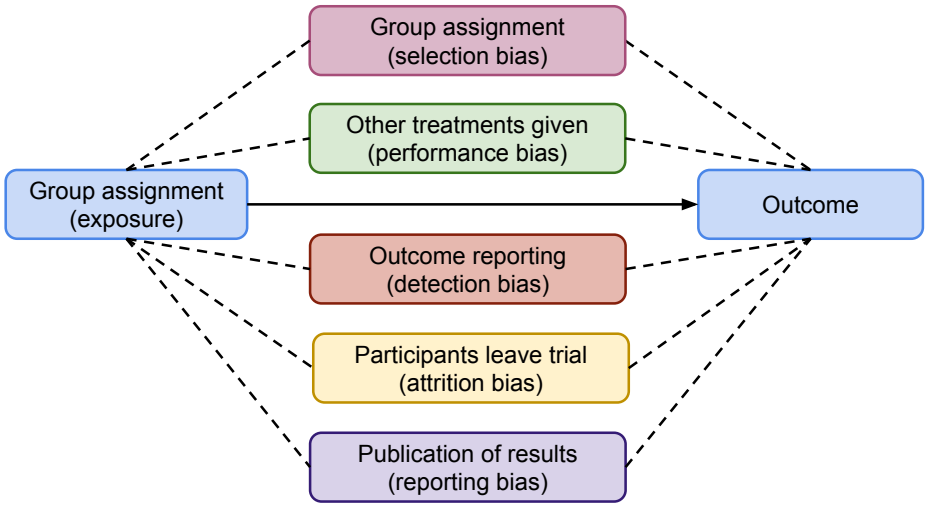


Figure 2.3: Risk of bias illustration.

The success of the RCT study design depends on successful randomisation of participants to the study groups. This depends on both the generation of the random sequence used to assign participants to groups, and the allocation of the participants to groups using these random assignments. If either one of these is insufficient the randomised assumption is violated and the trial may suffer from selection bias.

Sequence generation Methods that may be used to generate random sequences include throwing dice, tossing a coin, or using a computer program that generates random numbers [42]. Methods that are not random include using the participant's date of birth, or their date of admission to hospital. These are inadequate because the assignment can be predicted if these details are known. There are three types of randomisation methods: simple, restricted and stratified [43]. Simple randomisation generates a random sequence without any constraints on the proportion of participants in each group. Dice, for instance, may be used to assign to two groups, with equal probability per group. Simple randomisation can use custom probabilities to generate groups of unequal sizes such as with a weighted die, and can also create assignments for more than two groups. Given a small sample the proportions of participants in each group may be quite different from the intended proportions, but for larger samples this will converge.

It can be beneficial to have equal or close to equal numbers of participants in each group, which simple randomisation is not able to guarantee. Restricted randomisation uses a constrained sequence generation method to fix the number of participants in each group. This is commonly done with block randomisation, where a block size m is stated and each consecutive sequence of m participants registered to the trial will be assigned to groups in a fixed proportion. There may be only a single block for a whole study such that at the end of the study the participants will have been assigned to study groups in fixed proportion. It can be useful to use several blocks so that before the recruitment to a study is finished any interim analysis can be performed with the correct proportion of participants in each group. Block randomisation has the disadvantage that the sequence is not entirely random. When assigning the last few participants in each block it may be possible to know with high probability to which group participants will be assigned, and the last assignment in each block is deterministic. This is a bigger issue for smaller block sizes because the assignments may be predicted more frequently.

As well as balancing the number of participants assigned to groups, it may be bene-

ficial to balance the frequency of baseline characteristics across groups. Baseline characteristics may affect the outcome independently of the intervention and so if they are not balanced across groups we cannot know if it is the intervention or baseline characteristic that is causing any differences in the outcome across groups. For larger samples baseline characteristics will be naturally balanced across groups. For smaller samples this can be forced using stratified randomisation. This method involves stratifying on the baseline characteristic(s) and using a restricted randomisation approach within each of these strata [43]. The main disadvantage is the potentially small block size within the strata which can reduce the concealment of the random sequence [43].

Allocation concealment Allocation of incoming participants to groups should be concealed during the randomisation process [4, 44]. It is important that the clinician does not know which group is being assigned to the participant, or they may interfere with the assignment. A common method of allocation is to use envelopes, which should be opaque and sealed so that staff are unable to see the assignments. This method is vulnerable to attempts to alter the assignments, and so envelopes should be sequentially numbered so that the investigators are unable to change which envelopes are given to each participant. Furthermore, it is recommended that investigators write the participant name on the envelope prior to opening it so that they cannot decide to change to another envelope after seeing the assignment.

Another method of allocation is the use of a central computer random number generator, where the clinician cannot see future assignments without requesting them from the system. Provided the clinician enrols a participant prior to retrieving the random assignment, the assignment cannot be altered because the order of random assignments generated from the central system must match the order of enrolment [44].

Blinding The participants of a study and the study staff should, if possible, not know the treatment group of participants [45]. For example, where an intervention is given by pill, an identical but placebo pill can be also given to the control group. Blinding is not always possible, such as for non-medicinal interventions.

A participant's or outcome assessor's reporting of subjective outcomes may be affected if they know which treatment they are receiving. Hence, a study with inadequate blinding of participants or outcome assessors may suffer from detection bias. For in-

stance, a participant who knows they are receiving an intervention rather than a placebo may exaggerate the improvement in their condition. A study with inadequate blinding of treatment providers or participants may suffer from performance bias. Treatment providers should be unaware of the participant's assignment so that they provide consistent care to all groups.

Risk of bias assessments

To assess the risk of bias of studies in a systematic review, Cochrane recommend assessing the individual domains of risk of bias, such as sequence generation, which we have described above [4]. A single judgement for each study is then made, combining the risks from each individual domain, which involves assessing the importance of each domain [4]. For example, as discussed above, inadequate sequence generation, allocation concealment and blinding exaggerate the effect estimates for subjective outcomes. If the outcome is objective, such as all cause mortality, these domains may be less important [4]. The risk of bias should be assessed separately for each outcome within studies where multiple outcomes are reported. These judgements are highly subjective as the reviewer must make a judgement of the risk of bias of each property from the text in publications [46].

The results of risk of bias assessments should be used in the analysis of findings. Cochrane recommends either restricting to studies with known low risk of bias, or performing a separate meta-analysis for each risk of bias value [4]. Another common method is to perform a sensitivity analysis whereby a meta-analysis is performed both with and without the studies with high / unknown risk of bias to assess the effect of these studies on the result [4].

A key barrier in the assessment of risk of bias is the insufficient reporting of study methods in trial reports. The CONSORT statement was created to improve the reporting of clinical trials [47] and specifies that information relevant to risk of bias should be described in a trial report. However, although this has improved reporting of methods relating to risk of bias, this is still often inadequate [48, 49]. Other sources such as trial protocols can contain information that is not reported in the study publications [50].

One study compared the risk of bias assignment when using just the publication compared with using the publication with additional information from trial protocols,

data collection forms and individual patient data, for studies of cancer interventions [51]. The individual patient data was used to assess the level of attrition in a study. Further information about sequence generation and allocation concealment was sought from the trial protocol. Using the additional information increased the number of studies classified as low risk of bias. This was largely due to a move from unknown (rather than high) risk of bias, to the low assignment. The number of studies assigned low increased from 44% to 69%, and 42% to 93% for sequence generation and allocation concealment, respectively [51]. As this work focused on cancer treatments where the outcome was all cause mortality, in both cases the risk of bias due to blinding was always assigned low. Combining these assignments to give a single value per study, and the number of studies with a low assignment increased from 24% to 67% when the additional information sources were included [51]. This demonstrates the potential increase in number of trials known to give high quality evidence (with low risk of bias) if other sources of information are used. Furthermore, this work is consistent with other studies that showed that poor reporting did not mean the methods of the study were poor [52–54], such that a study with a unknown risk of bias assignment may actually have low risk of bias.

Cochrane reviews have been shown to have better methodological quality than non Cochrane reviews [55–59]. Hopewell et al. showed that Cochrane reviews assessed risk of bias more often than non Cochrane reviews [58]. This is expected because risk of bias assessments have always been a mandatory section in Cochrane reviews. In this study nearly all Cochrane reviews ($95\% - 100\%$) assessed sequence generation, allocation concealment, blinding and missing data due to attrition, whereas this was much lower ($60\% - 69\%$) for non Cochrane reviews [58].

2.2 Automating risk of bias assessments

In this section we present our main objectives to automate risk of bias assessments in order to assist systematic reviews. While we seek to automate the risk of bias assessments we emphasise that this is not done to replace human reviewers in this task, but to provide a set of tools to assist the reviewers. We begin by describing our motivation for these objectives.

2.2.1 Motivation

Thorough systematic reviews are time consuming, often lasting up to three years and requiring two reviewers to assess each research article to minimise errors. One study estimated that 80% of risk of bias assessments took between 10 and 60 minutes to perform [60]. Furthermore, risk of bias judgements are imperfect. Studies have shown that reviewers often report different levels of risk of bias for the same studies [61–65]. This may happen, for instance, if a reviewer misses key sentences [64]. Automating aspects of the risk of bias assessments has the potential both to reduce the time required to perform a review and to reduce human error and subjectivity in the reviewing process.

At present, risk of bias assessments require the reviewer to read through the research articles to find the relevant parts of the text. They use this information to assign a risk of bias judgement to a clinical trial for each methodological property affecting risk of bias. However, identifying relevant sentences and predicting a risk of bias assignment from text in articles are tasks that text mining methods have the potential to perform automatically.

When performing a systematic review, the order the reviewer assesses research articles is arbitrary. It may be beneficial to identify the high quality articles (with low risk of bias) early in a review so that these can be prepared for inclusion in the analyses. Rapid reviews could also benefit greatly from this automated prioritisation [27, 66].

2.2.2 Assisting reviews: our three objectives

We now specify three concise objectives, to assist systematic reviews in different ways. Objective 1 involves predicting the relevance of sentences within articles to risk of bias, whereas objectives 2 and 3 involve predicting the risk of bias values of the articles themselves. Figure 2.4 illustrates how these objectives fit into the systematic review process.

Objective 1: Assisting reviewers by identifying relevant sentences

Risk of bias assessments require the reviewer to read through the research articles, finding the relevant parts of the text. We aim to assist this process by predicting a relevance score for each sentence, denoting the likelihood that a sentence contains relevant in-

formation with regards to a particular risk of bias property. For instance, a sentence containing the phrase "we used sealed opaque envelopes" is a relevant text segment for the allocation concealment risk of bias property. This text can be highlighted in the article to indicate that the reviewer should consider this part of the text when assessing the level of allocation concealment performed in this study. To reiterate, this sentence level learning is concerned with predicting the relevance of a sentence for a risk of bias property, rather than the level of risk due to the information in the sentence.

Objective 2: Ranking articles by risk of bias

At present, the order the reviewer assesses research articles is arbitrary. We aim to create a model that can rank articles in order of risk of bias, such that reviewers can assess articles describing studies with low risk of bias before those with high or unknown risk of bias. It is beneficial for the reviewer to identify the high quality articles early in a review so that these can be passed to the statistician, who then has a better view of the evidence because these high quality studies are more likely to feature in analyses within the review.

Furthermore, ranking articles by risk of bias would be useful for rapid reviews, as described previously [27]. Rapid reviews should follow the broad principles of a systematic review, but need to be performed under strict time and resource constraints, so it may not be possible to review all relevant articles. Therefore we can imagine that rapid reviews would benefit from a ranking of articles by study quality, which is not currently performed. This is explored in Chapter 3.

Objective 3: Reducing the number of reviewers required to assess articles

Typically, two reviewers assess each research article for risk of bias and a consensus decision is made between them. Ideally, we would be able to use a model to predict the risk of bias values perfectly using text mining, such that this could become a completely automated process. However, a more realistic scenario may be that we are able, for a subset of articles included in a review, to determine the value of a risk of bias property with *good enough* certainty, that only one human reviewer is required to assess these articles. In effect, one of the human reviewers is replaced by an automated process, and the time spent reviewing this property would be reduced for this subset of articles. We

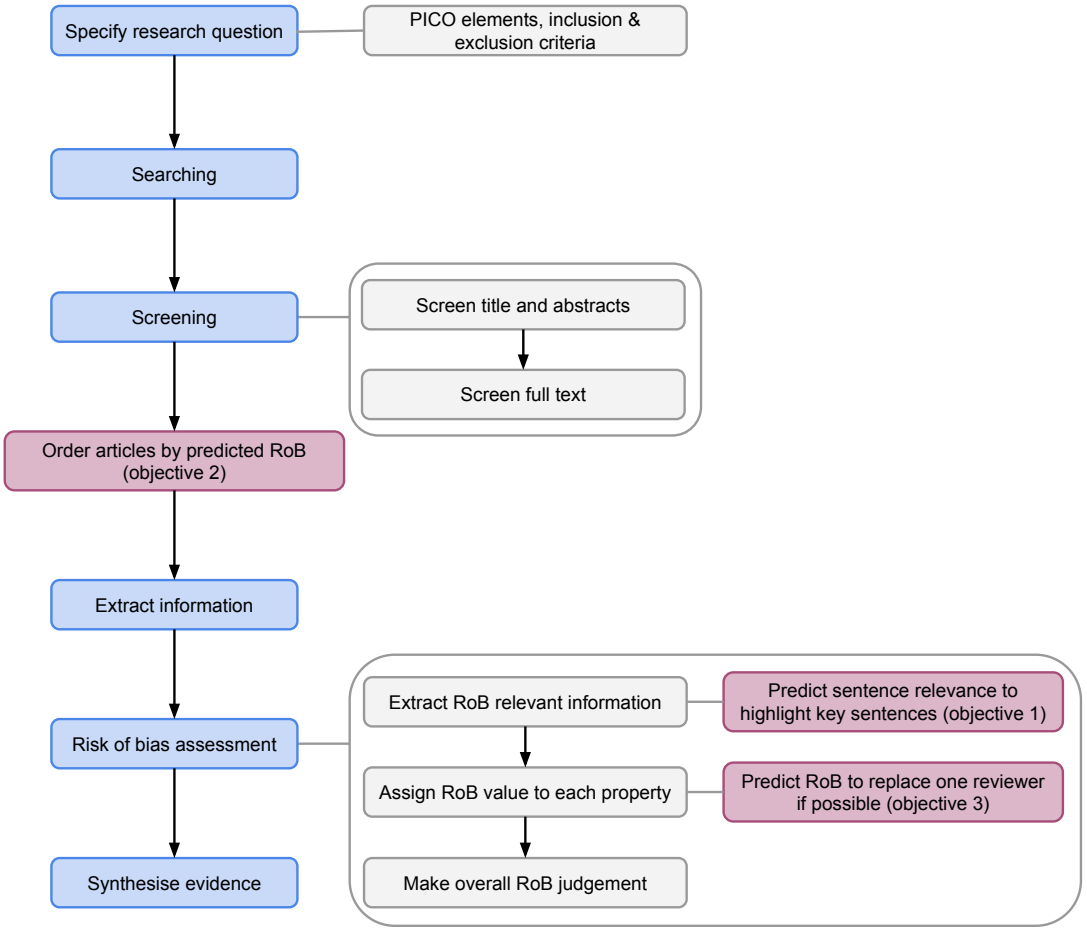


Figure 2.4: Illustration showing how our objectives fit into the systematic review process. Main activities are shown on the left in blue, sub-activities are shown in grey next to each main activity and our three objectives are shown in pink. PICO: population, intervention, control and outcome; RoB: risk of bias.

define *good enough* as articles where the assignment from our model has a probability of being correct that is higher than an estimated probability that a risk of bias judgement by a human reviewer is correct. In short, we try to be at least as reliable as a human reviewer.

This approach results in the following three possible outcomes for each risk of bias property, for a given article: 1) we are at least as certain as a human reviewer that this property has not been adequately carried out, 2) we are not as certain as a human reviewer for either assignment, and 3) we are at least as certain as a human reviewer that

this property has been adequately carried out. Articles in groups 1 and 3 would need to be assessed by only one reviewer, whereas those in group 2 would still need the typical two reviewers.

2.2.3 Related work

There has been much discussion in the literature about the potential to automate systematic reviews [6–9]. In this section we provide an overview of work that has sought to automate the tasks of searching, screening, extracting information, and risk of bias assessments in systematic reviews.

Searching online databases

The search for research articles in a systematic review requires careful construction of search criteria, to ensure all relevant articles are retrieved in the search. Ananiadou et al. [7] suggest improving this task using automatic query expansion, where a user inputs an initial search query and the retrieved research articles are used to generate additional terms to add to the search query. This helps to ensure that research articles are found even where they do not include the specific term specified in the original search query.

Ananiadou et al. [7] also propose using document clustering to assist the search task. Document clustering groups the research articles returned from the search into a set of concepts, to help the reviewer gain a better overview of the main topics of the articles.

Screening research articles

Research into automation for systematic reviews has largely focused on screening articles for inclusion in reviews [10–18, 67–70]. A recent systematic review of the use of text mining for this task identified 44 studies between 2006 and 2014 [10]. Work in this area has sought to automatically classify articles as relevant or not-relevant in order to reduce reviewer workload, for instance by replacing one of the two reviewers performing the screening task, or sought to rank articles by relevance for a particular systematic review in order to prioritise those articles that are most likely to describe a relevant study [10].

Automating screening is an interesting challenge because each review is highly individual such that whether an article is relevant or not depends on the particular protocol of a review. Therefore a unique model is needed to predict article relevance for each review. Furthermore, for new reviews (rather than update reviews) there are no examples of articles known to be relevant at the beginning of the review, with which to train a model. One approach that has been used to overcome this issue is active learning [14–17, 67]. This approach first learns an initial model using a small number of labelled examples, then one example is selected at a time to be manually labelled by the user, which is then added to the training set and the model is retrained. The idea is that by cleverly choosing examples with which to train the model, fewer examples are needed, and hence fewer articles need manually labelling by the reviewer. The examples to be labelled are chosen using a metric denoting how likely an example is to improve the performance of the model. For example, where a learner outputs a score between 0 and 1, a score near 0.5 may be deemed more uncertain than the extremities, and indicates that this example should be supplied to the learner so that the learner can perform better on this and similar examples.

The aim of an automated approach to citation screening is to reduce the amount of time reviewers spend screening citations themselves. The active learning approach just described aims to minimise the number of citations that need manually labelling, which indirectly reduces the amount of time spent manually labelling examples. This can be extended to instead directly reduce the amount of time spent performing the manual labelling. The time taken to review each citation varies due to differences in citation length and difficulty in classifying each as relevant or not relevant. Given two examples that may greatly improve model performance, if the first is expected to take longer to be manually labelled than the second, then the second example should be used. Wallace et al. [67] incorporate the expected time to review each citation into the example selection metric, and conclude that this improves the performance of their system.

Unlike new systematic reviews, when a review is updated there is already an existing set of example research articles available, that were used for the original review. Cohen et al. [13] show how these articles can be used to generate a classifier to predict whether the new articles should be included or excluded from a review. They seek to predict the inclusion decision based on the full text article, from the title and abstract only. Further work by Matwin et al. [11] and Cohen et al. themselves [68, 69] have

investigated the effect of different machine learning algorithms on the performance of automated screening.

There are common criteria that are often used to determine inclusion in a review, and these can be predicted using a supervised machine learning[1] approach. For example, often systematic reviews are restricted to studies of RCTs, but research articles are not adequately annotated in research databases such as PubMed, such that it is not possible to simply specify RCT in the search query. Hence, automatically classifying articles according to whether they are describing a RCT or other study design can assist the citation screening process. Cohen et al. use the noisy Medline RCT tags to train a support vector machine (SVM) to predict whether an article describes an RCT or not [71]. This is useful as the SVM generates scores that provide a measure of confidence rather than a binary classification. They suggest that this can be used to rank articles to prioritise them.

Nim et al [20] seek to classify sentences according to whether they contain information about the PICO elements (population, intervention, control and outcome) in the abstracts of research articles. This has the potential to be used during citation screening, to indicate to the reviewer where key information related to the inclusion criteria is described. They use conditional random fields (CRF) to predict which PICO elements each sentence describes. CRFs make use of label predictions at nearby examples, in this case nearby sentences within the citation, to predict the label for a particular sentence (in addition to the words this sentence contains). This may be effective because abstracts are reasonably structured with respect to the information they provide, often following the structure of: introduction, method, results and then conclusion.

Previous work has also sought to screen articles for scoping reviews [12]. Scoping reviews are more exploratory than standard systematic reviews, with the aim to help determine an appropriate precise research question and inclusion criteria for a subsequent review. Hence, the research question of a scoping review may be more general, and because of this often a large number of articles are returned in the initial search. Also, while it is highly important that all relevant studies are included in a systematic review, for a scoping review this is less important. Gaining a more general overview is more important for scoping reviews. This means that while a recall of 100% is needed during screening of standard systematic reviews, for scoping reviews this can be relaxed.

[1] Supervised machine learning uses a dataset to train (estimate the parameters of) a model.

Screening of scoping reviews needs to maintain a high recall, while also drastically reducing the number of articles in the review by removing those that are not relevant.

Shemilt et al. [12] use three methods to screen abstracts of scoping reviews – two ranking approaches and a classification approach. The first ranking approach is an automated method that uses automated term recognition to identify important terms and assign scores to each term denoting its importance. These terms and scores are then used to score title and abstracts using the scores as weights such that an article containing terms with higher scores are ranked higher. The second ranking approach uses manually curated lists of relevant and not relevant terms to score articles by a ratio of the number of relevant versus non relevant terms it contains. Classification is used to predict whether an article should be included or excluded in the scoping review. Their work was highly successful – of the two scoping reviews for which text mining was assessed by Shemilt et al. [12], workload was reduced by more than 88%. Screening for scoping reviews is an easier task compared to screening for standard systematic reviews, because of the relaxed requirement of perfect recall.

Kirichenko et al. present ExaCT, a tool to assist with the extraction of 21 study properties from full text research articles [19]. The study properties they extract are needed to perform a systematic review, and include: sample size, outcome, intervention and control. This tool performs a two step process. First, sentences containing this information are found using supervised machine learning, where a separate model is used to predict the occurrence of each property. Second, the specific part of these sentences describing each of these properties are identified. This second stage uses regular expressions[2] to find a specific pattern within a sentence for a property. The idea is that by reducing the size of the text from a full text article to just a few sentences that are thought to contain information about a property, a simple pattern matching method can be used to extract the relevant text. For example, a pattern that searches for a date may identify many dates in an article, such as the date the article was published. Given only sentences describing enrolment then any dates here are likely to be enrolment start or end dates.

[2]Regular expression are used to identify string sequences in text. For example, the regular expression $[0-9]^*$ would search for a sequence of numbers.

Risk of bias assessments

We know of only one research group that has investigated automating the prediction of risk of bias properties [21–23]. This work by Marshall et al. has two main aims, 1) to predict whether a trial is at low or not-low (which includes unclear and high) risk of bias and 2) to predict whether each sentence in an article describing a trial are relevant to risk of bias. This is done for the 6 risk of bias domains: sequence generation, allocation concealment, blinding (of participants and personnel), blinding (of outcome assessment) incomplete reporting of outcomes and selective outcome reporting.

The dataset consists of articles describing the results of clinical trials labelled with a risk of bias value for the 6 domains, and with a set of sentences labelled as relevant or not relevant for each risk of bias property. The labelling is derived from the Cochrane Database of Systematic reviews. In fact, the dataset creation approach they take is very similar to our approach, described in the next section. In brief, Marshall et al. use data on 5,400 systematic reviews from the Cochrane risk of bias tool, where each systematic review contains a set of citations corresponding to the studies used in the systematic review. The citations were matched with citations in the PubMed database in order to identify the publications for each study. The PDF articles of each matched citation were retrieved where possible, and each labelled with values for the risk of bias properties as recorded in the Cochrane data.

The Cochrane data also includes justifications of the risk of bias assignments given to each study, and these sometimes include quotations indicating the text that informed the risk of bias value assigned. Quotations were extracted from the justification and each study with a quotation was used in the sentence level dataset. The sentences containing the quotation were labelled as relevant and the sentences not containing the quotation were labelled as irrelevant. A comparison with our dataset construction approach is given in Section 2.3.2.

The article level data differed between their latest work [22] and preliminary work [23]. In their preliminary work [23], Marshall et al. only include the 2,200 articles in their dataset where at least one quotation (across the risk of bias domains) existed in the Cochrane data and the full text PDF article was retrieved. The quotation does not have to appear in the article, for this article to be included in the dataset. Also, all 2,200 articles were used for the article level learning irrespective of which risk of bias

property was quoted for each article. In their most recent work [22] they included all 12,808 articles where both a citation existed in the Cochrane data and where the full text PDF could be found. In both cases, the articles labelled with each risk of bias value may not contain the information used to inform the risk of bias judgement. This is because reviewers may look at multiple sources when assessing the risk of bias of a study.

Marshall et al. take an interesting multi-task learning approach where a single model is used for all 6 risk of bias domains, for each of the two aims. One model is generated to predict risk of bias for a research article (describing a trial) and a second is generated to predict the relevance of each sentence in a research article. Unigram and bigram [3] features are constructed from the article or sentence text for the article and sentence models respectively. In order to predict the 6 domains using a single model a copy of the feature vector (comprising the unigram and bigram features) is included in the model for each domain. The features in each of these vectors are only used to predict the domain to which they correspond. An additional copy of the feature vector is used to learn the predictions jointly across domains, and this is where the relationships of predictions across the domains are captured in the model.

When predicting the risk of bias from article text it may be beneficial to know which sentences contain information relevant to this prediction. Marshall et al. add this information to the article level learning task by supplementing this model with information from the sentence level model. Extra features that denote whether an n-gram [4] feature occurs in a relevant sentence (extracted from the sentence model) are added to the article model.

Performance of the article model was evaluated with accuracy. This assumes that the misclassification cost of each positive example (articles with low risk of bias) is equal to that of each negative example (articles with high/unclear risk of bias). Marshall et al. compared the performance of the article model with a simpler approach where a separate model is trained for each risk of bias domain. Using a multi-task formulation was found to improve the results compared to predicting the domains individually, for all domains except blinding of outcome assessment, although the authors note that this improvement was not significant. The accuracy of human reviewers was also evaluated

[3] Unigram features denote the occurrence of a single word, and bigrams denote the occurrence of a pair of adjacent words.

[4] In general, features denoting the occurrence of a set of n adjacent features are referred to as n-grams. In this case n-gram refers to unigrams (1-grams) and bigrams (2-grams) – the n-grams used in this work.

and this was found to be higher than both the simple and multi-task models.

The sentence model was evaluated in two ways, where the most relevant and three most relevant sentences were selected from an article (for each domain), referred to as top-1 and top-3 respectively. This was then compared against two alternative strategies. The first strategy used the text justifying a risk of bias assignment, stored in the Cochrane risk of bias database. The second strategy selected a random sentence from the article. Reviewers assigned scores to the sentences identified in the top-1 and top-3 approaches, and the two alternative strategies. These scores denoted how relevant the reviewer thought each sentence was to the domain, where the highest score denoted a highly relevant sentence. Their analyses found that the top-1 sentences were assigned the highly relevant score more often than the random approach, but less often than the text from the Cochrane database. The highest sentence score of the top-3 sentences was assigned the highly relevant score more often compared to the text from the Cochrane database, although this difference was not significant. This is an exciting result because it indicates that a very small subset of sentences can be selected and provided to the reviewer as relevant text for a particular risk of bias domain.

2.3 The RoBAL dataset

In order to work towards the three objectives specified in Section 2.2.2, we construct a dataset so that we can perform supervised machine learning to predict the three properties. We call this the RoBAL (Risk of Bias Article Labelling) dataset. RoBAL consists of a set of 1,467 full text articles, each assigned a value for at least one of the three risk of bias properties – sequence generation, allocation concealment and blinding, where the value is supported by some part of the text in the article (summarised in Figure 2.5 and Table 2.3). The risk of bias values are either *low* or *not-low*. A value of *low* denotes that a particular property has a low risk of causing bias. A value of *not-low* denotes that a particular property either has a high risk of causing bias or that the value is unclear (because there is insufficient information in the article to determine the risk of bias). RoBAL also includes a binary label attached to sentences, denoting whether they contain relevant information with respect to a risk of bias property. Each sentence is *relevant*, *not-relevant*, or is unlabelled. We also have the title and abstract text of each research article in our dataset, retrieved from the PubMed database. RoBAL will be

available at http://www.datamining.org.uk (title and abstract only). We now describe the dataset construction process.

2.3.1 RoBAL dataset construction

We constructed RoBAL using data collected from the Cochrane Database of Systematic Reviews, and specifically from the Cochrane risk of bias assessment tool [36]. We used data on 1,399 systematic reviews from this tool, conducted between 2008 and 2011 (with 81, 389, 686 and 243 reviews in 2008, 2009, 2010 and 2011, respectively). This set consists of all intervention reviews that used the new (2008) Cochrane risk of bias assessment tool and reported assessments for at least two domains, up to and including issue 4, 2011. The Cochrane review groups of these reviews are shown in Table 2.1. As already mentioned, this data provides citations and information used to label the research articles. We also collected the full text PDF article from the world wide web.

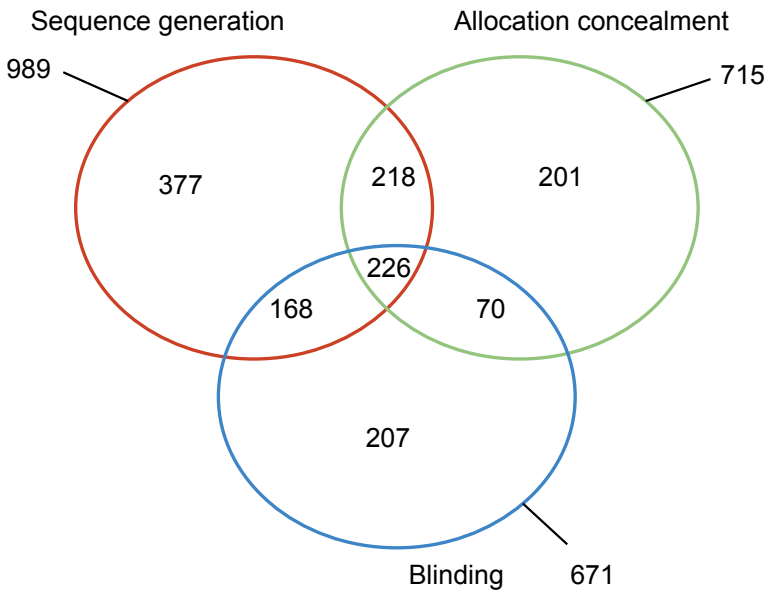


Figure 2.5: Venn diagram of article labelling in RoBAL.

Group name	Number	Group name	Number
Pregnancy and Childbirth (preg)	128	Drugs and Alcohol (addictn)	22
Airways (airways)	89	Effective Practice and Organisation of Care (epoc)	21
Neonatal (neonatal)	86	Colorectal Cancer (coloca)	19
Menstrual Disorders and Subfertility (menstr)	74	Bone, Joint and Muscle Trauma (muskinj)	19
Cystic Fibrosis and Genetic Disorders (cf)	63	Peripheral Vascular Diseases (pvd)	18
Acute Respiratory Infections (ari)	48	Dementia and Cognitive Improvement (dementia)	17
Hepato-Biliary (liver)	48	Skin (skin)	16
Gynaecological Cancer (gynaeca)	45	Depression, Anxiety and Neurosis (depressn)	15
Schizophrenia (schiz)	41	Infectious Diseases (infectn)	15
Neuromuscular (neuromusc)	39	Hypertension (htn)	14
Social, Psychological, and Educational Controlled Trials Register (sympt)	39	Inflammatory Bowel Disease (ibd)	13
Back (back)	31	Metabolic and Endocrine Disorders (endoc)	11
Musculoskeletal (muskel)	31	Fertility Regulation (fertilreg)	11
Wounds (wounds)	36	Incontinence (incont)	11
Eyes and Vision (eyes)	32	Breast Cancer (breastca)	10
Injuries (inj)	31	Lung cancer (lungca)	10
Anaesthesia, Critical and Emergency Care (anaesth)	29	Prostatic Diseases and Urologic Cancers (prostate)	7
Oral Health (oral)	28	Consumers and Communication (commun)	7
Heart (vasc)	26	Multiple Sclerosis (ms)	7
Tobacco Addiction (tobacco)	25	Epilepsy (epilepsy)	6
Ear, Nose and Throat Disorders (ent)	24	Haematological Malignancies (haematol)	6
HIV/AIDS (hiv)	24	Childhood Cancer (childca)	5
Upper Gastrointestinal and Pancreatic Diseases (uppergi)	24	Movement Disorders (movement)	4
Developmental, Psychosocial and Learning Problems (behav)	23	Occupational Safety and Health (occhealth)	2
Stroke (stroke)	23	Public Health (pubhlth)	2
Renal (renal)	22	Sexually Transmitted Diseases (std)	2

Table 2.1: Number of reviews in each Cochrane review group in original data source (and corresponding group code).

Figure 2.7 shows the data flow from the Cochrane Database of Systematic Reviews to the articles included in our dataset, and Figure 2.6 shows the number of reviews or articles at each stage. The 1,399 systematic reviews in the Cochrane database contain 18,167 references to research articles describing the clinical trials in these reviews. From this set of research articles our dataset creation process described below results in a set of 1,467 research articles.

To construct RoBAL there are three main steps: 1) Collating the data from the Cochrane risk of bias tool, 2) collecting full text PDF articles from the world wide web, and 3) labelling research articles with risk of bias values, and their constituent sentences with relevance labels. A flow diagram illustrating the data construction process is given in Figure 2.7.

Collating Cochrane risk of bias data

The Cochrane risk of bias data consists of a set of systematic reviews, each with a set of citations of articles that reviewers assessed for each clinical trial in the review. Each trial also has a *low*, *high* or *unclear* value assigned to each risk of bias property for each outcome in the trial. A value of *low* for blinding, for instance, means that blinding was adequately performed in this study such that the risk of bias is low. A value of *unclear* indicates that there was insufficient information in the article to determine the risk of bias. These judgements are supported by text descriptions, often including direct quotations from articles or a comment stating that no information was found in the article.

The risk of bias labels are entered by hand into the Cochrane risk of bias tool, and

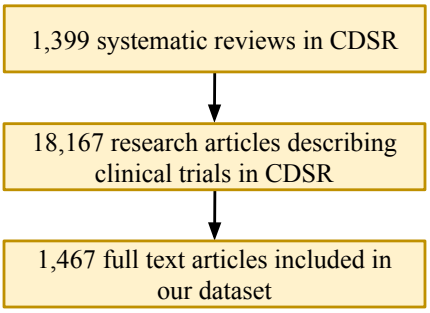


Figure 2.6: Data flow diagram.

hence vary across systematic reviews. However, most of the labels were one of a small set, and we mapped these to our three risk of bias properties as follows. The labels 'Blinding?' and 'Blinding' were mapped to the blinding property, 'Adequate sequence generation?', 'Adequate sequence generation', 'Sequence generation?' and 'Sequence generation' were mapped to sequence generation, and 'Allocation concealment?', 'Allocation concealment' and 'Method of allocation Concealment' were mapped to allocation concealment.

In the case where a trial reported multiple values for a given risk of bias property

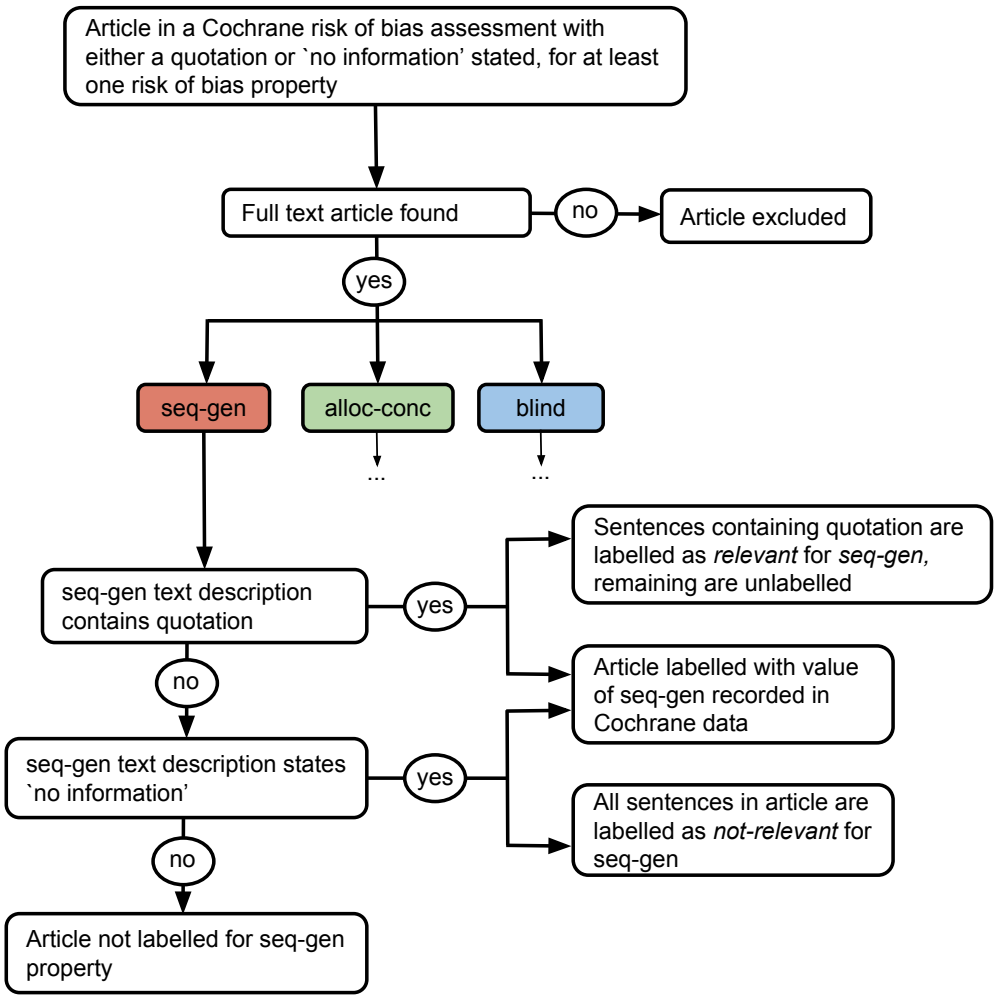


Figure 2.7: Dataset construction flow diagram.

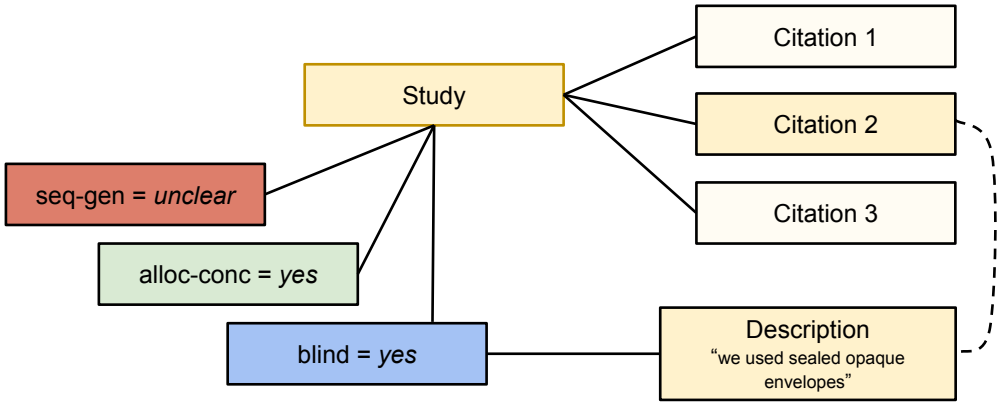


Figure 2.8: Data from the Cochrane risk of bias tool. Dotted line represents a relationship that we need to infer. Citation 2 is included in our dataset whereas citations 1 and 3 are excluded because no link with a risk of bias property could be inferred.

(because the trial assesses more than one outcome) then the first value encountered when processing the Cochrane risk of bias data files was used. In the set of systematic reviews we used from the Cochrane risk of bias tool, this was only ever the case for the blinding property.

The Cochrane data do not specify which articles contain the information that informed the risk of bias judgement. We use the text descriptions to infer this. First, articles containing quoted text contain information used to make the judgement. For instance, Figure 2.8 shows an example where a study has a quotation for the blinding property, which is found in the article content of citation 2. We then infer that the blinding judgement was made using information in this particular article. Articles where 'no information' was stated do not contain any information, and we can infer that the lack of information is the reason for this choice of label value. For instance, an article may have the label *unclear* for blinding and 'no information' in the text description because all research articles cited for this study in this review have been found to contain no relevant text, such that an assignment for the property value to *low* or *high* could not be given. We only include articles in RoBAL when either 'no information' is stated, or a quotation is found in the article text.

We combine the *high* and *unclear* labels in the Cochrane data to give a binary variable with values *low* and *not-low*. We justify this on the basis that a reviewer generally wants to identify the high quality studies, such that the articles of *high* and *unclear*

risk of bias can be grouped together. This is also beneficial to maximise the number of examples in each class which is reduced as the number of classes increases.

We use the quotations and 'no information' statements to label sentences as *relevant* or *not-relevant*. A sentence is *relevant* if it contains a quotation supplied for this article in the Cochrane risk of bias assessment. A sentence is *not-relevant* if it is within an article associated with a study where 'no information' was stated. Otherwise, a sentence is unlabelled. We cannot determine the label of the unlabelled subset because when a reviewer provides a quotation during a risk of bias assessment, they are likely to choose only exemplary text rather than to include quotations for all relevant text in an article. Hence, these unlabelled sentences may contain relevant information.

We use two regular expressions to extract quotations from the description field, which identify single and double quotes, respectively. We also searched for quotations where a reviewer had skipped some of the text inside a quote, using the '. . . ' notation. We dealt with this by treating the text either side of the '. . . ' as separate quotations, such that, when labelling using these two quotations the same sentence would be labelled.

Collecting full text PDF research articles

Due to the time-consuming nature of creating a dataset by hand, we instead use an automated process. This process retrieves the full text PDF articles from online sources, for citations given in the Cochrane risk of bias assessments. Figure 2.9 gives an overview of the sources of information we use to collect the PDF articles. The PubMed database is useful because it contains links to full text research articles. We use this database in a two-stage process.

The data from the first source, the Cochrane risk of bias tool, has been described above. For each citation retrieved from the Cochrane data we search the PubMed database for a matching citation, in order to find the PubMed identifier of this refer-

Rule	Exact match	Partial match
Rule 1	title	
Rule 2	page start, page end	title
Rule 3	page start, page end, volume, issue	

Table 2.2: Matching Cochrane citations to PubMed citations.

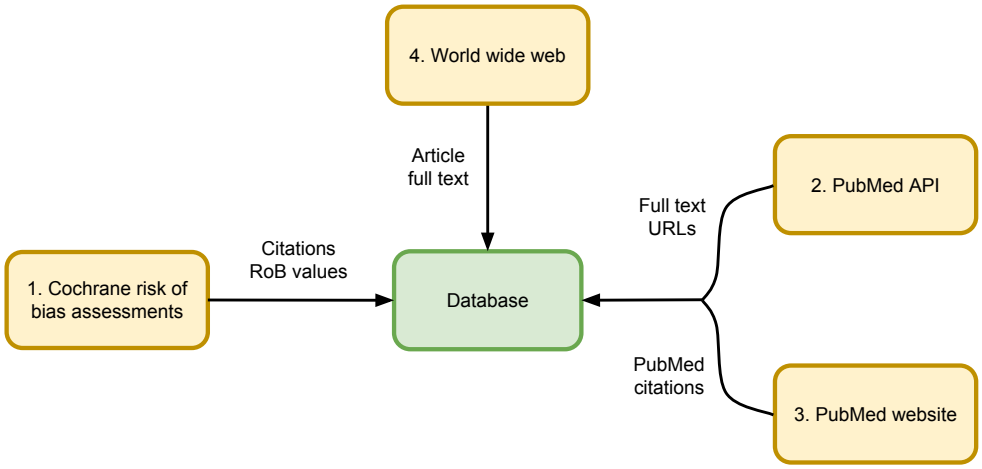


Figure 2.9: Online data sources for collecting PDF research articles. First the citations and labels are retrieved from the Cochrane data files. PubMed is then used to find the PubMed ID for each citation by matching the Cochrane citation with a PubMed citation using the PubMed API. The PubMed website is then scraped to identify possible URLs which may contain or link to the full text. We then use these URLs as starting points to search the world wide web for the full text PDF articles.

ence. This is done using the PubMed application programming interface (API)[5]. Our initial search uses the title of the research articles only, with the additional search query constraints that 1) the publication type is not a letter and not a comment and 2) the language is English. This may return several articles, which we compare to the original citation details of the Cochrane risk of bias tool to find one corresponding to this reference. We use a series of rules to determine whether a match has been found, shown in Table 2.2. Each rule assesses the similarity of particular fields and uses either exact matching or partial matching (or a combination of both) to compare the citation details from the Cochrane data with the results from PubMed. Partial string matching is required because the data in the Cochrane tool is entered by hand and hence is often noisy. Where only partial string matching is used we require tougher constraints on other fields. For example, a match is found if the titles match exactly, but if there is only a partial match between titles then we require the start and end page numbers to match also.

[5]This API allows the PubMed database to be queried using a computer programme rather than only being able to access this data through loading webpages on the PubMed website.

Partial string matching was performed with the Smith-Waterman algorithm [72]. This algorithm performs local alignment string matching, where the best match of a given string, s_1, is found within a potentially much longer string, s_2. This contrasts to global alignment methods that try to match two strings entirely. The Smith-Waterman algorithm allows for a degree of noise in the text, rather than requiring that s_1 is exactly matched within s_2. Two types of noise are allowed, a mismatch or a gap. Prior to running the algorithm an award is set for a character match, and a penalty is set for a mismatch and gap, which tend to be positive and negative values respectively. The algorithm uses these to determine which alignment is preferred. For instance, if a gap incurs a high penalty, these would not occur in the alignment unless the benefits of using a gap outweighed this penalty (because a large subsequent sequence of characters were then successfully aligned). Also, we can calculate an alignment score denoting the degree to which s_1 was locally aligned within s_2, the sum of the awards and penalties for a given local alignment. The following example compares the string 'double-blind' with 'double b ind'. We use awards of 1 for a match and -1 for both a mismatch and gap.

s_1			d	o	u	b	l	e		b		i	n	d	
s_2	...		d	o	u	b	l	e	-	b	l	i	n	d	...

This comparison has 10 character matches, one mismatch (- versus a space) and one gap where s_1 is missing the character *l*. We can see how adding the gap improves the match overall because a gap incurs the award of -1, but this aligns the 'ind' at the end of the strings which has an award of 3. The score for this alignment is $10-2=8$.

In this work we use the same awards as the example above, assigning 1 for a match and -1 for both a mismatch and gap. To determine whether s_1 is sufficiently matched in s_2 we use a relative threshold of the alignment score compared to the total possible alignment score for a given s_1. The total possible score is the length of s_1 and we use a threshold of 0.8, such that an 80% match is sufficient to determine that $s1$ has been found in s_2. These settings are the same for all further tasks using partial string matching.

After citations have been identified in PubMed we use the PubMed article pages (such as http://www.ncbi.nlm.nih.gov/pubmed/24071462) to provide a set of links to

the full text. We use the following recursive method where each page linked to from PubMed is parsed to attempt to find the full text PDF article. First we attempt to parse the page as a PDF document. If this is successful then this is likely to be the PDF of the research article. If this is not successful then we process the page to find HTML links, which may link to the PDF, or to a page that itself links to the PDF. We find links with a URL that either contains the word 'PDF' or 'full', or a link whose text is similar to the title of the research article, using partial string matching. We limit this search to a link depth of three (including the direct link from the PubMed website).

After retrieval of the PDF articles the full text content is extracted using the adobe PDFbox text extraction tool (version 1.8.6). We check that this research article corresponds to the citation by attempting to find the title and abstract in the text of this PDF. This is a non-trivial task because text extracted from PDF documents is inherently noisy. A text extractor (such as the Apache PDFBox[6] extractor that we have used) may be unable to recognise columns so that two columns are extracted as a single column. Characters may be read incorrectly or marks on the page may be passed as characters. The text in page headers and footers, and the data from tables and figures, may be included into the main body of text. We therefore check that the title and abstract are found in the PDF text using partial string matching as described above. For articles where a quotation is assigned we also check that the quotation is found in the article, also using partial string matching. If the quotation could not be found then this article is not included in the dataset.

Labelling research articles

We segment the article text into sentences using the PTBTokenizer of the Stanford CoreNLP Java package (version 3.4.1). Sentences in each article with a quotation are parsed to identify those containing the quotations attached to each study (again using partial string matching). We save each article in a Javascript object notation (JSON) file, with the full text content of the PDF article stored as an array of sentences. Each quoted sentence has a quotation attribute that relates this sentence to the specific quotation it contains. Each JSON file also contains the citation data we retrieved from PubMed including the PubMed title and abstract, and the risk of bias property values.

[6]Apache PDFBox available at http://pdfbox.apache.org.

Property	Property value	Number	Proportion
BLIND	YES	7,120	0.415
	NO	4,705	0.275
	UNCLEAR	5,316	0.310
SEQ_GEN	YES	6,846	0.408
	NO	679	0.041
	UNCLEAR	9,263	0.552
ALLOC_CONC	YES	6,057	0.338
	NO	1,342	0.075
	UNCLEAR	10,498	0.587

Table 2.3: Number of studies with a value of each property, value pair in our original dataset (18,167 studies).

2.3.2 Comparison with related work

The approaches used by ourselves and Marshall et al. to generate a labelled dataset are very similar [21–23]. Both use data from the Cochrane risk of bias tool to provide labelled data, use full text articles extracted from PDF articles found on the world wide web, and use an automated procedure to identify quotations in the articles. Marshall et al. used data on 5,400 systematic reviews from the Cochrane data whereas we used data on 1,399 systematic reviews.

The article level dataset of preliminary work by Marshall et al. used only the 2,200 articles where a quotation was attached to at least one risk of bias property and the PDF article could be retrieved [23]. The latest work by Marshall et al. used the 12,808 articles where a PDF could be retrieved [22]. In contrast, we only include circa 1,500 articles where we have evidence that the risk of bias assignment was made using a particular article, where either a quotation was found in the article text or a reviewer stated that no information could be found. This meant that our dataset contained different articles for each risk of bias property depending on which quotations were supplied and identified in the article text for each property. Requiring quotations to be identified in the article text caused many to be excluded for two reasons. First, an article cited in the Cochrane data may not have been the source of the quotation as it is common for reviewers to look at many sources. Second, due to the noisiness of PDF extractions, it may not be possible to locate the quotation in the extracted text (using the automated process we used).

Furthermore, to ensure that the article corresponds to the citation we check that the title and abstract from the PubMed database is located in the article text. We use partial string matching to identify text within articles in order to allow for noise created during the PDF text extraction, whereas Marshall et al. use exact string matching. In addition to using the full text, we also collect the title and abstracts from PubMed, whereas Marshall et al. focus their work solely on making predictions using the full text. Our sentence labelling uses sentences known to have a *not-relevant* label (because the article has been described as containing no information), where as Marshall et al. use unlabelled sentences as the *not-relevant* examples.

While we focus on three risk of bias domains (sequence generation, allocation concealment and blinding), Marshall et al. look at 6 risk of bias properties, using blinding of outcome assessor and blinding of participants and personnel as separate domains and additionally investigating incomplete reporting of outcomes and selective outcome reporting.

2.4 Machine learning methods

In Chapters 3 to 5 we use machine learning methods to work towards the objectives stated above. In this section we give an overview of these methods and our notation.

2.4.1 Learning predictive models

Text mining is an established field where predictions are made from text data. We take a supervised machine learning approach, where we use a dataset (described in Section 2.3) to train models in order to make predictions for new, unseen examples. In this thesis we use common machine learning terminology. The dataset is composed of a set of examples, also referred to as instances. We refer to the parameters of a model (also known as the independent variables) as features. The variable being predicted (also known as the dependent variable) is referred to as the class or label. Hence each instance consists of a set of features and a label value. Commonly the features and label are referred to as the input and outputs of the model, respectively. The terms ‘learn’ and ‘train’ are used interchangeably and refer to the estimation of a models parameters using any given algorithm.

An algorithm may learn a simple linear model, commonly used in epidemiology, such as logistic regression (as used in Chapter 5). Other algorithms may learn models less commonly used in this field, including decision trees and naive Bayes (as used in Chapter 3). Each algorithm differs in the set of assumptions of the models it learns, referred to as the model's inductive bias. For instance, the inductive bias of logistic regression is the assumption of linearity – that the label is inferred from a linear combination of the features.

We use two types of models to make predictions from the article text. The first, used for objective 1, predicts the relevance of each sentence of an article using the words it contains. The second, used for objectives 2 and 3, predicts the risk of bias of a study from the words contained in an article. We call the models used to predict sentence relevance and article risk of bias sentence models and article models respectively. In line with the domain-based nature of a risk of bias assessment, we generate separate models for each risk of bias property: sequence generation, allocation concealment and blinding.

A flow diagram illustrating the process from PDF article to model predictions is shown in Figure 2.10. The text is extracted from the PDF article and this is segmented into sentences. Each article model takes the article text as input, in the form of a 'bag of words' representation. In this representation each variable is the number of times a word occurs in the article. Each sentence model takes the text of a single sentence as input, again using a bag of words representation, where each feature is the number of times a word appears in the sentence.

As is common practice in machine learning, we evaluate model performance using 10-fold cross validation, where the data is split into 10 equal sized parts called folds. A model is then trained using 9 folds (90% of the data) and tested on the remaining fold (10% of the data), and this is repeated 10 times with different 90/10 splits of the dataset. We stratify the labels across folds such that each fold contains approximately equal numbers of examples of each class. Cross validation avoids overoptimistic estimates of performance, which can arise when the model is trained and tested on the same data. We evaluate the performance on the test sets using ROC analysis, which involves averaging the curves of the individual folds to generate a single ROC curve. This is discussed in the next section.

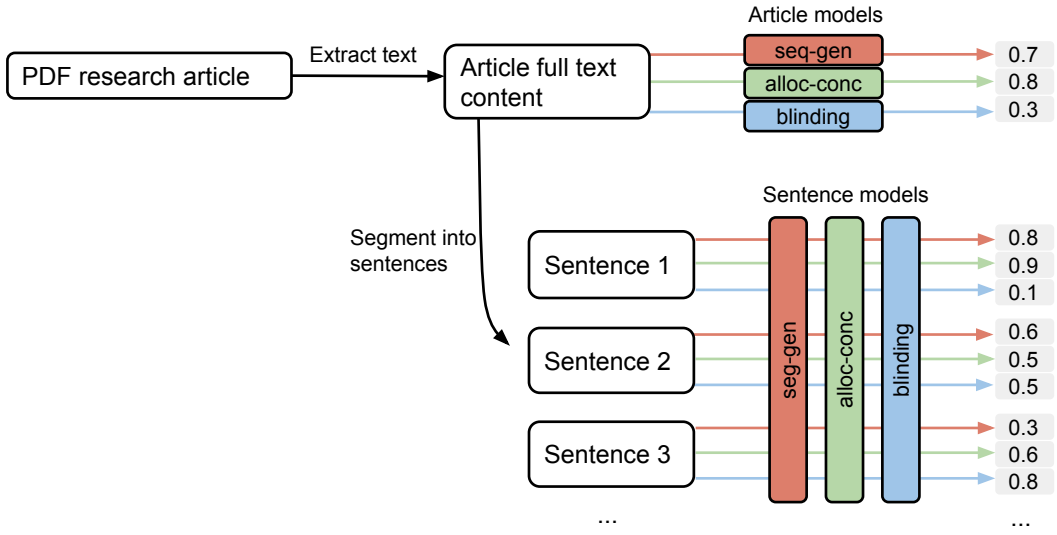


Figure 2.10: Flow diagram from research article to model predictions.

2.4.2 Evaluating models with ROC analysis

ROC analysis is used to evaluate the ranking or classification performance of a machine learning model. Many types of machine learning models are first and foremost ranking methods. For example, a logistic regression model assigns a score between zero and one to each example, which can then be used to rank these examples in order of descending score. An example ranking (given labels 0 and 1 where 0 denotes the positive class and 1 the negative class, and where a lower score predicts the example is more likely to be positive) is:

Label	0	0	1	0	0	1	1	1	0	1
Score	0.05	0.1	0.15	0.2	0.55	0.7	0.75	0.8	0.85	0.9

Ranking methods can be used as a classifier by specifying a classification threshold at some point along the ranking. This classification threshold says that all examples before this point on the ranking should be predicted as positive and all those after should be predicted as negative. For example, the vertical lines in the ranking below depict two example thresholds, that could be used to classify the examples.

Label	0	0	1	0	0	1	1	1	0	1
Score	0.05	0.1	0.15	0.2	0.55	0.7	0.75	0.8	0.85	0.9
	0.0				0.5					

The threshold at $score = 0.0$ classifies all examples as negative, whereas the threshold at $score = 0.5$ classifies 4 as positive (3 positives and 1 negative) and 6 as negative (2 positives and 4 negatives). Each threshold point on a ranking has a corresponding contingency table, that contains the number of positive examples correctly and incorrectly classified as positive and negative, and the number of negative examples correctly and incorrectly classified and negative and positive. Respectively, the contingency tables for the 0.0 and 0.5 thresholds shown above are:

		Predicted label	
		0	1
Actual label	0	0	5
	1	0	5

		Predicted label	
		0	1
Actual label	0	3	2
	1	1	4

Receiver operating characteristic (ROC) curves show both the ranking and classification performance of ranking models. These plots are used to assess the performance of the models visually [73]. A ROC curve is a plot of true positive rate on the y-axis against false positive rate on the x-axis. The false positive rate (equivalent to $1-$specificity), is the number of negative examples incorrectly classified as positive. The true positive rate (also called recall or sensitivity), is the number of positive examples correctly classified as positive.

The ROC curve of our example ranking is given in Figure 2.11. Each point on the ROC curve represents a position in the ranking, and a potential classification threshold. For example, the points of the two thresholds given above are at (0, 0) and (0.2, 0.6) on the ROC curve. The first classifies no examples as positive, hence the true positive and false positive rates are both zero. The second classifies 4 of the 5 positive examples as positive giving a true positive rate of 0.8, and 2 of the 5 negative examples as positive giving a false positive rate of 0.4.

The set of examples used in the above ranking (and ROC curve) can be thought of as a sample from a larger population of examples. A ROC curve generated from a sample is known as an empirical ROC curve. We can also generate a ROC curve for the population, referred to as an analytical ROC curve. In fact, while an empirical ROC curve can be generated by sampling a set of examples from the population and ranking them, as we show in Chapter 4, empirical ROC curves can also be generated by sampling the analytical ROC curve. An example analytical curve is given in Figure 2.12. While an empirical curve usually has linear segments that may be vertical or horizontal, an analytical curve tends to be smoother. The reason for this becomes clear in the next section.

Metrics in ROC space

The true positive and false positive rates just discussed are two metrics that are conveyed in ROC space. There are several other metrics that are also shown in ROC space. Our work involves binary classification tasks where we are predicting *relevant* or *not-relevant* for sentence examples, or *low* or *not-low* risk of bias for article examples. We

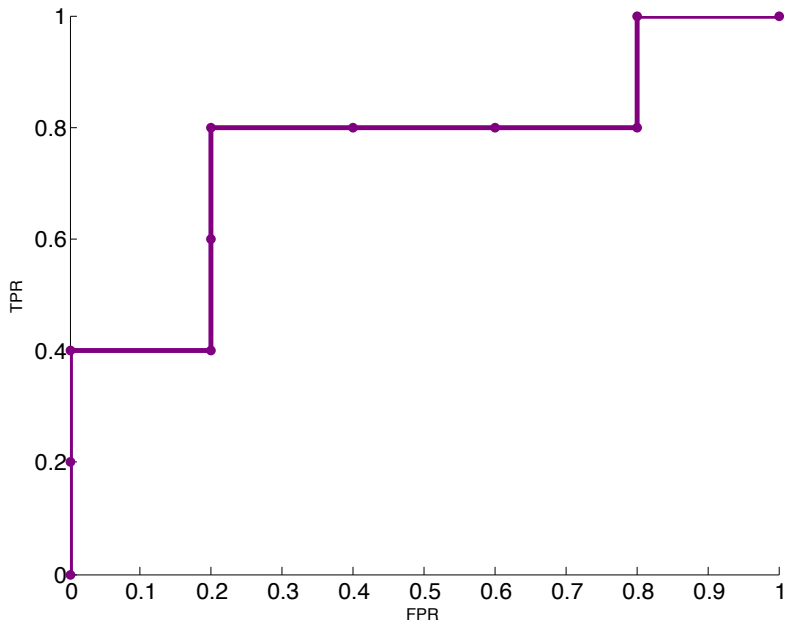


Figure 2.11: Example empirical ROC curve.

now specify this problem setting formally, before introducing a set of key metrics relevant to the work in this thesis, together with our notation (consistent with the notation of [74]).

We assume a two-class classification problem with instance space $\mathscr{X}$ [7]. The positive and negative classes are denoted by 0 and 1, respectively. We fix the *relevant* sentence label and *low* article label as the positive classes. We make this choice because these are the labels we are interested in identifying. The learner outputs a score $s(x) \in [0,1]$ for each instance $x \in \mathscr{X}$, such that lower scores express a stronger belief that x belongs to class 0. To be clear, the positive class is denoted by 0 and a lower score (nearer to 0) denotes an example is more likely to be positive.

The score probability distributions and cumulative probability distributions are denoted by f_k and F_k for class $k \in \{0,1\}$. When we refer to the analytical case f_k is a probability density function, and when we refer to the empirical case f_k is a probability mass function. Given a threshold at score t the true positive rate (TPR), also called sen-

[7] The instance space is the input to the model – the set of features used to predict the label.

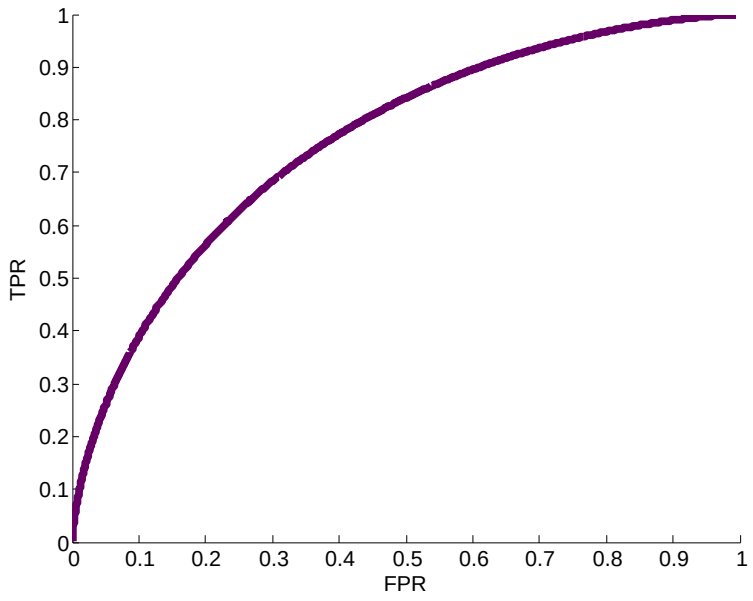


Figure 2.12: Example analytical ROC curve.

sitivity or positive recall, is $P(s(x) \leqslant t|k=0) = F_0(t)$ and the false positive rate (FPR) is $P(s(x) \leqslant t|k=1) = F_1(t)$. The true negative rate, also called specificity or negative recall, is $1 - F_1(t)$. $F_0(t)$ and $F_1(t)$ are monotonically non-decreasing with increasing t, and this has some notational advantages. In the empirical case the true and false positive rates can be given in terms of the number of examples:

$$F_0(t) = \frac{TP_t}{TP_t + FN_t} \tag{2.1}$$

$$F_1(t) = \frac{FP_t}{TN_t + FP_t} \tag{2.2}$$

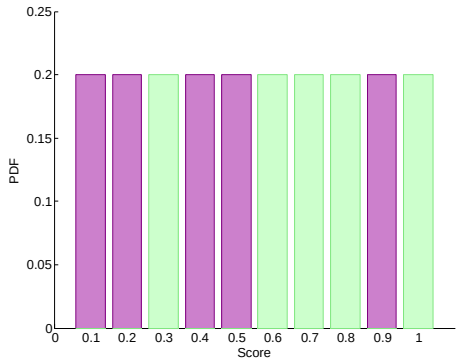
where TP_t and FP_t are the number of positives correctly classified as positive, and negatives incorrectly classified as positive, respectively, and TN_t and FN_t are the number of negatives correctly classified as negative, and positives incorrectly classified as negative, respectively. These values correspond to the four cells of a contingency table:

		Predicted label	
		0	1
Actual label	0	TP_t	FN_t
	1	FP_t	TN_t

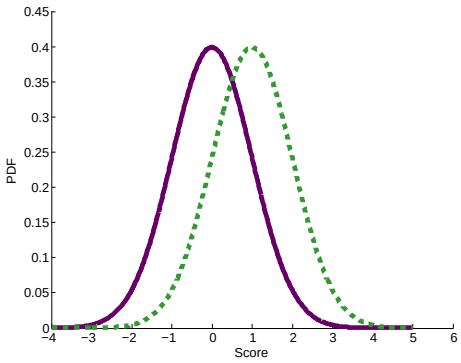
We can see from the definitions of the true and false positive rates that each ROC curve can be represented as two score distribution functions, one for each class. To be more precise, when the ROC curve is empirical the probability distribution is a histogram and when the ROC curve is analytical the probability distribution is a probability density function. Example probability mass and probability density functions for our example empirical and analytical ROC curves respectively are given in Figures 2.13a and 2.13b. These are only examples, as many score densities produce the same ROC curve, because the scores are not fixed for a ROC curve – it is the relative densities at each score and the order of these across scores that matters. For example, in Figure 2.13a, we could imaging shifting the first bar at score 0.1 to score 0. This would not change the ROC curve because the relative ordering of each bar in the ROC curve is the same. This is equivalent to the fact that for a ranking of examples of an empirical ranking it is the relative scores that determine the ranking rather than the actual score

values. This idea is used in Chapter 4 where we introduce a method called rate-oriented sampling.

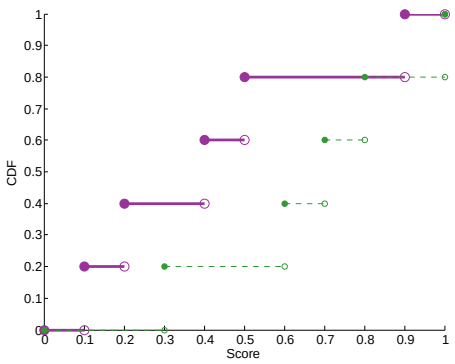
Example cumulative distribution functions are shown in Figures 2.13c and 2.13d for our empirical and analytical ROC curves of Figures 2.11 and 2.12. Since the probability



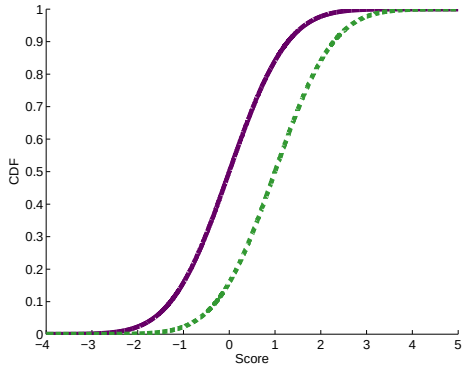
(a) Probability mass function for empirical ROC curve of Figure 2.11.



(b) Probability density function for analytical ROC curve of Figure 2.12. Normal distributions with $\mu = 0, \sigma = 1$ for positive class (purple solid) and $\mu = 1, \sigma = 1$ for negative class (green dashed).



(c) Cumulative distribution function for empirical ROC curve of Figure 2.11 and probability mass function of Figure 2.13a.



(d) Cumulative distribution function for analytical ROC curve of Figure 2.12 and PDF of Figure 2.13b.

Figure 2.13: Example probability distribution functions and cumulative distribution functions. Purple corresponds to the positive class and green corresponds to negative class.

distribution of the analytical ROC curve is a continuous probability density function, the CDF is also continuous. In contrast, as the probability distribution of an empirical curve is a discrete probability mass function, the CDF is right continuous with left limits (shown in Figure 2.12 as a series of horizontal segments with filled and hollow circles at the left and right ends, respectively). This occurs because when increasing the score from zero to one the CDF only changes when we reach a score that has a non-zero probability, jumping by the probability at this score.

While a ROC curve can be inferred using just the class score densities, many metrics cannot be inferred from these densities alone. We also need to know the proportion of examples in each class, known as the class distribution. We denote the proportion of positives and negatives by π_0 and π_1 respectively. The number of positive and negative examples are denoted n_0 and n_1 respectively, and the total number of examples is denoted n, such that $n = n_0 + n_1$, $\pi_0 = n_0/n$ and $\pi_1 = n_1/n$. A dataset with an equal number of examples in each class ($\pi_0 = \pi_1 = 0.5$) has a uniform class distribution. We now define three key metrics, the predicted positive rate, accuracy, and the area under the ROC curve (AUC). The predicted positive rate and accuracy are measures of classification performance, whereas the AUC is a measure of ranking performance.

The score probability distribution of the mixed distribution is denoted by f and given by:

$$f(t) = \pi_0 \cdot f_0(t) + \pi_1 \cdot f_1(t) \tag{2.3}$$

The cumulative distribution of the mixed probability distribution is denoted by F and given by:

$$F(t) = \pi_0 \cdot F_0(t) + \pi_1 \cdot F_1(t) \tag{2.4}$$

Again, the mixed probability distribution is a probability density function in the analytical case and a probability mass function in the empirical case. The mixed cumulative distribution is continuous in the analytical case and right continuous with left limits in the empirical case.

The mixed cumulative distribution is also the proportion of positive predictions at threshold t known as the predicted positive rate, which we abbreviate to the rate. In the discrete case, the rate can also be given by:

$$r(t) = \frac{TP_t + FP_t}{TP_t + FP_t + TN_t + FN_t} \tag{2.5}$$

Accuracy *acc* at threshold t is the proportion of examples that have been correctly classified, and can be formulated as a weighted average of positive and negative recall, weighted by the class distribution:

$$acc(t) = \pi_0 F_0(t) + \pi_1(1 - F_1(t)) \tag{2.6}$$

In the discrete case accuracy can also be given by:

$$acc(t) = \frac{TP_t + TN_t}{TP_t + FP_t + TN_t + FN_t} \tag{2.7}$$

The area under the ROC curve (AUC) is the true positive rate averaged over all false positive rates:

$$AUC = \int_0^1 F_0 dF_1 = \int_{-\infty}^{+\infty} F_0(t) f_1(t) dt \tag{2.8}$$

The following table gives the values of the four classification metrics we have introduced, for our example ranking when the threshold is set to each position along the ranking. TPR and FPR are the true and false positive rates, respectively, and rate is the predicted positive rate.

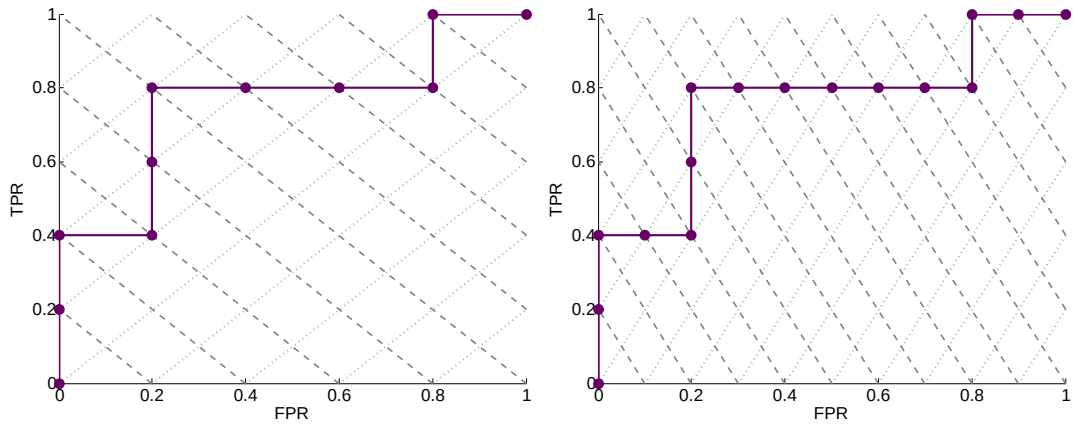
Label	0	0	1	0	0	1	1	1	0	1
Score	0.05	0.1	0.15	0.2	0.55	0.7	0.75	0.8	0.85	0.9
TPR	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	1	1
FPR	0	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	1
Accuracy	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{7}{10}$	$\frac{6}{10}$	$\frac{5}{10}$	$\frac{6}{10}$	$\frac{5}{10}$
Rate	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{5}{10}$	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{9}{10}$	1

We note how ROC curves are not sensitive to changes in class distribution because the true positive and false positive rate metrics each pertain to a single class. Hence the AUC is also not sensitive to changes in class distribution. Accuracy and rate, however, are sensitive to changes in class distribution. This can be seen using isometrics in ROC space.

Isometrics

Metrics such as the rate can be depicted in ROC space using isometrics – points on ROC space that have the same value for a given metric [75]. For example FPR isometrics are vertical lines, and TPR isometrics are horizontal lines. The predicted positive rate and accuracy also have isometrics in ROC space. Examples of these isometrics are shown in Figure 2.14. The isometrics shown in Figure 2.14a correspond to a dataset with a uniform class distribution, whereas those in Figure 2.14b correspond to a dataset with twice the number of negatives compared to positives. Rate isometrics have slope $\frac{-\pi_1}{\pi_0}$ and accuracy isometrics have slope $\frac{\pi_1}{\pi_0}$, such that they become steeper as the proportion of negatives increases, as shown in Figure 2.14. Accuracy and rate isometrics have gradients 1 and -1 respectively, when the class distribution is uniform.

The dependence on the class distribution can be seen intuitively as follows. Each example in a ranking has equal weight, irrespective of the class, for both accuracy and rate. If there are twice the number of negatives to positives, then half the number of negatives have the same weight as all the positives. For accuracy, classifying half the negatives correctly and no positives has the same worth (the same accuracy) as classifying all the positives correctly and no negatives. For rates, the same number of examples are classified as true when classifying half the negatives as positive and no positives as positive, compared to classifying all positives as positive and no negatives as positive.



(a) Uniform class distribution (b) Double the number of negative to positives

Figure 2.14: Accuracy (green dotted) and rate (gray dashed) isometrics in ROC space.

Consensus curves

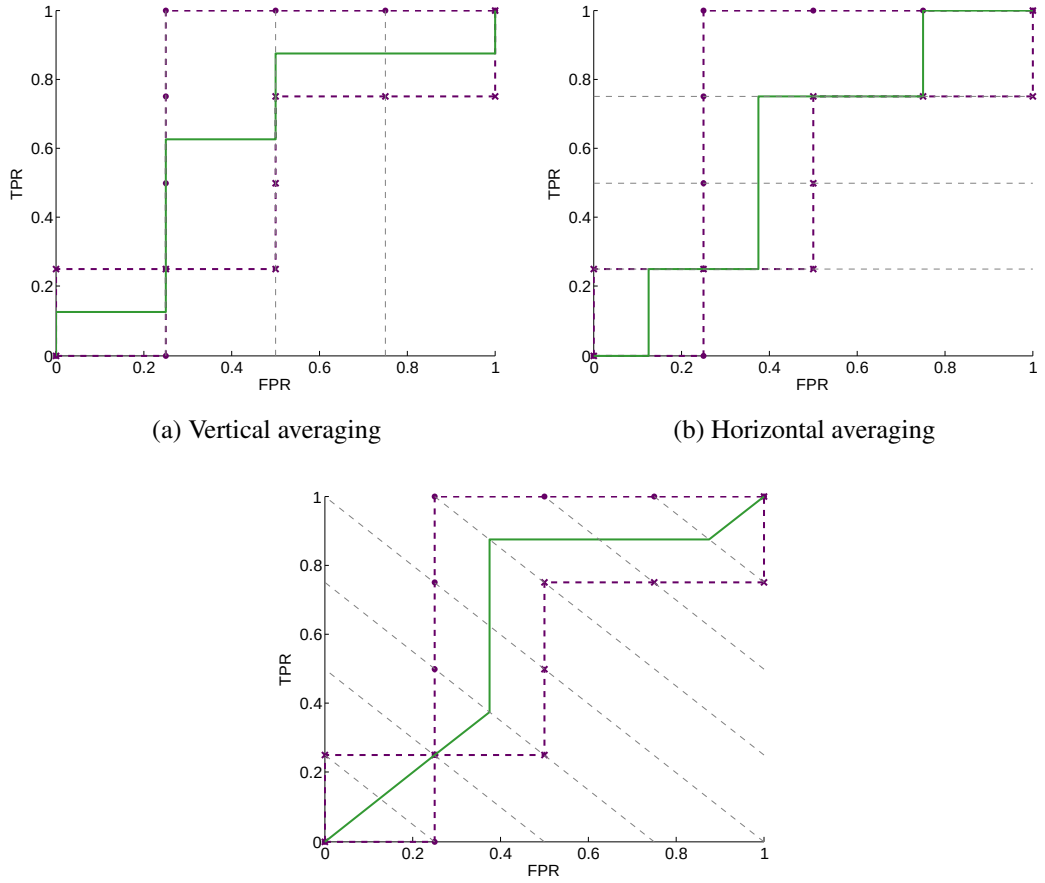
In general, when several sample ROC curves are generated, such as with *m*-fold cross validation, they can be used to produce an average ROC curve, called a consensus curve [73]. Common approaches include vertical and horizontal averaging. Vertical averaging calculates the average true positive rate while fixing the false positive rate (Figure 2.15a). Horizontal averaging is a similar approach that instead calculates the average false positive rate while fixing the true positive rate (Figure 2.15b). Another approach we call *rate-averaging*, also referred to as pooling [76], where the average of the true and false positive rates at each rate are calculated and then used to generate a single curve (Figure 2.15c).

We also note that consensus curves are also generated for meta-analyses of diagnostic test accuracy studies. We discuss these together with a discussion of confidence bounds in this context, in Section 4.7.

2.5 Summary

In this chapter we have summarised the systematic review process, and the issue of bias due to the methodological quality of the clinical trials. We have introduced our key objectives of: 1) identifying relevant sentences within research articles, 2) ranking articles by risk of bias and 3) reducing the number of assessments the reviewers need to perform by hand. We have reviewed related work that has used text mining to assist or automate systematic reviews. We have described our dataset and the method we used to create it. Lastly, we have introduced key methods we use in subsequent chapters to train and evaluate machine learning models.

Chapters 3 to 5 address our specified objectives. Firstly, in Chapter 3 we present a novel metric to evaluate ranking models with particular constraints, and apply this to the task of ranking articles for rapid reviews.



(a) Vertical averaging

(b) Horizontal averaging

(c) Rate-averaging (assuming uniform class distribution)

Figure 2.15: Existing approaches to generating consensus curves. Two ROC curves (dashed purple) are averaged to create a single consensus curve (solid green).

Chapter 3

Rate-constrained ranking for rapid reviews

In this chapter we show how rapid reviews can be formulated as a task with particular constraints, which we call rate-constrained ranking tasks. We introduce the rate-weighted AUC (rAUC), a metric to evaluate the performance of ranking models for rate-constrained ranking tasks. Much of the work presented in this chapter is published in [27].

3.1 Rapid reviews – a new approach

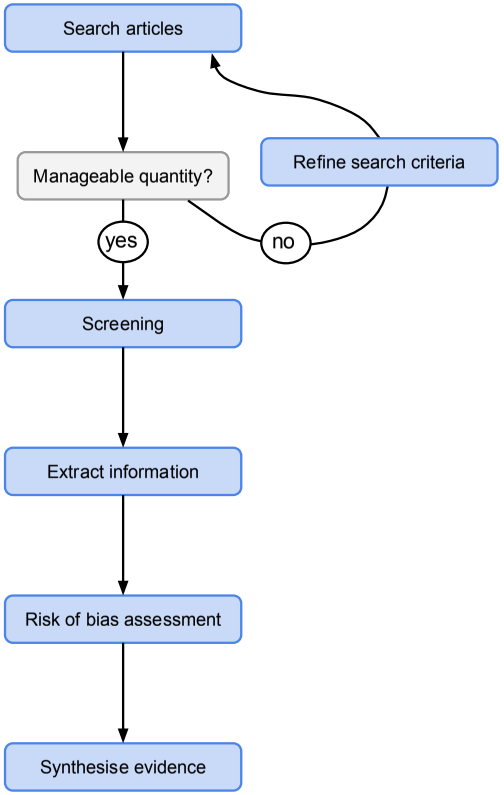
As described in Section 2.1.1, a rapid review needs to be performed under strict time and resource constraints, so it may not be possible to review all relevant articles. Currently, a rapid review is performed by human reviewers who search online medical research databases for articles reporting clinical trials of a particular research question [34]. In order to retrieve a set of articles that can be reviewed in the allotted time, the reviewer may iteratively refine the search query until the number of articles is deemed manageable. For example, a reviewer may restrict the search to articles in English language only. This process is illustrated in Figure 3.1 (left).

As is the case for standard systematic reviews, an important consideration when performing a rapid review is the quality of each study. Low-quality studies are more likely to give a biased estimate of the research question and may need to be excluded

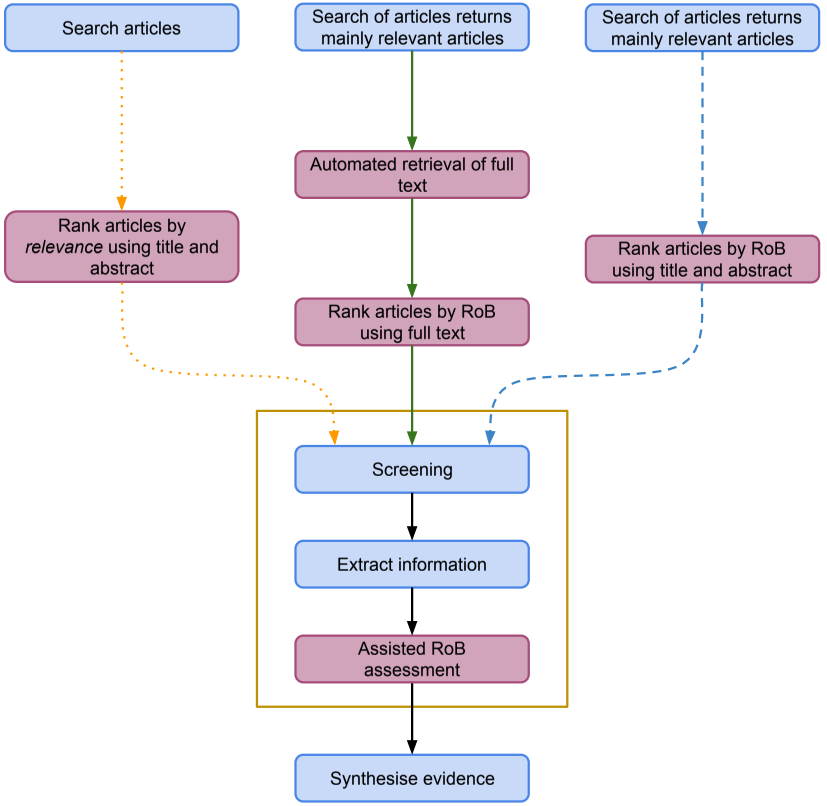
from the review or considered with caution [77]. If not all relevant articles can be included in the review then the reviewer would prefer those describing studies with low risk of bias to have priority. Therefore, we should aim to maximise the number of high-quality articles assessed, given the particular time constraints of the review. The iterative search method described above is a rather crude approach that does not consider article quality, and can be thought of as a classification of articles as *included* or *excluded* from the review. We suggest that this can be greatly improved by instead learning a model for estimating the article's study quality, and using the model's scores to rank the studies under review, such that the most reliable research is assessed first. The reviewers can then simply review the articles in decreasing order of estimated quality until they run out of time. There is no need to classify the articles as *included* or *excluded* before beginning the review.

Figure 3.1 (right) shows three alternatives of our proposed approach, all of which rank the articles instead of iterating the search. Each of these approaches ranks the articles and then, one article at a time (within the yellow box in this figure) the tasks of screening, extracting information and the assisted risk of bias assessment are performed. Performing these tasks one article at a time means that the reviewer can simply continue until they run out of time. In contrast reviewers using the current approach need to make sure they have time to screen all the articles then extract information from all the articles, and so forth, such that they reach the evidence synthesis stage in the allotted time.

Use of the three alternatives in Figure 3.1 (right) depends on other progress in this area. At the moment the search tools available cannot be queried precisely enough and this means that the search results contain many irrelevant articles. Therefore it would currently be most helpful to rank the articles by predicted relevance, as illustrated in the left path (dotted yellow arrows) of Figure 3.1 (right). As described in Section 2.2.3, automation of the search and screening stages are active areas of research, such that in the future it is likely that the search will improve and articles returned will be mostly relevant to the review. In this case we can rank the articles returned by risk of bias, using the title and abstract returned in the search to predict the risk of bias scores of each article, as shown in the right branch (dashed blue arrow) of Figure 3.1 (right).



(a) Typical rapid review search method with iterative searching.



(b) Proposed approaches. Left (yellow dotted): rank articles by predicted relevance; centre (green solid): rank articles by predicted RoB using full text; right (blue dashed) rank articles by predicted RoB using title and abstract.

Figure 3.1: Current rapid review approach and possible alternatives.

The centre path of Figure 3.1 (right) shows our idealized approach, which also requires an automated retrieval of full text articles. At present the title and abstracts can be automatically retrieved using the PubMed application programming interface (API), but there is no easy method to automate the retrieval of full text articles. In the future this should improve such that the full text articles can be used to predict risk of bias and rank the articles.

In this chapter we assume that the search returns mostly relevant articles and full text articles can be retrieved automatically, such that we are using the approach in the centre path of Figure 3.1(right). We note however, that all three strategies given in Figure 3.1(right) can be specified as a rate-constrained ranking task.

3.2 Rate-constrained ranking for rapid reviews

Our approach ranks research articles by study quality using a machine learning model, such that the reviewer can assess articles from high to low study quality until they run out of time. This suggests that a good model is one that exhibits good ranking behaviour with respect to study quality, with particular emphasis on the proportion of articles that can reasonably be processed. This proportion is not known before the review is performed, because we cannot know exactly how many articles a reviewer (or reviewer team) will be able to review. However, the total amount of time available for a review and the number of articles returned from the initial search query is typically known. Given an estimate of the time it will take a reviewer to assess a single article, the proportion of articles in the search results that is expected to be processed can be inferred. In terms of binary classification this proportion is the rate, the proportion of examples classified as positive by a model (formally the predicted positive rate), defined in Section 2.4.2. If the rate is known precisely, finding the best model is straight-forward.

Figure 3.2 illustrates this with two hypothetical ROC curves where neither curve dominates the other. The two dashed lines show two example rate values that could be inferred for a rapid review. We can see that the rate value affects which model is chosen. The (solid) green model is chosen when $rate = 0.5$ and the (dashed) blue model is chosen when $rate = 0.3$, as these models have the highest recall (F_0) at these respective points on the ROC curves. We assess the models using recall because we would like to assess the highest number of high quality articles in the allotted time. We refer to

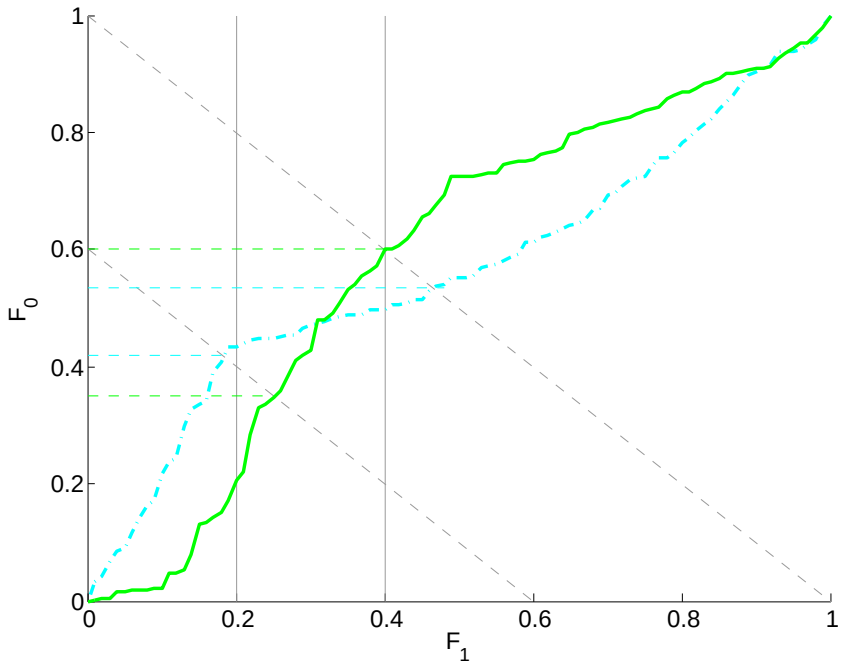


Figure 3.2: Two hypothetical ROC curves (x-axis: false positive rate, y-axis: true positive rate), example rate isometrics (diagonal lines with $rate = 0.5$ (top) and $rate = 0.3$ (bottom); the slope of -1 indicates a uniform class distribution) and example partial AUC bounds (vertical lines).

these tasks as rate-oriented – the threshold is specified by a rate. This contrasts to the standard method where the threshold is specified by a score, and the score predicted by the model is compared to this to determine the classification of an example.

In this rapid review task however, the rate inferred depends on the time needed to review a single article, and this is not known precisely. Articles vary in length and difficulty and hence it is only possible to estimate a probability distribution across the rate, rather than specify a single value. When choosing a model to rank articles for a review, we can use this probability distribution to assess the performance of the models, with consideration for the probability that the reviewer will stop at each point in the ranking (because they are out of time). Rather than finding the known (estimate of) recall at a particular position, as shown in Figure 3.2, we can only calculate the expected recall, given these probabilities. For instance, a model might have recall r_j at position j, and recall r_k at position k. If the number of articles that will be screened is twice as

likely to be j compared with k, then r_j will have twice the weight as r_k when calculating the expected recall.

We call this type of task rate-constrained – rate-oriented tasks where the threshold rate is not known precisely before the ranking is processed. As we have alluded to, an appropriate measure of rate-constrained ranking would average the true positive rate across each value of rate, weighted by its probability. In this chapter we develop such a measure.

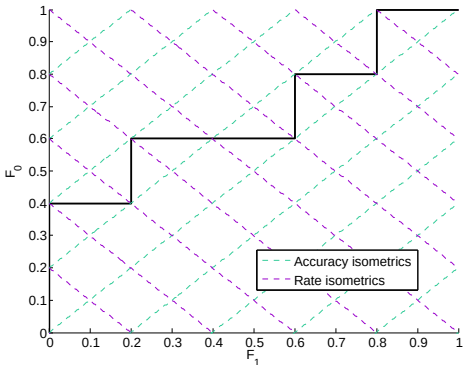
3.3 The rate-weighted AUC (rAUC)

Common formulations of the AUC are given as an expectation of the true positive rate across all false positive rates or thresholds (Equation 2.8). It is not possible to apply a weight across rates using these formulations, because they are given in terms of expectations over F_1 and t, rather than the rate. The following section derives the AUC as an expectation across rates, such that the derived formula can be altered to weight the AUC with respect to the rate.

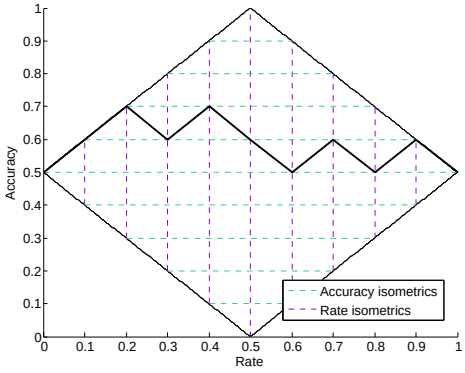
As introduced in Section 2.4.2, accuracy isometrics in ROC space are lines of constant accuracy with slope π_1/π_0 [75]. Similarly, rate isometrics are lines of constant rate with slope $-\pi_1/\pi_0$. Examples are shown in Figures 3.3a and Figure 3.3c for uniform and non-uniform class distributions, respectively.

Definition 3.1. Rate-accuracy space is a plot of rate on the x-axis and accuracy on the y-axis. Rate-recall space is a plot of rate on the x-axis and recall on the y-axis. Where positive recall is used, rate-recall space is denoted rate-$F_0(r)$ space. Where negative recall is used, rate-recall space is denoted rate-$(1-F_1(r))$ space.

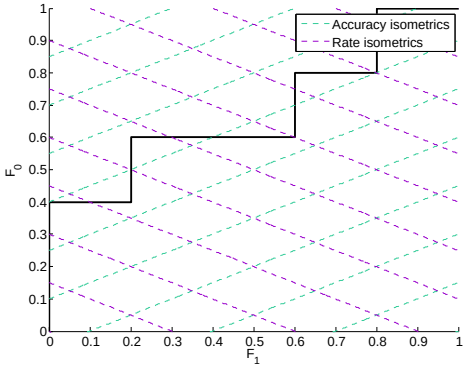
We translate the ROC curve to rate-accuracy and rate-recall spaces using a linear transformation, such that the AUC can be calculated in this space instead. The ROC curve of Figure 3.3a is transformed into the rate-accuracy curve shown in Figure 3.3b, and the rate-recall curves shown in Figures 3.3e and 3.3f, for positive and negative recall respectively. We can see that the transformations into rate-accuracy and rate-recall spaces result in unreachable areas. The upper bounds of the rate-accuracy and rate-recall curves correspond to the ROC curve of a perfect classifier, and the lower bounds to that of a pessimal classifier.



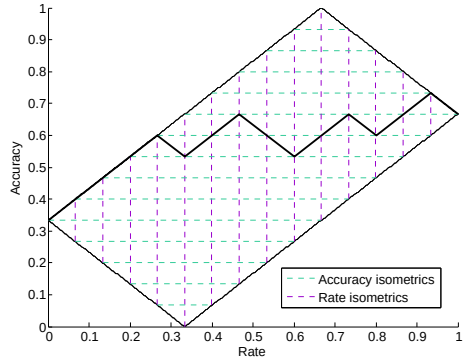
(a) Example ROC curve with rate and accuracy isometrics for $\pi_0 = \pi_1 = \frac{1}{2}$.



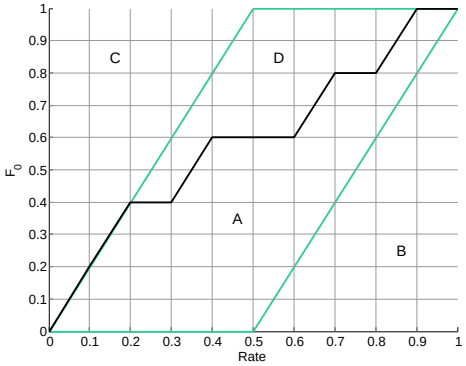
(b) Rate-accuracy curve corresponding to ROC curve shown in Figure 3.3a.



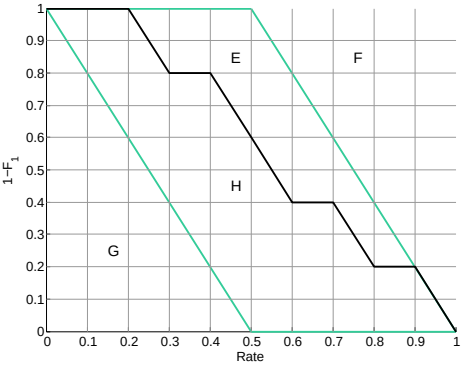
(c) Example ROC curve with rate and accuracy isometrics for $\pi_0 = \frac{2}{3}$, $\pi_1 = \frac{1}{3}$.



(d) Rate-accuracy curve corresponding to ROC curve shown in Figure 3.3c.



(e) Rate-recall curve for the positive class of ROC curve shown in Figure 3.3a.



(f) Rate-recall curve for the negative class, of ROC curve shown in Figure 3.3a.

Figure 3.3: Example ROC curves, rate-accuracy curves and rate-recall curves.

Definition 3.2. The lower bounds in x-y space are given by a function $f_{min}(x)$ specifying the minimum possible value of y at each value of x. The upper bounds in x-y space are given by a function $f_{max}(x)$ specifying the maximum possible value of y at each value of x.

We now focus on rate-accuracy space, but a similar derivation can be given for rate-recall space (given in Theorem 3.5). In rate-accuracy space, the lower and upper bounds of accuracy at rate r are given by:

$$acc_{min}(r) = |\pi_1 - r| \qquad acc_{max}(r) = 1 - |\pi_0 - r| \tag{3.1}$$

These are derived from Equation 2.6 and the fact that acc_{min} corresponds to points with $F_0 = 0$ when $r \leqslant \pi_1$ and points with $F_1 = 1$ when $r \geqslant \pi_1$, and acc_{max} corresponds to points with $F_1 = 0$ when $r \leqslant \pi_0$ and points with $F_0 = 1$ when $r \geqslant \pi_0$.

Clearly, a ROC curve can only cross each rate isometric at a single point, which allows us to reformulate the AUC in terms of accuracy and rates in order to apply a weight across rates. Accuracy difference acc_{dif} is the difference in the accuracy value of the ROC curve with the minimum possible accuracy value for a given rate:

$$acc_{dif}(r) = acc(r) - acc_{min}(r) \tag{3.2}$$

Theorem 3.3. The AUC is equal to the normalised accuracy difference across all rates $r \in [0,1]$:

$$AUC = \frac{1}{K_{acc}} \int_0^1 acc_{dif}(r)dr \tag{3.3}$$

where K_{acc} is constant for a fixed class distribution:

$$K_{acc} = \int_0^1 (acc_{max}(r) - acc_{min}(r))\, dr \tag{3.4}$$

Proof. The transformation of a ROC curve to rate-accuracy space requires only linear transformations from F_0 and F_1 into rate and accuracy, such that the relative areas under and above the curve within the transformed bounds of the original ROC space remains

the same. Therefore:

$$AUC = \frac{area}{K_{acc}} \tag{3.5}$$

where K_{acc} is the total area within the bounds of rate-accuracy space, and *area* is the absolute area under the rate-accuracy curve given by:

$$area = \int_0^1 acc_{dif}(r)dr \tag{3.6}$$

This concludes the proof. □

This reformulation of AUC in terms of rates allows us to introduce a rate-constrained generalisation.

Definition 3.4. The rate-weighted AUC of a ROC curve is the AUC weighted across the rates:

$$rAUC = \frac{1}{K_{acc,w(r)}} \int_0^1 w(r)acc_{dif}(r)dr \tag{3.7}$$

where $w(r)$ is a density over the rate and $K_{acc,w(r)}$ is given by:

$$K_{acc,w(r)} = \int_0^1 w(r)\left(acc_{max}(r) - acc_{min}(r)\right)dr \tag{3.8}$$

In rate-F_0 space, the lower and upper bounds of recall at rate r are given by:

$$F_{0,min}(r) = \max\left(0, \frac{r-\pi_1}{\pi_0}\right) \qquad F_{0,max}(r) = \min\left(1, \frac{r}{\pi_0}\right) \tag{3.9}$$

These are derived from Equation 2.6 and the fact that $F_{0,min}$ corresponds to points with $F_0 = 0$ when $r \leqslant \pi_1$ and points with $F_1 = 1$ when $r \geqslant \pi_1$, and $F_{0,max}$ corresponds to points with $F_1 = 0$ when $r \leqslant \pi_0$ and points with $F_0 = 1$ when $r \geqslant \pi_0$.

Theorem 3.5. The rAUC is equal to the normalised F_0 difference weighted across all rates. With a slight abuse of notation we use $F_k(r)$ to mean $F_k(F^{-1}(r))$.

$$rAUC = \frac{1}{K_{F_0,w(r)}} \int_0^1 w(r)\left(F_0(r) - F_{0,min}(r)\right)dr \tag{3.10}$$

where

$$K_{F_0,w(r)} = \int_0^1 w(r)\,(F_{0,max}(r) - F_{0,min}(r))\,dr \tag{3.11}$$

Proof. The transformation of a ROC curve to rate-F_0 space requires only a linear transformation of F_1 into rate, such that the relative areas under and above the curve within the transformed bounds of the original ROC space remains the same. Therefore:

$$AUC = \frac{area}{K_{F_0,w(r)}} \tag{3.12}$$

where $K_{F_0,w(r)}$ is the total weighted area within the bounds of rate-recall space, and *area* is the absolute weighted area under the rate-recall curve given by:

$$area = \int_0^1 w(r)(F_0(r) - F_{0,min}(r))dr \tag{3.13}$$

This concludes the proof. □

Clearly, we can derive an analogous result using negative recall ($1 - F_1(r)$) instead of positive recall ($F_0(r)$). The area under the rate-recall curve is the expected recall (positive or negative) given a uniform distribution across the rates. This makes the formulation of the rAUC in rate-recall space particularly interesting, as we can infer the relationship between $\mathbb{E}[F_0]$ – the quantity we intend to maximise in rate-constrained ranking – and the rAUC.

Rate-recall space, as shown in Figures 3.3e and 3.3f can be divided into 4 distinct regions, for both positive and negative recall (labelled A-D and E-H respectively). We use A both to label the region A and as the rate-weighted mass of this region (the area of this region weighted by the rates it contains).

Theorem 3.6. The rate-weighted expected true positive rate is related to the rAUC, given a distribution over the rates, by:

$$\mathbb{E}[F_0] = (1 - B - C) \cdot rAUC + B \tag{3.14}$$

where $C = \int_0^{\pi_0} w(r) \left[\frac{\pi_0 - r}{\pi_0}\right] dr$ and $B = \int_{\pi_1}^1 w(r) \frac{r - \pi_1}{\pi_0} dr$.

Proof. Rate-F_0 space is bounded by $r = 0$, $r = 1$, $F_0 = 0$ and $F_0 = 1$, such that the total weighted mass of this area $\int_0^1 w(r)dr = 1$, hence $A + B + C + D = 1$. As $rAUC = \frac{A}{A+D}$, it follows that:

$$\begin{aligned}\mathbb{E}[F_0] =& \frac{A+B}{A+B+C+D} \\ =& A + B = rAUC \cdot (A+D) + B \\ =& (1 - B - C) \cdot rAUC + B\end{aligned} \tag{3.15}$$

Area C is the triangular region bounded by the lines $r = 0$, $F_0 = 1$ and $F_0 = \frac{r}{\pi_0}$. The weighted mass of C is given by:

$$C = \int_0^{\pi_0} w(r) \frac{\pi_0 - r}{\pi_0} dr \tag{3.16}$$

Area B is the triangular region bounded by the lines $r = 1$, $F_0 = 0$ and $F_0 = \frac{r-\pi_1}{\pi_0}$. The weighted mass of B is given by:

$$B = \int_{\pi_1}^1 w(r) \frac{r - \pi_1}{\pi_0} dr \tag{3.17}$$

This completes the proof. □

B and C depend only on the class and weight distributions, which implies that the relationship between $\mathbb{E}[F_0]$ and rAUC depends only on these and not the shape of the ROC curve. Therefore, maximising $\mathbb{E}[F_0]$ is equivalent to maximising $\mathbb{E}[rAUC]$, which means that rAUC is a suitable metric to evaluate models for rate-constrained ranking.

3.3.1 Algorithm to calculate the rAUC of an empirical ROC curve

We now use rate-accuracy space to compute the rAUC. A similar algorithm could be implemented in rate-recall space (of either positive or negative recall). Algorithm 1 estimates the rAUC from an empirical ROC curve, where the number of positive n_0 and negative n_1 instances is known ($N = n_1 + n_0$). This algorithm is similar to the standard AUC $O(N)$ algorithm [73] where the ROC space is processed one vertical (or horizontal) slice at a time.

The rate-accuracy curve in Figure 3.4 corresponds to the ranking $\{001[011]0101\}$, where [ ] denotes examples with the same score, which we call tied examples. As can be seen in this figure, the area under the ROC curve in rate-accuracy space is composed of a series of vertical slices of width $\frac{1}{N}$, each corresponding to an instance. Ties sections may also exist, which correspond to a set of examples with the same score. Each rate-accuracy curve AUC section corresponding to a set of tied examples may have: A) vertical segments corresponding to tied negative examples, B) vertical segments corresponding to tied positive examples, and C) a ties triangle. We divide ties sections into these three component areas in order to easily calculate the rate-weighted mass of this section, as we describe below. An example of a tied section with all three components is given in Figure 3.4, between rates 0.3 and 0.6. This example has three vertical segments because three examples are tied, and these sit below the dashed lines shown in this figure. The first two segments correspond to two negative examples and the last corresponds to a positive example. Note how, although these examples are tied, when considering only the vertical segments and ignoring the ties triangle, the negative vertical segments precede the positive vertical segments in a ties section. This is shown on Figure 3.4 by the dashed line first corresponding to decreasing accuracy (for the negative vertical segments) and then to increasing accuracy (for the positive vertical segments). The tied section also has a ties triangle, sitting above the dashed lines in Figure 3.4 (consisting of T_A and T_B). A ties section only contains a ties triangle when it consists of both positive and negative examples.

The rAUC is calculated as a summation of the weighted mass of all vertical segments and ties triangles, normalised by the weighted mass of the whole rate-accuracy space. The algorithm we propose has four functions: *rAUC*, *SAUC*, *VAUC* and *TAUC*. The *rAUC* function is the main function that iterates through the ranking of instances counting the number of positive and negative instances with the same score (and so in the same tied section), and calling the *SAUC* function when a new score is reached.

The *SAUC* function calculates the weighted mass of a tied section of the rate-accuracy curve, by calling the VAUC and TAUC functions. The *VAUC* and ties triangle *TAUC* functions calculate the vertical segments (excluding ties triangle) and the ties triangle, respectively. Each section is treated as a tied section even though this may consist of only one example. The vertical sections of the negative instances are processed before those of the positives because, as already mentioned, when there is a ties

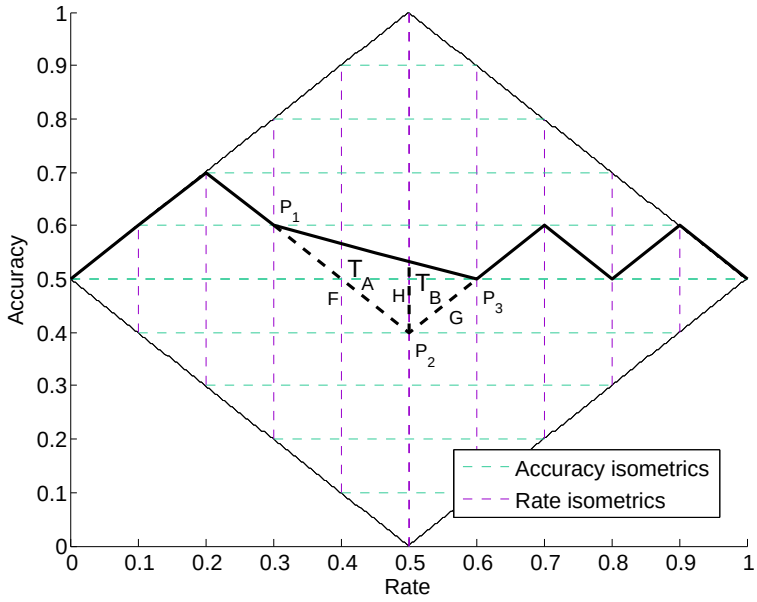


Figure 3.4: Example rate-accuracy curve with a ties section (set of instances with the same score). Lengths and angles used to calculate rAUC are labelled.

triangle the shape of the area under this triangle in rate-accuracy space is given by the area of the negative instances, followed by the positive instances in this ties section. For instance, we can see in Figure 3.4 that the position of the dashed line above the tied positive example would change depending on the number of negative examples in this tied section.

The *VAUC* function computes the mass of a vertical slice of the area under the curve, using two equations depending on whether the current instance is negative or positive. The accuracy difference equation is used, which is computed in terms of r and either F_0 or F_1 depending if the example is negative or positive respectively (as for instance, if the example is positive the value of F_1 stays constant). For instance, the absolute weighted mass of the vertical segment of negative examples is given by:

$$\begin{aligned} a_u &= \int_{r_l}^{r_u} w(r)\,(acc(r) - acc_{min}(r))\,dr \\ &= \int_{r_l}^{r_u} w(r)\,(2\pi_0 F_0 + \pi_1 - r - |\pi_1 - r|)\,dr \end{aligned} \tag{3.18}$$

where r_l and r_u are the lower and upper rates of this segment in the ranking.

The *TAUC* function computes the mass of the ties triangle. A ties triangle T is composed of 2 sub-triangles T_A and T_B where $T = T_A \cup T_B$. T_A and T_B adjoin on line H, where H is fixed along the rate isometric that passes through the right angled corner of T (see Figure 3.4).

We calculate the mass of a ties triangle by first finding the length H and the rate at

Algorithm 1 The rAUC algorithm. *scores*: list of scores of instances, in decreasing magnitude. *x*: list of class labels corresponding to the instances of *score*. n_0: number of positive instances. n_1: number of negative instances.

procedure RAUC($scores, x, n_0, n_1$)
  $\pi_0 \leftarrow n_0/(n_0+n_1)$; $\pi_1 \leftarrow n_1/(n_0+n_1)$; $N \leftarrow n_0+n_1$
  $a_u \leftarrow 0$; $TP \leftarrow 0$; $FP \leftarrow 0$
  $N^+_{ties} \leftarrow 0$; $N^-_{ties} \leftarrow 0$; $score_{ties} \leftarrow -1$
  for $i = 1$ to N **do**
    if $score_{ties} = scores(i)$ **then**
      if x_i is POSITIVE **then**
        $N^+_{ties} \leftarrow N^+_{ties} + 1$
      else
        $N^-_{ties} \leftarrow N^-_{ties} + 1$
      end if
    else
      $[FP, TP, a_u] \leftarrow SAUC(a_u, N^-_{ties}, N^+_{ties}, FP, TP, n_1, n_0)$
      if x_i is POSITIVE **then**
        $N^+_{ties} \leftarrow 1$; $N^-_{ties} \leftarrow 0$
      else
        $N^+_{ties} \leftarrow 0$; $N^-_{ties} \leftarrow 1$
      end if
      $score_{ties} \leftarrow score(i)$
    end if
  end for
  $[FP, TP, a_u] \leftarrow SAUC(a_u, N^-_{ties}, N^+_{ties}, FP, TP, n_1, n_0)$
  $a \leftarrow K(w, \pi_0, \pi_1)$
  $rAUC \leftarrow \frac{a_u}{a}$
  Return *rAUC*
end procedure

Algorithm 1 (continued)

procedure SAUC(a_u, N^-_{ties}, N^+_{ties}, FP,TP, n_1, n_0)
 if $N^-_{ties} \geqslant 1$ **then**
  $FP_{prev} \leftarrow FP$; $FP \leftarrow FP + N^-_{ties}$
  $a_u \leftarrow a_u + VAUC(FP_{prev}, TP, FP, TP, N, 0)$
 end if
 if $N^+_{ties} \geqslant 1$ **then**
  $TP_{prev} \leftarrow TP$; $TP \leftarrow TP + N^+_{ties}$
  $a_u \leftarrow a_u + VAUC(FP, TP_{prev}, FP, TP, N, 1)$
 end if
 if $N^+_{ties} \geqslant 1$ & $N^-_{ties} \geqslant 1$ **then**
  $a_u \leftarrow a_u + TAUC(FP, TP, N^-_{ties}, N^+_{ties}, N)$
 end if
 Return $[FP, TP, a_u]$
end procedure

procedure VAUC($FP_{prev}, TP_{prev}, FP, TP, N, label$)
 $start \leftarrow FP_{prev} + TP_{prev}$
 $end \leftarrow FP + TP$
 $f_1 \leftarrow \frac{FP_{prev}}{nTotalMinus}$
 $f_0 \leftarrow \frac{TP_{prev},}{nTotalPlus}$
 for $i = start$ to end **do**
  if $label = 0$ **then**
   $FP_{prev} \leftarrow FP_{prev} + 1$; $f_1 \leftarrow \frac{FP_{prev}}{nTotalMinus}$
   $a_u \leftarrow a_u + \int_{\frac{i}{nTotal}}^{\frac{i+1}{nTotal}} (w(r)(2\pi_1 f_0 + \pi_1 - r - |\pi_1 - r|)dr$
  else
   $TP_{prev}, \leftarrow TP_{prev}, +1$; $f_0 \leftarrow \frac{TP_{prev},}{nTotalPlus}$
   $a_u \leftarrow a_u + \int_{\frac{i}{nTotal}}^{\frac{i+1}{nTotal}} (w(r)\ (2\ (r\ -\ \pi_1\ f_1) +\ \pi_1\ -\ r\ -\ |\ \pi_1\ -\ r\ |\ )\ dr$
  end if
 end for
 Return a_u
end procedure

each corner of T, labelled P_1, P_2 and P_3 in Figure 3.4. Length H is given by:

$$H = \frac{sin(a) \cdot G}{sin(c)} \tag{3.19}$$

Algorithm 1 (continued)

procedure TAUC($FP, TP, N^-_{ties}, N^+_{ties}, N$)
$r_1 \leftarrow (FP + TP - (N^-_{ties} + N^+_{ties}))/N$
$r_2 \leftarrow (FP + TP - N^+_{ties})/N$
$r_3 \leftarrow (FP + TP)/N$
calculate H using Eqns $3.19 - 3.23$
$T_A = \int_{r_1}^{r_2} w(r) \cdot H \cdot \frac{r - r_1}{r_2 - r_1} dr$
$T_B = \int_{r_2}^{r_3} w(r) \cdot H \cdot \left(1 - \frac{r - r_2}{r_3 - r_2}\right) dr$
$T = T_A + T_B$
Return T
end procedure

where angle a, the angle of T at point P_3, is then given by:

$$a = tan^{-1}\left(\frac{F}{G}\right) \tag{3.20}$$

Angle b, the angle of T_B at point P_2 is fixed at 45. It follows that angle c, the remaining angle of T_B, is given by:

$$c = 180 - (45 + a) = 135 - a \tag{3.21}$$

The lengths F and G are given by:

$$F = \sqrt{\left((acc_1 - acc_2)^2 + (r_2 - r_1)^2\right)} \tag{3.22}$$

$$G = \sqrt{\left((acc_3 - acc_2)^2 + (r_3 - r_2)^2\right)} \tag{3.23}$$

where acc_i and r_i are the accuracy and rate at point P_i in Figure 3.4, respectively.

The weighted mass of the ties triangle is then the summation of T_A and T_B, which are given by:

$$T_A = \int_{r_1}^{r_2} w(r) \cdot H \cdot \frac{r - r_1}{r_2 - r_1} dr \qquad T_B = \int_{r_2}^{r_3} w(r) \cdot H \cdot \left(1 - \frac{r - r_2}{r_3 - r_2}\right) dr \tag{3.24}$$

where r_1, r_2 and r_3 are the rates at P_1, P_2 and P_3 respectively. These equations use the fact that the height of a ties triangle at a particular rate r is proportional to the distance between the minimum and maximum rates of this triangle, r_1 and r_2 for T_A and r_2 and r_3 for T_B. For T_A the height increases from r_1 to r_2 whereas for T_B the height decreases from r_2 to r_3.

The rAUC of a ROC curve is computed in $O(N)$ time. Algorithm 1 appears more lengthy compared to the standard AUC algorithm that is calculated in ROC space because each step across rate-accuracy space corresponds to a negative or positive instance and the height of the curve changes within this step. The standard AUC algorithm makes a step only when the instance is (for example) positive and (given this instance is not tied with another) the height of the ROC curve is constant within this step. The change in height at each step in rate-accuracy space means that the mass of the positive and negative vertical sections (and ties triangle) can only be calculated after the ties section has ended, hence the *SAUC* function is needed to do this.

3.4 Comparison of rAUC with other metrics

We have introduced a new ranking measure, the rAUC, that is able to account for constraints across rates, but several other metrics also exist to evaluate ranking models. In this section we give an overview of related metrics, and provide formal comparisons of these with the rAUC.

The AUC, defined in Section 2.4.2, is a popular choice to assess the performance of ranking models. Intuitively, the AUC estimates the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance, and thus represents ranking performance across the entire dataset. Historically, the AUC has often been used as a measure of ranking performance without consideration for the particular task at hand. However, when the performance of a learner in particular regions of ROC space has more importance than other areas for a particular task, the AUC is not an appropriate choice.

Alternatives to the AUC have previously been suggested to allow differential importance across true positive or true negative rates, for empirical [78,79] and analytical [80] ROC curves. These studies propose a partial AUC (pAUC) metric to restrict the evaluation of the AUC to a range of false positive or true positive rate values. The pAUC

measure is appropriate when it is required that either the true positive or false positive rates fall in a particular range. It is interesting to note that this is proposed for evaluation of medical diagnostic tests [78]. Diagnostic tests assess how well a particular test is able to detect a disease, and for tests that output a continuous value a ROC curve can be created to evaluate this test. A researcher, for instance, may be interested in areas of ROC space with low false positive rates because of high monetary costs incurred when the false positive rate is higher [78].

We suggest that the pAUC metric could be generalised using weights rather than bounds (as we have used for the rAUC), which may be more appropriate where there is a non-uniform probability distribution across either the true or false positive rate. Furthermore, a recent variant of the AUC called the half-AUC was proposed [81], and evaluates the AUC in only half of the ROC space, either where true positive rate is less than true negative rate or true positive rate is greater than true negative rate, giving two distinct regions (either side of the descending diagonal) that can be assessed.

Early retrieval tasks are those where examples near the top of the ranking are more important, as these examples are more likely to be processed. Several metrics have been suggested for early retrieval tasks, where evaluation focuses on the top of the rankings. Precision@k gives the precision of the top k results of a ranking, thus weighting each example uniformly within this section of the ranking. For binary classes this is akin to cumulative gain, which calculates the total number of positive examples up to (and including) a particular position in the ranking [82]. Precision@k is the cumulative gain at k, relative to the value of k itself.

Normalised discounted cumulative gain (NDCG) [82, 83], is one of several metrics that give decreasing weights to examples along the ranking, as will be discussed in Section 3.4.2. Others include robust initial enhancement (RIE) [84], the Boltzmann-enhanced discrimination of ROC (BEDROC) [85], concentrated ROC (CROC) [86] and sum of the log ranks (SLR) [87]. As we shall see in Section 3.4.2, the instance weights used by these approaches all share the characteristic that they translate into monotonically decreasing rate weights. This is often not appropriate for rate-constrained ranking tasks. For example, in our rapid review task it may be more likely that the reviewer stops processing the examples midway through the ranking, compared with at the beginning of the ranking.

3.4.1 Experimental comparisons

We used 5 UCI datasets [1] (vote, autos, credit-g, breast-w and colic) to generate a set of models using 3 learning algorithms (naive Bayes, decision trees and one-rule). We chose a binary variable for each dataset as the label, and learnt 10 models with each dataset/model pair using bootstrap samples of 54% of the data, resulting in 150 generated models. We computed the AUC and NDCG metrics for each of these models. We also computed the rAUC using 5 beta distributions as the weights across rates, with alpha and beta (α, β) values: $(3,19)$, $(7,15)$, $(11,11)$, $(15,7)$, and $(19,3)$, shown in Figure 3.5. We use beta distributions because they are constrained to values between zero and one and rates are also constrained in this way. The α and β parameters were chosen such that the modes of the beta distributions were equal distance apart across the rates. We use NDCG with log base 10.

Figure 3.6 shows the AUC and NDCG values, compared with the rAUC values, for each model. Each model is shown by 5 points with a single AUC / NDCG value and variable rAUC value (for each of the 5 rate distributions of Figure 3.5). The variance of the rAUC for each ROC curve across the 5 beta distributions ranges from 0 to 0.260 for these datasets.

[1] UCI is a machine learning repository including freely available datasets.

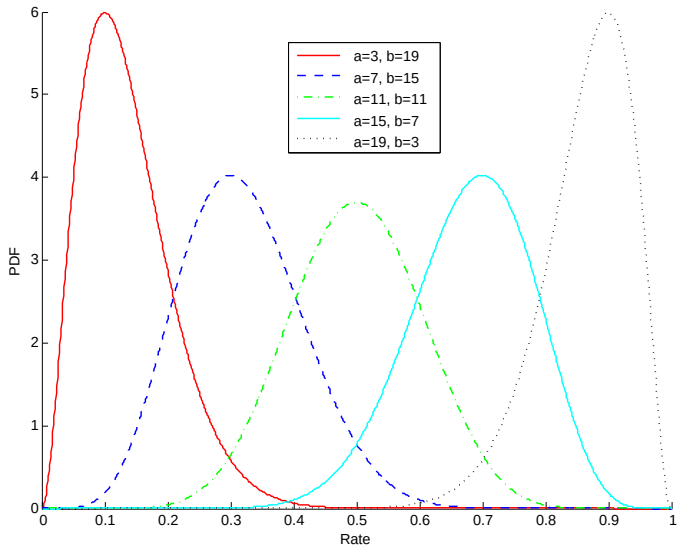
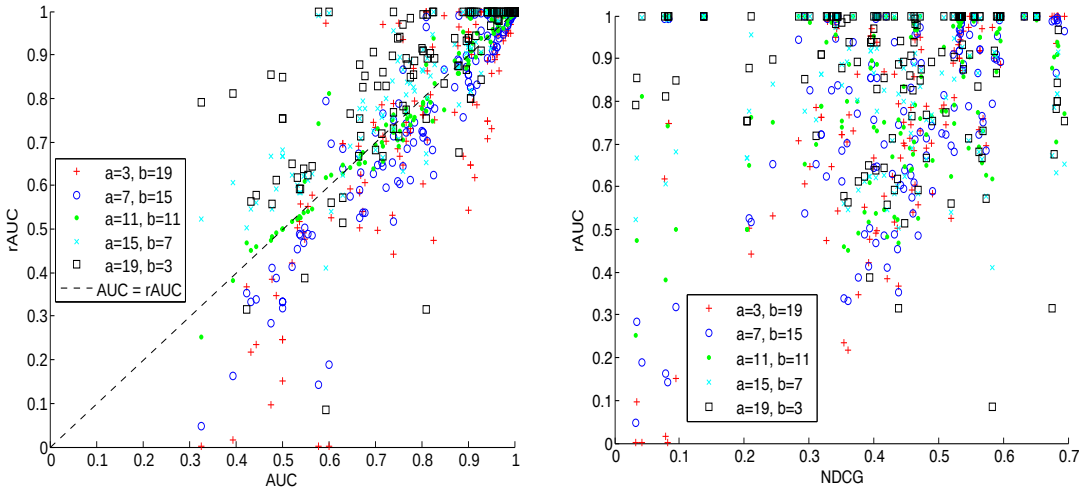


Figure 3.5: Beta distributions across the rate.



(a) AUC and rAUC

(b) NDCG and rAUC

Figure 3.6: Comparison of metrics for 150 models generated with various learners, datasets and distributions over rates.

For model selection, a monotonically increasing relationship between rAUC and another metric would indicate no difference, because the same model would be chosen using either metric. Spearman's rank correlations between the model rankings using each rate distribution are given in Table 3.1. A completely monotonically increasing relationship (with no ties) would have a Spearman's rank correlation of 1, and this decreases towards zero as monotonicity reduces.

The correlation of the rAUC with the AUC varied between 0.872 and 0.975, depending on the rate distribution, as illustrated by the positive association shown in Figure 3.6a. We should note that a proportion of the generated models have very high AUC values, and therefore very high rAUC values for most rate distributions (see Figure 3.6a). To correct for this inflation of the correlation values, Table 3.1 also shows reduced correlations when restricted to models with $AUC \leqslant 0.95$. The correlation of the rAUC with the NDCG varied between 0.018 and 0.565, depending on the rate distribution, as illustrated by the weak associations shown in Figure 3.6b. Correlations between the NDCG and rAUC metrics decrease dramatically when the mode of the beta distribution increases, as expected.

The correlation between rAUC metrics using rate distributions that weight different portions of ROC space is low in general. For instance, the rAUC values using rate

	$\alpha = 3$ $\beta = 19$	$\alpha = 7$ $\beta = 15$	$\alpha = 11$ $\beta = 11$	$\alpha = 15$ $\beta = 7$	$\alpha = 19$ $\beta = 3$
NDCG	0.565	0.438	0.235	0.092	0.018
AUC	0.872	0.951	0.975	0.927	0.886
AUC $\leqslant 0.95$	0.725	0.902	0.961	0.829	0.703
$\alpha = 3\ \beta = 19$		0.923	0.791	0.676	0.610
$\alpha = 7\ \beta = 15$			0.931	0.823	0.764
$\alpha = 11\ \beta = 11$				0.964	0.925
$\alpha = 15\ \beta = 7$					0.982

Table 3.1: Spearman's rank correlations comparing the rankings of the 150 models, ranked using the rAUC (with rate distributions of Figure 3.5), NDCG and AUC.

distributions with $\alpha = 3$, $\beta = 19$ and $\alpha = 19$, $\beta = 3$ have a Spearman's rank correlation of 0.610. This highlights the importance of using a rate distribution with an appropriate degree of uncertainty, as if it is incongruous with the true probability distribution a suboptimal model may be chosen.

3.4.2 Comparing the weights of NDCG and rAUC

NDCG is given by:

$$NDCG = \frac{1}{K} \cdot \sum_{i=1}^{n} \frac{1}{log_b(i+1)} rel_i \tag{3.25}$$

where $rel_i \in [0,1]$ is the label of the example at rank i, which can be continuous or binary and denotes the relevance of the example. K is the maximum possible DCG for a ranking of size n:

$$K = \sum_{i=1}^{n} \frac{1}{log_b(i+1)} \tag{3.26}$$

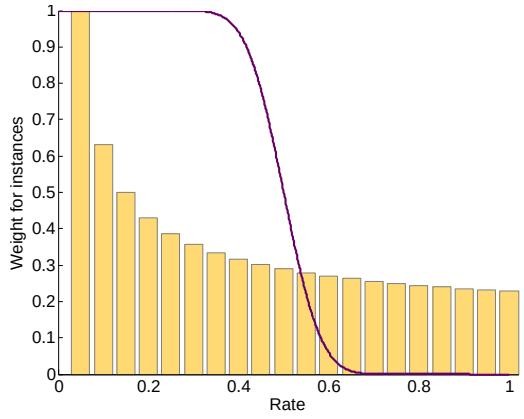
NDCG weights each point in the ranking according to the probability that this instance will be processed. In contrast, the rAUC weights each point in the ranking according to the probability this point will be the threshold index, such that processing will terminate at this point in the ranking. These formulations are closely related, since the probability that an instance at position i is processed is the probability that an instance at i or after position i is the threshold index. For example, if a person is processing 20

articles, the probability they will review the article at rank position 10 equals the probability they will stop processing articles at a position between articles 10 and 20. Hence, the relationship between the two weighting methods is given by:

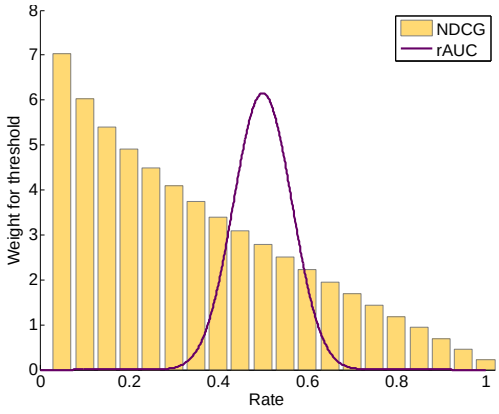
$$w_{instance}(i) = 1 - CDF_{w_{threshold}}(i) \tag{3.27}$$

where i is the position in the ranking and CDF denotes the cumulative distribution function.

The instance weights of NDCG are shown in Figure 3.7a, and the equivalent threshold weights are shown in Figure 3.7b. We can see that the weight of each threshold index decreases as we move further down the ranking. This is a key restriction of the NDCG (and related metrics), as it is not always the case that a ranking is more likely to be processed up to the rank positions nearer the top, as in our motivating example. Figure 3.7b also shows an example density across thresholds, using a beta distribution, where processing is most likely to stop at a rate of 50%. The corresponding instance weights are shown in Figure 3.7a. This weight distribution assigns several of the top-ranked instances the highest weight, something which is not possible with NDCG.



(a) Weights across instances, representing the likelihood an instance will be processed.



(b) Weights across thresholds, representing the likelihood a rate will be the threshold position (the instance at this rate will be the last to be processed).

Figure 3.7: NDCG discrete weights (using log base 2) assuming 20 instances and rAUC continuous weights using beta distribution ($\alpha = 30$, $\beta = 30$). Weights across instances in left figure are equivalent to weights across thresholds in right figure, respectively.

3.5 Application to screening for rapid reviews

We demonstrate the rAUC using our motivating example described in Section 3.1 – ranking research articles for rapid reviews of clinical trials. We formulate this task in terms of a rate-constrained ranking problem. To reiterate, the articles are ranked by estimated study quality, and the objective is to maximise the number of high-quality articles the reviewer assesses given the rate constraints. In this setting, the rate is the proportion of articles that the team reviews, which is not known precisely. The search will return A articles and the reviewers are allotted T minutes to complete the review. We use elicitation to determine appropriate parameters for the rate distribution, a method commonly used in epidemiology to establish feasible parameters for a distribution where there is no data from which to infer this. For simplicity, we consider the case of only one reviewer, who estimated the minimum (t_0) and maximum (t_1) time per article, t, the number of minutes they will on average expect to take to assess a single article. Also for simplicity, we use the *blinding* risk of bias property as a measure of quality. We assume our idealised situation given in the centre path of Figure 3.1b, where we have a set of relevant full text articles and we would like to rank them by predicted risk of bias.

We model t as a inverse beta distribution (with bounds $[\frac{T}{A}, \infty]$), having 0.95 probability of being in the range $[t_0, t_1]$. The rate (the proportion of articles that are reviewed) is given by: $r = \frac{T}{A \cdot t}$. This relationship with t infers a beta distribution across the rates. This assumes that a reviewer will not finish processing all the articles within the time allocated such that $r \in [0, 1]$.

We suppose a hypothetical and realistic rapid review where the search returns $A = 2,500$ articles and a reviewer is given 120 person hours ($T = 7,200$ minutes) in which to perform the review. We imagine that the reviewer states they will take between 10 and 45 minutes to assess a single article, which we use to specify two quantiles of t ($0.025 = CDF_t(0,10)$ and $0.975 = CDF_t(0,45)$) which we convert to equivalent quantiles of r ($0.975 = CDF_r(0,0.288)$ and $0.025 = CDF_r(0,0.064)$). We use the *beta.select* function of the *LearnBayes* R package [88] to find the α and β parameters with these quantiles, giving $\alpha = 6.23$ and $\beta = 32.80$ (shown in Figure 3.8).

We learn several models and then evaluate these using the weight distribution across rates that we have just specified, to determine which should be used to rank the 2,500 articles in the rapid review. We use a dataset consisting of 315 full-text articles reporting

the results from randomised controlled trials, each labelled with a binary value denoting whether blinding has been adequately carried out (*low* vs *not-low*), as described in the article. These are a subset of articles with a label for blinding given in Figure 2.5, from an earlier version of this dataset. There were an approximately equal number of articles of each class. We created a set of preliminary models using a bag of words representation (with unigrams), and evaluate these using 10-fold cross validation. To be clear, we learn and evaluate models using our labelled dataset and use these results to determine which model is expected to give a higher performance given the specific rapid review defined above (with 2,500 articles).

We generated consensus ROC curves for 3 learning algorithms: random forest, naive Bayes and support vector machine (SVM) (with a linear kernel), shown in Figure 3.9. We used *rate-averaging* to generate our consensus curves, as described in Section 2.4.2. This is appropriate for our rate-constrained task as the points of the consensus curves are the average performance given a particular rate constraint.

The random forest, naive Bayes and SVM models gave a mean rAUC (AUC) of 0.689 (0.636), 0.781 (0.639), and 0.639 (0.570), respectively, across the 10 folds. A two-tailed paired t-test of the AUC values of each model across the 10 cross validation folds, found no difference between the random forest and naive Bayes models ($P = 0.884$). A t-test using the rAUC values found the naive Bayes model is better than the random

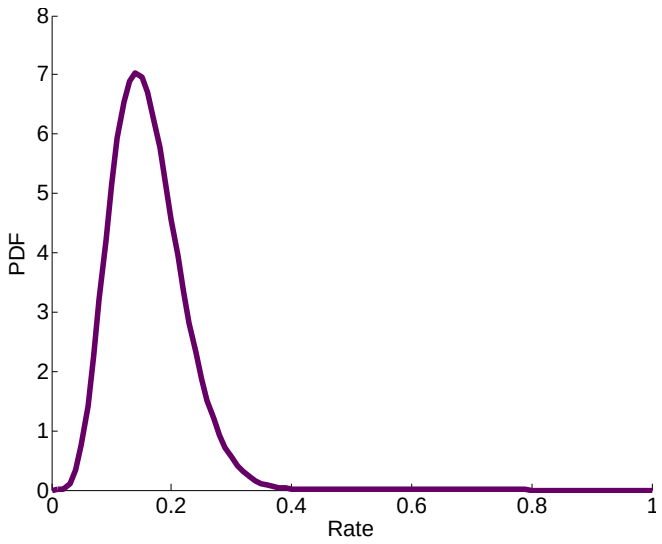


Figure 3.8: Beta distribution of weights across rates for rapid review.

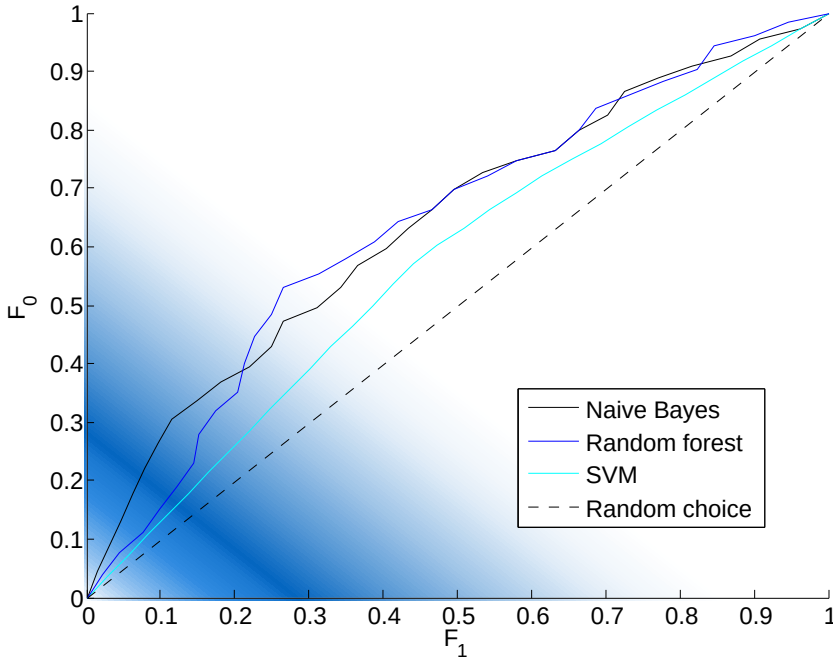


Figure 3.9: Consensus ROC curves (using rate-averaging) predicting the blinding risk of bias value of research articles. The blue shading depicts the beta distributed weighting across rates. The mode, at $rate = 0.14$ has the highest weight, and at this point on the ROC curve (and nearby points) the naive Bayes model has a higher recall compared to the random forest and SVM models.

forest model for this rate distribution ($P = 0.021$). The random forest and naive Bayes models clearly dominate the SVM model such that the SVM model would be inferior for any rate distribution. However, we have shown that while the random forest and naive Bayes models are similar in terms of ranking performance across the entire ranking, the naive Bayes model is much better than the random forest when considering which rates are more likely for this particular rapid review.

We thus clearly see that the weight distribution for rate-weighted AUC can be derived directly from the parameters of the rapid review task, in a way that could not be achieved with metrics such as the pAUC.

3.6 Summary

In this chapter we have introduced a new ranking measure, the rate-weighted AUC (rAUC), to better reflect model performance when the task is constrained by a probability distribution across the predicted positive rate, which we refer to as the rate. The AUC is equivalent to the rAUC given a uniform distribution across the rates. Furthermore, if the rate is fixed then models can be compared by simply comparing the recall at the point on the ROC curve with this rate. We have derived the rAUC from rate-accuracy space, and introduced rate-recall space as a visualisation of model performance. Furthermore, the rAUC is a linear transformation of rate-weighted expected recall (both the positive and negative respectively), given fixed class and rate distributions. We have described an $O(N)$ algorithm to calculate an estimate of the true rAUC using a data sample.

Our experiments have shown large variability of the rAUC as the rate distribution varies. A comparison with NDCG found low correlations indicating that when the likelihood that processing will stop at a particular position in the ranking is lower nearer the top of the ranking than elsewhere, NDCG may be inappropriate. Furthermore, a comparison with the AUC shows that often the rAUC prefers different models. Finally, we have also demonstrated how this approach can be usefully applied to real world tasks, using the example of ranking research articles for rapid reviews of clinical trials.

In addition to ranking articles for rapid reviews, there are many other tasks that are rate-constrained, with uncertainty across the rates. In general these tasks are rate-oriented with the additional property that the rate is not known precisely prior to processing examples along the ranking. This occurs because these tasks are restricted to a fixed budget of a resource such as time or money, where the exact expenditure for each instance is not known precisely. The aim of a rate-constrained ranking task is to maximise the expected true positive rate given the uncertainty across the rates. Another example is telephone sales, which is restricted by the allocated number of person hours, such that when ranking a database of customers to determine those most likely to show interest, it is not known exactly how many customers will be contacted as the time per phone call is variable.

Chapter 4

Rate-oriented confidence bounds

In this chapter we continue the rate-oriented theme and present a novel method of generating confidence bounds around ROC curves. We call this method rate-oriented pointwise confidence bounds. These bounds are particularly appropriate for rate-oriented ranking tasks, including those that are rate-constrained. We derive our approach from first principles and demonstrate its effectiveness experimentally. We start by providing an overview of existing approaches to creating confidence bounds around ROC curves. The work in this chapter has been published in [28].

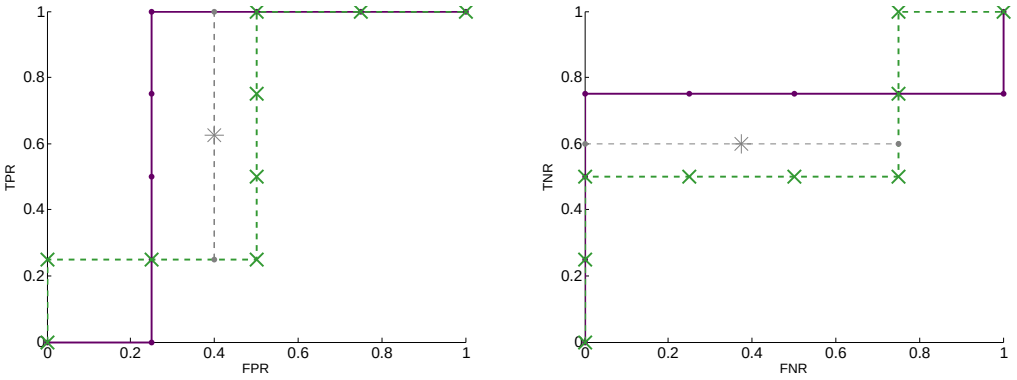
4.1 Approaches to create ROC confidence bounds

As described in Chapter 2, ROC curves are informative visualisations of model performance that show the ranking performance at different regions of a ranking, or the performance of a scoring classifier at each possible choice of operating point. ROC curves are often used to determine if one model is better than another, and confidence bounds provide a measure of the uncertainty such that this can be determined, for a specified confidence level. Recall from Section 2.4.2 that several ROC curves can be combined to produce a single average curve, called a consensus curve. The variation between the individual curves can be used to estimate a confidence around the consensus ROC curve. Several methods have been proposed to generate confidence bounds, mainly parametric approaches such as vertical [89] or threshold [90] averaging. While we focus on these approaches here, in Section 4.7 we describe other approaches that have been proposed,

such as those for meta-analyses of ROC curves for systematic reviews of diagnostic tests.

The vertical averaging method to generate a consensus curve, as described in Section 2.4.2, can be extended to generate a confidence band by calculating the standard deviation across the true positive rate at each point on the curve, and using this to generate a confidence interval. A similar procedure can also be performed with horizontal averaging. These are simple approaches to implement but have several shortcomings. Firstly, the false and true positive rates are metrics over which we have little control, such that it is difficult to set a threshold at a particular value. It is therefore preferable to evaluate a ROC curve with respect to a metric with which setting the threshold is simple in practice.

Secondly, vertical and horizontal averaging are not invariant to swapping the classes, such that if the x-axis and y-axis of ROC space become the false and true negative rate respectively, equivalent points will have different confidence bounds. The example in Figure 4.1 shows that vertical averaging in the original ROC space is equivalent to horizontal averaging in the swapped space. This is because the swapped space is a line mirroring of the original space along the descending diagonal. Finally, depending on the distributional assumptions of points at each false (or true) positive rate value,



(a) Original ROC space (b) 'Swapped' space, in terms of the negative class

Figure 4.1: Vertical and horizontal averaging correspondence when 'swapping the classes'. Swapped space is a line mirroring of original space along descending diagonal. Two gray points are averaged to consensus point (starred) with vertical averaging in original space (left). To do this in swapped space (right) horizontal averaging must be used.

the confidence bounds may not be constrained to the bounds of ROC space, such that $tpr \in [0,1]$ and $fpr \in [0,1]$ (where tpr and fpr are the true and false positive rates respectively).

Threshold-averaging is similar to vertical (and horizontal) averaging but instead fixes the score and averages over each cloud of points in ROC space with the same score (shown in Figure 2.15). This has the advantage that we can easily use thresholds set at a particular score, classifying each example by whether its score is below or above this threshold value. However, how best to generate confidence bounds for a set of points that are not constrained to a single dimension is not obvious. Fawcett et al. suggest averaging separately across false and true positive rates [90], but this creates a rectangular shaped bound for each score where a smoother bound would seem more natural.

4.2 Overview of our approach

To address the shortcomings of existing methods, we specify a set of properties we would like our confidence bounds to satisfy. Firstly, the generated confidence bounds should be invariant to swapping the classes, as described above. Secondly, the confidence bounds should be constrained to sit within the bounds of ROC space at all points along the lower and upper confidence bounds. Thirdly, we are particularly interested in generating confidence bounds around ROC curves that evaluate models of rate-oriented tasks. These tasks may also be rate-constrained such as the example of ranking articles for rapid reviews described in Chapter 3.

We suggest that when a task is rate-oriented, the consensus curve should be generated in a rate-oriented manner, such that each consensus point along the curve is generated with respect to a particular rate. The rate-averaging approach described in Section 2.4.2 is one such method. Furthermore, the comparison of several models should use confidence intervals also created at each rate value, which we call a *rate-oriented* approach, such that they can be compared with respect to the rate.

Our aim is to generate confidence intervals for a consensus curve at each rate value, such that at significance level σ the consensus curve of a set of new samples generated from this consensus curve (or precisely from the set of ROC curves from which the consensus curve was generated) pass between the lower and upper confidence limits at

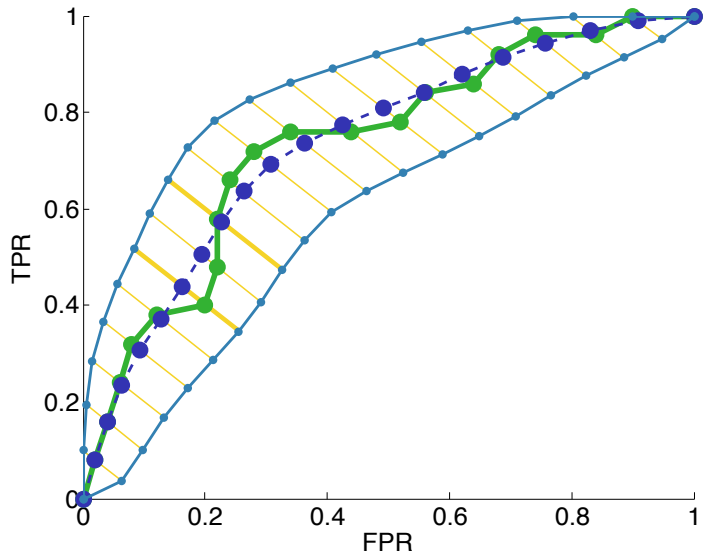


Figure 4.2: Rate-oriented confidence bound illustration showing a consensus curve (green solid) generated with rate-averaging, and the upper and lower confidence bounds. Each yellow line denotes the 95% confidence interval at a particular rate. Confidence intervals at $rate = 0.3$ and $rate = 0.4$ are emboldened. The smoothed consensus curve (blue dashed) is also generated with our confidence bounds approach. We have interpolated between the upper and lower bounds of the confidence intervals to create a confidence bound around the whole curve.

a given rate value, with probability $1 - \sigma$. The series of confidence intervals creates a confidence *bound* around the consensus curve. We call these *point-wise* confidence bounds in line with [91] in order to differentiate from the common meaning of ROC confidence bands, where the confidence refers to the proportion of whole curves sitting entirely inside the confidence band. Where we discuss methods that are solely used to generate a bound around the whole curve, we explicitly refer to these as bands.

An illustration of our rate-oriented point-wise confidence bounds is given in Figure 4.2. This figure shows a consensus curve and the generated point-wise confidence bounds. Each confidence interval sits along a rate isometric, and in this example we have generated 19 confidence intervals at rate intervals of 0.05. The upper and lower confidence bounds are composed of the upper and lower bounds of the confidence intervals, respectively. We have interpolated between the confidence intervals to approximate the bounds of intermediate rates. As we shall see, our approach also infers a new smooth

consensus curve, shown in this figure by the blue dashed curve.

Confidence bounds are useful to determine where in ROC space a consensus curve is likely to sit. In particular, our rate-oriented point-wise confidence bounds allows us to see the expected performance of a model as the rate varies. For example, in our example in Figure 4.2 we can see that at $rate = 0.3$ the recall (tpr) is expected to be between around 0.34 and 0.52. Increasing the rate to 0.4 however means that a recall between 0.47 and 0.67 is expected. A researcher may use this information to determine the best rate value to use as a threshold, according to the level of performance they require. Of course, while here we used recall any other classification metric that can be inferred from ROC space (such as accuracy) can be assessed in this rate-oriented manner.

4.3 Introducing ROC tables

ROC tables are a tabular form of empirical ROC curves. Formally, a ROC table such as that shown in Table 4.1, is a matrix with m rows and n columns, containing the results of independent tests using m samples, such as m-fold cross validation. It is the evaluation on the test data that is independent, not necessarily the training data, as is the case for cross validation with overlapping data in each training set. We discuss the implications of the degree of training set dependence in Section 4.8. We now introduce notation pertaining to ROC tables specifically, which extends the notation given in Section 2.4.2 for single rankings.

	k					
$y_{i,k}$	1	2	3	...	n-1	n
Sample 1	0	0	1	...	1	1
Sample 2	0	1	1	...	0	1
Sample 3	0	0	0	...	0	1
Sample 4	0	1	0	...	1	1
pos_k	4	2	2	...	2	0

Table 4.1: Example ROC table, with $m = 4$ samples and n columns, and pos_k positive examples in each column.

$s_{i,k}$	$s_{i,1}$	$s_{i,2}$	$s_{i,3}$
Sample 1	1	2	2
Sample 2	1	1	1
Sample 3	1	2	3
Sample 4	1	1	2

Table 4.2: Example $s_{i,k}$ values (number of positive examples up to column k in a sample) for example ROC table (left).

Each cell of a ROC table contains the label $y_{i,k} \in [0,1]$ of the example at position k along the ranking of sample i, where the examples of each sample are ranked by increasing score. A segment of consecutive positions in a ranking having the same score are assigned a fractional label to account for this – the average of the labels in this segment, calculated as:

$$y'_{q..q'} = \frac{1}{1+q'-q} \sum_{j=q}^{q'} y_j \tag{4.1}$$

where q and q' are, respectively, the start and end of the position range with equal score. The number of positives and negatives in a ranking are denoted by n_0 and n_1 respectively, such that $n = n_0 + n_1$.

The number of positives across samples at column k in the ROC table, denoted pos_k, is given by:

$$pos_k = \sum_{i=1}^{m} (1 - y_{i,k}) \tag{4.2}$$

Examples of pos_k are given in Table 4.1. The number of positives up to position k of row i in the ROC table, which we refer to as the *true positive value* (as opposed to the true positive rate) and denote by $s_{i,k}$, is given by:

$$s_{i,k} = \sum_{j=1}^{k} (1 - y_{i,j}) \tag{4.3}$$

Examples of $s_{i,k}$ are given in Table 4.2 for the ROC table shown in Table 4.1. The number of positives up to position k across all samples in the ROC table, denoted s_k, is given by:

$$s_k = \sum_{j=1}^{k} pos_j = \sum_{i=1}^{m} s_{i,k} \tag{4.4}$$

Recall (of the positive class) is the proportion of positive examples correctly classified as positive, at a given point on the ROC curve (also known as the true positive rate). We specify this in terms of a row of a ROC table. The recall, $tpr_{i,k}$, of sample i with operating point at position k is given by:

$$tpr_{i,k} = \frac{s_{i,k}}{n_0} \tag{4.5}$$

We denote an unsorted list of n items as $a_1, a_2 \ldots a_n$ and a sorted list as $a_{(1)}, a_{(2)} \ldots a_{(n)}$.

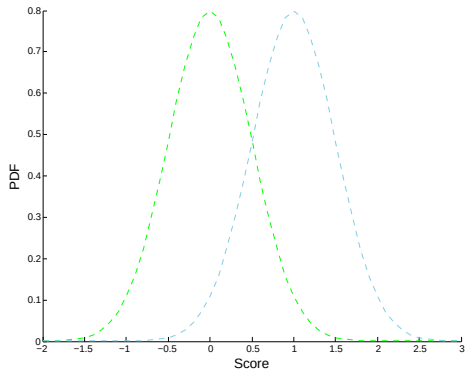
4.4 Rate-first sampling: a method to generate sample ROC curves

As discussed in Section 2.4.2, a ROC curve is an alternative representation of the cumulative distribution functions of two classes. Given the score probability distributions of each class, sample rankings and hence sample ROC curves can be generated using this distribution. For instance, we can sample a score from the mixed probability distribution and then sample a label using the probabilities of each class at this score, for each example in the new ranking. We call this score-first sampling, shown in Table 4.3 left. Example score densities are shown in Figure 4.3a, and the corresponding ROC curve is shown in Figure 4.3b. Figure 4.3b also shows sample ROC curves generated by sampling the score density functions, and are seen to vary about the true analytical ROC curve.

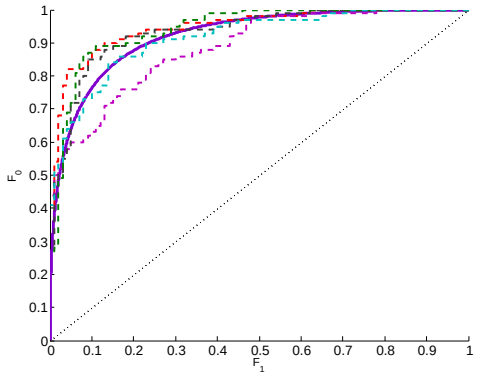
Given a ROC curve, we can sample this directly instead of using the score densities, by sampling across the rate. The gradient $g = \frac{f_0}{f_1}$ on the ROC curve is the class likelihood ratio, from which we can calculate the class probabilities at this point on the curve. The

Score-first:	Rate-first
Repeat n times: Sample score $s_j \sim f$ Sample label $y_j \sim bernoulli(\pi_{0,s_j})$ Rank labels by score s_j	Repeat n times: Sample rate $r_j \sim uniform(0,1)$ $\pi_{0,r_j} \leftarrow$ calculated from gradient at r_j on ROC curve Sample label $y_j \sim bernoulli(\pi_{0,r_j})$ Rank labels by rate r_j and introduce ties due to linear ROC segments

Table 4.3: Two sampling approaches. Left: score-first approach. Right: rate-first approach. π_{0,s_j} is the probability of a positive at score s_j (which is the score at position j in a ranking).



(a) Example score PDFs, normally distributed. Positive class (green) with $\mu_0 = 0$, $\sigma_0^2 = 0.5$; negative class (blue) with $\mu_1 = 1$, $\sigma_1^2 = 0.5$



(b) Analytical ROC curve (purple, solid) and 5 empirical ROC curves (dashed) generated from sampling the score distributions

Figure 4.3: Example analytical score densities with corresponding analytical ROC curve, and example empirical ROC curves generated by sampling the true score densities.

probability of a positive at rate r_j, denoted π_{0,r_j}, is given by:

$$\begin{aligned} \pi_{0,r_j} &= \frac{\pi_0 f_0}{\pi_0 f_0 + \pi_1 f_1} \\ &= \frac{\pi_0 g}{\pi_0 g + \pi_1} \end{aligned} \tag{4.6}$$

where rate r_j is the rate at position j in a ranking.

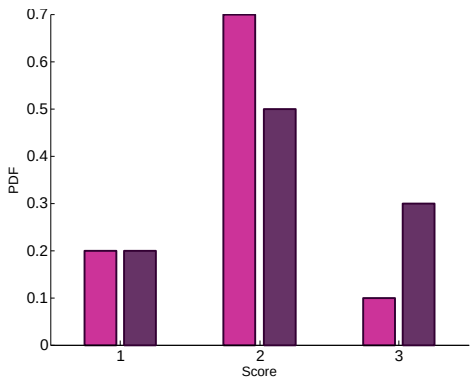
We can then sample a label using this class probability. We do not need to know the scores because the rate also determines the order of the examples in the ranking, and the ROC curve determines the class probabilities at each rate. We do, however, need to know which examples were sampled from the same linear segments of a ROC curve, as these samples should be tied in the generated ranking, rather than ordered by rate. We call this the 'rate-first' approach, given in Table 4.3 (right). This approach is similar to inverse transform sampling where a value of $CDF(x)$ is sampled uniformly (between 0 and 1), which infers a score, and then the class distribution at this score can be used to sample a label. This is because (as introduced in Section 2.4.2) the rate is the cumulative distribution of the mixed probability distribution.

An example of inverse transform sampling is shown in Figure 4.4, for a discrete

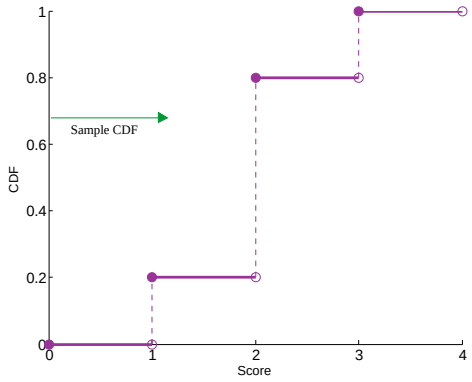
distribution with three score values. In this example the CDF is sampled at 0.68 which samples $score = 2$. Sampling uniformly across rates corresponds to sampling uniformly across the CDF (the y-axis of Figure 4.4b), and these approaches are equivalent to sampling the scores according to the mixed probability distribution function. However, in the rate-first approach we can go directly from sampling a rate to sampling a label, without the need to know the score. Tied examples generated by sampling from the same linear segment of a ROC curve corresponds to sampling the same score with the score-first approach or inverse transform sampling. For example, in Figure 4.4 sampling CDF values of 0.6 and 0.4 would both sample $score = 2$ and so these examples would be tied in the generated ranking.

4.5 Generating confidence bounds

In this section we give our approach to generating rate-oriented point-wise confidence bounds. This uses the rate-first sampling approach just described. We begin by describing a simple approach of inferring confidence bounds, used as a baseline in our experiments (described in Section 4.6).



(a) Score probability mass function of binary classes (pink: positive class; purple: negative class)



(b) Score mixed CDF (right continuous with left limits)

Figure 4.4: Illustration of inverse transform sampling. Score 2 is sampled, and then a class is sampled with probability 0.35 and 0.25 (assuming a uniform class distribution).

4.5.1 Baseline method

We use a simple parametric approach as a baseline method. This method is similar to previous approaches such as vertical averaging, but we fix the rate (the predicted positive rate) rather than the false positive rate, in line with our aims. We calculate the mean and variance of recall across samples and, after making an assumption of the underlying distribution across the ROC points of each sample at each rate, calculate the 95% confidence intervals. Here we use positive recall as a distance measure along rate isometrics in ROC space, but any metric that varies linearly along rate isometrics could also be used (such as negative recall or accuracy). An example is given in Figure 4.5.

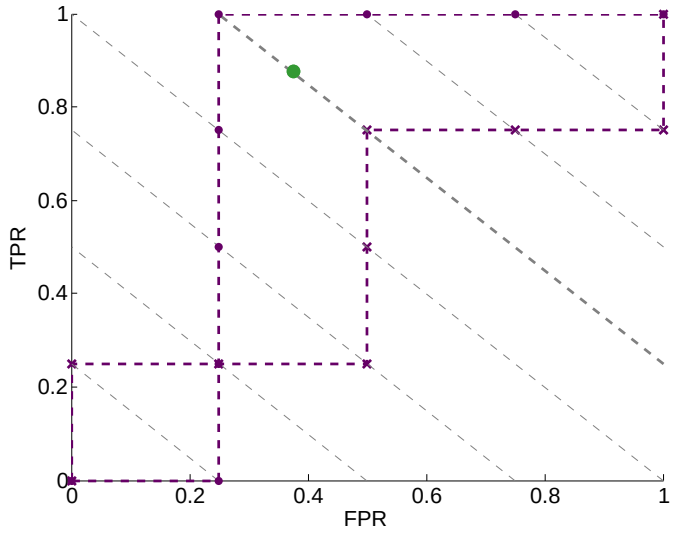


Figure 4.5: Illustration of baseline approach to generating confidence bounds. The rate $\frac{5}{8}$ is indicated by the emboldened rate isometric. The recall of the two curves at this rate is 1.0 and 0.75, and the mean recall is 0.875 (indicated by the green point). The variance is then calculated as $\frac{1}{2-1}\left[(0.75-0.875)^2+(1-0.875)^2\right]=0.0313$.

The variance across samples at position k along the ranking is given by:

$$sd_k^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(tpr_{i,k}-\overline{tpr}_k\right)^2 \tag{4.7}$$

where m is the number of samples, $tpr_{i,k}$ is the recall for sample i at position k and $\overline{tpr}_k$ is the mean recall across the samples, at position k. The standard error (the standard

deviation of mean sample recall) at each position k along the ranking is given by:

$$\sigma_k = \frac{sd_k}{\sqrt{m}} \tag{4.8}$$

To be clear, we are interested in the sample mean because we would like an interval for future consensus curves generated from a set of m ROC samples, rather than the individual ROC samples themselves.

In order to infer a confidence interval we need to assume a particular distribution across the recall of consensus curves at each position k. Assuming a normal distribution the confidence intervals are given by $\overline{tpr}_k \pm 1.96 \cdot \sigma_k$.

We also test this method using a beta distribution, which is bounded by $[0,1]$ such that we can constrain our confidence intervals to the bounds of ROC space. The bounds of F_0 at each rate, as introduced in Section 3.3 are given by:

$$F_{0,k,min} = max\left(0, \frac{r-\pi_1}{\pi_0}\right) \tag{4.9}$$

$$F_{0,k,max} = min\left(1, \frac{r}{\pi_0}\right) \tag{4.10}$$

To use the beta distribution we rescale, at each position k, each sample recall value from the range $max\left(0, \frac{r-\pi_1}{\pi_0}\right) \ldots min\left(1, \frac{r}{\pi_0}\right)$, where $r = \frac{k}{n}$ is the rate at k, to the range $0 \ldots 1$. We calculate the mean and standard deviation of these scaled recall values at each position k, directly from the mean and variance calculated on the unscaled recall values:

$$\mu_{beta,k} = \frac{\bar{F}_{0,k} - F_{0,k,min}}{F_{0,k,max} - F_{0,k,min}} \tag{4.11}$$

$$\sigma^2_{beta,k} = \frac{\sigma^2_k}{(F_{0,k,max} - F_{0,k,min})^2} \tag{4.12}$$

The parameters α_k and β_k are then calculated using standard definitions of the mean and variance of a beta distribution:

$$\mu_{beta,k} = \frac{\alpha_k}{\alpha_k + \beta_k} \tag{4.13}$$

$$\sigma^2_{beta,k} = \frac{\alpha_k \beta_k}{(\alpha_k + \beta_k)^2(\alpha_k + \beta_k + 1)} \tag{4.14}$$

Such that:

$$\alpha_k + \beta_k = \frac{\mu_{beta_k}(1 - \mu_{beta_k})}{\sigma^2_{beta,k}} - 1 \tag{4.15}$$

$$\alpha_k = \mu_{beta,k} \cdot (\alpha_k + \beta_k) \tag{4.16}$$

$$\beta_k = \frac{\alpha}{\mu_{beta_k}} - \alpha_k \tag{4.17}$$

The lower and upper recall confidence intervals are then found in the range 0 to 1 by using this beta distribution to find the values where $cdf = 0.025$ and $cdf = 0.975$ respectively. We then rescale this back to the original scale:

$$F_0(x,k) = x \cdot (F_{0,k,max} - F_{0,k,min}) + F_{0,k,min} \tag{4.18}$$

4.5.2 Overview of the rate-oriented point-wise confidence bounds approach

We assume a random process that generates ROC tables of size $n \cdot m$ from the score probability distributions. Let us denote by $S_{i,k}$ the random variable of the sum of the number of positives from position 1 to position k. Formally, for any fixed true positive value s at this position, with n_0 and n_1 all fixed, we want to estimate:

$$p\left(S_{i,k} = s \,|S_{i,n} = n_0\right) = \frac{p\left(S_{i,k} = s, S_{i,n} = n_0\right)}{\sum_{s'} p\left(S_{i,k} = s', S_{i,n} = n_0\right)} \tag{4.19}$$

We condition on the class distribution to reflect the fact that a data sample has a finite number of examples with a certain number of each class. This also corresponds to the fact that ROC curves must pass through the points $(0,0)$ and $(1,1)$. We present two different methods, a parametric and a bootstrap approach. We derive the probability distribution across the number of positives up to a position, k, in a sample, and use this to infer these two approaches. The bootstrap approach is particularly useful where the distributional assumptions of the parametric approach are invalid.

Importantly, our approach is naturally invariant to swapping the classes. In ROC space when classes are swapped the x-axis becomes the false negative rate $(1 - tpr)$,

and the y-axis becomes the true negative rate ($1-fpr$). The rates in the swapped space are given by $r'(t) = \pi_0(1-tpr) + \pi_1(1-fpr) = 1-r(t)$. Hence, for each set of points along a rate isometric in the original space, there is a corresponding rate isometric in the 'swapped' space along which this set of points also lie. The confidence bounds along these corresponding rate isometrics will have equivalent confidence intervals.

4.5.3 Parametric approach

We find the probability distribution across the number of positives from the first position to a position k in the ranking, $S_{i,k}$. We first derive an analytical solution (Theorem 4.1), and then provide an empirical version that can be computed directly using the ROC curve. At this point we fix i as we refer only to a single sample, so that $S_{i,k}$ is denoted S_k and $S_{i,n}$ is denoted S_n. To be clear, we are using a 'true' ROC table / consensus ROC curve, and sample this to generate a new sample ranking (one row of a new ROC table). After this, we give an extension to generate whole sample ROC tables.

Theorem 4.1. Let the score densities, f_0 and f_1, and the number of examples of each class in the sample, n_0 and n_1, be fixed. Then:

$$\begin{aligned} &p(S_k = s, S_n = n_0) \\ &\quad = \int_0^1 \left[binom(s, k-1, \pi_0^{<r'}) \cdot (1-\pi_0^{=r'}) + binom(s-1, k-1, \pi_0^{<r'}) \cdot (\pi_0^{=r'}) \right] \\ &\qquad\qquad \cdot binom(n_0 - s, n-k, \pi_0^{>r'}) \cdot p(R_k = r')dr' \end{aligned} \tag{4.20}$$

where

$$\pi_0^{<r'} = \frac{\pi_0 F_0(t)}{\pi_0 F_0(t) + \pi_1 F_1(t)} \tag{4.21}$$

$$\pi_0^{>r'} = \frac{\pi_0(1-F_0(t))}{\pi_0(1-F_0(t)) + \pi_1(1-F_1(t))} \tag{4.22}$$

$$\pi_0^{=r'} = \frac{\pi_0 f_0(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)} \tag{4.23}$$

$t = F^{-1}(r)$, $p(R_k = r) = beta(r, k, n-k+1)$, R_k is the random variable denoting the rate from which the example at position k was sampled. r' is the rate that is sampled whereas

k is the position in the new sample, such that r' has a corresponding k' where $r' = \frac{k'}{n'}$ and k has a corresponding r where $r = \frac{k}{n}$. $binom(k_b, n_b, p_b)$ is the binomial distribution for k_b successes in n_b trials, with probability of success p_b, and $beta(x, a, b)$ is the probability of value x for beta distribution with $\alpha = a$ and $\beta = b$.

Proof. To compute the left hand side of Equation 4.19 it is sufficient to compute:

$$p\left(S_k = s, S_n = n_0\right) \tag{4.24}$$

The probability of $S = s$ and $S_n = n_0$ in the new sample depends on which rate it was sampled from, such that:

$$p\left(S_k = s, S_n = n_0\right) = \int_0^1 p\left(S_k = s, S_n = n_0 \mid R_k = r'\right) \cdot p(R_k = r')dr' \tag{4.25}$$

The order statistic states that when sampling n values uniformly within the range 0..1 and sorting these examples, the probability that an example at position k was sampled from a rate r' is beta distributed with $\alpha = k$ and $\beta = n - k + 1$ [92]. Therefore, $p(R_k = r')$ of Equation 4.25 is the beta density.

The other component of Equation 4.25 is the probability of s positives up to a position k, given the example at this position is sampled from a particular rate r'. There are two cases where value s is the number of positives up to a position k: 1) $s-1$ positives occur before position k and the example at k is a positive, or 2) s positives occur before position k and the example at position k is a negative. In either case there must also be $n_0 - s$ positives after position k to ensure that the class distribution is correct.

The examples before position k can be sampled independently, with probability of a positive given by Equation 4.21. The examples after position k can also be sampled independently, with probability of a positive given by Equation 4.22. The independence between samples is valid because we are sampling a set of *unordered* examples. Intuitively, the position of a particular point on a ROC curve is independent of the order of the examples that preceeds it, it is only the number of positives and negatives that matters. This means that the probabilities of the set of examples before and after position k are binomially distributed, which infers:

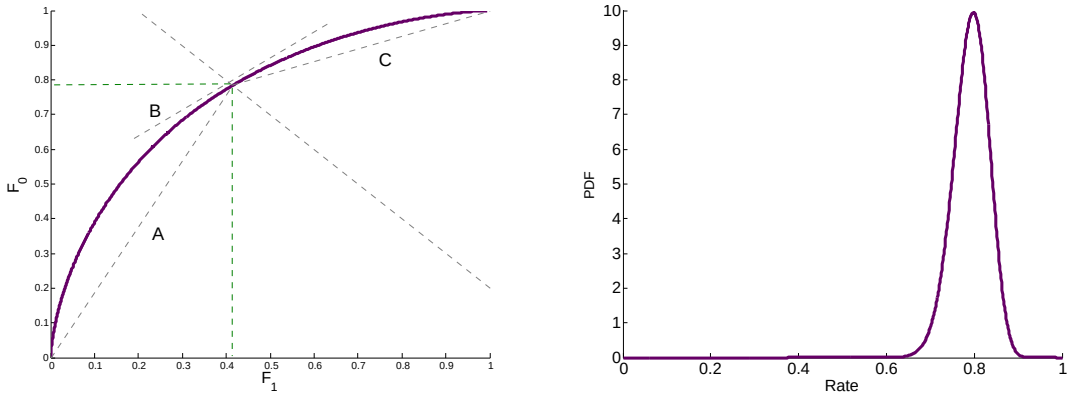
$$
\begin{aligned}
& p\left(S_k = s, S_n = n_0 | R_k = r'\right) \\
&= \left[p\left(\sum_{i=1}^{k-1} (1-y_i) = s \right) p(y_k = 1) + p\left(\sum_{i=1}^{k-1} (1-y_i) = s-1 \right) p(y_k = 0) \right] \\
&\qquad \cdot p\left(\sum_{i=k+1}^{n} (1-y_i) = n_0 - s \right) \qquad (4.26) \\
&= \left[binom(s, k-1, \pi_0^{<r'}) \cdot (1 - \pi_0^{=r'}) + binom(s-1, k-1, \pi_0^{<r'}) \cdot (\pi_0^{=r'}) \right] \\
&\qquad \cdot binom(n_0 - s, n-k, \pi_0^{>r'})
\end{aligned}
$$

Using Equation 4.26 in Equation 4.25 concludes the proof. □

To reiterate a key point – an example at position k has rate r for a newly sampled ROC table, and we can imagine this table is sampled from an initial ROC curve. The rate r' from which it is sampled on the 'true' ROC curve is probabilistic, corresponding to $p(R_k = r')$ in Equation 4.20. The class probabilities used to generate the example at position k are determined by the class distribution at the rate r' from which this example was sampled.

Imagine, for instance, that we wish to generate a ranking by sampling the ROC curve in Figure 4.6a. Imagine also that we are now sampling the example at position 16 of 20 such that $r = 0.8$. Figure 4.6b shows the beta distribution used to sample a rate, with parameters $\alpha = k = 16$ and $\beta = n - k + 1 = 5$, and mode 0.8. From this distribution rate $r' = 0.6$ (shown on Figure 4.6a) may be sampled, and it is this position on the ROC curve to which the probabilities before, at and after of Equations 4.21 - 4.23 refer.

An important aspect of Theorem 4.1 is that the sampling probabilities before, at and after rate r' (Equations 4.21 - 4.23) can be computed solely using the ROC curve (assuming the class distribution is also known). Recall from Section 4.4 that Equation 4.23 can be calculated from the gradient at r' on the ROC curve. We can also infer the values of Equations 4.21 and 4.22 from the ROC curve. Equation 4.21 is equivalent to the average probability of sampling a positive across all rates before r', and this can be inferred from the gradient of the straight line from point $(0,0)$ to the point at r' on the ROC curve (shown in Figure 4.6a). Similarly, Equation 4.22 can be inferred from



(a) ROC curve from which new ranking is sampled

(b) Distribution across rates, $p(R_k = r)$

Figure 4.6: Illustration of rate sampling. Sampling position $k = 16$ of a sample with 20 examples, such that the mode of Beta distribution is 0.8. This distribution is used to sample the original ROC curve (left), at rate of 0.6. The gradient at this rate is used to sample the 16th position in the new sample. Line A has gradient $\frac{F_0}{F_1}$. Line B has gradient $\frac{f_0}{f_1}$. Line C has gradient $\frac{1-F_0}{1-F_1}$.

the gradient of the straight line from the ROC curve point at r' to the point $(1,1)$ (also shown in Figure 4.6a). For example, for the rate sampled on the ROC curve in Figure 4.6a $F_0(t) \approx 0.77$ and $F_1(t) \approx 0.42$, and the gradient of the straight line connecting this point with the origin is ≈ 1.8 and this infers $\pi_0^{<r'}$.

Theorem 4.1 gives the analytical calculation but we cannot use this directly in practice, as we have empirical ROC curves / ROC tables rather than the score densities. Firstly, our empirical ROC tables have discrete rates such that in the discrete case the integral of Equation 4.25 is changed to a summation. We implement this as an average of the joint probability, for a set of rates of the CDF of the beta distribution (the sampling distribution for this k) at each 0.01 interval:

$$p\left(S_k = s, S_n = n_0\right) = \sum_{t=1}^{99} p\left(S_k = s, S_n = n_0 \mid R_k = F_{beta}^{-1}(0.01 \cdot t)\right) \cdot p(R_k = F_{beta}^{-1}(0.01 \cdot t)) \tag{4.27}$$

such that we sample the rates at each 0.01 interval of the CDF of the beta distribution

(with $\alpha = k$ and $\beta = n - k + 1$). This CDF models the probability that an example at position k is sampled by each rate (according to the order statistic), such that the rates that are more likely to be sampled are given more weight when estimating this probability.

We also require discrete versions of Equations 4.21- 4.23 that can be used with an empirical ROC table. That is, when we have a ROC table from which we would like to sample in order to create a new ROC table. These are given in Equations 4.31- 4.33:

$$\pi_0^{<r'} = \frac{1}{r' \cdot n \cdot m} \left[s_{\lfloor r' \cdot n \rfloor} + d \cdot pos_{\lceil r' \cdot n \rceil} \right] \tag{4.28}$$

$$\pi_0^{=r'} = \frac{1}{m} pos_{\lceil r' \cdot n \rceil} \tag{4.29}$$

$$\pi_0^{>r'} = \frac{1}{(1-r') \cdot n \cdot m} \left[m \cdot n_0 - s_{\lceil r' \cdot n \rceil} + (1-d) \cdot pos_{\lceil r' \cdot n \rceil} \right] \tag{4.30}$$

where $d = r' \cdot n - \lfloor r' \cdot n \rfloor$ is the relative distance of the rate between positions $\lceil r' \cdot n \rceil$ and $\lfloor r' \cdot n \rfloor$. These probabilities could be retrieved from the rate-averaged consensus ROC curve or from the ROC table.

For example, Table 4.4 shows an example ROC table with 2 sample rankings of length 4 where $\pi_0 = 0.5$. Imagine we are sampling this ROC table to generate a new sample ranking, and we are currently generating the example at position $k = 2$. We sample a rate $r' = 0.3$ using the beta distribution with $\alpha = k = 2$ and $\beta = n - k + 1 = 3$. As there are four examples in each ranking, the examples reside at the rates $\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ and 1. This means that the rate $r' = 0.3$ is part way between the first and second examples, and this distance is calculated as $d = r' \cdot n - \lfloor r' \cdot n \rfloor = 0.3 \cdot 4 - \lfloor 0.3 \cdot 4 \rfloor = 0.2$. Therefore:

$$\pi_0^{<r'} = \frac{1}{0.3 \cdot 4 \cdot 2} \left[2 + \frac{1}{5} \cdot 1 \right] = \frac{11}{12} \tag{4.31}$$

$$\pi_0^{=r'} = \frac{1}{2} \cdot 1 = \frac{1}{2} \tag{4.32}$$

$$\pi_0^{>r'} = \frac{1}{(1-0.3) \cdot 4 \cdot 2} \left[2 \cdot 2 - 3 + \left(1 - \frac{1}{5} \right) \cdot 1 \right] = \frac{9}{28} \tag{4.33}$$

k	1	2	3	4
Sample 1	0	0	1	1
Sample 2	0	1	0	1
r'_k	0.25	0.5	0.75	1
s_k	2	3	4	4
pos_k	2	1	1	0

Table 4.4: Example ROC table with two samples and 4 examples per sample.

Extension to generate a whole ROC table

The probabilities of each S_k value computed in Theorem 4.1 correspond to only a single row of the ROC table. We need the distribution across the number of positives up to position k of all samples in the ROC table. We now use $S_{1,k}$ to refer to the number of examples from position one to k in one new sample row of a ROC table. We use S_k to refer to the number of positives from position 1 to position k in the new sample, across all rows in the ROC table. For each S_k value we need:

$$p\left(S_k = s \,|\forall i \in 1\ldots m : \sum_{j=1}^{n}(1-y_{i,j}) = n_0\right) \tag{4.34}$$

Computing this exactly is computationally intractable, as for each possible s at a position k the probability is given as the summation of the probabilities of all possible combinations of values at position k that sum to this value. We instead approximate the confidence intervals using the estimated variance of this distribution. The mean and variance of the distribution of one sample up to position k are given by:

$$\mu_{1,k} = \sum_s p\left(S_{1,k} = s \,|S_{1,n} = n_0\right) \cdot s \tag{4.35}$$

$$\sigma^2_{1,k} = \sum_s p\left(S_{1,k} = s \,|S_{1,n} = n_0\right) \cdot (s-\mu_{1,k})^2 \tag{4.36}$$

where 1 denotes that these functions correspond to a single sample. We assume each row is identically distributed such that the mean and variance of s at position k of the ROC table are given by:

$$\mu_k = \sum_{i=1}^{m} \mu_{i,k} = m \cdot \mu_{1,k} \quad (4.37) \qquad \sigma_k^2 = \sum_{i=1}^{m} \sigma_{i,k}^2 = m \cdot \sigma_{1,k}^2 \quad (4.38)$$

At each k we restrict to only the possible values of S_k, rescale these to between zero and one, and use a scaled beta distribution to model this distribution and estimate the confidence intervals. We calculate the mean and variance across S_k values at each position k, where the S_k values have been rescaled to the range $[0,1]$:

$$\mu_{k,\beta} = \frac{\mu_k - minS_k}{maxS_k - minS_k} \quad (4.39)$$

$$\sigma_{k,\beta}^2 = \frac{\sigma_k^2}{(maxS_k - minS_k)^2} \quad (4.40)$$

where $maxS_k = m \cdot maxS_{1,k}$ and $minS_k = m \cdot minS_{1,k}$ and:

$$minS_{1,k} = max(0, n_0 - n + k) \quad (4.41) \qquad maxS_{1,k} = min(k, n_0) \quad (4.42)$$

We use these to parameterise a beta distribution and infer a confidence interval, which we then rescale to the original scale.

4.5.4 Bootstrap approach

We now introduce a rate-oriented point-wise confidence bounds approach using bootstrapping. This is the bootstrapped equivalent of the parametric approach just described. Bootstrap samples are generated using rate-sampling, and the upper and lower bounds are set at each rate such that they contain 95% of the bootstrapped ROC curve samples.

We generate 2,000 bootstrapped ROC tables each with m samples. Each sample is generated independently using the rate-first sampling approach, as follows.

The rates are sampled uniformly and sorted:

$$r_1, r_2 \ldots r_n \xrightarrow{sort} r_{(1)}, r_{(2)} \ldots r_{(n)} \quad (4.43)$$

The probability distribution at each rate is found by:

$$\pi_{0,r} = \frac{1}{m} pos_{\lceil r \cdot n \rceil} \quad (4.44)$$

We then use this probability to generate a label at k:

$$l_k \sim binom(\pi_{0,r}) \tag{4.45}$$

In this way we generate a set of 2,000 bootstrap ROC tables (generating $2,000 \cdot m$ samples in total).

This sampling procedure does not ensure that each sample has the correct class distribution. This is needed so that the confidence intervals generated from these samples reflect that at rates 0 and 1 we are certain the curve passes through the points $(0,0)$ and $(1,1)$ in ROC space, respectively. A simple approach to restrict to a fixed class distribution discards all samples where the class distribution is not correct. However, this approach is only feasible when the number of examples is low, as otherwise samples are rarely generated with the correct class distribution and this method becomes too slow.

We propose another approach that can be used with a larger number of examples, where we adjust the rate and the number of true and false positives at each position in order to correct the class distribution. We call this the rate-adjustment approach. The rates of the bootstrap ROC tables are equally distributed along the ranking, as shown in Figure 4.7. For each sample individually we adjust these rates and the true positive values at each position, by scaling each position according to a correction factor, a value for each sample and class that rescales the 'width' of each example in the ranking to correct the class distribution. This adjustment is illustrated in Figure 4.7, and shows how the effect is to stretch or narrow the examples along the ranking.

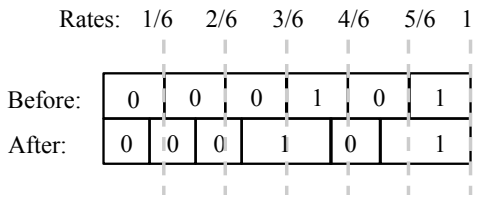


Figure 4.7: Illustration of rate adjustment to correct class distribution.

We use the bootstrapped ROC tables with the corrected true positive values, to estimate the confidence bound of the true ROC curve. For each ROC table, and at each

position k along the ranking, we calculate the average recall across the samples:

$$\overline{tpr}_k = \frac{1}{m \cdot n_0} \sum_{i=1}^{m} s_{i,k} \tag{4.46}$$

Each position k in the ranking has a set of average recall values, one for each sample ROC table. This now corresponds to the probability density function we stated in Equation 4.19. The proportion of bootstrap ROC tables with recall value between $\overline{tpr}_k$ and $\overline{tpr'}_k$ gives an estimate of the probability that the recall at this position is between these values, given this sample has a particular class distribution.

The confidence interval for position k is obtained from the mean recall values, $\overline{tpr}_k$, of the bootstrapped ROC tables as follows. For each position k we take the $\overline{tpr}_k$ value of each ROC table, sort these values in ascending order, and select the 2.5% and 97.5% percentiles as the lower and upper endpoints of the 95% confidence interval. This gives a series of recall-rate pairs for the lower and upper limits of the confidence interval at each position k. A confidence bound can be created by interpolating between these points.

4.6 Experiments

The following experiments use a known ROC curve to generate samples for which we create confidence bounds, specified by normally distributed score density functions with mean 0 and 1 for the positive and negative class respectively, and a variance of 1. These score distributions, and the corresponding ROC curve are shown in Figure 4.8. Our tests use ROC tables with 10 samples and 50 examples per sample.

We evaluate whether the generated confidence intervals meet our aims, where at significance level σ new samples generated from this consensus curve pass between the lower and upper confidence limits at a given rate value, with probability $1 - \sigma$. Given a single sample ROC table and its confidence bounds, we generate 1,000 new sample ROC tables from this sample. We count, at each rate, the number of consensus curves (of these samples) the confidence interval contains. A true 95% confidence interval at a given rate, should contain the consensus curve of new samples 95% of the time.

The results are shown in Figure 4.9. The results of the baseline parametric ap-

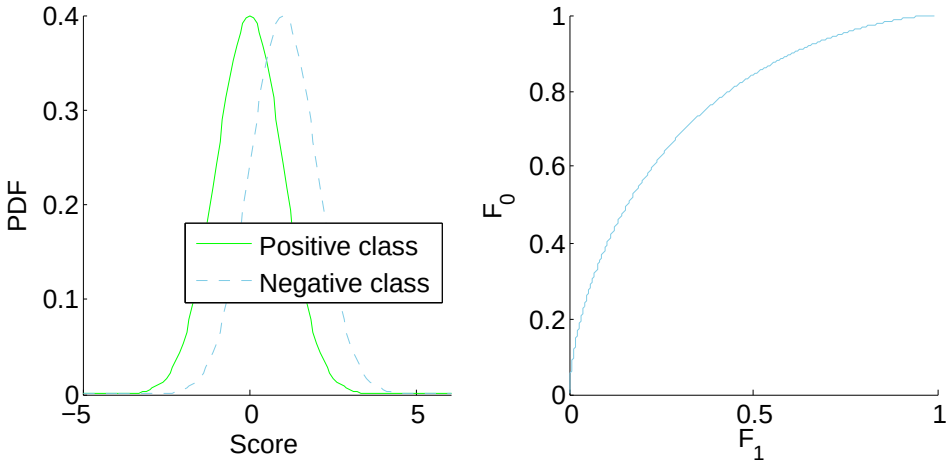
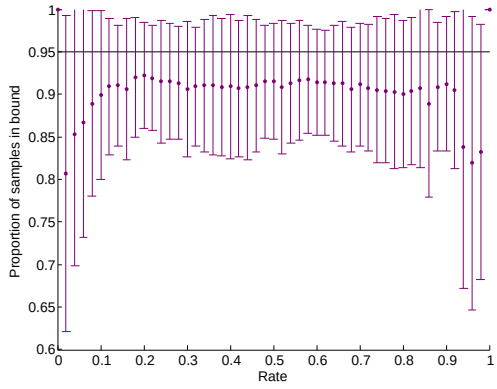


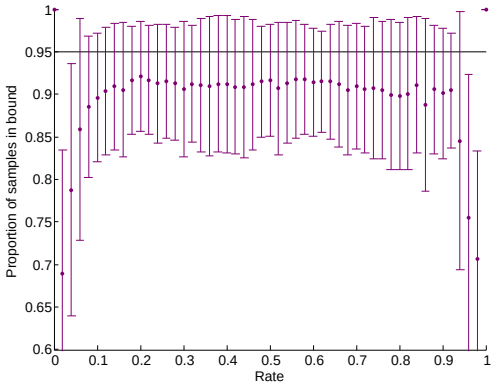
Figure 4.8: Score probability densities for two classes (positive class: $\mu = 0$, $\sigma^2 = 1$; negative class: $\mu = 1$, $\sigma^2 = 1$), and corresponding ROC curve.

proaches (Figures 4.9a and 4.9b) are highly variable. The rate-oriented point-wise parametric approach (Figure 4.9c) reliably generates confidence bounds with close to 95% confidence, except at the extremes. This indicates that the assumption that the number of positives up to a particular position in the ranking is beta distributed is not valid at the extremes (only).

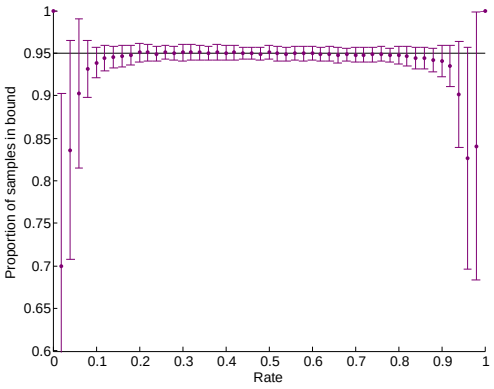
The rate-oriented point-wise bootstrap approaches are also much more effective compared to the baseline results (Figures 4.9d-4.9f). They are a little conservative, particularly at the extremes of the distribution, due to the nature of bootstrap sampling, where the variation between bootstraps may be too low to calculate strict confidence intervals (for instance, where the lower and upper bounds of the 95% limits are the same as those for the 94 or 96% limits). For example if a bootstrap sample contained only one example then the values at the 95% bounds would also be the same values as for the 0% or 100% limits. This also justifies the shape of the graph in Figure 4.9f. Here the variance of the score distribution of the negative class is reduced to 0.2. This increases the range of rates near to rate zero where the bounds are conservative because for these rates the probability of a positive is near to one so there is little variation in recall across the bootstrap samples. We also note that the rate-adjustment approach to correcting the class distribution produces an interesting relationship between the rate and the proportion of samples in bound, as shown in Figure 4.9e, and this needs further investigation.



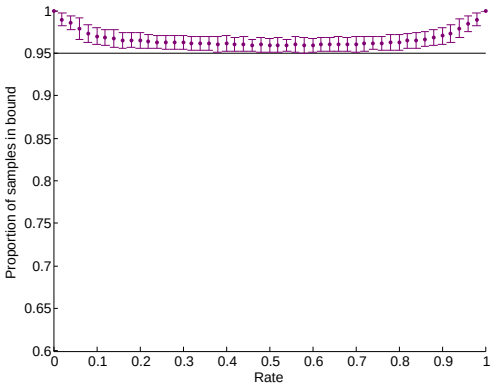
(a) Results of baseline with normal assumption



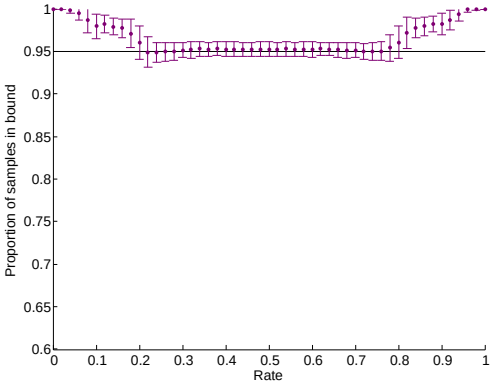
(b) Results of baseline with beta assumption



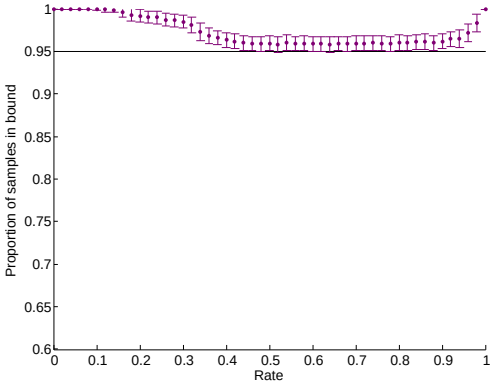
(c) Results of parametric approach



(d) Results of bootstrap approach (using discarding approach to correct class distribution)



(e) Results of bootstrap approach (using rate-adjustment approach to correct class distribution)



(f) Results of bootstrap approach (using discarding approach to correct class distribution) for score distributions with: $\mu_0 = 0$, $\sigma_0^2 = 1$, $\mu_1 = 1$, $\sigma_1^2 = 0.2$

Figure 4.9: Mean (standard deviation) of the proportion of 1000 new samples (sampled from ROC table) within confidence interval at each rate, across 100 tests.

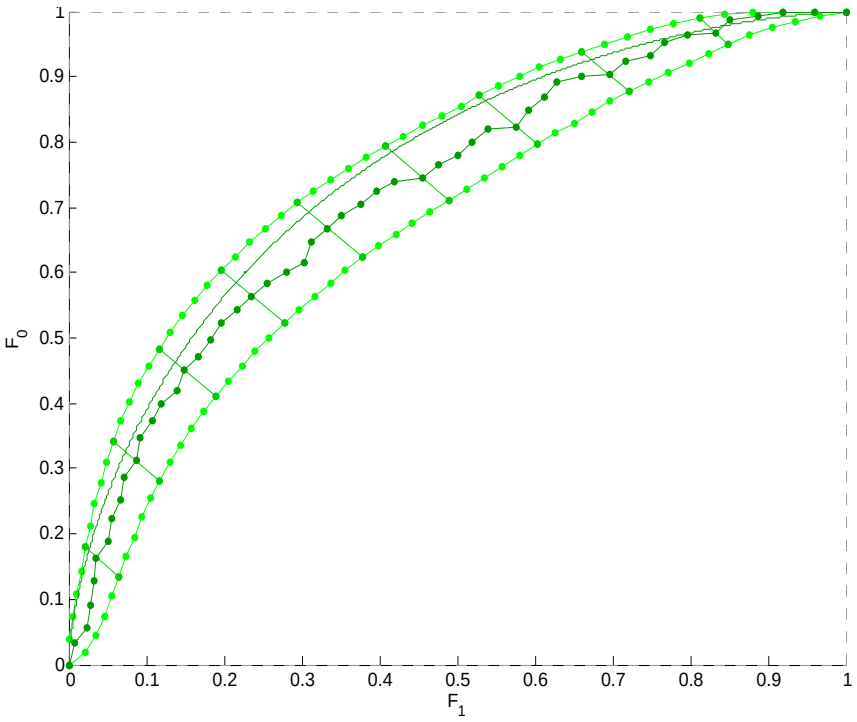


Figure 4.10: Example ROC curve and confidence bounds using rate-oriented confidence bounds analytical approach. Each point on the upper and lower bound shows the upper and lower point of the confidence interval at a particular rate. We show the confidence intervals at a selection of rates shown for illustration. We have interpolated between the points to show where the point-wise bounds are likely to sit at rates for which the confidence intervals were not calculated.

Figure 4.10 shows an example ROC curve generated using our parametric approach. The equivalent rate-recall curve is shown in Figure 4.11, with vertical confidence intervals.

4.7 Related work

In Section 4.1 we discussed two parametric approaches to generating confidence bounds – vertical (or horizontal) and threshold averaging. Here we discuss other proposed approaches.

A non-parametric approach called fixed width bands [93, 94] works by displacing the whole ROC curve up and left, and down and right, to create an upper and lower

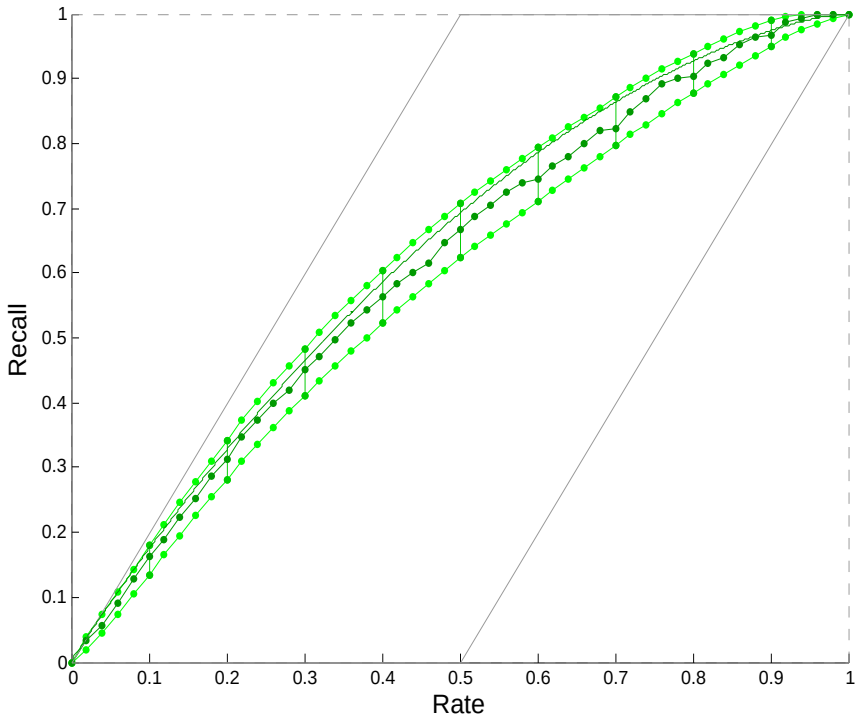


Figure 4.11: Equivalent rate-recall curve [27] and confidence bounds for ROC curve shown in Figure 4.10. Grey lines indicate bounds of rate-recall space.

confidence band respectively. The curve is displaced along the gradient $-\sqrt{(}n_0/n_1)$ (chosen as an approximation of the standard deviation ratios of the two classes). Rate isometrics have a gradient $-n_0/n_1$ such that if we changed the displacement gradient to the gradient of the rate isometric this could be used as a rate-oriented approach. However, the size of displacement is constant along the ROC curve which does not constrain the confidence bounds to ROC space. Furthermore, this is an approach for calculating the confidence around the whole curve, but in this work we are interested in point-wise confidence bounds instead.

Tilbury et al. suggest a non-parametric approach, derived from first principles [95]. Their approach divides the whole ROC space into a grid and calculates the probability that each cell contains the true ROC curve, given the sample curve that has been provided. A confidence bound can then be inferred by enclosing all cells that have a probability above a specified inclusion threshold. Another non-parametric approach uses kernel methods to generate point-wise bounds [96].

It is also interesting to mention methods that have been proposed specifically in the medical field, to generate confidence bounds on ROC curves to evaluate diagnostic tests [97]. As already mentioned in Section 3.4, diagnostic tests assess how well a particular test is able to detect a disease. If the test outputs a binary value indicating the prediction (disease / no disease), then the results of the diagnostic test give the number of tests where participants were correctly identified with or without the disease, or incorrectly identified with or without the disease. These are the numbers of true positives, true negatives, false positives and false negatives respectively. Hence, each diagnostic test (with a binary test value) infers a point on the ROC curve.

When several studies have performed the same diagnostic tests a meta-analysis may be performed. This includes inferring the average (consensus) point on the ROC curve and a confidence bound around this point, referred to as a meta-analysis in ROC space. Two common methods to do this are the hierarchical summary ROC model and the bivariate random effects model [97]. These approaches are point-wise, used for binary diagnostic tests that infer a single point in ROC space. As with standard meta-analyses, a meta-analysis in ROC space also needs to take into account the precision of the estimates and the between study heterogeneity (differences in the estimates due to differences in study design). For instance, a diagnostic test result with greater precision should be given more weight in a meta-analysis compared to those with lower precision. Hence, while these approaches are point-wise like ours, they involve estimation techniques specifically designed for the meta-analysis of diagnostic tests.

4.8 Summary

We have described a new approach to generate confidence bounds, which we call rate-oriented point-wise confidence bounds. To our knowledge there is no approach in the literature to infer rate-oriented confidence bounds. Macskassy el al. [98] claim that rate-averaging makes the strong assumption that the rates are estimating the same point in ROC space, and this is not appropriate. However, other approaches make similar assumptions across a different metric, such as the false positive rate in vertical-averaging.

Our main aim was to address some important weaknesses of other existing methods. Calculating the consensus and confidences bounds at each rate is practical as rate is a measure over which we have control in practice. On the other hand, vertical (or

horizontal) averaging fix the false positive rate (true positive rate) and average across the true positive rate (false positive rate), but these metrics are not under our control so are of little use in practice. Score-averaging creates confidence bounds around clouds of points, and how best to do this is an open problem. Rate-averaging does not have this problem because it constrains to a single dimension.

Our approach is also invariant to swapping the classes, and we suggest that this property is sensible when generating confidence bounds. The confidence of a point on the ROC curve should not depend on which class is labelled as positive. Furthermore, our bounds have the advantage that they are smooth, due to the sampling across rates we perform as part of our method.

Our secondary aim was to find appropriate bounds for assessing models used specifically for rate-oriented tasks. Using a rate-oriented approach ensured that the performance (and confidence interval) shown at a rate is an estimate for this particular rate.

We have analytically derived the probability distribution of the number of positives up to each position in the ranking, and then used this to develop two methods, a parametric and a bootstrap approach. The parametric approach gave confidence bounds having very close to the 95% confidence, except at the extremes. The bootstrap approach did generate satisfactory bounds at the extreme but also had greater variance around the 95% confidence level. Therefore, we suggest that when the performance at the extremes of the ROC curve are of little importance, the parametric approach should be used, but where this is not the case the bootstrap approach can be used instead.

The ROC tables used as input to this confidence bounds method are composed of a set of rankings, each generated by evaluating a model using an independent data sample. Ideally, assuming we are evaluating a learning algorithm, the data used to both train and test a set of models should be independent. A set of m models are trained using m independent training samples of size p, and tested using m independent test samples of size q. The ROC table is then an m by q table containing the rankings of the test samples. This would give a valid estimate of the performance of a learning algorithm trained on a set of p independent samples and then tested on a set of q independent test samples. Typically, the amount of data available is limited, such that it is not possible to use independent training sets, and in this case cross validation is often used. Cross validation has independence of the test sets but the training sets share data and so are not independent. When evaluating a learning algorithm, the confidence bounds esti-

mated using cross validation may be conservative (i.e. narrower) because there is less variation in the training data across the folds, than would occur if the training data were independent. We may expect the variability in the models to decrease with increasing number of folds, as this increases the overlap of the training data across folds.

In Chapter 3 we introduced the rate-weighted AUC (rAUC) [27], a general measure where the distribution of weights along the ranking can be chosen for the specific task at hand. We also described existing metrics that all evaluate rankings with respect to the rate including; NDCG [82] in information retrieval, RIE [84], the BEDROC [85], CROC [86] and SLR [87]. When assessing tasks that use these metrics in ROC space, we suggest it is most appropriate to generate rate-averaged consensus curves with rate-oriented point-wise confidence bounds.

In the next Chapter we present our analyses and results for assisting systematic reviews using ranking and classification methods. We show the performance of models in ROC space with rate-oriented point-wise confidence bounds and the associated smooth consensus curves, and compare model performance using these bounds.

Chapter 5

Predicting risk of bias

In this chapter we present our methods and results for the three objectives introduced in Chapter 2. These objectives are to: 1) identify relevant sentences within research articles, 2) rank articles by risk of bias and 3) reduce the number of assessments the reviewers need to perform by hand.

5.1 Statistical and machine learning methods

We use logistic regression to create sentence level and article level models. The sentence level models predict whether each sentence contains relevant information for a risk of bias property. The article level models predict the risk of bias value as described in the text of an article. In line with the domain-based nature of a risk of bias assessment, we implement this individually for each risk of bias property: sequence generation, allocation concealment and blinding. The dependent variable for a sentence level model is the binary variable with values *relevant* or *not-relevant*, indicating whether a sentence contains information relevant to a particular risk of bias property. The dependent variable for an article level model is the binary variable with values *low* or *not-low*, describing whether a particular property has a low risk of causing bias, as described by the contents of the article. Each independent variable is the number of occurrences of a word in a sentence or article respectively (known as a 'bag of words' representation with unigram features).

Logistic regression provides predictions in terms of a score that denotes the proba-

bility of each particular label. We use logistic regression as it has the following attractive features. Firstly, the parameters of logistic regression have a clear interpretation. Given the logistic model $y = 1/(1 + e^{\beta_1 x_1 + \beta_0})$, a one unit increase in the independent variable x_1 corresponds to a β_1 change in the log odds of y. Secondly, logistic regression is known to produce scores that are well calibrated [99]. Scores are calibrated if, for example, given a set of articles that all have a score of 0.8, we can expect 80% of these articles to have a label of *not-low* (assuming a high score denotes more likely to be *not-low*, as specified in Section 2.4.2). This means that we can use these scores as probabilities that an article (or sentence) belongs to a particular class [99, 100].

We used the Weka machine learning package [101] to perform the following preprocessing of the features, commonly performed in text mining tasks. We converted the terms to lower case such that, for instance, the words 'random' and 'Random' correspond to a single parameter in the model. We performed word stemming using Porter's algorithm [102]. This reduces words to the word stem, removing the variable endings of words. For example, the words 'blinded' and 'blinding' are reduced to the same stem, 'blind'. This means that similar words are converted into a single parameter in the model. The frequency of the word stem is the total frequency across all of its word variants and this often improves the estimation of these model parameters.

We removed common words such as 'the' and 'some', known as stop words from the set of features as these are unlikely to be predictive and vastly increase the number of features. We use the standard stop word list in the Weka package. We removed words that occur less than 5 times in the dataset and words of one or two characters in length. All remaining words were included in the models.

We used the Weka machine learning package [101] to learn the logistic regression models with stochastic gradient descent (SGD), using the Weka SGD algorithm. Stochastic gradient descent is an iterative algorithm where the parameters are updated sequentially, in the direction that minimises the log loss of the logistic regression model. Each iteration is called an epoch. There are two main parameters for the Weka SGD algorithm, the learning rate and the number of epochs. The learning rate determines how much the model parameters should change on each update, and the number of epochs states the number of iterations to be performed. For instance if we use 100 epochs then each parameter will be updated 100 times, once in each epoch. For the sentence level learning we use a learning rate of 0.001 and 2000 epochs. For the article level learning

we use a learning rate of 0.0001 and 4000 epochs. We reduce the learning rate and increase the epochs for the article level learning because in general it is preferable to do this where the running time remains feasible. The article level task has a smaller number of examples with which to train each model and so this it is much faster to train these models compared to the sentence level models. The Weka SGD method has an additional parameter λ, used for regularised logistic regression, and as we do not use regularised logistic regression we set $\lambda = 0$.

5.2 Methods and illustrative results

Our methods used to train and evaluate models depends on the specific objectives. We now describe the methods and results for each objective in turn.

5.2.1 Objective 1: Identifying relevant sentences

As described in Section 2.2.2, this objective aims to rank sentences in order of relevance, for each risk of bias property. Each sentence in our dataset is one of three types, with respect to a particular risk of bias property: *relevant*, *not-relevant* or unlabelled. Using this labelling there are two choices of dataset we can use to train the parameters of the logistic regression model. The first dataset uses the sentences known to be *relevant* as positive examples and *not-relevant* as negative examples, and does not include unlabelled sentences. Here we are trying to train a model that can distinguish *relevant* sentences from *not-relevant* sentences. We refer to this as the *relevant/not* labelling approach. The second option is to use the *relevant* sentences as the positive examples, and both the *not-relevant* and unlabelled sentences as the negative examples. Here we would be trying to separate the *relevant* sentences from the rest. We refer to this as the *relevant/rest* labelling approach.

Our aim is to separate *relevant* sentences from *not-relevant* sentences, but it is not clear which dataset is preferable to train a model to do this. While the *relevant/not* data clearly represents the *relevant* versus *not-relevant* notion more appropriately, the *relevant/rest* dataset has the advantage of a much larger sample size (see Table 5.1). Previous work by Marshall et al. [23] used the *relevant/rest* labelling approach to separate *relevant* from *not-relevant* sentences, which assumes (as they note) that the unlabelled

sentences are all *not-relevant*, and this is unlikely to be the case. The *relevant/not* labeling approach assumes that the subset of examples labeled as *relevant* and *not-relevant* are representative of the remaining sentences in the dataset, for which the labels are not known.

We compare the use of the *relevant/rest* and *relevant/not* datasets for separating *relevant* sentences from *not-relevant* sentences. To do this we perform 10-fold cross validation with three different setups, where we train and test on the different labelling approaches. Firstly, we train models with the *relevant/rest* dataset, and test these models also with the *relevant/rest* dataset (test A). We then train models with the *relevant/not* dataset and test these models also with the *relevant/not* dataset (test B). We compare the results using these two datasets, to give an indication of the predictive ability of each dataset. We also train models using *relevant/rest* sentences and test using the *relevant/not* dataset (test C). This allows us to assess how well a model learnt with the *relevant/rest* dataset can separate the *relevant* sentences from the *not-relevant* sentences, even though the unlabelled data are included in the *relevant/rest* dataset. We compare the evaluation on the *relevant/not* dataset, when estimating the parameters with both the *relevant/not* (test A) and *relevant/rest* (test C), to determine which has higher performance when trying to separate *relevant* sentences from *not-relevant* sentences. Another possible test is to train models using *relevant/not* sentences and test using the *relevant/rest* dataset, but since we are interested in the performance when separating *relevant* sentences from *not-relevant* sentences this is not necessary.

Assessing performance for objective 1

Results of the comparisons of tests A, B and C are given in Table 5.1, the average number of parameters in each model is given in Table 5.2, and the ROC curves of the models generated in test B and test C are shown in Figure 5.1. The number of features for test B is much lower than for tests A and C because this test does not use the unlabelled sentences that constitute a large proportion of the sentences (as shown in Table 5.2).

We focus on the comparison of tests B and C because these both evaluate the models using the *relevant/not* and it is this labelling in which we are most interested. All results indicate very good ranking performance, and this can be seen on the ROC curves as they

		seq-gen	*alloc-conc*	*blind*
Number of sentences	*unlabelled*	243 477	129 155	148 934
	not-relevant	14 989	59 390	24 190
	relevant	1667	514	1156
Mean AUC (SD)	A) *relevant/rest*	0.974 (0.008)	0.981 (0.009)	0.974 (0.007)
	B) *relevant/not*	0.987 (0.003)	0.986 (0.011)	0.991 (0.006)
	C) train *relevant/rest*, test *relevant/not*	0.978 (0.008)	0.983 (0.009)	0.980 (0.007)
P-value	A vs B [1]	< 0.001	0.229	< 0.001
	B vs C [2]	0.005	0.462	0.001

Table 5.1: Results for sentence level: predicting the relevance of each sentence with regards to a risk of bias property. [1] P value using two-tailed unpaired t-test to compare the AUC values across the 10 folds of cross validation (data are not matched), [2] P value using two-tailed paired t-test to compare the AUC values across the 10 folds of cross validation (data are matched).

pass near to the point (0,1) in ROC space. We evaluate the models using the area under the ROC curve (AUC) metric, because for this objective we are concerned with how well our models are able to rank sentences by relevance. We are concerned with ranking rather than classification because we seek to provide an ordering to the reviewer such that they can see the most relevant sentences in an article. For example, the ranks allow sentences to be highlighted with different colours or shades in an electronic version of the article. Table 5.1 gives the numbers of sentences that are *relevant*, *not-relevant* and unlabelled, for each risk of bias property, and the results as the mean AUC across the 10 folds of cross validation. We compare the results of Test B and C using a two-tailed paired t-test that compares the AUC evaluated on the models of the 10 folds. We use an unpaired t-test to compare the results of Tests A and B because these tests evaluate the models with different sets of sentences.

Training and evaluating models using the *relevant/not* data produces better performance compared with training and testing models using the *relevant/rest* dataset, for two of our three labels (tests A vs B in Table 5.1). This may be because the *relevant/rest* data is noisier as it has some relevant sentences labelled as *rest* rather than *relevant*. This can have two effects. Firstly, it is more difficult for the model to separate the *relevant*

examples from the *rest* examples. Secondly, when evaluating the test data, the relevant sentences that have been incorrectly labelled as *rest* would be evaluated incorrectly. Comparing the performance between training using the *relevant/rest* dataset and training using the *relevant/not* dataset, while testing both using the *relevant/not* labelling, we again found that the model trained with the *relevant/not* labelling method gave a better performance for two of the three properties (test B compared with test C).

These tests have indicated that the *relevant/not* labelling should be used to learn models to predict sentence relevance. These models gave very high ranking performance, with mean AUC values across the 10 folds higher than 0.985 for all three properties. This can be interpreted as follows. Given a randomly selected *relevant* sentence, and a randomly selected *not-relevant* sentence, the probability that the *relevant* sentence would be ranked more highly than the *not-relevant* sentence is higher than 0.985.

	Sentence models			Article models		
	A	B	C	All	Title only	Title and abstract
seq-gen	14845.6 (26.49)	2372.2 (9.95)	14845.6 (26.49)	12176.7 (50.98)	98.9 (4.25)	1463.2 (7.60)
alloc-conc	12059.1 (30.31)	5539.4 (18.50)	12059.1 (30.31)	9907.6 (53.97)	60.1 (2.47)	1144.3 (13.31)
blind	11408.7 (33.28)	3004.2 (9.09)	11408.7 (33.28)	9352.9 (26.84)	64.5 (2.01)	1090.8 (8.16)

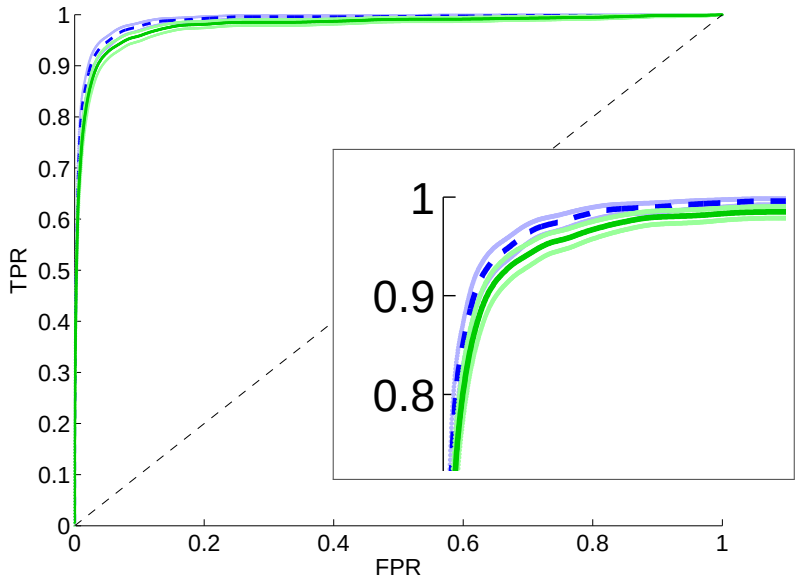
Table 5.2: Mean number of features (standard deviation) across cross validation folds.

5.2.2 Objective 2: Ranking articles by risk of bias

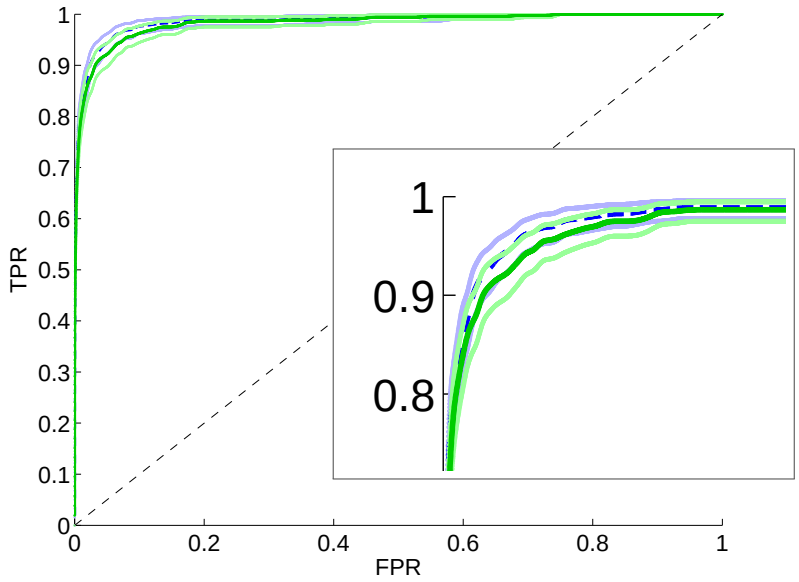
Objective 2 aims to rank articles by risk of bias, by training a logistic model to predict the risk of bias value of each article. The scores output by the model are used to rank articles by predicted risk of bias. As discussed in Chapter 3 this task is particularly relevant for rapid reviews, where it is important the high quality studies are prioritised.

As well as using the full text article, we also test this objective using the title and abstract from PubMed only. We generate models using: 1) the full text content of the

[1]This method requires a constant number of examples of each label (*low*, *not-low*) in each fold, so we add examples to make N constant, and use random selection of examples to correct these frequencies, for instance by removing a randomly selected positive example and duplicating a randomly selected negative example in a particular fold.

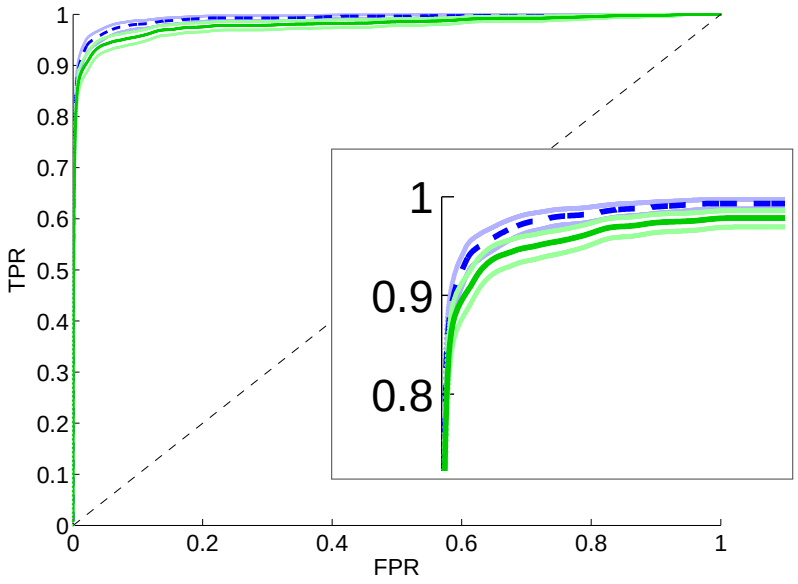


(a) Sequence generation



(b) Allocation concealment

Figure 5.1: ROC consensus curves for sentence level learning comparing results using the *relevant/not* dataset (test B; blue dashed curve), compared with the train *relevant/rest*, test *relevant/not* results (test C; green solid curve), for each risk of bias property Generated with parametric rate-oriented method[1] [28], with associated point-wise confidence bounds.



(c) Blinding

research articles, 2) the article title from the PubMed database and 3) the article title and abstract from the PubMed database. A comparison of these should help determine whether the effort required to retrieve the full text is offset by the improvement in performance when using it to predict risk of bias, in comparison to just using the text available in the PubMed database.

We compare the average AUC across the 10 folds with random models, to assess whether the ranking performance of the models is better than random. We do this using permutation testing, where we take the true labels of the original cross validation test folds, and randomly permute the order in each fold to give a random ranking with the same number of positives and negatives. We calculate the average AUC of these 10 test folds. This is repeated 1000 times and we give the proportion of times that an AUC greater than that of our models is found with these random rankings.

We also evaluate the performance of ranking articles using a combined score of sequence generation, allocation concealment and blinding. This is important because when ranking articles, we would like to prioritise those with a *low* overall risk of bias. We propose a general strategy where a weighted average is used to combine the individual scores of the risk of bias properties. We illustrate this with equal weights, which assumes that the three properties are equally important for predicting an articles risk of

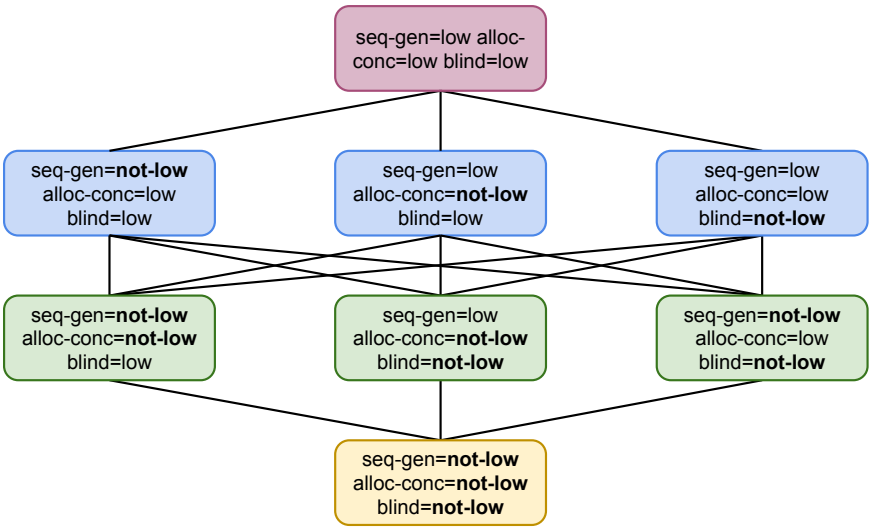


Figure 5.2: Hasse diagram of the partial ordering of property values when combining into a single score. Overall risk of bias increases as the number of properties with a value of *not-low* increases.

bias. The order of examples is then given by the partial order shown in Figure 5.2. This partial order states that, for example, an article with two of the three properties having value of *low* should be ranked more highly than an article with only one property with a value of *low*.

Assessing performance for objective 2

We again use the AUC to evaluate the ranking performance of these models, given in Table 5.3 as the average AUC across the 10 folds of cross validation. The mean (SD) number of parameters across the 10 folds for each model is given in Table 5.2. The ROC curves of the models generated using the full text and the title and abstract only are shown in Figure 5.3. The models using the full text had mean AUC > 0.72. The models using the PubMed title had mean AUC > 0.67. The models using the PubMed title and abstract had mean AUC > 0.68. All models are better than random (all permutation P values < 0.001). Models using the article content are able to rank articles better than when using only the title, or title and abstract from the PubMed database, for the sequence generation and allocation concealment properties only. We could not find a difference between using the title and using the title and abstract, although this may be

due to a lack of power because our sample size is small. It is interesting to note that the title models are predictive even with a relatively small number of features (as detailed in Table 5.2).

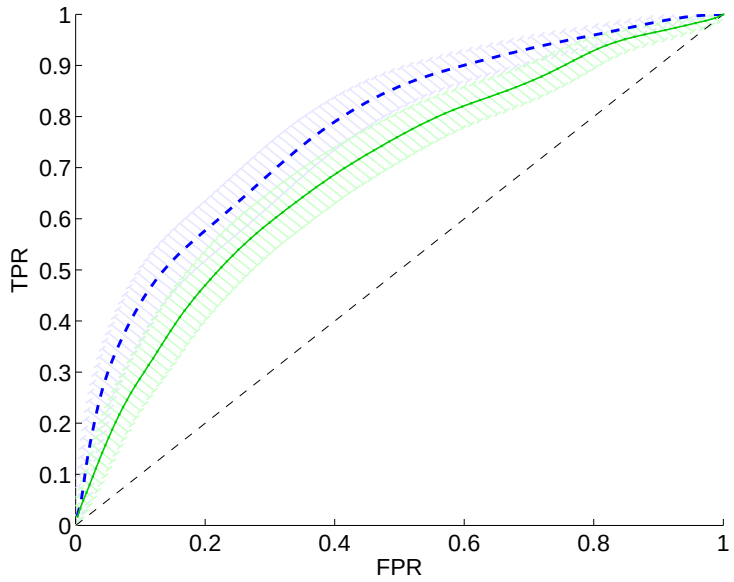
To evaluate the performance when ranking according to the combined score, we use the 226 articles in our dataset that have labels for all properties (shown in Figure 2.5) to test the models. This is necessary so that we can evaluate the predicted ranking against the ranking using the known true labels of the articles, which is inferred from the values of the three properties. We train a model for each property individually using all examples with a label for this property, that were not included in the test set. For instance, the dataset includes 991 articles for sequence generation and we use the remaining 766 articles (after removing the 226 in the test set) to train the sequence generation models. In order to estimate the uncertainty about the estimated ranking performance, we randomly divide the test set into 3 partitions, and treat these as three separate samples. If we were to generate a single model using the test set we would not know how this ROC curve would vary across test sets with different examples.

We calculate the AUC of each test set using pairwise comparisons. In order to do this we assign each example a class that denotes its level on the partial order. The classes are denoted l_0, l_1, l_2, l_3 where l_i corresponds to the examples with *i low* labels. For example, an example with *blind=low*, *seq-gen=low* and *alloc-conc=not-low* has two *low* labels and is assigned class l_2.

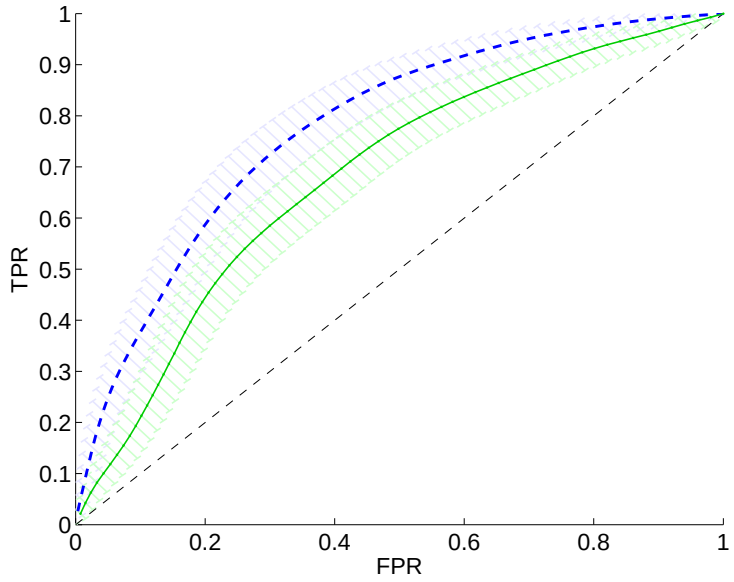
The pairwise AUC value can then be calculated as follows. Given 6 examples with ranking $l_2, l_1, l_3, l_1, l_0, l_0$ (ranked by increasing score), we start by comparing the first

Dataset	*seq-gen*	*alloc-conc*	*blind*
	AUC (standard deviation)		
1.Article content	0.769 (0.051)	0.777 (0.034)	0.726 (0.051)
2.PubMed title	0.682 (0.053)	0.690 (0.072)	0.675 (0.063)
3.PubMed title and abstract	0.692 (0.037)	0.685 (0.047)	0.694 (0.065)
P values[1]: comparison of performance using feature sets 1, 2 and 3			
1 vs 2	0.001	0.004	< 0.001
1 vs 3	< 0.001	0.002	0.206
2 vs 3	0.672	0.741	0.497

Table 5.3: Ranking performance using different datasets and P values comparing these models using a paired two-tailed t-test. [1] P values using two-tailed paired t-test to compare the AUC values across the 10 folds of cross validation.

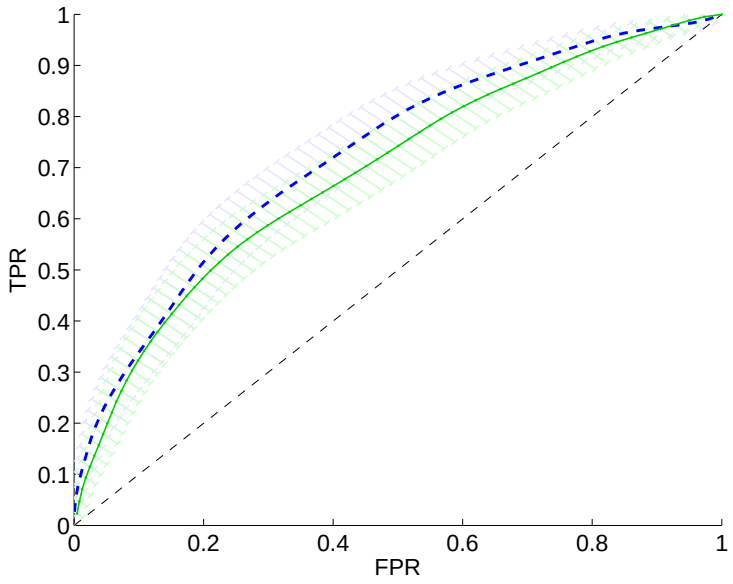


(a) Sequence generation



(b) Allocation concealment

Figure 5.3: ROC consensus curves for predicting article risk of bias. Generated with analytical rate-oriented method [28], with associated point-wise bounds. Using logistic regression. Blue dashed: Article contents (model 1); green solid: PubMed title and abstract (model 3).



(c) Blinding

example with class l_2, with all other examples. This example is correctly ranked before l_1, l_1, l_0, l_0 but incorrectly ranked before l_3, hence has four correct comparisons out of five. The example at position 2 with class l_1 is then compared against all examples with a lower ranking than itself, except the other example with the same class. This example is correctly ranked before l_0, l_0 but incorrectly ranked before l_3, hence has two correct comparisons out of three. This continues until the last example is reached. The pairwise AUC is the proportion of these comparisons that are ranked in a correct order. In this example $AUC = \frac{11}{13}$.

	AUC	
	multi-class	all *low* versus rest
Test set 1	0.707	0.686
Test set 2	0.655	0.829
Test set 3	0.666	0.767
Mean (SD)	0.676 (0.027)	0.761 (0.072)

Table 5.4: AUC for combined ranking, for the three test sets. Multi-class AUC is the pairwise AUC when treating each level of the partial order as a separate class. *low* versus rest is the AUC of the ROC curve when using three *low* values (top level of partial ordering) as the positive class, and all other labellings as the negative class.

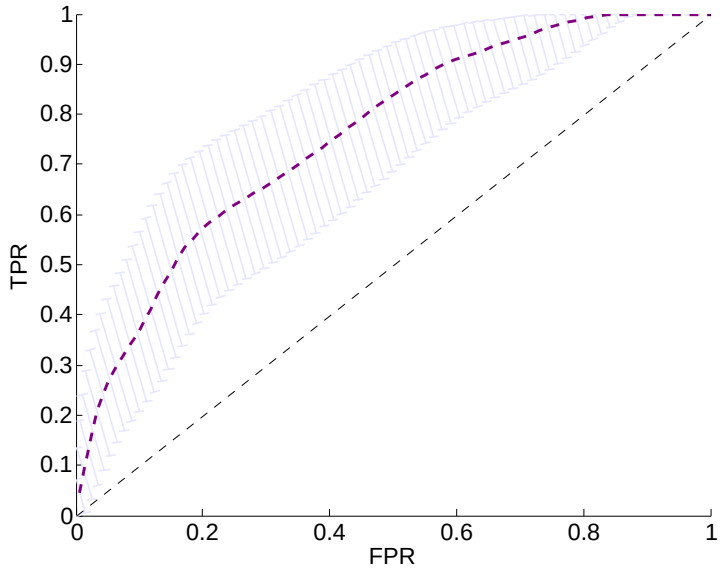


Figure 5.4: ROC consensus curve for *low* versus rest evaluation of combined ranking. Three *low* values (top level of partial ordering) as the positive class, and all other labellings as the negative class.

The results for the combined ranking are shown in Table 5.4. The three partitions of the test set, when treating all partial order levels as separate classes, have mean $AUC = 0.676$ ($SD = 0.027$). It is not possible to generate a ROC curve for the multi class ranking because these graphs require binary classes. Figure 5.4 shows the ROC generated using the 1-versus-rest ROC curve generation approach, for the three *yes* labels (the top of the partial order) versus all other label groups. Here the positive class corresponds to examples with *three yes labels* and the negative class corresponds to examples with *less than three yes labels*. The ranking has mean $AUC = 0.761$ ($SD = 0.072$) using this evaluation approach. This is higher than the multi-class AUC (where each level of the partial order is treated as a separate class), which indicates that distinguishing between the very high quality articles (with all *low* labels) and those that are not is an easier task compared with learning to distinguish between all levels of the partial order.

5.2.3 Objective 3: Reducing the number of assessments the reviewers need to perform by hand

Objective 3 aims to reduce the reviewer work load by identifying articles that can be classified as *low* or *not-low* with high enough certainty, so that only a single reviewer is needed to assess these articles by hand. We suggest that the certainty is high enough when the model's assignment is at least as likely to be correct as an assignment by a human reviewer. When this is the case it may be reasonable to replace a human reviewer by this model prediction. As already mentioned, the logistic regression model produces a well calibrated score s for each article, such that s can be interpreted as the probability that the article has a risk of bias value of *not-low* (because higher scores indicate an example is more likely to be *not-low*). We can compare the scores assigned by a model to two fixed probability thresholds t and $1-t$, where t is an estimate of the proportion of human assignments that are correct. Articles are classified as *not-low* for a property if $s \geqslant t$ and as *low* if $s \leqslant 1-t$. This assumes that the human reviewer makes the same proportion of mistakes with *not-low* and *low* articles respectively.

We apply the thresholds t and $1-t$ to the logistic regression models that were generated for objective 2, to convert these ranking models into classifiers needed for this objective. To determine the value of t we use results of previous work by Lensen et al. [64] and Hartling et al. [61]. These works analysed the degree of concordance of risk of bias assignments given by reviewers who have assessed the same studies. Lensen et al. found disagreements (number of disagreement/number of comparisons) of 11/123, 26/123 and 41/123 for sequence generation, allocation concealment and blinding, respectively. Hartling et al. [61] found disagreements of 8/28, 19/46 and 20/31 for sequence generation, allocation concealment and blinding, respectively. We calculate the average proportion of disagreements across these studies and properties to give an estimate of the proportion of reviewer disagreements of 26.4%. For these articles we know that one assignment is incorrect and the other is correct. For the other article assignments where both reviewers agree we cannot know whether they are both correct or both incorrect. Hence the proportion of assignments that are definitely incorrect (because the assignments disagree) is 13.2%. We use this proportion as a proxy for the proportion of assignments that are incorrect, which would be the case if we assume that if two reviewers agree then they are both correct. Therefore, a probability that is

higher than 0.868 would be better than the certainty of a human reviewer, and we set the threshold value t to 0.868.

The lower threshold, $1-t = 0.132$, denotes the score below which we are at least as certain as a human reviewer that an article has an assignment of *low*, according to the model prediction. The upper threshold, $t = 0.868$, denotes the score above which we are at least as certain as a human reviewer that an article has an assignment of *not-low*, according to the model prediction. A score between 0.132 and 0.868 indicates that the model could not predict the label with as much certainty as a human reviewer, and these articles should be assessed as usual by two reviewers.

Assessing performance for objective 3

Table 5.5 shows the number of articles our models classify as *low* or *not-low* using these score thresholds. All models were able to classify more than 33% of articles as either *low* or *not-low* with a certainty at least as high as a manual reviewer. We suggest that only one human reviewer is needed to assess these articles manually. In the next section we discuss and assess the model calibration.

Assessing calibration for objective 3

As described in Section 5.1, logistic regression is known to produce scores that are well-calibrated. Objectives 1 and 2 involve ranking examples using the score and therefore score calibration is not necessary. Instead we only care about the order of the scores assigned to the examples. In contrast, objective 3 seeks to classify examples and we do this using score thresholds that assume the scores are calibrated. For example, we take our upper threshold to mean the lower bound of the probability that we are at least as

	Predicting *not-low* $score \geqslant 0.868$		Predicting *low* $score \leqslant 0.132$		
	% articles	Precision	% articles	Precision	Total %
seq-gen	16.9 (2.42)	0.821 (0.114)	20.9 (2.77)	0.838 (0.072)	38.2 (3.14)
alloc-conc	15.5 (2.92)	0.874 (0.090)	9.9 (2.77)	0.816 (0.181)	35.5 (4.45)
blind	7.6 (2.91)	0.803 (0.145)	14.8 (2.62)	0.810 (0.101)	33.4 (6.58)

Table 5.5: Mean number of articles (standard deviation) and precision (standard deviation) across 10 folds (using sentence model).

certain as a human reviewer that the risk of bias assignment is *not-low*. Therefore, for this objective it is important to determine the quality of the calibration.

We assess this with reliability diagrams, shown in Figure 5.6. Reliability diagrams are plots showing the score output by the model on the x-axis and the calibrated scores on the y-axis. We used the COREleam R package to perform isotonic calibration to create reliability diagrams [99]. Isotonic calibration is a method of generating the calibrated scores using a ROC curve. This method has two steps. First, the ROC curve is converted into a convex ROC curve, if this is not already the case. Second, the calibrated score, which is simply the probability of a *not-low* example at a particular position of a ROC curve, is calculated using the gradient (and class distribution) as we discussed in Section 4.4. The first step is necessary because the scores must be decreasing along the ranking (by definition), and if the calibrated scores (step 2) were calculated on a non-convex ROC curve this would not be the case.

An example of isotonic regression is shown in Figure 5.5 and Table 5.6. The ROC curve corresponds to three sets of examples each with a particular score. This curve is concave because the second segment (with two positives and one negative) has a higher proportion of positive examples compared to the first segment (with one positive and one negative). The concavity is removed by combining the first and second segments of the ROC curve to create a convex ROC curve, shown by the dashed line in Figure 5.5. The calibrated scores are the empirical probabilities of a negative in each segment of the convex ROC curve.

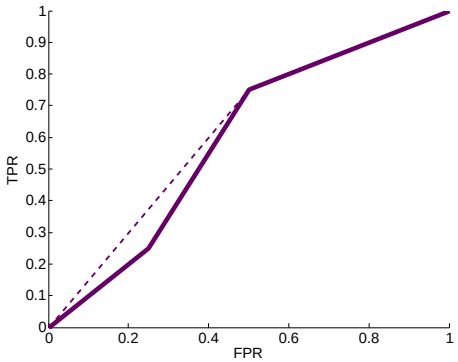
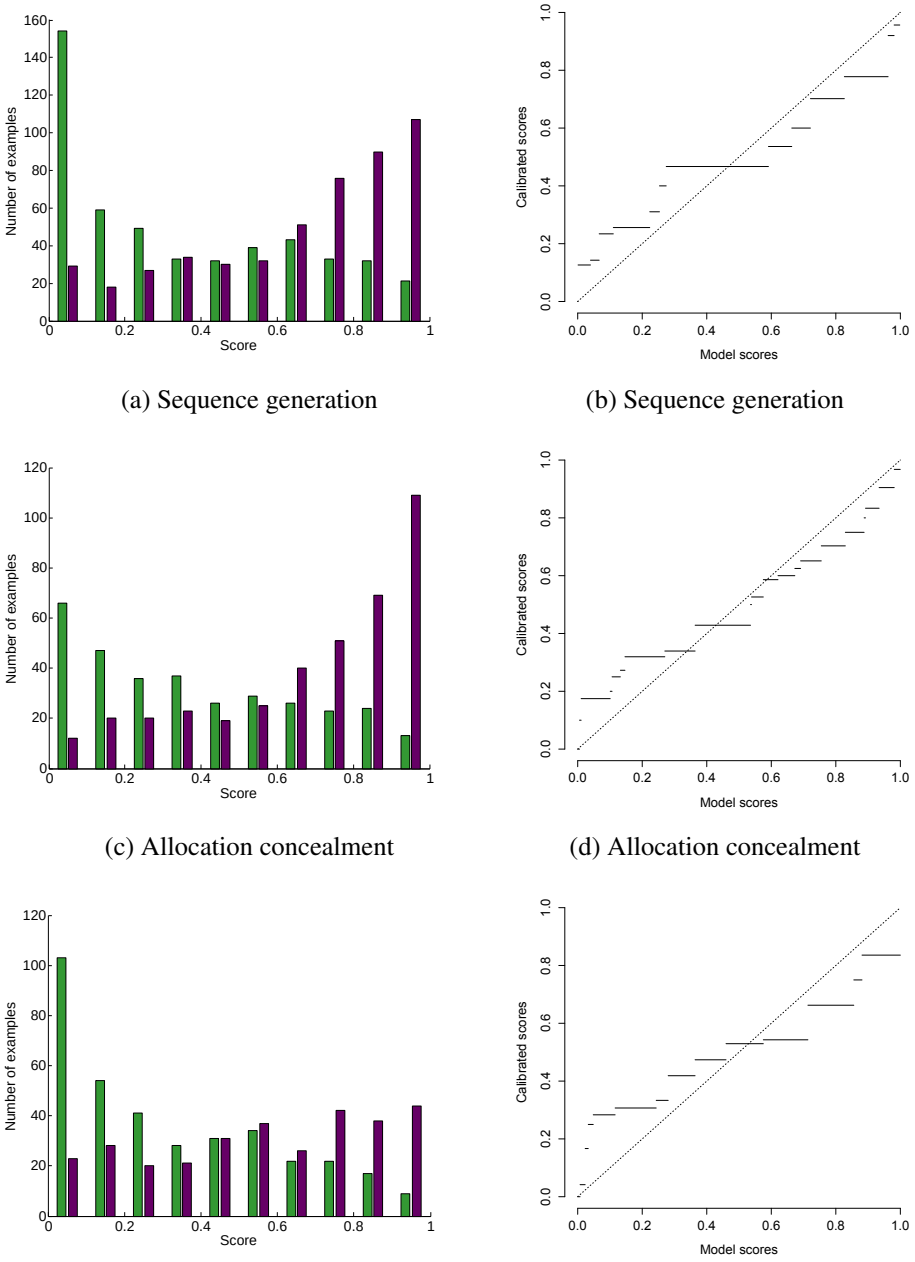


Figure 5.5: Example of isotonic calibration.

Labels	[0 1]	[0 0 1]	[0 1 1]
Model scores	0.7	0.8	0.9
Calibrated scores		$\frac{2}{5}$	$\frac{2}{3}$

Table 5.6: Example ranking.



(a) Sequence generation (b) Sequence generation

(c) Allocation concealment (d) Allocation concealment

(e) Blinding (f) Blinding

Figure 5.6: Left: Score distributions predicted by logistic regression models, green: *low*; purple: *not-low*). Right: Assessing calibration with reliability diagrams [103] (generated with isotonic regression).

Convexifying the ROC curve creates a set of 'bins' containing examples that may have different predicted scores but have been assigned the same calibrated score. Reliability diagrams show each 'bin' as a horizontal line on the graph. If the scores output by the model are calibrated then we could expect the plot to sit along the ascending diagonal. Visual assessment of the calibration maps shows that the bin lines do sit across or close to the ascending diagonal. We do however notice a slight logistic shape, where the bin lines close to (0,0) tend to sit above the ascending diagonal, whereas those near to (1,1) tend to sit below the ascending diagonal. This indicates that the model scores near to scores zero and one tend to be more extreme than the calibrated score.

5.2.4 Effect of changes in class distribution

The scores output by a model are only calibrated if the proportion of *low* and *not-low* articles remains constant between the data used to train the model, and the new data on which predictions are made. The reason can be shown with an example. Given a set of 10 articles all with a calibrated score of 0.2. As a high score indicates that the *not-low* label is more likely, this means the probability of *not-low* is 0.2. Hence, we would expect that 8 examples have *low* risk of bias and 2 have *not-low* risk of bias. Now imagine duplicating the *low* articles, such that we have 18 articles. As these are duplicates the model would again output 0.2 for all 18 articles. However, now there are 16 *low* articles and 2 *not-low* articles, so the proportion of *not-low* articles is $\frac{1}{9}$ and the scores are no longer calibrated.

A simple adjustment to a score s will give a score s_{cal} that is calibrated according to the new class distribution:

$$s' = \frac{s \cdot scalar_{neg}}{s \cdot scalar_{neg} + (1-s) \cdot scalar_{pos}} \tag{5.1}$$

where:

$$scalar_{pos} = \frac{\pi'}{\pi} \tag{5.2}$$

$$scalar_{neg} = \frac{1-\pi'}{1-\pi} \tag{5.3}$$

where π is the proportion of positives in the data and the new data is denoted by $'$.

For example, we can use this to adjust for the change of class distribution for $s = \frac{2}{10}$

in the example above, where $scalar_{neg} = \frac{2}{18} \div \frac{2}{10} = \frac{5}{9}$ and $scalar_{pos} = \frac{16}{18} \div \frac{8}{10} = \frac{10}{9}$:

$$\begin{aligned} s' &= \frac{s \cdot scalar_{neg}}{s \cdot scalar_{neg} + (1-s) \cdot scalar_{pos}} \\ &= \frac{\frac{2}{10} \cdot \frac{5}{9}}{\frac{2}{10} \cdot \frac{5}{9} + (1 - \frac{2}{10}) \cdot \frac{10}{9}} = \frac{1}{9} \end{aligned} \tag{5.4}$$

This adjusted score corresponds to the correct probability after the change of class distribution. In practice however, this rescaling can be difficult because while the class distribution used to train a model with supervised machine learning is known (because we have labelled data), the class distribution of new data may not be known.

In objective 3 we use two score thresholds at $s = 0.868$ and $s = 0.132$. These denote the probabilities above and below which we are at least as sure as a human reviewer that the prediction made by our model is correct. These thresholds are fixed across all class distributions. When the class distribution changes the adjustment of the calibrated scores means that the examples falling below and above the thresholds may change. In our case, new data has a lower proportion of *low* articles (with lower score on average) and a higher proportions of *not-low* (with higher score on average) than the data on which the models were trained. This means that when adapting the calibration of our article models to new data we can expect less scores assigned below the lower threshold (predicting *low*, at $s = 0.132$) and more scores assigned above the upper threshold (predicting *not-low*, at $s = 0.868$).

5.2.5 Inference using rate-oriented point-wise confidence bounds

In Chapter 4 we introduced rate-oriented point-wise confidence bounds, and we can use these in this setting to determine rates at which the classification performance of ROC curves differ. Given two ROC consensus curves, R_1 and R_2 at a particular rate r, each with a 95% confidence interval with lower bound $F_{0,r,l}(R_i)$ and upper bound $F_{0,r,u}(R_i)$. This means that $p\left(F_{0,r}(R_i) \geqslant F_{0,r,l}(R_i)\right) = 0.975$ and $p(F_{0,r}(R_i) \leqslant F_{0,r,u}(R_i)) = 0.975$, such that:

$$\begin{aligned} &p(F_{0,r}(R_1) \geqslant F_{0,r,l}(R_1) \wedge F_{0,r}(R_2) \leqslant F_{0,r,u}(R_2)) \\ &\qquad = p(F_{0,r}(R_1) \geqslant F_{0,r,l}(R_1)) \cdot p(F_{0,r}(R_2) \leqslant F_{0,r,u}(R_2)) \\ &\qquad = 0.975 \cdot 0.975 \approx 0.95 \end{aligned} \tag{5.5}$$

If the confidence intervals of R_1 and R_2 at a rate r do not overlap and the expected $F_{0,r}$ (the consensus point) is greater for R_1 than R_2, then we can say that with $\geqslant 0.95$ probability the consensus curve of R_1 has higher classification performance when using rate r as the classification threshold.

The ROC curves of our article level models (in Figure 5.3) show that the ROC curves mostly have overlapping confidence intervals along the rate-oriented point-wise bounds. Only the ROC curve of the sequence generation property (shown in Figure 5.3a) has regions where the confidence intervals do not overlap, at around rates 0.25 and rate 0.7. This indicates that with 95% probability, the consensus curve generated from the full text has better classification performance compared with the consensus curve generated from the title and abstract from the PubMed database, at these rates.

We note that in general point-wise confidence bounds have limitations when seeking to detect differences in classification performance between models. This is because each confidence interval denotes where a consensus curve, the average point across folds, is likely to sit in ROC space. Hence we are comparing averages across folds rather than comparing on a fold by fold basis. This is akin to the difference between an unpaired and paired t-test – a paired t-test compares the performance of each fold of test 1 with the same fold of test 2, whereas an unpaired t-test compares the performance of all folds of test 1 with all folds of test 2. When confidence intervals of our bounds overlap it could be the case that the performance of each fold is higher for test 1 compared with test 2, when comparing folds in a paired fashion. In this case a paired t-test of classification performance (such as recall or accuracy) may be able to detect a difference in performance, whereas this difference is not detectable when viewing confidence bounds in ROC space. In short, the paired t-test has greater power. However, rate-oriented bounds provide a visualisation in ROC space of the overall comparison of models across all rates, and hence is a useful tool to compare models.

5.3 Analysis of predictors

Table 5.7 shows the top 20 predictors of the sentence models, having the highest coefficient values in the logistic regression models. We note that some word stems end with an *i* because the stemming process includes changing a *y* suffix to *i* when the preceeding character is not a vowel and is not the first character of the word. The coefficients

denote the increase in log odds of *not-relevant* for a 1 standard deviation increase of the number of times a word occurs in a sentence. We notice that all but one of our top predictors have a negative coefficient, where sentences containing these words (and more of them) are more likely to be *relevant* for a risk of bias property. Positive coefficients are not common because the occurrence of words are not predictive of *not-relevant*, it is the absence of a set of words (those with negative coefficients) that predict this. For example, the occurrence of the word stem *randomli* predicts that a sentence is *relevant* for sequence generation. The absence of the *randomli* stem is predictive of a *not-relevant* sentence.

As in conventional epidemiology, our predictors may be identified because they are confounded with other words that are relevant to risk of bias, rather than being relevant themselves. However, these are still valid predictors. For example, the *studi* word stem predicts a sentence is relevant to the risk of bias due to sequence generation, but this may be because often authors refer to a randomised study and hence *studi* may be a predictor because of this association.

As expected, word stems related to the methods of generating random number sequences are predictive of sentence relevance for the sequence generation property, such as *computer-gener*, *randomli* and *stratifi*. The more often one of these words occurs in a sentence, the more likely the sentence is to be *relevant* to sequence generation. Interestingly, several of the top words for sequence generation are types of drugs. Drug names may be good predictors because they may indicate whether the outcome is subjective or objective, and the risk of bias due to blinding is affected by this. For instance, *vildagliptin*, *metformin* and *insulin* are predictors or relevant sentences for sequence generation, and these drugs are used to treat diabetes, which is assessed objectively through blood tests.

The word stems *randomli* and *randomis* are indicators that a sentence is relevant to the risk of bias due to allocation concealment. Allocation can only be concealed if the assignment sequence is random, such that it cannot easily be predicted by the study personnel making the allocations. There are several words relating to the methods used to ensure concealment of allocation, that are also predictors of sentence relevance for this property. For example, using *seal*ed *envelop*es, using a *central* location or a *telephon*e system to perform the allocation. Also, making sure the allocations are *ident*ical across groups in a trial so personnel cannot identify the study group assigned.

Sequence generation		Allocation concealment		Blinding	
toothbrush	-8.71	envelop	-8.73	mask	-8.07
computer-gener	-7.64	assign	-7.28	capsul	-7.91
randomli	-7.56	randomis	-6.84	unawar	-7.58
studi	-7.01	code	-6.15	doubleblind	-5.50
block	-6.81	particip	-5.79	indistinguish	-5.23
comput	-6.50	alloc	-5.65	match	-5.11
vildagliptin	-6.48	central	-5.39	blindli	-4.86
assign	-6.42	ident	-5.38	awar	-4.84
morphin	-6.09	randomli	-5.04	anaesthesiologist	-4.71
patient	-5.80	seal	-4.68	assign	-4.65
alloc	-5.75	blind	-4.64	doubl	-4.51
subject	-5.69	telephon	-4.07	pain	4.17
insulin	-4.99	independ	-4.02	hernia	-4.05
random-numb	-4.80	packag	-4.00	laparoscop	-3.91
envelop	-4.67	close	-3.93	fluvoxamin	-3.88
sonic	-4.64	labetalol	-3.86	24-week	-3.60
number	-4.64	hydralazin	-3.66	incision	-3.58
stratifi	-4.51	bottl	-3.61	label	-3.57
group	-4.39	involv	-3.58	patient	-3.52
metformin	-4.28	sponsor	-3.56	pharmaci	-3.31

Table 5.7: Top 20 word stem predictors of sentence relevance and normalised coefficients (such that each coefficient denotes the increase in log odds of *not-relevant* for a 1 standard deviation increase of the number of times a word occurs in a sentence).

The top predictors of sentence relevance for the blinding property include, as expected, word stems variants of the word *blind* – *doubleblind*, *blindli*. Words that relate to whether the assignment is known are also important, such as *unawar*e, *awar*e, *mask*ed, and *indistinguish*able. Unexpectedly, the word *pain* has a positive coefficient for *blinding*, which is difficult to explain.

The predictors of risk of bias values of articles also contain many words that appear sensible predictors, shown in Table 5.8. For example, *computergener*ated is a predictor for *low* risk of bias due to sequence generation. Terms referring to the practice of allocation concealment are predictors for *low* risk of bias due to this property. For example *ident*ical, *numbered*, *opaqu*e, *seal*ed, and *envelop*es. As expected, word stems that would be used if blinding were performed are predictors for *low* risk of bias due to blinding. Examples are *blind*, *double-blind*, and *placebo*.

Across all three risk of bias properties far more words that appear relevant are used

Sequence generation		Allocation concealment		Blinding	
631	-0.62	envelop	-0.85	placebo	-0.90
computergener	-0.61	opaqu	-0.73	potassium	0.60
approv	-0.60	grate	-0.69	explan	-0.52
www	-0.60	power	-0.64	acknowledg	-0.48
opaqu	-0.60	3depart	0.54	suppli	-0.47
envelop	-0.59	bulletin	0.52	blind	-0.46
95%	-0.57	jone	-0.52	double-blind	-0.42
alloc	-0.55	seal	-0.49	interf	0.41
inclus	-0.50	mann–whitnei	-0.46	broken	-0.40
faster	0.50	committe	-0.46	[4]	0.39
spss	-0.49	smoke	0.44	pharmaci	-0.39
fund	-0.49	januari	0.43	elsevi	-0.38
otherwis	-0.49	ident	-0.43	182	0.37
discourag	-0.48	england	-0.42	undesir	-0.36
mention	0.48	preclud	0.42	316	-0.36
conceal	-0.48	epidemiologi	-0.41	advantag	0.36
bmj	-0.45	2002	-0.41	withdraw	-0.35
655	0.45	remaind	-0.40	55%	0.35
randomis	-0.44	emerg	-0.40	277	0.35
predomin	-0.44	symposium	0.40	horm	-0.34

Table 5.8: Top 20 word stem predictors of article risk of bias and normalised coefficient (such that each coefficient denotes the increase in log odds of *not-low* risk of bias for a 1 standard deviation increase of the number of times a word occurs in the article).

in the sentence level models compared to the article level models. There are some terms in the article level predictors for which the most likely explanation of their use as a predictor is simply chance. For example, the terms *277* and *[4]*. This is consistent with the level of performance we achieved and may be due to the number of examples we have for this task (which is much smaller than for sentence learning) or the difficulty of the learning tasks. We consider this further in the discussion section below.

It is common to use a regularised model with term frequency-inverse document frequency (TF-IDF) feature transformations for text mining problems. Regularisation is used to prevent overfitting, by constraining the size of the parameters in the model. This is often beneficial when a dataset has a large number of features and a relatively small number of examples, as is often the case for text mining tasks. TF-IDF is a feature transformation that aims to better represent the importance of a word in the dataset. It does this by offsetting the number of times a word appears in a document by the

	Sequence generation		Allocation concealment		Blinding	
	AUC	P value	AUC	P value	AUC	P value
Sentence models	0.989 (0.003)	0.007	0.850 (0.014)	0.598	0.992 (0.005)	0.021
Article models	0.789 (0.046)	0.012	0.796 (0.061)	0.020	0.744 (0.048)	0.010

Table 5.9: Results with regularisation and TF-IDF transformation. P value for a paired two-tailed t-test comparing 10 folds of cross validation with original sentence level results (test A in Table 5.1) and article level results (test 1 in Table 5.3), respectively.

proportion of documents within which the word occurs. We did not use regularisation or TF-IDF, and it may be the case that these approaches help to reduce the occurance of potentially irrelevant features in our top lists shown in Tables 5.7 and 5.8.

We rerun our sentence and articles models using regularisation and TF-IDF transformations to determine if this improves the parameter estimation and hence the performance of our models. We use the default regularisation parameters of the SGD Weka class (l1 regularisation with $\lambda = 0.0001$). The results, given in Table 5.9, show that using these settings does give a small improvement in the AUC, for 5 of the 6 results. We provide the top 20 features of these models in Table A.1 and Table A.2, for the sentence and article models, respectively. A number of the potentially irrelevant top features are no longer in the top 20 features of each model. The sequence generation top feature in the article model was '631' and this no longer appears in the top features listed. The following features are no longer top features in the blinding article model; '[4]', '182', '316', '277', '55%'. While regularisation and TF-IDF do improve model predictions, the TF-IDF transformation means that the parameters of the model are no longer interpretable.

5.4 The Systematic Review Assistant – a prototype

We have created a prototypical tool, available at http://www.datamining.org.uk, to demonstrate how the objectives can be implemented in practice. This tool allows users to run a prediction generator to make predictions for each article. Scores for each article and each sentence are generated, predicting the risk of bias value and relevance, respectively, for each uploaded article. The reviewer can then view the articles supplemented

with the predictions, as shown in Figure 5.7. The sentences are highlighted by relevance with respect to each risk of bias property. For example the sentence highlighted in red in Figure 5.7 has been predicted as relevant for sequence generation (the colours are codes as indicated by the buttons in the top right of the window). This sentence describes assignment of participants to groups. A stronger highlighting colour (higher opacity) indicates higher relevance. The articles are assigned scores denoting the risk of bias as described in this thesis, for each risk of bias property.

We created the Systematic Review Assistant prototype using Java Servlets with a MongoDB database. We chose MongoDB because it gives us a great deal of flexibility during development of the prototype. This is because MongoDB is a noSQL database, where there are no tables with a fixed schema (a pre-specified set of columns), as in an SQL database. Instead, a MongoDB contains collections, and each collection contains a set of documents. A collection is equivalent to an SQL table and a document is the equivalent to a row in an SQL table. Unlike SQL tables, collections have no particular structure, such that the fields of documents in a single collection may differ. This means we can easily change a collection's fields by changing the Java code that saves and retrieves the data from the database, without the need to update the database directly.

5.5 Discussion

We have shown that we can rank sentences by predicted relevance (for each risk of bias property) with high ranking performance (AUC > 0.98). This is useful to assist reviewers by indicating which parts of the article text are particularly relevant to risk of bias. We were able to rank articles according to risk of bias with AUC > 0.72. Ranking articles by risk of bias means that the reviewer is able to assess the articles from predicted *low* to predicted *not-low* risk of bias.

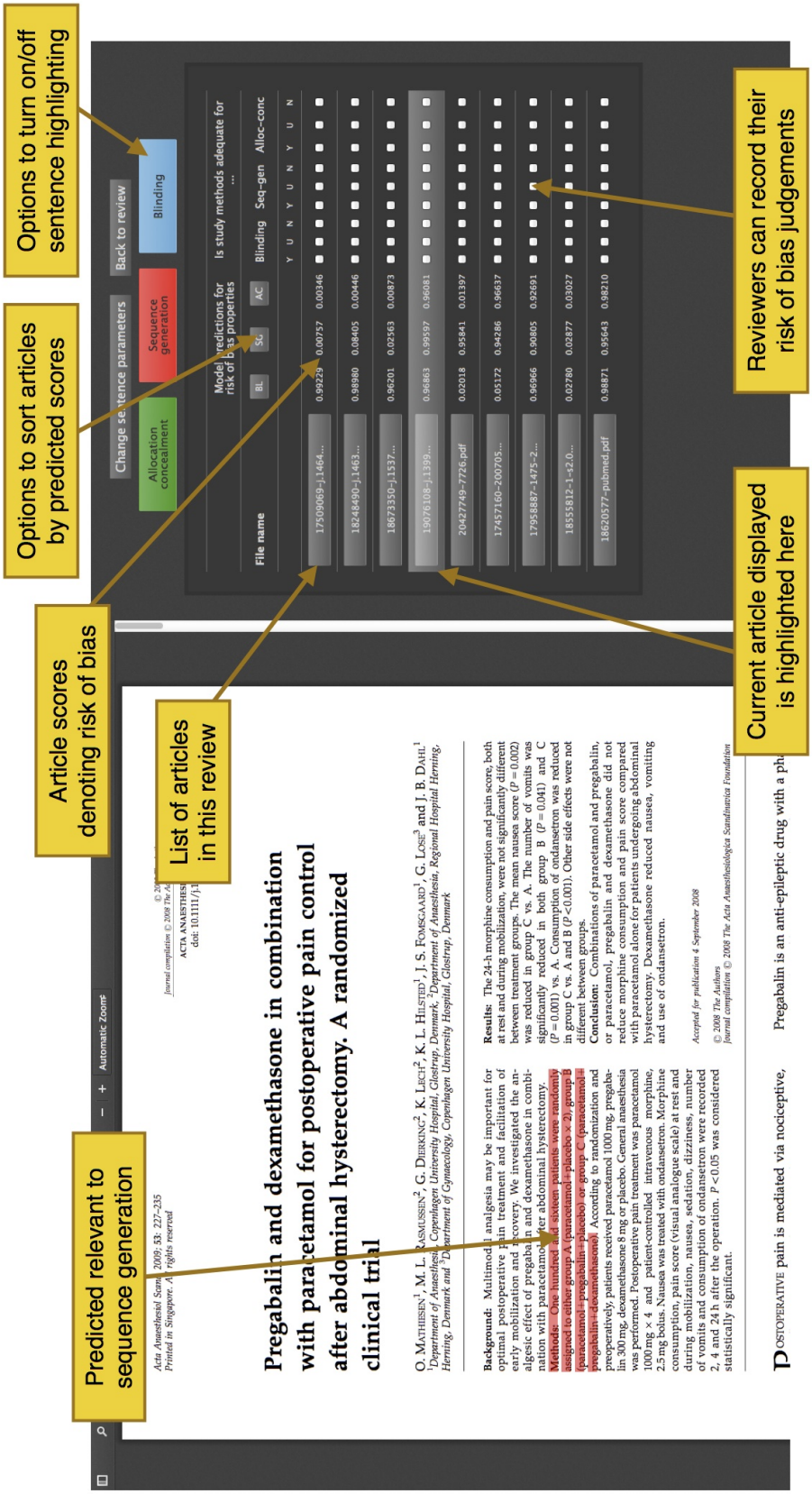


Figure 5.7: Prototypical tool for risk of bias assessments.

We found a small decrease in performance when using only the article title and abstract from PubMed, compared to using the full text extracted from the article PDF document for sequence generation and allocation concealment (t-test $P \leqslant 0.001$ and $P = 0.002$ respectively). The full text content will often contain more information about risk of bias compared to the title and abstract alone. However, this benefit may be offset by noise from extraction of the article content from PDF documents and the volume of content within articles that is irrelevant to risk of bias. Conversely, while the PubMed abstract may not contain as much information about risk of bias compared to the full text, it may have less noise because this text is a concise summary of the full text article retrieved from the PubMed database. Retrieving PubMed data is quick and straight forward, whereas obtaining the full text of research articles requires more effort and text extractions from PDF documents are noisy. Hence, the increase in performance from using the full text may not be worth the cost of its retrieval.

Our results indicate that it is possible to use text mining to reduce the reviewer work load, by identifying the articles that have been classified with a certainty higher than that of human reviewers. We suggest that these articles only need to be manually assessed by one reviewer. On average more than 33 % of research articles can be labelled as *low* or *not-low* with higher certainty than that of a human reviewer, offering the potential to reduce the amount of time required by human reviewers.

The sentence level learning achieved much higher performance (in terms of AUC) compared with the article level learning. This may be because the article level task is more difficult. We can imagine that establishing the risk of bias values depends on combinations of words and the interaction between them in the article text. For example, the word *envelope* does not itself determine whether an article describes a study with low risk of bias due to allocation concealment, as it is whether the envelopes are *sequentially numbered*, and *opaque* that determines this. In contrast, if a sentence simply contains the word *envelope* it is likely to be *relevant* to the allocation concealment property hence learning sentence relevance is a simpler learning task.

5.5.1 Limitations

Our work has the following limitations. The RoBAL dataset was inferred from the risk of bias assessments of a subset of Cochrane reviews (described in Section 2.3.1). This

subset contains unequal numbers of reviews across topics, shown in Table 2.1 in terms of Cochrane groups. Some groups (such as the Behavioural Medicines group or the Complementary Medicine Field group) were not present in this review subset at all. The performance of our models may differ across review topics. For instance, it may be the case that our models perform better on articles about topics that were prevalent in our dataset compared to those that were rare or not present. The Cochrane reviews from which our dataset were derived were performed between 2008 and 2011. The extent to which risk of bias is reported in articles describing clinical trials and the way it is described may change with time. For instance, reviews performed in 2015 may include articles describing clinical trials published after 2011, and this may mean the RoBAL dataset on which our models were trained is not representative of these articles.

The limited size of our dataset (between 671 and 989 per risk of bias property) may have restricted the performance of our models. We only included articles if the title and abstract could be found in the text extracted from the PDF articles, such that articles with poor text extractions are less likely to be included in our dataset because noise within the text means that the title and abstract may not be found. Therefore, it is likely that our dataset is less noisy than study articles on average. Furthermore, we only include articles in our dataset where a quotation was supplied or no information was stated, and so it is possible this sample is unrepresentative of articles describing clinical trials. We use labels inferred from data from Cochrane risk of bias assessments such that these labels may not be the same as directly annotated labels. Lastly, previous work has indicated discordance between reviewers who assess the same article and this indicates that the labels we have used, from the Cochrane risk of bias assessments, may not always be correct.

An automated approach is limited by the degree of reporting in trial publications, as although the CONSORT statement specifies that information relevant to risk of bias should be described in a trial report, this is often not the case [48]. However, it is known that trial protocols can contain information that is not reported in the study publications [50], hence risk of bias information could potentially be extracted from these protocols.

5.5.2 Comparison with related work

As mentioned in Chapter 2, to our knowledge only recent work from one research group has investigated text mining for risk of bias assessments [21–23]. We now describe the similarities and differences between their work and the work presented in this chapter.

Firstly, while the aims of our work are similar because we both make predictions of article level risk of bias and sentence level relevance, there are also important differences. For the article level learning Marshall et al. aim to classify articles as *low* and *not-low*. While we also train our article level models to predict this binary variable, we propose two distinct aims for the article level learning which focus on assisting systematic reviews. These aims are 1) to rank articles by risk of bias and 2) to determine which articles have been predicted with enough certainty by a model that only a single human reviewer is needed rather than the standard two reviewers. These aims correspond to a ranking task and a classification task using custom thresholds, respectively. For the sentence level learning Marshall et al. aim to classify sentences as *relevant* and *not-relevant*. While we also train our sentence models to predict this binary variable, our aim is to achieve high ranking performance such that *relevant* sentences are ranked before *not-relevant* sentences, rather than to classify sentences as *relevant* or *not-relevant*.

The larger dataset allows Marshall et al. to investigate the use of more complex models. While we learn to predict the risk of bias of each domain individually, Marshall et al. learn these together with a multi-task learning approach. Furthermore, they investigate the use of sentence predictions to improve the article level predictions by adding extra parameters to the article model that describes which words are contained in relevant sentences. In contrast, we learn models for the sentence level and article level independently. A similarity of our modelling approaches is the use of linear models.

Marshall et al. evaluate their models as classifiers (using metrics such as accuracy), whereas we evaluate our models using metrics that are appropriate for each of our three objectives, including the AUC to evaluate ranking performance for objectives 1 and 2. Our work includes an experimental comparison of performance using the full text compared with using only the title and abstract from the PubMed database, whereas Marshall et al. only assess performance using the full text. It is difficult to compare empirical results because we use different evaluation approaches. However, conclusions of both works are positive regarding the potential for using text mining approaches for

risk of bias assessments.

Part II

Assisting hypothesis selection

Chapter 6

Background

In this part of the thesis we present a novel method that uses a hypothesis-searching approach to screen for causal associations in a potentially large hypothesis space. The approach has been submitted for publication [26]. In this chapter we provide an overview of current approaches that search for hypotheses in epidemiology, discuss the issues of causality and confounding in observational epidemiology and provide an overview of an approach called Mendelian randomisation that can help elucidate causal effects.

6.1 Hypothesis-free approach to hypothesis selection

As mentioned in Chapter 1, epidemiology is typically hypothesis-driven, using prior knowledge to specify a hypothesis to be tested. However this can bias epidemiological research to hypotheses where there is a prior belief that an association exists. Also, the analyst's research interests and preconceptions about the composition of causal pathways may affect the hypotheses they decide to test. This means that some 'popular' hypotheses may be tested many times by different research teams, while others may never be tested.

An alternative approach is to use hypothesis-free methods to identify hypotheses to test further. This approach is valuable where no strong prior knowledge exists to indicate which hypotheses should be investigated. Instead of a researcher deciding which hypothesis to test, this is done systematically using an automated process to search a potentially large number of hypotheses. The set of hypotheses that are tested is known as

the hypothesis space [104]. For instance in a genome-wide association study (discussed in Section 6.2.1) the hypothesis space is the set of linear associations of a particular trait such as BMI, with all (typed) genetic loci. The hypothesis-free approach tests all hypotheses in the hypothesis space in order to select those to be followed up with a hypothesis-driven analysis on a different dataset. The prior knowledge used to choose hypotheses for hypothesis-driven analyses is generated by this data analysis.

6.2 Current hypothesis-free approaches

Hypothesis-free approaches are being increasingly used to find associations in individual level data. Here we discuss three such methods: the genome-wide association study (GWAS), the environment-wide association study (EWAS) and the phenome-wide association study (pheWAS).

6.2.1 The genome-wide association study (GWAS)

Candidate gene studies test the association of a specific region in the genome with a phenotypic trait. Prior knowledge is used to suggest the candidate locations in the genome that may be causally associated with a trait. The researcher then tests this location to determine if the trait is associated with genetic variation at this location.

Historically, the results of candidate gene studies have been shown to be largely non-replicable [24, 25]. There are several factors contributing to this. First, there is a large number of genetic variants that could potentially become a candidate for any particular phenotypic outcome but only a relatively small number of these may actually have an effect on the outcome. This means that unless the priors used to choose a hypothesis are strong far more often than not a study will test an association that is null. This means that over all candidate gene studies the type 1 error rate – the proportion of 'null' associations incorrectly found to be 'significant' – is high. Second, typically these studies use inadequate significance thresholds (to determine if an effect has been found) and small sample sizes, which further increase the type 1 error rate [105, 106]. Finally, all these issues are further compounded by publication bias, where 'significant' results are published whereas those identified as 'null' remain unpublished.

Hypothetically we can imagine 500,000 genetic variants are tested separately in different candidate gene studies, of which 100 affect a particular trait. We would expect that an $\alpha = 0.05$ threshold would identify around $499,900 \times 0.05 = 24,995$ associations incorrectly. The number of true associations identified would depend on the test power. In the best case all 100 true associations would be identified, and the proportion of results identified as an association (according to significance testing) that actually are false associations, called the false discovery rate (FDR) is approximately $\frac{24995}{100+24995} = 0.996$. Assuming only positive results are published, due to publication bias, this is then the proportion of published associations that are false. If we have a better prior about the hypotheses likely to be true then the FDR could be reduced because the relative proportion of true associations of those tested would increase.

The genome-wide association study (GWAS) approach has provided an effective screening step, used to identify genetic markers to then follow up with candidate gene studies. The GWAS approach tests all typed markers (typically around 500,000) for an association with a phenotype. This approach uses a Bonferroni adjusted threshold, where the threshold P value used to determine if an identified association is likely to be real is adjusted to account for the number of tests performed. The Bonferrroni adjusted threshold t_b is given by $t_b = \frac{t}{n}$ where t is the original threshold and n is the number of tests performed. Commonly, $t = 0.05$ such that the Bonferroni adjusted threshold is $t_b = \frac{0.05}{n}$. The Bonferroni adjusted threshold of GWAS is typically set as 5×10^{-8}, to account for the approximate number of tests performed in a single GWAS study (which assumes 1 million SNPs such that $\frac{0.05}{1{,}000{,}000} = 5 \times 10^{-8}$). GWAS studies therefore require large enough samples to give sufficient power to detect associations using this threshold [107]. GWAS give more replicable results for two reasons [105, 108]. Firstly, the more stringent significance threshold means that 'significant' associations are less likely to be false associations. Secondly, when GWAS are published in the literature they show not only these 'significant' associations, but also the tests performed that did not show an association. This means that GWAS are not affected by publication bias because null results are also published.

While the stringent threshold means that 'significant' associations are less likely to be false associations, this also means that true associations that do not meet significance using this threshold are missed. To date, the genetic variation known to affect a particular trait explains only a small proportion of the estimated variation attributable to

genetics. This has been referred to as the missing heritability, and is likely to be in part due to the stringent threshold used for GWAS [1] [109]. For example, the largest GWAS to date for body mass index identified 97 loci meeting genome wide significance ($P < 5 \times 10^{-8}$) [110]. These loci account for $\sim 2.7\%$ of variation in BMI. BMI is affected by both genetic variation and environmental factors, and it has been estimated that genetic variation accounts for 40 - 70% of the variance of BMI [111, 112], much higher than the variation explained by the identified loci.

6.2.2 Environment-wide association study (EWAS)

Hypothesis-searching has been applied to the task of identifying associations between phenotypic variables, using the environment-wide association study (EWAS) approach [113–115]. For instance, one study [114] tested the association of a collection of environmental factors with lipid levels.

EWAS search for associations between observational variables (excluding genetic and epigenetic variables), and this is a major limitation of this study design. Whilst observational associations are useful for establishing risk factors to determine who in a population is at risk of developing a disease, this type of association does not provide evidence of causal effect between two traits. The observational estimates may be biased estimates of causal effects due to confounding. We discuss the issues of causal inference and confounding between phenotypes in Section 6.3. The EWAS design can be improved by using an approach called Mendelian randomisation to identify associations that are likely to be causal.

6.2.3 Phenome-wide association study (PheWAS)

The phenome-wide association study (pheWAS) approach tests the association of a potentially large number of phenotypes with a small set of genetic variants [116]. This is similar to the GWAS approach which also tests associations between genetic variants and phenotypes. The difference is that pheWAS focus on a small number of genetic variants and a large number of phenotypes, whereas the opposite is true for GWAS.

[1] Other sources of genetic variation other than allele dosages may also contribute to the missing heritability such as epigenetic variation and interactions between genetic variants.

Initially the pheWAS approach was used with electronic health record (EHR) data. An early example used the International Classification of Disease (ICD) codes to construct a dataset of phenotypes [117]. They tested the association of these phenotypes with 5 SNPs that had previously been associated with a set of diseases including multiple sclerosis and rheumatoid arthritis. Their analysis replicated several known associations between SNPs and diseases, and also identified other novel associations.

PheWAS has also been performed using cohort studies, and this tends to be less restrictive compared with the use of EHR data because these studies often have a more diverse range of phenotypes. One recent study sought to identify pleiotropy, where a genetic variant is associated with more than one phenotype. They tested the association of 80 SNPs with 1,008 phenotypes, using the National Health and Nutrition Examinations Surveys (NHANES) dataset [118]. This dataset includes three different surveys with participants of 3 different ethnicities. Hence the analysis could be performed separately within these partitions, and associations identified only when replication was achieved.

While the PheWAS approach can help to identify phenotypes associated with the same genetic variant, the results can tell us little about the relationship between these phenotypes. An association between a genetic variant and two phenotypes may occur because: 1) the genetic variant affects one phenotype that then affects the other phenotype, 2) the genetic variant independently affects both phenotypes, or 3) the phenotypes are correlated. The first and second mechanisms are known as vertical and horizontal pleiotropy and are explained further in Section 6.4. The third mechanism is due to the close relationship between many traits such that correlations exist between them. For instance, there are multiple traits related to bone content such as bone mineral density and bone mineral content, and these are highly related.

We may be particularly interested in the first mechanism, to determine whether there is a causal association between two phenotypes. In Chapter 7 we introduce a novel approach that extends pheWAS in order to investigate the causal relationship between phenotypes. We do this using an approach called Mendelian randomisation, to overcome the issues of causality and confounding present in observational studies. First we provide an overview of these issues and the Mendelian randomisation approach.

6.3 Causality and confounding in observational epidemiology

Observational phenotypes in epidemiology are highly correlated. One study performed pairwise correlations across a set of phenotypic variables and found that 54% were correlated at a $\alpha = 0.05$ significance level, compared to 5% expected by chance alone [30]. This is problematic because an association between an exposure and an outcome may be because they are both associated with a third trait, a confounding factor. Figure 6.1 shows an example of confounding, where an association between drinking alcohol and lung cancer may be found because both alcohol and lung cancer are associated with smoking. Smoking confounds the association between alcohol and lung cancer – the association is real but neither causes the other.

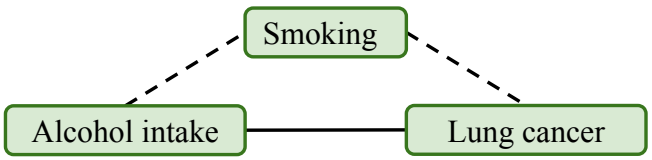


Figure 6.1: Example of confounding.

If all confounders are known and they are perfectly measured then it is possible to determine if there is an association between an exposure and outcome that is not due to confounding. Commonly this is done by adjusting for the confounding factors in the model. For instance, when testing the linear association of exposure E with outcome O we may use two models: $O = \beta_E E + \beta_0$ and $O = \beta_E E + \beta_C C + \beta_0$, where C is a confounder of the association. If the association between exposure and outcome, given by the coefficient β_E attenuates to the null when C is included in the model, then it is likely that the association is due to confounding. However, in practice it is not possible to know whether a model accounts sufficiently for confounding factors. The confounding factors may not be known, and if they are known then they may not be measured in a particular cohort. Furthermore, measurement error in observed confounders may cause residual confounding [30, 119].

6.4 Inferring causality with Mendelian randomisation

Observational analyses can infer associations between two traits, but cannot infer a causal relationship between an exposure and an outcome. A randomised controlled trial (RCT) on the other hand, can provide evidence of causality. In this study design participants are randomly assigned to intervention or placebo groups so that the treatment they are given (or placebo) is not associated with confounding factors. Any association between intervention received and outcome is then known to be due to an effect of the intervention on the outcome. An RCT however is not always feasible, as it may not be ethical to give (or withhold) a particular intervention, they are typically expensive, and they often needs to run for years to see an effect on the outcome.

Mendelian randomisation can help researchers infer causation by using an instrumental variable (IV) constructed from genetic variants [120, 121]. Instead of testing an association using an observed exposure we can use variation in the genome that is associated with the exposure. The genome, having been created at conception, precedes the outcome such that the direction of causality of any association is from genetic exposure to outcome. Furthermore, following from the two (approximate) laws of Mendelian genetics, confounding is in principle avoided. Mendel's first law (the Law of Segregation) states that the probability that any particular allele is passed from parent to offspring is independent of environment, such that genetic variants are not normally associated with factors that often confound associations between observed traits. Mendel's second law (the Law of Independent Assortment) suggests that the genetic variants are inherited independently of each other, such that an association between a genetic variant and a trait cannot (in principle) be due to an association with another genetic variant that confounds the association. We note however, that after Mendel proposed these laws it was found that the second law was not always true because genetic confounding can occur, and we discuss this in more detail below.

Mendelian randomisation is similar in principle to the randomisation approach of RCTs. As already discussed, the RCT study design randomly assigns participants to study groups so that the intervention they receive (the exposure) is not associated with confounding factors. Similarly, according to Mendel's second law each locus on the genome is (in principle) inherited independently to all other loci, such that they are effectively randomised with respect to each other [122].

In order for a genetic variant to be a valid instrument for an exposure, three instrumental variable assumptions need to hold [122]. Firstly, the IV should be associated with the exposure. Secondly, the IV should not be associated with factors that confound the association between the exposure and the outcome. Thirdly, the IV should be independent of the outcome, given the exposure and confounding factors. These assumptions are illustrated in Figure 6.2. Assumption 3 holds because there is no path between the genetic variants and the outcome that is not also through the exposure.

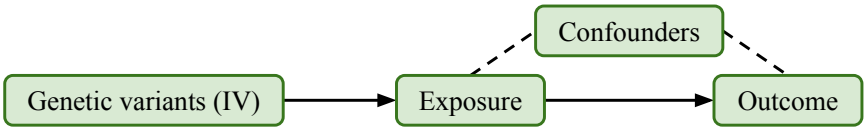


Figure 6.2: Instrumental variable assumptions. Solid arrows depict relationship we are testing. Dashed arrows depict confounding.

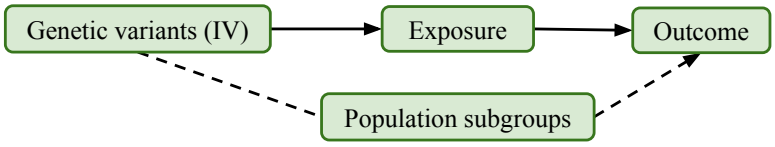
The third assumption may be violated by genetic confounding through population stratification, linkage disequilibrium and horizontal pleiotropy. These are illustrated in Figures 6.3a-c, and are each different ways in which alternative paths are generated from IV to outcome that do not pass through the exposure. Population stratification occurs when subsets of the population have a different genotype, and the outcome also varies according to these population subsets. This creates an association between the IV and the outcome. Population stratification is often caused by ethnic differences and this can be seen if associations between an exposure and outcome are not present when analysing within ethnic groups [123].

Linkage disequilibrium refers to the association between nearby genetic variants on the genome. This occurs because, contrary to Mendel's second law, alleles on the genome are not passed on independently. In particular, nearby regions on the genome are more likely to be inherited together. This can create an alternative pathway from genetic IV to the outcome if the genetic variant with which the IV is also associated, is associated with the outcome (not through the exposure). Correlation between genetic variants is less common than correlation between observed phenotypes [30] such that confounding through linkage disequilibrium is far less likely than confounding in observational epidemiology.

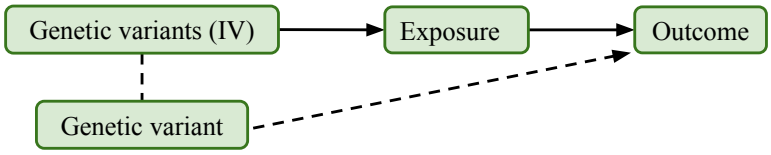
Pleiotropy refers to the effect of a genetic variant on multiple phenotypes, and, as al-

ready mentioned, there are two types – horizontal and vertical pleiotropy [124]. Vertical pleiotropy occurs when a genetic variant affects a trait and this in turn has an effect on another trait. This type of pleiotropy is the essence of the aims of Mendelian randomisation, to determine if one trait affects another as depicted in Figure 6.2. Horizontal pleiotropy occurs when a genetic variant affects multiple traits on different pathways. This is problematic if a genetic variant affects the exposure, but also a second trait that also affects the outcome, as shown in Figure 6.3c. This alternate pathway does not pass through the exposure and this invalidates the IV assumptions. In the rest of this part of the thesis, when we refer to the issue of pleiotropy we are referring specifically to horizontal pleiotropy.

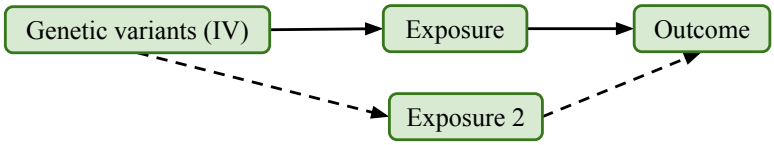
The occurrence of these confounding sources – population stratification, linkage disequilibrium and horizontal pleiotropy – cannot be directly tested for. However, there are approaches that are able to examine whether they are likely to exist and affect results to any meaningful extent, such as those we describe in Section 7.2.3.



(a) Population stratification



(b) Linkage disequilibrium



(c) Horizontal pleiotropy

Figure 6.3: Types of confounding in Mendelian randomisation analyses. Solid arrows depict relationship we are testing. Dashed arrows depict confounding.

A robust approach to Mendelian randomisation is to estimate the association of ge-

netic variants directly with the outcome. This can be done with a simple linear regression model of the outcome on the genetic variants. This provides a valid test of whether an exposure causes an outcome and only depends on the three core instrumental variable assumptions described above [125]. In order to estimate the size of the effect of an exposure on an outcome, the exposure phenotype must also be used in the analysis (or have been used in a subsequent analysis as is the case of two sample Mendelian randomisation [121]). For instance, a two stage least squares analysis first estimates the effect of the genetic IV on the exposure, and then estimates the effect of the exposure on the outcome using the predicted exposure values from the first stage.

When the exposure is used in the analysis the instrumental variable assumptions may be invalidated in other ways. For example, if the instrument affects the outcome through the exposure phenotype at other time points than those included in the analysis then this invalidates the third instrumental variable assumption, because this means there is a path from the instrument to the outcome that does not pass through the exposure. For instance, if body mass index (BMI) at age 2 and at age 25 both affected coronary heart disease the published allele score for BMI cannot identify the independent effects of BMI at these time points [126]. Furthermore, researchers must impose stronger assumptions to estimate the size of the effect of the exposure on the outcome, referred to as point identifying assumptions. For example, epidemiologists have commonly assumed constant treatment effects or no effect modification for continuous outcomes, or no effect modification for binary outcomes [127]. Researchers can investigate the validity of the point identifying and core instrumental variable assumptions if multiple genetic variants are associated with the exposure. If two or more variants affect an exposure through different causal pathways, and the core instrumental variable assumptions and either of the point identifying assumptions hold, the variants should estimate the same size of causal effect of the exposure on the outcome [122].

6.5 Summary

In this chapter we have discussed current hypothesis-searching approaches in epidemiology. We have given an overview of the issues of causality and confounding in observational epidemiology and discussed how Mendelian randomisation can help to determine the causal effect between two traits. In the next chapter we introduce a new

approach that searches for causal associations in a potentially large set of hypotheses, using Mendelian randomisation. We demonstrate this approach by searching for the causal effects of BMI.

Chapter 7

Methods

In this section we illustrate how Mendelian randomisation can be used to investigate the effects of an exposure with a large set of outcomes, to identify outcomes potentially causally influenced by an exposure. As with GWAS, this is a screening approach where identified associations need to be validated through replication studies. We illustrate this method with an example application – searching for the causal effects of BMI.

7.1 PheWAS with causal inference: a new approach to identify potentially causal hypotheses

We present a general framework to search for causal associations of a exposure with a potentially large number of outcomes, using Mendelian randomisation. While Mendelian randomisation is becoming increasing used to test for causal effects, to date (to our knowledge) there have been no hypothesis-free analyses using this method. Figure 7.1 shows an outline of our approach. We use a two-stage approach to identify potentially

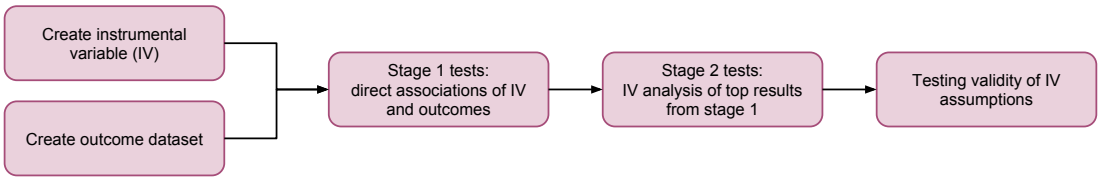


Figure 7.1: Pipeline of the Mendelian randomisation hypothesis generation approach.

causal relationships, and then follow this with investigations into the validity of the IV assumptions. The first stage uses linear regression to test the association between a genetic IV and the outcomes directly. This identifies a set of outcomes that may be causally affected by the exposure, but does not estimate the effect of the exposure on the outcomes. We use an arbitrary threshold of $P < 0.05$ to select outcomes to take forward to the second stage. The second stage estimates the effect of the exposure on the outcome. Finally, we investigate the validity of the IV assumptions, to assess the reliability of our estimates calculated in the second stage analysis.

This approach is scalable because the number of tests grows linearly in the number of outcomes. Given a single outcome N tests are performed, if this is increased to 100 then $100 \times N$ tests would be performed.

Whilst we give a general framework here, the specific methods may vary between studies. These details include how the instrumental variable is constructed. The IV may simply be the allele dose of a single SNP (the count of a particular allele at this locus), or it may be a score calculated as the sum of the allele doses of a set of SNPs known to be associated with the exposure. Furthermore, it is also possible to include weights to create a weighted score that is the sum of the allele dosages, weighted by the size of the effect of each SNP on the exposure. This is only possible when the effect sizes have been calculated in a different study to the one used for the Mendelian randomisation analysis, as otherwise the score may overfit to this data. In our example application the IV is a weighted-score using the 32 SNPs known to be associated with BMI (our exposure), and more details of this are given in Section 7.2.2.

The number and strength of SNPs known to be associated with the exposure also determines the degree to which the IV assumptions can be assessed, the last step in our pipeline. In the following analysis we use tests that make comparisons across SNPs or subsets of SNPs (see Section 7.2.3). Therefore it is clear that this is only possible where multiple SNPs are available. Furthermore, there may be insufficient levels of power when using SNPs or a set of SNPs individually, such that these tests will lack power.

We now apply this general pipeline to a specific example – identifying the causal effects of BMI.

7.2 Example application: searching for the causal effects of BMI

BMI has a close relationship with many traits and is associated with diseases such as type 2 diabetes [128] and cardiovascular disease [129]. An association between BMI and a phenotype may be due to confounding or because the phenotype affects BMI, rather than because BMI affects the phenotype. For instance, clinical trials have shown BMI is affected by behavioural factors such as diet and exercise [130, 131]. We build upon previous Mendelian randomisation studies that have investigated the causal effects of BMI on inflammation, cancer, age at menarche, diabetes, atherosclerosis risk and blood pressure and hypertension [132–140] and bi-directional studies that have analysed the effects of BMI and a second exposure such as C-reactive protein [136, 141], serum uric acid [142, 143], vitamin-D [144] and fetuin-A [145].

In this section we describe the methods used for this analysis. Figure 7.2 shows the generic pipeline (Figure 7.1) with annotations detailing the specific methods of this analysis.

7.2.1 Study population

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a prospectively collected pregnancy cohort that recruited pregnant women with expected delivery dates between April 1991 and December 1992 from Bristol, UK (see [146–148] for the study

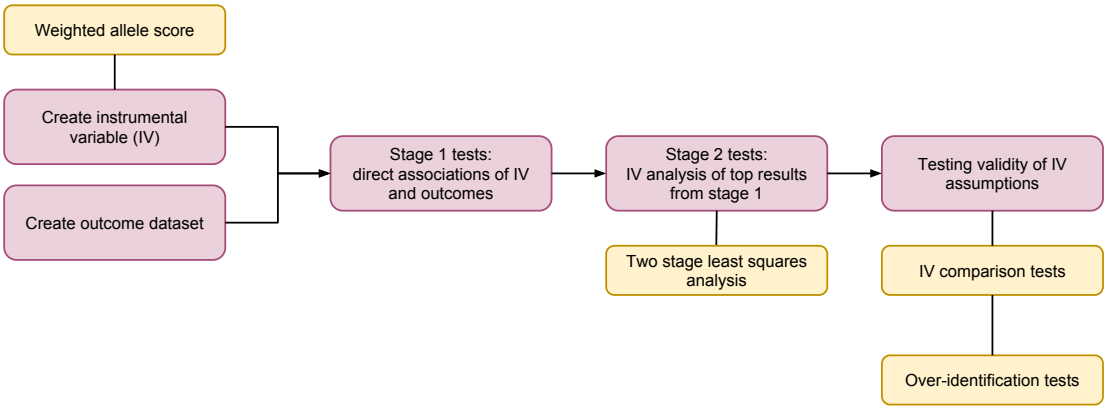


Figure 7.2: Specific pipeline for our example: finding the causal effects of BMI.

details). Ethical approval was obtained from the ALSPAC Law and Ethics Committee and local research ethics committees.

A total of 9,912 ALSPAC children were previously genotyped using the Illumina HumanHap550 quad genome-wide single nucleotide polymorphism (SNP) genotyping platform by the Wellcome Trust Sanger Institute, Cambridge, UK and the Laboratory Corporation of America, Burlington, NC, USA. We now describe the quality control steps previously performed on this genotypic data. Individuals were excluded from further analysis on the basis of five criteria that each indicate quality of the data. Firstly, individuals with incorrect sex assignments when comparing the recorded sex to that found by genotyping. Secondly, individuals with a very low or very high proportion of heterozygous gene loci, where the two alleles of a locus are different (using thresholds of < 0.320 and > 0.345 for the Sanger data and < 0.310 and > 0.330 for the LabCorp data). Thirdly, individuals with a disproportionate level of missingness ($> 3\%$ of SNPS) because this indicates that the SNPs may be incorrectly typed.

Fourthly, individuals having similarity on the genome that indicate individuals may be related (have a shared common ancestor) even though they may be unaware of this. This is called cryptic relatedness and is determined using a measure of genetic relatedness between individuals that indicates the likelihood that these individuals are related, called identity by descent (IBD). Individuals with $IBD > 10\%$ are removed from the dataset. It is important that individuals in the study are not related to each other as this may bias the result such that it is not representative of the population as a whole [149]. Lastly, individuals of non-European ancestry[1] were removed in order to restrict our analysis to individuals of white European ethnic origin to reduce the potential for population stratification, which could confound associations between the BMI allele score and the outcomes. The resulting data set consisted of 8,365 individuals and 488,311 autosomal SNPs.

Three criteria were used to assess SNPs for quality control. Firstly, SNPs with a minor allele frequency of $< 1\%$ were removed because this reduces the power to detect associations such that false positive associations are more likely. Secondly, SNPs with a call rate of $< 95\%$ were removed, because if values cannot be inferred for a

[1]Non-European ancestry was detected by a multidimensional scaling analysis seeded with HapMap 2 individuals, EIGENSTRAT analysis revealed no additional obvious population stratification and genome-wide analyses with other phenotypes indicate a low lambda

large proportion of the sample for a particular SNP then this indicates the typing of this SNP may be difficult and hence have poor quality. Lastly, only SNPs which passed an exact test of Hardy-Weinberg equilibrium ($P > 5 \times 10^7$) were considered for analysis, because deviations from HWE indicate there may have been errors in the genotyping process [149].

The autosomal variants were imputed with MACH 1.0.16 Markov Chain Haplotyping software, using CEPH individuals from phase 2 of the HapMap project (HG18) as a reference set (release 22). After quality control assessment and imputation the data set consisted of 8,365 non-related children of European descent with 2,608,006 SNPs available for analysis. The number of SNPs increases because the imputation procedure is able to impute SNPs that were not initially genotyped, using the related structure of the genome due to linkage disequilibrium. Of these 8,365 we removed 244 individuals with no data for all outcomes giving a sample size of 8,121.

7.2.2 Exposure and outcomes

BMI allele score exposure We created an allele score of the BMI variants, constructed using a weighted sum of 32 loci known to be associated with BMI (listed in Table B.1). The weights were generated from the effect size of BMI associated SNPs found in a large GWAS [150]. This GWAS did not include the ALSPAC study [151]. We constructed the score in terms of the number of BMI-increasing alleles so that a higher score corresponds to a higher BMI:

$$score(i) = \sum_{l \in loci} \begin{cases} d_{l,i} \times e_l & \text{if } d_{l,i} \text{ is BMI increasing} \\ (2 - d_{l,i}) \times e_l & \text{otherwise} \end{cases}$$

where d is the number of BMI-increasing alleles of individual i such that $0 \leqslant d \leqslant 2$, and e_l is the effect size of loci l, scaled relative to the effect of *FTO* which has the largest effect size of these loci. We standardised the BMI allele score to have a mean of zero and standard deviation of one.

Phenotypic outcomes We compiled a set of 172 continuous variables from the ALSPAC dataset, comprising a range of variables recorded between birth and 15 years old, including primary measures (from questionnaires or focus clinics) and also derived

variables. The dataset was compiled by selecting a set of complete clinic assessment based data files from the ALSPAC cohort, each corresponding to a separate measurement event.

The intention is to include a random subset of available clinic measures, rather than select variables where we have an a priori interest or evidence in their association with BMI. The data files were processed in turn to reduce the size of the dataset by manually removing variables where multiple similar variables were found. We did this by including a composite score measure where available (and removing its component phenotypes) or keeping only one measure from each similar group of variables. This delivered a diverse range of 160 arbitrarily selected variables at a range of time points to give a rich outcome dataset (given in Table B.3).

We also included a selected set of outcomes, as we need to ensure that the dataset contains variables both with and without previous evidence of an association with BMI, such that we can validate our screening approach. We therefore included the following 12 outcomes (in addition to the 160 that were randomly selected), previously suggested to be associated (perhaps causally) with BMI: glucose [138], insulin [138], leptin [136], age at menarche [133], systolic blood pressure [132] and C-reactive protein [136, 141], intelligence and attainment measures (Wechsler Intelligence Scale for Children (WISC), Diagnostic Analysis of Nonverbal Accuracy (DANVA), and literacy scores (2 phenotypes)) [152], lung function [153] and the Home Observation for Measurement of the Environment (HOME) score [154]. Further details are given in Table B.2. The number of participants with a value for each outcome varies across outcomes. Where outcomes were available at multiple time points we used the most complete measure. We removed values of each outcome that were coded as missing, and refer to this as the original dataset.

7.2.3 Statistical methods

We performed all analyses using Stata v11.2 (StataCorp LP, 2009; College station, TX, USA) [155]. We follow the general pipeline given in Figure 7.1, such that our main analysis to search for associations was a two-stage process. The first stage involved a large-scale analysis to screen for associations of the BMI allele score with all outcomes in our dataset. The second stage followed up 'top' associations identified in the first

stage, with an IV analysis and sensitivity analysis assessing the degree of pleiotropy.

Stage 1 tests

In the first stage we began by transforming the outcome variables in order to harmonise this dataset, such that a single analytical approach can be applied in the subsequent BMI score screening step. We used a rank-based inverse normal transformation to ensure all outcomes were normally distributed, and standardised these to give distributions with a mean of zero and standard deviation of one. We tested the associations of the BMI score with all transformed outcomes, using univariate linear regression analysis, with robust standard errors (the robust option). We ordered the resulting associations by P value to rank the associations from strongest to weakest (where a rank of 1 denotes the strongest result). The rank position gives an indication of the relative strength of associations of the allelic score with the outcome variables. We identified associations of outcomes and allele score with, as an illustration, a nominal $P < 0.05$ and took these forward for further tests in the second stage analysis.

In addition to the P values of these tests we report Bonferroni adjusted P values calculated by multiplying the P values by 160 to account for the number of tests performed. We exclude the validation set from these calculations as we have selected these phenotypes based on prior knowledge. The Bonferroni adjusted P values are conservative because the Bonferroni correction assumes the tests are independent but this is not the case for the outcomes in our dataset. We determine the proportion of our top results that are expected to be false positives, the false discovery rate (calculated as the expected number of results with P value < 0.05 by chance alone (160×0.05) divided by the number of results found with a P value < 0.05). We also report alternative permutation P values – the probability that an outcome at rank i, would be found at rank j where $j \leqslant i$ given there is no association between the BMI score and all outcomes. These are estimated using permutation testing, and we again exclude the validation set from this analysis. We permute the values of each outcome variable across participants and repeat the stage 1 analysis, performing linear regression of the BMI score on each outcome and generating a ranking of associations. We repeat this 5,000 times to derive an empirical distribution of the rank position of each outcome. For each outcome at rank i we report the proportion of these tests where the outcome is found at rank j

where $j \leqslant i$. This gives a P value that accounts for the tests with the other 159 outcomes in the dataset.

Stage 2 tests

In the second stage of our analysis we tested each stage 2 outcome (that had an association with $P < 0.05$ in stage 1) with a formal instrumental variable analysis using two-stage least squares regression (the Stata *ivregress* command). Although the BMI allele score is a risk factor for lifelong BMI we did not observe lifelong BMI and so instead use this score to estimate the effect of BMI at a single time point, at age 8, on the outcomes. We log-transformed then standardised BMI at age 8 so that its distribution was approximately normal with a mean of zero and standard deviation of one. We used the original outcome dataset (rather than the inverse normal transformed version) and transformed any variables with skewed distributions (identified visually) to give distributions that were approximately standard normal (with mean of zero and standard deviation of one). We converted outcomes with distributions that were not normal and not right skewed to binary variables with approximately equal numbers in each group. We used linear and logistic regression for the second stage of the instrumental variable analysis for normally distributed and binary outcomes respectively. Finally, we also tested the associations of observational (log-transformed) BMI at age 8 using the same protocol with the 172 outcome variables, for comparison.

Testing validity of the instrumental variable assumptions

As discussed in the previous chapter, Mendelian randomisation tests require the IV assumptions to be satisfied. These assumptions are; 1) the genetic IV is associated with the exposure (observational log BMI at age 8), 2) the genetic IV only affects the outcome through its effect on the exposure, and 3) the genetic IV is independent of all factors confounding the association between BMI and the outcomes. We used univariate linear regression (the Stata regress command) to test the strength of the BMI allele score as an instrument for observational BMI, where a larger F-statistic implies greater power [156]. We cannot directly test assumptions 2 and 3, but we can look for evidence that the assumptions do not hold. If the 3 core instrumental variable assumptions hold,

and either of the point identifying assumptions hold[2], the estimated effect of BMI on an outcome should be consistent across different variants. We explored this in two ways, each of which compares the results when different instrumental variables are used. The first compares the results using two independent instrumental variables: 1) *FTO* (the SNP most strongly associated with BMI), and 2) the remaining 31 variants (we refer to this as the 31-allele score). *FTO* explains around a quarter of the variance of BMI explained by the other 31 combined (31 variants were associated with BMI $r^2 = 0.0215$, *FTO* was associated with BMI $r^2 = 0.0055$). Furthermore, we tested the strength of the associations of each instrument to ensure they were both strongly associated with BMI.

The second approach we used compared the effect estimates across the variants individually. We estimate these individual effects by performing a single IV regression (for each outcome), with separate IV's in this model for each of the 32 genetic variants. This is known as an over-identification test[3], because when there are more instruments (genetic variants) than dependent variables (exposures), an instrumental variable analysis is referred to as over-identified. We then test for differences in these effects using Hansen tests [159]. The null hypothesis states that there is no evidence of differences in the IV effect estimates between different variants. Thus rejection of this test suggests there are differences between the estimates based on each of the variants. This may suggest that the instrumental variable assumptions do not hold, for example if the effects of the variants are not solely mediated through BMI.

Sensitivity analyses We present results using the original data as our main analyses, where the sample size for each test varies depending on the outcome. This creates a potential for differences in P values to be caused by differences in sample size, or bias due to missing data. In order to assess the impact of this we repeated our analysis using an imputed dataset. We compared the ordering of the outcome variables by P value across the original and imputed versions using Spearman's rank correlation.

The imputed dataset consisted of all 8,121 individuals and 172 variables in the orig-

[2]Estimating an effect is referred to as point identification and an extra assumption is needed to be able to do this. These assumptions are either constant treatment effect or monotonicity for continuous outcomes, and no effect modification or monotonicity for binary outcomes. For example, the assumption of constant treatment effect says that the causal effect of the exposure on the outcome does not change across the population. For further details we refer the reader to [127, 157, 158].

[3]Over-identification tests were performed with continuously updating estimator (CUE) via generalized method of moments (GMM).

inal dataset. We used multiple imputation using chained equations (ice command in Stata), to impute missing values for all variables, and generated 20 imputation data sets [160]. We used predictive mean matching (match option) for non-normal (or log-normal) variables because it does not assume normality, to prevent extrapolation beyond feasible values. To inform the imputation we included additional socio-economic position (SEP) variables which may help to explain missingness: household social class, maternal education, smoking during pregnancy, and ethnicity. The purpose of this is to satisfy the missing at random (MAR) assumption of the imputation method; the probability of missingness does not depend on the missing data conditional on the observed data. We included the BMI allele score and all outcomes in our imputation, to inform the prediction of each outcome. The large number of variables in our dataset should also help to satisfy the MAR assumption, as the variable set should include variables predictive of both the variables and missingness of the variables [161].

7.3 Summary

In this chapter we have presented a novel approach to search hypotheses for potentially causal relationships to follow up. We have described the methods of a proof-of-principle analysis to search for the causal effects of BMI. In the next chapter we present the results of this work.

Chapter 8

Results

In this chapter we present the results of our proof-of-principle analysis to search for the causal effects of BMI, the methods of which were described in Chapter 7.

8.1 Crude associations

The association between the BMI allele score and observed BMI across childhood strengthens with age and stabilizes at around age 10 (Table 8.1). A standard deviation (SD) increase in BMI allele score was associated with a 0.163 SD increase in log BMI at age 8 (95% confidence interval (CI): 0.14, 0.19, F=140.66). Furthermore, we found little evidence of associations with common socio-economic confounders compared with many strong associations for observational BMI at age 8 (Table 8.2). These tests support the notion that the BMI allele score may be a valid instrument for life-long BMI.

8.2 Results of stage 1 and stage 2 tests

Our stage 1 tests found the BMI allele score was associated with 21 outcomes, using an unadjusted $P < 0.05$ threshold (Table 8.3). Of these, 14 outcomes were from the 160 outcomes we randomly included in our dataset (test of proportions $P = 0.030$), compared to 8 expected by chance alone (160×0.05, making the conservative assumption that all outcomes are uncorrelated). Hence we would expect 6 of the 14 identified

Number with measured BMI	Mean age at BMI measurement	SD increase of log BMI during childhood per 1 SD increase of BMI allele score with 32 SNP variants			SD increase of log BMI during childhood per 1 SD increase of *FTO* allele dosages			SD increase of log BMI during childhood per 1 SD increase of BMI allele score with 31 SNP variants (*FTO* removed)		
		SD change	95% CI	F-statistic[1]	SD change	95% CI	F-statistic[1]	SD change	95% CI	F-statistic[1]
6,601	7.48 wks	0.016	-0.01, 0.04	1.65	-0.018	-0.04, 0.01	2.12	0.025	0.00, 0.05	4.22
6,282	40.50 wks	0.032	0.01, 0.06	6.36	-0.029	-0.05, 0.00	5.19	0.047	0.02, 0.07	14.10
5,797	1.69 yrs	0.036	0.01, 0.06	7.42	-0.041	-0.07, -0.02	9.66	0.056	0.03, 0.08	18.78
5,430	3.71 yrs	0.068	0.04, 0.09	25.29	-0.004	-0.03, 0.02	0.10	0.075	0.05, 0.10	31.39
6,076	7.57 yrs	0.142	0.12, 0.17	125.00	0.049	0.02, 0.07	14.88	0.133	0.11, 0.16	109.75
5,087	8.68 yrs	0.163	0.14, 0.19	140.66	0.074	0.05, 0.10	28.18	0.146	0.12, 0.17	112.70
5,623	10.68 yrs	0.175	0.15, 0.20	175.27	0.090	0.06, 0.12	45.96	0.150	0.12, 0.18	128.80
5,116	12.80 yrs	0.170	0.14, 0.20	151.21	0.082	0.05, 0.11	35.28	0.150	0.12, 0.18	116.58
4,746	13.83 yrs	0.167	0.14, 0.19	134.07	0.074	0.05, 0.10	26.78	0.150	0.12, 0.18	107.05
4,174	15.45 yrs	0.158	0.13, 0.19	106.35	0.071	0.04, 0.10	21.60	0.142	0.11, 0.17	84.49
2,665	17.04 yrs	0.176	0.14, 0.21	83.94	0.086	0.05, 0.12	20.14	0.153	0.12, 0.19	63.13

Table 8.1: Associations of BMI allele score with BMI across childhood.

Abbreviations: BMI, body mass index; CI, confidence interval; SD, standard deviation; wks, weeks; yrs, years

All BMI variables are log transformed. Stata ivregress command used with robust standard errors (robust option).

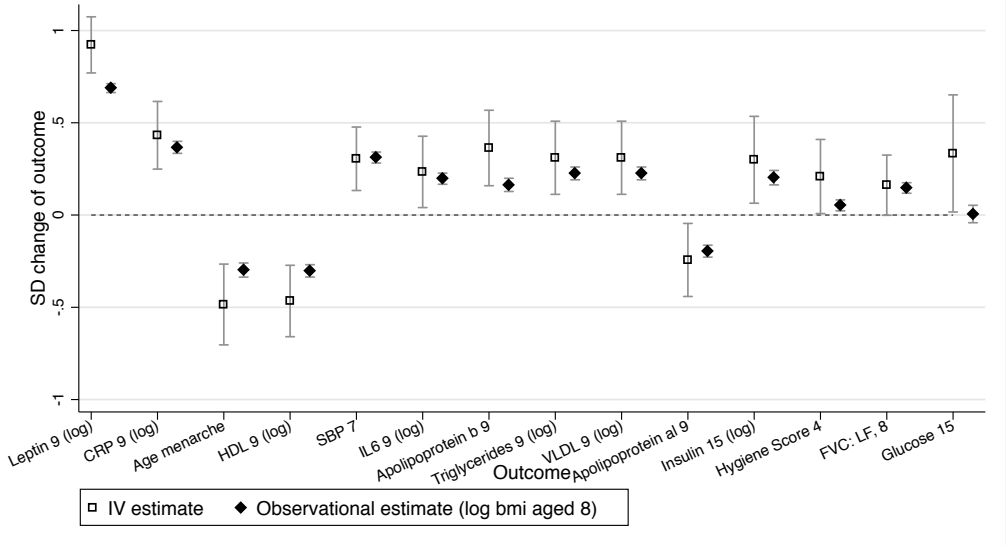
BMI calculated as weight (kilograms) divided by height (metres squared).

Based on the original data of 8,121 participants (variable sample size per BMI measurement).

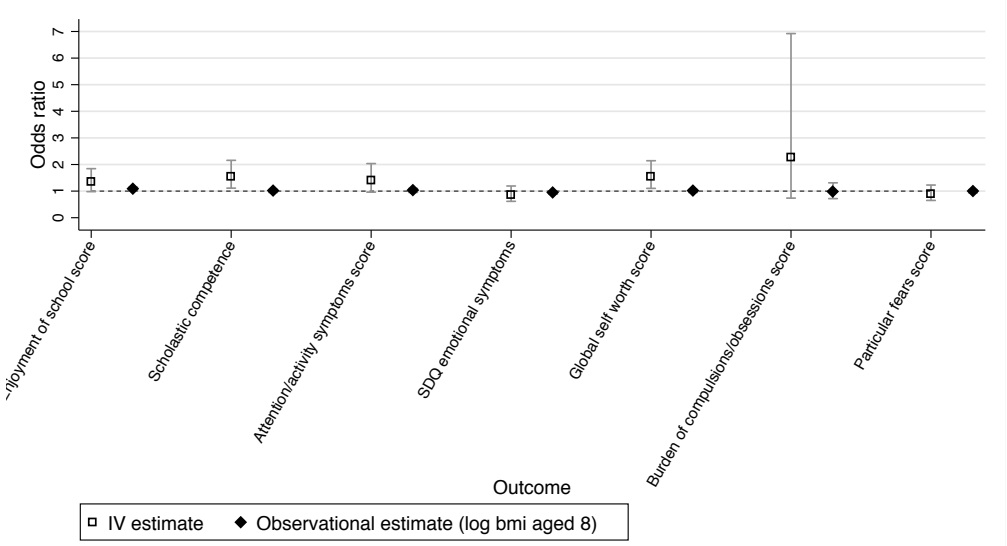
[1] F-statistic: Measure of strength of association between exposure and outcome.

outcomes to be true associations (false discovery rate of 0.571). We found stronger associations than would be expected by chance, illustrated by the QQ plot in Figure 8.2, suggesting that BMI affects many outcomes. After Bonferroni correction only HDL at age 9 was found with a P value below $P < 0.05$ whereas using the permutation P values we found 8 associations with $P < 0.05$. In comparison, we found 57 stage 1 associations with $P < 0.05$ using observational BMI at age 8. Of these, 48 were from the 160 randomly included in our dataset (test of proportions $P \leqslant 0.001$), compared to 8 expected by chance alone. The instrumental variable effect estimates (stage 2 results) are given in Table 8.4, Figure 8.1a and Figure 8.1b (and observational estimates are also provided for comparison).

The stage 1 direct tests identified several known associations, such as with leptin and age at menarche. The two-stage least squares IV analysis estimated that a 1 SD increase in log BMI at age 8 increased leptin at age 9 by 0.92 SD (95% confidence interval (CI): 0.77, 1.07). A 1 SD increase in log BMI age 8 was associated with a 201.7 day earlier age at menarche (95% CI: 112.3, 291.1). We also identified novel effects of BMI. For instance, a 1 SD increase in log BMI at age 8 increased the odds of having a global self-worth score $\geqslant 20$ by 54% (95% CI: 1.10, 2.14). We list all outcomes in our dataset in Table B.3 and the results of the stage 1 tests in Table B.4, so that readers can view results where the CI includes the null value. The observational estimates were nearer the null than the IV estimates, and we found differences between the IV and observational estimates for 6 phenotypes, using the Wu-Hausman test (Table 8.4).



(a) Continuous outcomes



(b) Binary outcomes

Figure 8.1: A comparison of the observational and instrumental variable estimates.
Continuous outcomes: The standard deviation change of outcome for a 1 SD increase of log BMI aged 8. IV estimate of effect using two-stage least squares regression of log BMI at age 8 as the exposure, with robust option. Observational estimates are the SD change of the outcome for a 1 SD increase in log BMI at age 8.

Binary outcomes: Odds ratio between groups of outcomes, for a 1 SD change of log BMI aged 8. Observational estimates are the odds ratio between outcome groups. Categories for binary variables given in Table B.6.

Graphical illustration of the results in Table 8.4.

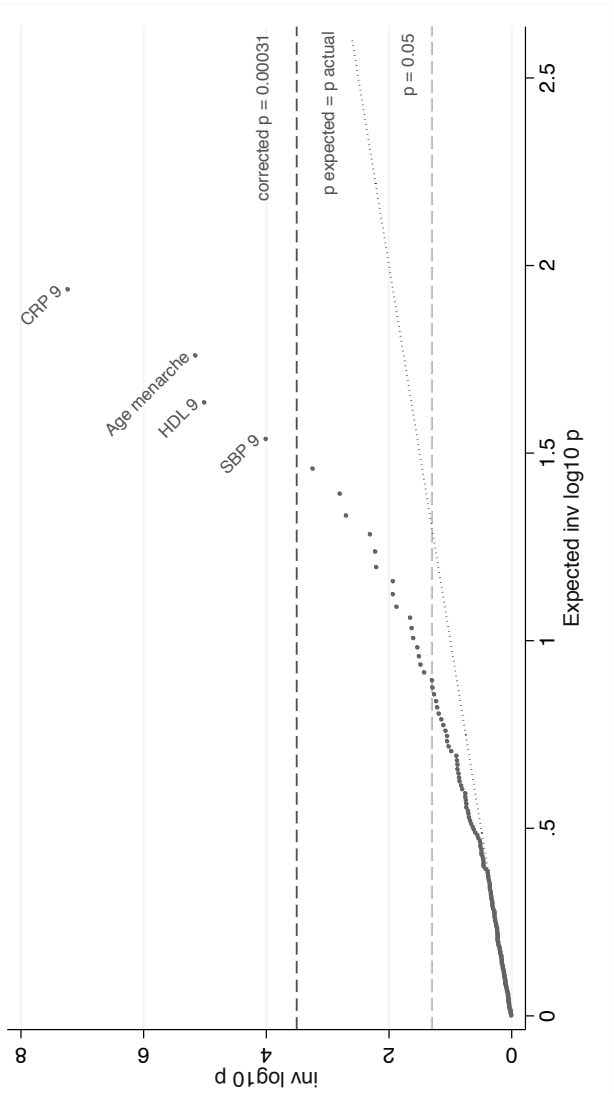


Figure 8.2: QQ-plot of the associations between the BMI allelic score and the 172 outcomes.
Association of log BMI age 8 with outcomes, of the stage 1 tests. Using the original dataset with variable number of individuals for each outcome. Tests performed with the Stata regress command and robust option. Top result leptin is not shown as P value too small. Corrected P = 0.00031 line: The Bonferroni corrected P = 0.05, accounting for the 160 tests (excluding validation set) performed. P expected = actual line: The expected trajectory assuming the P values are uniformly distributed.

	SD change of **BMI allele score** for a 1 SD change of confounder				SD change of **log BMI age 8** for a 1 SD change of confounder			
Potential confounder	Sample size	SD change	95% CI	P value [4]	Sample size	SD change	95% CI	P value [4]
Maternal education [1]								
Less than O-level			Reference				Reference	
O-level	7,306	-0.02	-0.08, 0.04	0.871	4,735	-0.05	-0.13, 0.03	0.001
A-level		0.03	-0.04, 0.09			-0.05	-0.13, 0.03	
Degree or above		-0.03	-0.11, 0.04			-0.16	-0.25, -0.07	
Sex								
Female	8,121		Reference		5,087		Reference	
Male		0.00	-0.05, 0.04	0.828		0.13	0.07, 0.18	< 0.001
Household social class [2]								
Class I (professional)			Reference				Reference	
Class II		0.01	-0.06, 0.08			0.10	0.02, 0.18	
Class III (non-manual)	6,929	-0.01	-0.08, 0.07	0.776	4,554	0.09	0.00, 0.19	0.213
Class III (manual)		-0.01	-0.11, 0.08			0.10	-0.02, 0.21	
Class IV/V (manual)		0.02	-0.11, 0.14			0.07	-0.09, 0.23	
Parity [3]								
0			Reference				Reference	
1	7,408	-0.03	-0.08, 0.02	0.850	4,741	-0.03	-0.10, 0.03	0.789
2		0.01	-0.05, 0.07			0.01	-0.07, 0.08	
Mother smoking in pregnancy	7,484		Reference		4,788		Reference	
No								
Yes		0.02	-0.03, 0.08	0.444		0.18	0.11, 0.25	< 0.001

Table 8.2: Association of BMI allele score and observational BMI at age 8 with potential confounders of BMI

Abbreviations: SD, standard deviation; CI, confidence interval; BMI, body mass index

[1] Maternal education: O-level (ordinary level) exams are taken in different subjects usually at age 1516 at the completion of legally required school attendance, equivalent to today's General Certificate of Secondary Education); A-level (Advanced-level) exams are taken in different subjects usually at age 18) [2] Household social class: The mother recorded the occupation of both herself and her partner in a questionnaire at 32 weeks gestation, which were used to allocate them to social class groups using the 1991 Office of Population, Censuses and Surveys classification; the highest class of the mother and her partner was used in analysis. [3] Parity: Obtained from obstetric records. Mothers with parity of two or more were grouped into a single category. [4] P values for linear trend given if more than two ordinal categories.

Rank	Outcome variable (original data with variable N)	SD change of inverse normal transformed outcome for a 1 SD change of BMI allele score				
		Sample size	SD change	95% CI	P value (adjusted P value) [1]	Permutation P value [2]
1	Leptin, 9 [3] [4]	4,249	0.138	0.11, 0.17	<0.001	-
2	CRP, 9 [3] [4]	4,250	0.083	0.05, 0.11	<0.001	-
3	Age menarche [3]	2,946	-0.083	-0.12, -0.05	<0.001	-
4	HDL, 9 [4]	4,250	-0.067	-0.10, -0.04	<0.001 (0.002)	0.008
5	SBP, 7 [3]	6,013	0.049	0.02, 0.07	<0.001	-
6	IL6, 9 [4]	4,240	0.053	0.02, 0.08	0.001, (0.091)	0.010
7	Enjoyment of School Score, 4	5,807	0.041	0.02, 0.07	0.002, (0.255)	0.022
8	Self Esteem: Scholastic Competence, 8	5,222	0.042	0.02, 0.07	0.002, (0.323)	0.025
9	Apolipoprotein B, 9	4,250	0.043	0.01, 0.07	0.005, (0.788)	0.028
10	Triglycerides, 9 [4]	4,250	0.042	0.01, 0.07	0.006, (0.970)	0.037
11	VLDL, 9 [4]	4,250	0.042	0.01, 0.07	0.006, (1)	0.048
12	Apolipoprotein al, 9	4,250	-0.038	-0.07, -0.01	0.012, (1)	0.046
13	Insulin, 15 [3] [4]	2,859	0.047	0.01, 0.08	0.012	-
14	Attention/activity symptoms score, 11	4,541	0.037	0.01, 0.07	0.013, (1)	0.055
15	SDQ emotional symptoms score, 6	5,748	-0.030	-0.06, 0.00	0.022, (1)	0.063
16	Hygiene Score, 4	6,231	0.028	0.00, 0.05	0.024, (1)	0.070
17	Self Esteem: Global Self Worth Score, 8	5,214	0.031	0.00, 0.06	0.025, (1)	0.079
18	FVC: lung function, 8 [3]	5,276	0.030	0.00, 0.06	0.030	-
19	Burden of compulsions/obsessions score, 7	5,684	0.028	0.00, 0.05	0.031, (1)	0.080
20	Particular fears score, 7	5,734	-0.028	-0.05, 0.00	0.033, (1)	0.086
21	Glucose, 15 [3]	2,862	0.041	0.11, 0.17	0.038	-

Table 8.3: Ranking by association strength (P value) of the stage 1 tests with p<0.05: Outcome associations with BMI allele score for original dataset Abbreviations: BMI, body mass index; CI, confidence interval; SD, standard deviation; VLDL, very low density lipoprotein; IL6, interleukin 6; SBP, systolic blood pressure; HDL, high density lipoprotein; VLDL, very low density lipoprotein; SDQ, Strengths and Difficulties Questionnaires; FEF, forced expiratory flow; LF, lung function; FVC, forced vital capacity. Full names of variables are given in Table B.3. All outcomes are transformed to normal distributions using a rank-based inverse normal transformation. Exposure and outcome variables are standardised. Outcome as dependent variable, BMI allele score as independent variable. Full variable names given in Table B.5. [1] Adjusted P values are adjusted for the 160 tests performed using the Bonferroni correction: $pcorrected = poriginal \cdot 160$. Adjusted P values greater than 1 are rounded to 1. Outcomes in validation set are excluded from Bonferroni correction. [2] Permutation P values are generated with permutation testing. Null hypothesis: The outcome variable at rank i in this table would be ranked lower than i if no association was found with the allelic score. We exclude outcomes in validation set such that such that, for instance i=1 for HDL because all three outcomes ranked higher than HDL are in the validation set. [3] Variables that are in the 'validation set', that were chosen for inclusion using apriori knowledge of their association with BMI. Adjusted P value are not given for these are they are not part of the main outcome dataset. [4] Log transformed outcomes, such that distributions approximately normal.

Rank	Outcome variable (original data with variable N)	Sample size	IV estimate of SD change of outcome for a 1 SD change of log BMI age 8 [1]			SD change of outcome for a 1 SD change of log BMI (observational, age 8)			
			Test statistic	95% CI	P value (adjusted P value[2])	Test statistic	95% CI	P value (adjusted P value[2])	P value [5]
Linear regression of continuous, normally distributed outcomes. Test statistic is the mean difference (SD) per 1 SD greater log BMI age 8 or percentage difference per 1 SD greater log BMI age 8 for log transformed outcomes [3]									
1	Leptin, 9 [3 4]	3,381	0.922	0.77, 1.07	<0.001	0.688	0.66, 0.71	<0.001	0.002
2	CRP, 9 [3 4]	3,382	0.432	0.25, 0.62	<0.001	0.367	0.33, 0.40	<0.001	0.484
3	Age menarche [4]	2,186	-0.485	-0.70, -0.27	<0.001	-0.298	-0.34, -0.26	<0.001	0.080
4	HDL, 9 [3]	3,382	-0.466	-0.66, -0.27	<0.001 (<0.001)	-0.303	-0.34, -0.27	<0.001 (<0.001)	0.096
5	SBP, 7 [4]	4,641	0.305	0.13, 0.48	0.001	0.312	0.28, 0.34	<0.001	0.940
6	IL6, 9 [3]	3,372	0.234	0.04, 0.43	0.018 (2.852)	0.197	0.17, 0.23	<0.001 (<0.001)	0.709
9	Apolipoprotein B, 9	3,382	0.363	0.16, 0.57	<0.001 (0.080)	0.163	0.13, 0.20	<0.001 (<0.001)	0.048
10	Triglycerides, 9 [3]	3,382	0.310	0.11, 0.51	0.002 (0.341)	0.226	0.19, 0.26	<0.001 (<0.001)	0.392
11	VLDL age 9 [3]	3,382	0.310	0.11, 0.51	0.002 (0.343)	0.226	0.19, 0.26	<0.001 (<0.001)	0.393
12	Apolipoprotein al, 9	3,382	-0.243	-0.44, -0.05	0.016 (1)	-0.196	-0.23, -0.16	<0.001 (<0.001)	0.636
13	Insulin, 15 [3 4]	2,285	0.299	0.06, 0.53	0.013	0.202	0.16, 0.24	<0.001	0.434
16	Hygiene Score, 4	4,335	0.209	0.01, 0.41	0.042 (1)	0.052	0.02, 0.08	0.001 (0.086)	0.118
18	FVC: lung function, 8 [4]	4,869	0.162	0.00, 0.32	0.052	0.147	0.12, 0.18	<0.001	0.855
21	Glucose, 15 [4]	2,288	0.334	0.02, 0.65	0.039	0.006	-0.04, 0.05	0.819	0.012
Logistic regression of binary outcomes. Test statistic is the odds ratio between outcome groups [6] for a 1 SD increase in log BMI age 8									
7	Enjoyment of School Score, 4	5,808	1.344	0.98, 1.84	0.066 (1)	1.075	1.01, 1.14	0.025 (1)	0.700
8	Self Esteem: Scholastic Competence, 8	5,223	1.548	1.11, 2.16	0.010 (1)	1.004	0.95, 1.06	0.896 (1)	0.036
14	Attention/activity symptoms score, 11	4,542	1.399	0.96, 2.03	0.079 (1)	1.022	0.95, 1.10	0.570 (1)	0.166
15	SDQ emotional symptoms score, 6	5,749	0.856	0.62, 1.19	0.357 (1)	0.947	0.88, 1.01	0.095 (1)	0.543
17	Self Esteem: Global Self Worth Score, 8	5,215	1.536	1.10, 2.14	0.011 (1)	1.002	0.95, 1.06	0.932 (1)	0.046
19	Burden of compulsions/obsessions score, 7	5,685	2.258	0.74, 6.92	0.154 (1)	0.970	0.67, 1.27	0.842 (1)	0.507
20	Particular fears score, 7	5,735	0.893	0.65, 1.23	0.487 (1)	0.992	0.93, 1.05	0.792 (1)	0.519

Table 8.4: IV stage 2 results and observational estimates for outcomes with P<0.05 in stage 1, using original dataset.

Abbreviations: BMI, body mass index; CI, confidence interval; SD, standard deviation; IV, instrumental variable; VLDL, very low density lipoprotein; IL6, interleukin 6; SBP, systolic blood pressure; HDL, high density lipoprotein; SDQ, Strengths and Difficulties Questionnaires; FEF, forced expiratory flow; LF, lung function; FVC, forced vital capacity.
Full names of variables are given in Table B.3.

Figure 8.1a and Figure 8.1b shows these results graphically.

The ranks are based on tests of association of the BMI allele score with the outcome directly, for the 22 associations with a $p < 0.05$. We then perform IV analysis of these results, and it is the IV estimate that is given in this table.

Exposure and outcome variables are standardized. Outcome as dependent variable, BMI allele score as independent variable.

[1] Continuous outcomes: Using Stata ivreg2 command (robust option) and BMI allele score as instrumental variable for log BMI age 8. First stage predicting log BMI at age 8 with the BMI allele score, and the second stage performs an unadjusted association of these log BMI age 8 predictions with the outcome. Binary outcomes: Using regress command (robust option) for first stage, and logistic command (vce(robust) option) for the second stage, to associate outcome with the predicted values of log BMI age 8 from the first stage.

[2] Adjusted P values: Adjusted for the 160 tests performed using the Bonferroni correction: $p_{corrected} = p_{original} \times 160$. Adjusted P values greater than 1 are rounded to 1. Outcomes in validation set are excluded from Bonferroni correction.

[3] Log transformed outcomes, such that distributions are approximately normal.

[4] Variables that are in the 'validation set', that were chosen for inclusion using apriori knowledge of their association with BMI. Adjusted P value are not given for these are they are not part of the main outcome dataset.

[5] P value of Wu-Hausman test, comparing the effect estimates using the allelic score with the effect estimates using observational log BMI age 8. This uses a test for endogeneity (endog argument of ivreg2 Stata command).

[6] Categories for binary variables given in Table B.6.

Outcome variable (original data with variable N)	IV estimate using 31 SNPs excluding *FTO* SNP 1			IV estimate using *FTO* SNP 1 only			
	Test statistic	95% CI	P value [3]	Test statistic	95% CI	P value [3]	Hansen P value [5]
Linear regression of continuous normally distributed outcomes. Test statistic is the mean difference (SD) per 1 SD greater log BMI age 8 or percentage difference per 1 SD greater log BMI age 8 [2]							
Leptin, 9 [2 4]	0.851	0.68, 1.02	<0.001	1.193	0.85, 1.54	<0.001	0.051
CRP, 9 [2 4]	0.447	0.24, 0.66	<0.001	0.375	0.01, 0.74	0.042	0.735
Age menarche [4]	-0.443	-0.70, -0.19	0.001	-0.631	-1.05, -0.21	0.003	0.449
HDL, 9 [2]	-0.517	-0.75, -0.29	<0.001	-0.271	-0.64, 0.10	0.153	0.275
SBP, 7 [4]	0.290	0.10, 0.48	0.002	0.385	-0.05, 0.83	0.086	0.694
IL6, 9 [2]	0.300	0.08, 0.52	0.008	-0.027	-0.41, 0.36	0.891	0.136
Apolipoprotein B, 9	0.281	0.05, 0.51	0.018	0.678	0.24, 1.12	0.003	0.098
Triglycerides, 9 [2]	0.262	0.04, 0.49	0.023	0.496	0.09, 0.90	0.016	0.307
VLDL age 9 [2]	0.261	0.04, 0.49	0.023	0.497	0.09, 0.90	0.016	0.303
Apolipoprotein al, 9	-0.275	-0.51, -0.04	0.021	-0.122	-0.50, 0.26	0.533	0.509
Insulin, 15 [2 4]	0.214	-0.05, 0.48	0.116	0.634	0.08, 1.19	0.026	0.163
Hygiene Score, 4	0.241	0.02, 0.46	0.032	0.045	-0.44, 0.53	0.858	0.477
FVC: lung function, 8 [4]	0.226	0.04, 0.41	0.016	-0.116	-0.49, 0.25	0.540	0.096
Glucose, 15 [4]	0.423	0.06, 0.78	0.021	-0.018	-0.56, 0.52	0.949	0.161
Logistic regression of binary outcomes. Test statistic is the odds ratio between groups for a 1 SD increase in log BMI age 8							
Enjoyment of School Score, 4	1.349	0.95, 1.92	0.096	1.319	0.66, 2.64	0.435	0.897
Self Esteem: Scholastic Competence, 8	1.665	1.15, 2.41	0.007	1.122	0.54, 2.34	0.759	0.517
Attention/activity symptoms score, 11	1.534	1.01, 2.32	0.043	0.916	0.40, 2.07	0.832	0.882
SDQ emotional symptoms score, 6	0.816	0.56, 1.18	0.277	1.072	0.52, 2.22	0.852	0.815
Self Esteem: Global Self Worth Score, 8	1.723	1.19, 2.50	0.004	0.926	0.45, 1.93	0.837	0.357
Burden of compulsions/obsessions score, 7	2.241	0.60, 8.44	0.233	2.255	0.14, 36.76	0.568	0.734
Particular fears score, 7	0.864	0.61, 1.23	0.418	1.041	0.52, 2.10	0.910	0.901

Table 8.5: Testing for invalidity of IV assumptions: associations of two instrumental variables for log BMI at age 8; using 31 SNPs (excluding *FTO* SNP) and only the *FTO* SNP respectively.

Abbreviations: BMI, body mass index; CI, confidence interval; SD, standard deviation; IV, instrumental variable; VLDL, very low density lipoprotein; IL6, interleukin 6; SBP, systolic blood pressure; HDL, high density lipoprotein; SDQ, Strengths and Difficulties Questionnaires; FEF, forced expiratory flow; LF, lung function; FVC, forced vital capacity.

IV estimate calculated with ivregress and robust option (for robust standard errors).

Categories for binary variables are given in Table B.6.

Full names of variables are given in Table B.3.

Figure 8.3a and Figure 8.3b shows these results graphically.

[1] *FTO* SNP is rs1558902.

[2] Log transformed outcomes, such that distributions approximately normal.

[3] P values are not adjusted for the multiple tests performed.

[4] Variables that are in the 'validation set', that were chosen for inclusion using a priori knowledge of their association with BMI.

[5] Hansen P value comparing effect estimates using 31 SNP score and *FTO*.

8.3 Evidence of violation of IV assumptions

The 31 SNP score and *FTO* allele were both strong instruments for log BMI at age 8. A 1 SD increase of the 31-SNP score was associated with a 0.146 standard deviation increase in log BMI at age 8 (95% CI: 0.12, 0.17, $F = 112.70$). A 1 SD increase of *FTO* was associated with a 0.074 standard deviation increase in log BMI at age 8 (95% CI: 0.05, 0.10, $F = 28.18$). We found little evidence of pleiotropy, linkage disequilibrium or population stratification as the tests with the *FTO* and 31-SNP scores were highly consistent (Figure 8.3a and Figure 8.3b and Table 8.5). We found evidence using the Hansen tests of differences between the estimated effects of BMI using each instrument for 5 outcomes, such as apolipoprotein AI, apolipoprotein B, insulin, leptin and the emotional symptoms score (Table 8.6). This may be due to chance, or alternatively may suggest that the genetic variants related to BMI have pleiotropic or heterogeneous effects on these outcomes.

Table 8.1 shows the associations of *FTO* and the 31-allele score respectively, with BMI across childhood. We found evidence of an inverse association of *FTO* with BMI in early childhood, as previously suggested [162]. For instance, an increase of 1 BMI increasing *FTO* allele was associated with a 0.059 decrease of log BMI at age 1 year 8 months (95% CI: -0.096, -0.022). In contrast, the 31-allele score positively was associated with BMI at all ages measured.

8.4 Sensitivity analyses

The outcomes had varying numbers of missing values (as shown in Figure B.1), which means there were differences in statistical power across outcomes. However, the ranking of our main analysis is highly correlated with the ranking of the imputation dataset (Spearman's rank correlation of 0.919 ($P < 0.001$)).

Outcome variable (original data with variable N)	CUE[1]			
	Test statistic	95% CI	P value[2]	Hansen P value[2]
Linear regression of continuous normally distributed outcomes. Test statistic is the mean difference (SD) per 1 SD greater log BMI age 8 or percentage difference per 1 SD greater log BMI age 8 [3]				
Leptin, 9 [3 4]	0.934	0.74, 1.13	<0.001	0.019
CRP, 9 [3 4]	0.410	0.22, 0.60	<0.001	0.606
Age menarche [4]	-0.508	-0.82, -0.20	0.001	0.077
HDL, 9 [3]	-0.447	-0.65, -0.24	<0.001	0.388
SBP, 7 [4]	0.314	0.12, 0.50	0.001	0.456
IL6, 9 [3]	0.250	0.07, 0.43	0.006	0.885
Apolipoprotein B, 9	0.339	0.04, 0.64	0.028	0.010
Triglycerides, 9 [3]	0.245	0.04, 0.45	0.020	0.455
VLDL age 9 [3]	0.245	0.04, 0.45	0.020	0.454
Apolipoprotein al, 9	-0.302	-0.53, -0.07	0.010	0.005
Insulin, 15 [3 4]	0.781	0.28, 1.28	0.002	0.013
Hygiene Score, 4	0.299	0.08, 0.52	0.007	0.265
FVC: lung function, 8 [4]	0.111	-0.05, 0.27	0.180	0.670
Glucose, 15 [4]	0.312	-0.01, 0.63	0.059	0.877
Linear regression of binary outcomes. Test statistic is the change in probability that outcome has value 0 for a 1 SD increase in log BMI age 8				
Enjoyment of School Score, 4	0.008	-0.24, 0.25	0.949	0.309
Self Esteem: Scholastic Competence, 8	0.075	-0.10, 0.25	0.408	0.134
Attention/activity symptoms score, 11	0.174	-0.08, 0.43	0.184	0.702
SDQ emotional symptoms score, 6	0.066	-0.20, 0.33	0.621	0.046
Self Esteem: Global Self Worth Score, 8	0.087	-0.11, 0.28	0.374	0.097
Burden of compulsions/obsessions score, 7	0.035	-0.06, 0.13	0.475	0.587
Particular fears score, 7	-0.071	-0.28, 0.14	0.500	0.283

Table 8.6: Overidentification tests of IV using CUE.

Abbreviations: BMI, body mass index; CI, confidence interval; SD, standard deviation; IV, instrumental variable; VLDL, very low density lipoprotein; IL6, interleukin 6; SBP, systolic blood pressure; HDL, high density lipoprotein; SDQ, Strengths and Difficulties Questionnaires; FEF, forced expiratory flow; LF, lung function; FVC, forced vital capacity.

Categories for binary variables given in Table B.6.

Full names of variables are given in Table B.3.

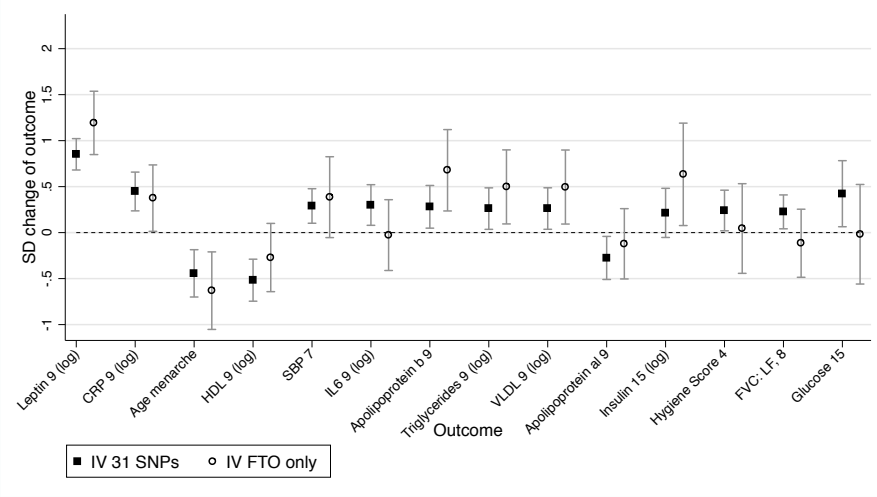
Tests use all 32 SNPs separately in the model.

[1] CUE: continuously updating estimator, with robust standard errors (robust option).

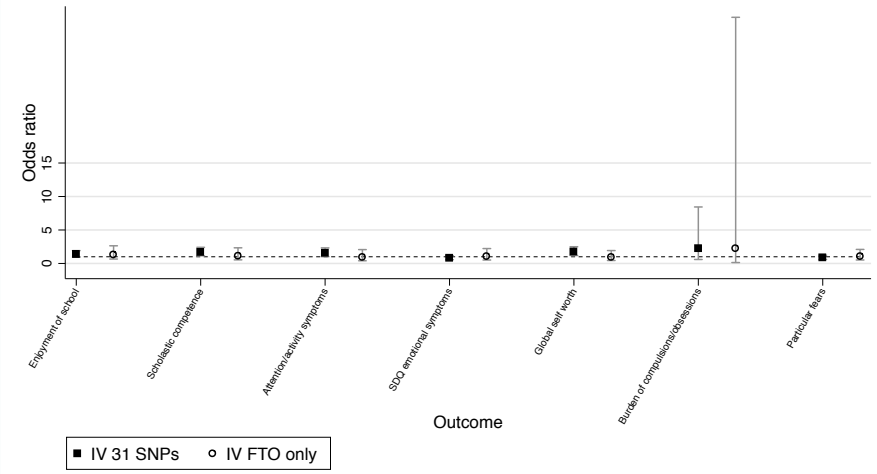
[2] P values are not adjusted for the multiple tests performed.

[3] Log transformed outcomes, such that distributions approximately normal.

[4] Variables that are in the 'validation set', that were chosen for inclusion using a priori knowledge of their association with BMI.



(a) Continuous outcomes



(b) Binary outcomes

Figure 8.3: Testing invalidity of IV assumptions: associations of two instrumental variables using distinct SNP subsets.
Comparison between the SNP subsets: (1) 31 SNPs (excluding *FTO* SNP) and (2) the *FTO* SNP only. IV estimate of effect using two-stage least squared regression of log BMI at age 8 as the exposure. Binary outcomes: Odds ratio between outcome groups, for a 1 SD change of log BMI aged 8. Graphical illustration of the results in Table 8.5. Categories for binary variables given in Table B.6.

8.5 Discussion

We end this section with a discussion of the main findings of our analyses, the limitations of this work and the issues of causal inference in Mendelian randomisation analyses.

8.5.1 Main findings

Epidemiologists have struggled to produce robust replicable evidence of the causal effect of risk factors [163]. Population geneticists have been extremely successful in using hypothesis-free approaches to produce replicable associations. We have shown that it is possible to use a similar hypothesis-free approach, using Mendelian randomisation to highlight the strongest effects of an exposure in a large sample of individuals.

We used a Mendelian randomisation analysis to screen for potentially causal effects. Our stage 1 analysis tested the association of the BMI allele score directly with each of the outcomes. Identifying known effects with this approach validates the use of this score as an instrument for life-long BMI. The BMI allele score was most strongly associated with leptin, which is produced in adipose tissue and is involved in satiation [164]. This result is consistent with previous research [136]. Also consistent with previous studies, we identified effects of BMI on the following metabolic traits: glucose, insulin, interleukin-6, systolic blood pressure [132], low and high density lipoprotein cholesterol, triglycerides, and C-reactive protein [135, 136, 138, 141, 165–167]. The BMI allele score was also strongly inversely associated with age at menarche. Previous observational studies have reported that age at menarche is inversely associated with BMI [168], and a recent study also used Mendelian randomisation to argue that BMI affects age at menarche [133]. We found a novel positive effect of BMI on a global self-worth score. We did not replicate the novel associations, as our aim in this work is to carry out a proof-of-principle analysis demonstrating our hypothesis-free searching method. We found more associations using BMI aged 8 than with the BMI allele score. This may be due to reverse causation (because some of the outcomes actually affect BMI) or due to confounding, and this highlights the benefit of a Mendelian randomisation analysis [30]. Alternatively, this could be due to the lower power of tests with the allele score, compared to observational BMI.

The observational estimates were consistently closer to the null than the IV estimates. This may be due to the winner's curse because in stage 1 we rank all 172 estimates of the allele score on the outcomes, such that the highest ranked are more likely to be higher than their respective true values because of the random variation of these sample estimates about their true values. The stage 2 estimates (also ranked by the stage 1 estimates) may also be affected by this winners curse because the stage 1 (direct) and stage 2 (2sls) models are highly related. The estimates using the observational BMI exposure may not be affected to the same extent as they are ranked by the results using the BMI allele score, with which it is not perfectly correlated. Alternatively, the IV estimates may be more extreme than the observational estimates because the allelic score is a measure for life-long BMI, so the effect on outcomes may be larger due to the cumulative effects of BMI across the life course.

The validity of our results depends on whether the instrumental variable assumptions hold, such that the genetic variants only affect the outcomes through BMI (the exposure). We ensured our BMI allele score was a strong instrument for BMI, and was not associated with common confounding variables such as sex and potential socioeconomic confounders such as household social class. In contrast, BMI at age 8 was associated with confounders consistent with the social patterning reported previously [169–171].

A number of mechanisms could invalidate the second instrumental variable assumption: that the genetic variants only affect the outcomes through BMI at age 8. These include genetic induced confounding through horizontal pleiotropy (where a locus affects several outcomes directly [121]), population stratification and linkage disequilibrium, each of which could add a causal path from the IV to an outcome which was not mediated by BMI. The effect estimates when using two independent instruments (*FTO* and the remaining 31 variants) were consistent providing evidence against pleiotropy because it is unlikely that two independent instruments suffer the same pleiotropic effects. However, we found evidence of heterogeneity when testing the 32 SNPs individually using Hansen tests, for 5 of the 21 'top' results. This may indicate either the core or point identifying IV assumptions are invalid. Any biases introduced by violations of these assumptions may be amplified due to the low power of the individual SNPs. This is because these weak instruments account for only a small proportion of the variance of BMI, such that their effect through BMI is small compared with the strength of the

association through one of these alternative pathways [126].

The stage 1 analysis estimated the association of the BMI allelic score and each outcome, rather than providing an estimate of the effect size, as is estimated in the stage 2 IV analysis. While an estimate of the effect size is generally preferable, the stage 1 tests are important to consider because they only depend on the core IV assumptions, whereas the stage 2 tests also require point identifying assumptions. Also, some of the ways the IV assumptions can become invalidated are circumvented in the stage 1 tests. This is because the stage 1 test requires the exposure to be defined, but a variable representing this exposure is not actually used in the test of association. This is useful because the BMI allelic score is a measure of lifelong risk of increased BMI, but we only have measures of observational BMI at a set of discrete time points, rather than a composite measure representing observational lifelong BMI. The use of BMI at a single time point, at age 8, is a valid exposure if all pathways through BMI at all other ages (prior to measurement of the outcome) also pass through BMI at age 8. If this is not the case (and this is likely) then the IV assumptions are false. For example, we found a different effect of the *FTO* SNP and 31 SNP allele score on BMI in early childhood, which indicates that these variants affect BMI through different pathways. Any pathway from the genetic variants to an outcome through BMI at an age other than age 8 would invalidate the instrumental variable assumptions. This is not a problem for our stage 1 tests, as we need only specify the exposure as 'lifelong BMI up to the point of outcome measurement'. Furthermore, this removes the issue of measurement error in the observed exposure variable since it is not actually used in the model.

8.5.2 Study limitations

We now discuss some further limitations of our analysis. We tested only for linear relationships and hence it is possible that non-linear relationships exist. We used an inverse rank normal transformation, which may not be appropriate for numeric outcomes with only a small number of values, as the rank within each set with the same value is randomly generated, and this may add noise to the data. Ranking results means that we should expect the true strength of associations to be less than we reported, due to the winner's curse. This means that the effect sizes are not reliable and need replicating in a hypothesis-driven manner. However, conventional epidemiological studies also suffer

this due to flexibility in study design, where several methods may be used in turn to examine a particular relationship and the strongest result reported [172].

The size of an effect estimate may be reduced due to developmental compensation (or canalisation) where a foetus may develop to protect itself from the adverse effects of a particular polymorphism that is expressed during foetal development. This protection may continue throughout the life course such that a high BMI will have fewer health implications, and our reported associations may be reduced [120,121]. Dynastic effects, where the outcome trait of the child is also affected by the parental exposure caused by the parental genotype, can also affect the size of an effect estimate. For example, parents that are genetically predisposed to higher BMI (the parental exposure) may rarely encourage their child to exercise (outcome trait of the child). This means that the amount of exercise of a child will be associated with the parent's genotype and hence also the child's genotype – the association between exercise and the child's BMI is confounded by the correlated genome of parent and child.

Our dataset included the most complete version of each repeated measure, which was usually at the earlier time point. While this may improve the statistical power of our tests, this benefit may be offset by the reduction of power because associations are often less pronounced at younger ages (as shown in Table 8.1 for BMI). Mendelian randomisation analyses have low statistical power compared to conventional observational analyses, because genetic variants typically only explain a small proportion of an exposure's variance. Although we used a combined allelic score to maximize the power from the genetic predictors, some associations may not have been detected due to a lack of power. Furthermore, performing multiple tests reduces the statistical power as we need to account for the number of independent tests we performed. The varying degrees of missingness of the outcome variables means: 1) it is possible the associations are biased if the outcome data are not missing at random (conditional on the variables in the model, i.e. observed BMI or the allelic score), and 2) the ranking may be affected by differences in power amongst the outcomes, including false negative results where the power is too low to detect an association. Using the time point with the largest available sample sizes for each trait reduces the risk of bias due to missingness.

Traditionally, hypothesis-driven studies, where many hypotheses are tested independently by several research teams, suffer the issues of multiple testing and selection bias from the researcher choosing which hypothesis to test and the methods to use, as well as

publication bias [106]. A large proportion of null findings are unpublished such that it is not possible to determine the true probability the reported result would occur by chance. This is a most problematic form of multiple testing because we cannot know how many and what associations have been tested. In contrast, by searching for hypotheses in a single study we are able to report the results of all analyses, including 'null' results, so that our work does not contribute to this publication bias. We have provided the results of all stage 1 tests in Table B.4. We presented unadjusted and Bonferroni corrected P values, and estimated a false discovery rate of 0.571, such that 6 of the 14 associations we found with a P value < 0.05 (excluding our validation set) may indicate causal relationships between BMI and these outcomes. Given the high degree of confounding in observational data, the adjusted P values and false discovery rate are likely to be conservative estimates, because they both account for the number of independent tests, but the outcomes in our dataset are not independent. We also provide permutation testing P values that are an appropriate way to assess the results as this method implicitly accounts for the number of tests performed. A result is less likely to achieve a rank of 1 by chance alone as the number of tests increases. The P values of the Bonferroni and permutation testing were very different highlighting the conservative nature of a Bonferroni correction when outcomes are not independent. Appropriately adjusting P values (or equivalently using appropriate statistical thresholds) and reporting all results of a study both help to reduce reporting bias and improve the reliability of published research [163].

A Bonferroni correction should be used when concerned with the global hypothesis, such that the researcher wants to control the probability that at least one test is incorrectly shown to have an association by chance, known as a false positive finding [173]. Bonferroni corrections use a more stringent threshold such that while the number of false positive findings is lower, the number of 'true associations that are not identified (because their associated P values are above the Bonferroni corrected threshold) is higher. This is not helpful for hypothesis-searching studies because we may then miss potentially important associations. Also, the cost of a false positive association is lower in hypothesis-searching studies compared with traditional epidemiological studies because the results will be followed up with a further analysis rather than claimed to be a definitive result. In a hypothesis-searching study the researcher may be happy to follow up n tests knowing that $m\%$ of these may be false positives, such that it may be more

appropriate to control the false discovery rate. This false discovery rate can be adjusted by changing the P value threshold.

8.5.3 Determining causality

The observational associations reported by previous EWAS studies may be caused by bias or confounding and do not provide reliable evidence of causation. We have used Mendelian randomisation to search for true, causal relationships. Our analyses with observational BMI across childhood found a much larger number of strong associations, a distinction that has been previously reported [30]. EWAS studies may be worthwhile to test observational relationships, which can then be followed up with a Mendelian randomisation analysis. However, observational associations may be weaker than the true causal effect because masked confounding – where a direct effect between two traits is not observed because their respective associations with a confounder conceals this association – and measurement error can move associations towards the null [174]. The pheWAS approach has been previously used to identify associations between a set of genetic variants and a set of phenotypic variables [116]. Our approach extends the pheWAS approach in order to identify potentially causal associations. While pheWAS test the association of individual SNPs with observed phenotypes, we use an allelic score composed of variants known to be associated with a particular risk factor as an instrumental variable.

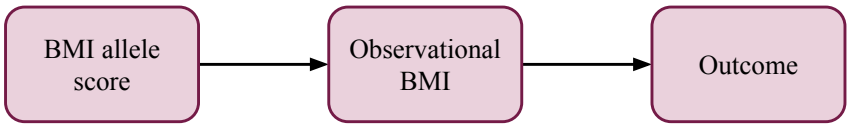
While our approach provides evidence of causality, the direction of this causal effect is less clear, as illustrated in Figure 8.4. Instead of our hypothesized relationship, it is possible that the allele score actually directly affects the 'outcome', which in turn affects BMI, our exposure variable. For instance, as leptin is involved in satiation it is possible that the BMI allelic score, or a subset of variants of which it is composed, affects BMI through leptin rather than vice versa. Currently, our understanding of the biological effects of these variants are often not sufficient to have certainty over the direction of the mechanism of action. Whilst it is not possible to directly test this, this can be investigated by comparing the effect estimates of independent instruments, as if two instruments affect the 'outcome' through the 'exposure' the estimated effect of BMI on an outcome should be consistent across different variants.

We believe hypothesis-free searching with Mendelian randomisation is a valuable

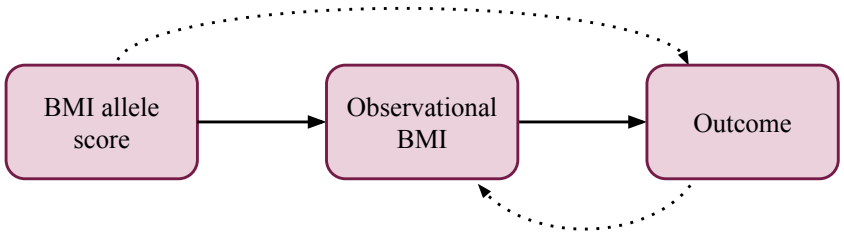
first step towards identifying causal effects, without specifying a particular hypothesis a priori. Where possible it would be informative to test the association of an allelic score of an 'outcome' (e.g. leptin) with observational BMI, to further elucidate causality through a bi-directional analysis [121]. To ensure associations found are not due to reverse causation (where the 'outcome' has a causal effect on the 'exposure') each allelic score should be composed of variants that have a strong association only with the 'exposure' in this bi-directional analysis [175].

8.6 Conclusions

In this and the preceding chapter we have introduced a general pipeline that can be used to search for causal effects of an exposure using Mendelian randomisation. We have demonstrated this hypothesis-searching approach using BMI as an exemplar. Unlike traditional hypothesis-driven approaches we test the association with a large, randomly



(a) The intended pathway we have investigated, where the BMI allele score is an IV for BMI and the variants affect the outcome solely through observed BMI in childhood.



(b) An alternative causal pathway, where the allelic score affects BMI indirectly through the outcome variable.

Figure 8.4: Graphs illustrating two possible causal pathways to explain associations of the BMI allele score with the outcomes. It is possible that these two pathways both occur for a given outcome, such that the graph would become cyclical. Abbreviations: BMI, body mass index.

selected set of phenotypes, rather than specifying a hypothesis to test a priori. We have found that observational BMI was associated with a large number of phenotypes, illustrating the problematic nature of observational tests due to the confounding prevalent between observational phenotypes. These associations in observational data do not indicate causality. In contrast, our genetic instrument was associated with fewer phenotypes because (subject to instrumental variable assumptions) its constituent alleles are not associated with confounding factors, and the causal direction can be investigated because the genome exists prior to the observed phenotypes. We used a set of positive controls to validate the use of this hypothesis-searching method.

This scalable and systematic approach can be repeated with Mendelian randomisation variables of other exposures, in order to gradually determine the causal structure of an otherwise complex network. As with GWAS, associations identified with this type of analysis would need further investigation to validate the relationship through replication studies and elucidate the direction of causality.

Chapter 9

Conclusions

In this thesis we have described two applications of data mining in an epidemiological context – assisting systematic reviews and assisting hypothesis selection. In this final chapter we summarise this work, and describe potential areas of future work.

9.1 Thesis summary

Part I of this thesis presents work investigating the use of text mining to automate elements of risk of bias assessments of systematic reviews. We identified and addressed three key objectives to assist risk of bias assessments and the systematic review process. These are to: 1) identify relevant sentences within research articles, 2) rank articles by risk of bias and 3) reduce the number of assessments the reviewers need to perform by hand.

To achieve these objectives we used text mining to make predictions using the text content of research articles describing clinical trials. The predictions corresponded to the following methodological properties of studies that affect the risk of bias: 1) the method of random sequence generation, 2) the concealment of participants allocations to the study groups, and 3) the method of blinding of participants and personnel. We learnt two types of models, at the sentence and article level. The sentence level models predict the relevance of each sentence for each risk of bias property. The article level models predict the risk of bias of a study due to each risk of bias property, as described in the text of an article.

We were able to rank sentences by relevance with very high ranking performance, so it is possible to show a reviewer which parts of an article are important for each risk of bias property. We have shown that it is possible to rank articles by risk of bias far better than random, such that the articles describing studies with low risk of bias can be prioritised during a systematic review. We found that it was possible to predict the risk of bias values with a certainty at least as high as a human reviewer for at least 33% of articles. This indicates great potential to reduce the time needed to manually assess research articles by hand.

We discussed how rapid reviews in particular could benefit from a method to rank articles from low to high risk of bias. This is because for this type of review only a subset of the articles may be reviewed due to time constraints, such that it is important to assess the best quality evidence first. This led us to define rate-constrained ranking tasks, of which ranking articles for rapid reviews is an example.

We presented a novel metric called the rate-weighted AUC (rAUC), to evaluate models used to rank examples for rate-constrained ranking tasks. In the rapid review context, the rAUC is a measure of ranking performance that accounts for the likely number of articles that will be reviewed in the allotted time. We showed that the rAUC should be used for rate-constrained ranking tasks as if other metrics such as nDCG or AUC are used to choose the best ranking model then a suboptimal model may be chosen.

We also presented a novel approach for generating confidence bounds around ROC curves, which we call rate-oriented point-wise confidence bounds. These bounds consist of a series of confidence intervals at specified rates along the ROC curve. Each confidence interval lies along a rate isometric and denotes the interval within which we expect, with 95% probability, the consensus curve of a set of new samples generated from this consensus curve to cross this rate isometric. These bounds are effective because they use the positions of nearby points on a ROC curve to infer where the ROC curve is likely to cross a given rate isometric. We suggest that rate-oriented point-wise confidence bounds are particularly appropriate for rate-oriented tasks, where a researcher is concerned with the performance of a model at particular rates.

Part II began with an exploration of currently accepted data mining applications in epidemiological research. One approach, the environment-wide association study (EWAS), is used to search for associations between observational variables. A weakness of this approach is that it is unable to determine whether associations between observed

traits are causal. Another approach, the phenome-wide association study (pheWAS), is used to identify pleiotropy on the genome, where genetic variation is associated with multiple phenotypes. However, this approach does not attempt to determine if a causal relationship exists between two observed traits that are associated with the same genetic variant. We introduced a novel approach that uses Mendelian randomisation to search for causal associations between a single exposure and a potentially large set of observational variables. This is a hypothesis-generating approach, to identify potentially causal associations that can then be followed up with a more thorough analysis.

We performed a proof-of-principle analysis to demonstrate our approach, searching for the causal effects of BMI. We found several associations already shown to be causally associated with BMI, providing validation for our approach. For example, the top association was with leptin, a hormone known to be involved in satiation. Our approach found far fewer associations compared to the EWAS alternative, because our method does not suffer from the confounding present between observational variables.

9.2 Future directions

In this section we present some key areas of future work.

9.2.1 Extending to other risk of bias domains

Our work exploring assisting systematic reviews focused on predicting three risk of bias domains – sequence generation, allocation concealment and blinding. The automated prediction of two other domains, selective reporting and incomplete outcome data, could also be investigated. This may require custom techniques, as we now discuss. Establishing the risk of bias due to selective reporting requires knowledge of the analyses researchers have performed so that they can be compared with the reported results, to determine whether some are unreported or whether the primary analysis has been changed. This information is often available in trial protocols, so an automated approach could attempt to extract this information from these documents and compare this with the analyses reported in a research article.

The risk of bias due to incomplete outcome data can be assessed by examining the loss to follow-up of a trial, the proportion of individuals that were originally regis-

tered for a trial but for which the outcome data is not available. As this is a numeric value automated prediction of this domain may therefore involve identifying the loss to follow-up (if reported) in a research article of a trial, and using a threshold to determine whether attrition is sufficiently low such that the risk of bias is also low.

9.2.2 Learning better models for risk of bias predictions

A limitation of our work is the small size of our risk of bias dataset (RoBAL), as described in Section 5.5.1. Whilst the performance of our models is encouraging, it could be improved further by creating a larger dataset with which to train the model parameters. Manual dataset creation is very time consuming. However, as this data is created during the risk of bias assessment process, data collection could be incorporated into a risk of bias assessment tool. This functionality could be integrated into an already accepted tool such as the Cochrane risk of bias tool, to minimise changes to the user process.

A larger dataset would make it possible to investigate the use of more complex models to predict risk of bias, which may improve performance. Given more data, we could divide the data into a tuning set and a hold-out set. The tuning data is used to test multiple models in order to decide which method is best for a particular task, using cross validation to evaluate the models. This could involve trying different algorithms (such as decision trees or naive Bayes), different feature sets or different parameters of a particular model. A hold-out dataset is needed because, when selecting a model with the best performance on the tuning data, the estimated performance may be optimistic. This is for two reasons. First, when ranking models by performance it is expected that the top performing models will be worse in the future, due to the winners curse. Second, if a model is tuned too much on the tuning set it may over-fit this data, performing better on the tuning data (even when evaluated with cross validation) than on new data, because it does not generalise well to unseen examples. The hold-out set provides data that is independent to the tuning data to evaluate the final model on unseen examples.

A larger dataset is also beneficial to increase the power to detect patterns in the data. This improves the estimated parameters because the features are less likely to be associated with our risk of bias properties by chance alone. This also means that more complex patterns may be identified. As discussed in Section 5.5, interactions between

words may also be important. In our work we used a bag of words representation with unigram features, which treats each word independently. Additional features can be used to model the relationship of a word with nearby words (using n-grams), or words within the same sentence. When introducing word interactions a larger dataset is needed so that there are sufficient examples of each word interaction to estimate the parameters of a model.

An obvious example of a word interaction is a negation, using words such as *not* or *no* that may precede other words. For instance, a study may state 'we could not blind participants'. In the current feature representation this negation cannot be represented, and when predicting the risk of bias value of an article containing 'not blind' the occurrence of the word blind would indicate that *low* risk of bias is more likely. While this example shows this will likely be important when predicting the risk of bias values of an article, there may also be cases where this is beneficial for determining sentence relevance. A word may add context that changes the predictivity of another word in the same sentence. For instance, the word *blind* may not be relevant to the *blinding* property when the sentence also contains words related to the eye, as this would indicate that the text is referring to eye blindness rather than the *blinding* risk of bias property. Interactions between words encoded in a dataset may be restricted to couples of adjacent words (known as bigrams), or they may include more distant relationships such as pairs of words in a sentence.

In addition to word interactions, we could investigate creating additional features using external information. An ontology could be used to map words to standardised terms and entity types. For example, drug names could be mapped to an indicator variable that denotes whether a drug is mentioned in an article, in addition to the original features for the specific drug names. Furthermore, it may be useful to manually construct custom look up tables that map terms to particular notions that may be relevant to risk of bias. For example, we know that whether an outcome is subjective or objective affects the risk of bias due to blinding. Hence, a feature denoting whether outcomes are subjective or objective may improve model performance.

The risk of bias due to sequence generation, allocation concealment and blinding are highly related. It may be beneficial to learn these properties together, such that information about the label of one property can also help to predict the label of another property. This is known as a multi-task learning approach. For example, a machine

learning model called a classifier chain [176] that learns the labels sequentially may be effective. For instance, sequence generation may first be predicted using a bag of words representation. Allocation concealment is then predicted using a bag of words representation and additionally the predicted sequence generation values. This means that the relationship between the risk of bias properties can be represented in the model. Marshall et al. have used a multi-task learning approach to predict risk of bias [22], as described in Section 2.2.3.

9.2.3 Assisting systematic reviews

Researchers have suggested that we are heading towards an integrated system to automate systematic reviews. However, the following issues make this a challenging aim at this time. Firstly, systematic reviews typically only include a small number of studies, which means it is important that the correct studies are included in the review and the study properties are accurately extracted from articles reporting these studies, so that correct inferences can be made. Also, currently not all the information needed to perform the review is directly available in research articles describing the studies. Reviewers may contact authors directly to gain specific details. Hence, we believe that an automated system for systematic reviews is not feasible at present.

Instead, we see a goal in the near future as an interactive system that focuses on assisting reviewers with their risk of bias assessments rather than replacing reviewers completely. Our prototypical tool provides a starting point, and this can be refined by carrying out user testing to identify changes that would be beneficial. We believe that future work should investigate the use of these assistive techniques in a practical setting.

Furthermore, while in this work we have focused on predicting three risk of bias properties, there is much other work that can be used to assist systematic reviews, as we described in Section 2.2.3. For instance, automatically identifying articles describing RCTs is very useful because at the moment it is not possible to find these using the search tools of databases such as PubMed, because this is not always indexed correctly. Also, the screening of research articles according to their relevance to the research question posed in a review is an important task. Systematic reviewers may benefit greatly if risk of bias predictions and tasks such as those just mentioned were integrated into a single system.

9.2.4 Automating hypothesis generation pipeline

We suggest that automation of elements of the hypothesis generation pipeline proposed in Section 7.1 would be highly beneficial to researchers. A set of tools could be implemented at each step of the pipeline, where the user can input the details specific to their particular research question. For example, a tool for creating the instrumental variable could ask the user to input the SNPs and the weights (if using a weighted score) and the tool could then retrieve the genetic data from a particular dataset (that the user has access to) and generate the allele score.

There is also the opportunity to automatically determine which SNPs should be used to construct a score, using an online database of GWAS results maintained by the National Human Genome Research Institute [177]. This database can be used to retrieve SNPs associated with an exposure of interest. Results can also include the effect sizes which may be used as weights to generate a weighted allele score.

9.2.5 Testing the MR-pheWAS approach with a larger dataset

The MR-pheWAS analysis we performed with the ALSPAC dataset may not have detected some associations because of a lack of power. Also, we included a relatively small number of variables in our outcome dataset, to test this approach. We intend to further test this approach using 500,000 participants in the Biobank UK dataset, and over 1,000 outcomes. This will greatly increase power to detect associations, and allow us to search for the causal effects of BMI across a wide range of traits in an adult population.

To conclude, in this thesis we have shown two applications of data mining techniques in epidemiology. The data-intensive approaches we have used allow for efficient exploration of potentially large amounts of data. The data-deluge experienced by epidemiologists will only increase, as more technologies emerge generating more and new types of data. Hence, development and exploration of data-intensive approaches is imperative if we are to use the available data to its full potential, to gain greater understanding of causal mechanisms affecting public health.

Appendices

Appendix A

Assisting systematic reviews

A.1 Supplementary tables

Sequence generation		Allocation concealment		Blinding	
random	-9.59453	envelop	-7.47791	blind	-9.13649
randomis	-8.72771	alloc	-6.49915	doubleblind	-6.70298
randomli	-7.31203	assign	-5.72496	ident	-6.54086
alloc	-6.79806	randomis	-5.48667	mask	-6.37214
toothbrush	-6.14166	random	-5.3923	open	-6.13433
vildagliptin	-5.74126	open	-5.03137	placebo	-5.88171
morphin	-4.45936	code	-4.93616	capsul	-4.40904
studi	-4.33722	central	-4.67483	unawar	-4.09867
patient	-4.18157	pharmaci	-4.61331	run-out	4.02426
insulin	-4.11151	seal	-4.3545	single-blind	-3.91737
block	-4.05597	ident	-4.15013	alloc	-3.86031
pioglitazon	-3.97791	particip	-3.89514	laparoscop	-3.56122
altern	-3.94826	prepar	-3.8115	repair	-3.53471
sonic	-3.7443	unawar	-3.52678	assign	-3.43134
disulfiram	3.68192	conceal	-3.50108	open-label	-3.41086
stratif	-3.59312	bottl	-3.4588	staff	-3.34353
a1c	-3.57684	personnel	-3.45346	drug	-3.33948
clonidin	-3.42949	peer	-3.28038	vitamin	-3.32485
finasterid	3.36928	nurs	-3.27809	perform	-3.26358
envelop	-3.36342	packag	-3.23059	renal	3.07396

Table A.1: Top 20 word stem predictors of sentence relevance and normalised coefficients, using TF-IDF features transformation with regularised logistic model.

Sequence generation		Allocation concealment		Blinding	
alloc	-0.81565	envelop	-0.85989	placebo	-0.9522
exclud	-0.57037	seal	-0.62387	blind	-0.57947
95%	-0.55694	power	-0.58893	month	0.48464
activ	-0.54465	januari	0.54545	indic	0.48205
approv	-0.5313	randomis	-0.50639	double-blind	-0.47148
seal	-0.5125	opaqu	-0.50441	review	-0.44726
bmj	-0.47563	grate	-0.48322	potassium	0.44648
steril	-0.47462	assign	-0.46926	record	-0.44491
societi	-0.4737	jone	-0.46643	code	-0.41714
prevent	0.4722	stratifi	-0.44116	explan	-0.40479
computer-gener	-0.46561	committe	-0.43763	affect	0.39516
introduct	-0.46232	group	0.42335	1989	0.38382
envelop	-0.45245	bulletin	0.42048	suppli	-0.37767
number	-0.43431	emerg	-0.4112	wai	-0.37735
receiv	0.43181	methodologi	-0.40954	order	-0.37555
jone	-0.42134	bristol	-0.40949	acknowledg	-0.37374
opaqu	-0.4179	ident	-0.39177	withdraw	-0.36513
18%	-0.41268	3depart	0.38862	1998	0.35299
random	-0.41121	characterist	0.38378	test	0.34979
depart	0.40801	oral	-0.37662	wilcoxon	0.34921

Table A.2: Top 20 word stem predictors of article risk of bias and normalised coefficient, using TF-IDF features transformation with regularised logistic model.

Appendix B

Assisting hypothesis selection

B.1 Supplementary figures

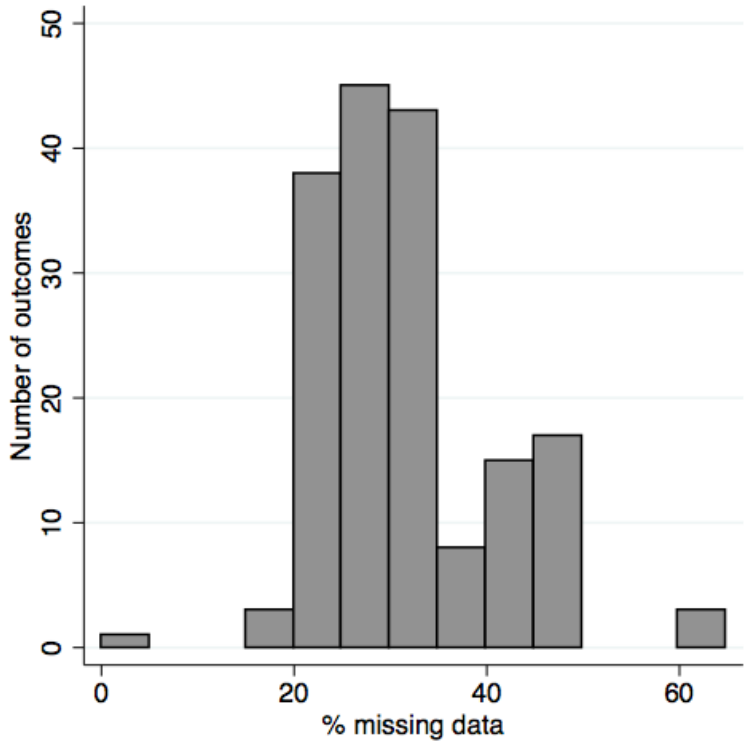


Figure B.1: Distribution of the percentage of missing data, in our 8,121 sample, across the 172 outcomes

B.2 Supplementary tables

SNP	Imputation quality (r^2)	SNP	Imputation quality (r^2)
rs10150332	0.9963	rs2867125	0.9997
rs10767664	0.9965	rs2890652	0.9888
rs10938397	0.9881	rs29941	0.9999
rs10968576	0.9995	rs3810291	0.7652
rs11847697	0.9688	rs3817334	0.9984
rs12444979	0.9975	rs4771122	0.9313
rs13078807	0.9966	rs4836133	0.9429
rs13107325	0.9972	rs4929949	0.9671
rs1514175	0.9984	rs543874	0.9965
rs1555543	0.9960	rs571312	0.9995
rs1558902	0.9967	rs713586	0.9993
rs206936	0.9875	rs7138803	0.9980
rs2112347	0.9945	rs7359397	0.9988
rs2241423	0.9997	rs887912	0.9972
rs2287019	0.9991	rs9816226	0.9556
rs2815752	0.9964	rs987237	0.9994

Table B.1: List of SNPs used to construct the BMI allele score. BMI associated SNPs found in the largest GWAS to date [150].

Dataset	Age at measurement	Data file name	Variables chosen
		Clinic blood measures	
Focus age 9 – bloods	9 yrs 6 mnths	f_9_lipids	All variables included. Includes leptin and CRP variables used in validation set
Focus age 15 –Fasting bloods	15y 6m	fasting15_v9_nodups	Glucose and insulin for validation set
		Clinic other measures	
MacArthur CDI: saying and understanding scores	15m	CDI	All used, removed duplicates.
Focus age 7	7y 6m	f07_3d	We only included systolic blood pressure from this dataset, as one of our validation variables.
Focus age 8	8y 6m	f08_3b	This dataset contains 861 variables, and a set of derived scores. We kepts only variables representing the main concepts of this dataset, by including the scores. DANVA, WISC and lung function variables used for validation set.
Focus age 9	9y 6m	f09_3b	We used the main scores from this dataset – the "total raw accuracy score" and the "total raw comprehension score", included in our validation set.
Haemoglobin levels	7y 6m	haemoglobin_focus	This only contains haemoglobin, earliest timepoint used.
Coefficient of variation of total energy intake	10 years	cv_energy10y2a_aln	This only contains one energy measure earliest timepoint used.
		Questionnaires	
Age at menarche	n/a	age_at_menarche_mar12	This dataset only contains age at menarche, used in validation set.
Derived from strength and difficulties questionnaire	6 years 9m	sdq81mns_kq	All used, removed age and duplicates.
Girl/Boyv Toddler questionnaire	1year 6m	kd_4b	We used only the home score from this dataset, in our validation set.
3 year FFQ Nutrient Intake, derived from food frequencies in My 3 Year old Boy/Girl	3 yr 2 mnths	kgnut3yr_v3	All variables included.
My Young 4 Year Old Girl/Boy	4y 6m	kk_2c	We used only the scores from this dataset, and removed all duplicates of these.
My Daughter/Son at School	6y 9m	kq_2c	We used only the scores from this dataset, and removed all duplicates of these.
My Daughter/Sons Wellbeing	7y 7m	kr_1b	We used only the scores from this dataset, and removed all duplicates of these.
Your Daughter/Son at 9	9yr 7m	ku_r2b	We used only the scores from this dataset, and removed all duplicates of these.
Schools – "The Developing Child", "Questionnaire for Class Teacher", "Questionnaire for Head Teacher"	11yr 1m	sefg_1b	We used only the score variables.

Table B.2: ALSPAC data files used to create the outcome dataset and the rules used to determine inclusion / exclusion of variables For further information of ALSPAC variables See [146, 147] and the ALSPAC website: http://www.bristol.ac.uk/alspac/researchers/resources-available/.

Variable	Description	Timepoint	Validation set
chol_9	focus @ 9, cholesterol, mmol/l	9 yrs 6 mnths	
trig_9	focus @ 9, triglycerides, mmol/l	9 yrs 6 mnths	
vldl_9	focus @ 9, very low density lipoprotein, mmol/l	9 yrs 6 mnths	
ldl_9	focus @ 9, low density lipoprotein, mmol/l 9 yrs 6 mnths		
hdl_9	focus @ 9, high density lipoprotein, mmol/l	9 yrs 6 mnths	
apoai_9	focus @ 9, apolipoprotein al, mg/dl	9 yrs 6 mnths	
apob_9	focus @ 9, apolipoprotein b, mg/dl	9 yrs 6 mnths	
crp_9	focus @ 9, c-reactive protein,mg/l	9 yrs 6 mnths	YES
leptin_9	focus @ 9, leptin, ng/ml	9 yrs 6 mnths	YES
adiponectin_9	focus @ 9, adiponectin, ng/ml	9 yrs 6 mnths	
il6_9	focus @ 9, interleukin 6, pg/ml	9 yrs 6 mnths	
glucosem_15	glucose(mmol/l), 15 year fasting bloods	15 yrs 6 mnths	YES
insulini_15	insulin (iu/l), 15 year fasting bloods	15 yrs 6 mnths	YES
hb_f7	Haemoglobin at f@7	7 yrs 6 mnths	
AGE_MENARCHE_YEARS_comp	Age at menarche	n/a	YES
f7sa021	Mean BP systolic: samples F@7	7 yrs 6 mnths	YES
f9sn702	Total raw accuracy score: Story F@9	9 yrs 6 mnths	YES
f9sn703	Total raw comprehension score: Story F@9	9 yrs 6 mnths	YES
kc_und	DV: Number of words child understands (out of 134) at 15m	1 yr 3 mnths	
kc_says	DV: Number of words child can say (out of 134) at 15m	1 yr 3 mnths	
kgalcohol	daily alcohol intake (g) from ffq at 3years version 3	3 yr 2 mnths	
kgcalcium	daily calcium intake (mg) from ffq at 3years version 3	3 yr 2 mnths	
kgcarbohydrate	daily carbohydrate intake (g) from ffq at 3years version 3	3 yr 2 mnths	
kgcarotene	daily carotene intake (microgrammes) from ffq at 3years version 3	3 yr 2 mnths	
kgcholesterol	daily cholesterol intake (mg) from ffq at 3years version 3	3 yr 2 mnths	
kgenergy	daily energy intake (kj) from ffq at 3years version 3	3 yr 2 mnths	

kgfat	daily fat intake (g) from ffq at 3years version 3	3 yr 2 mnths
kgn3	daily n-3 fatty acid intake (g) from fish only from ffq at 3years version 3	3 yr 2 mnths
kgdha	daily dha intake (g) from fish only from ffq at 3years version 3	3 yr 2 mnths
kgepa	daily epa intake (g) from fish only from ffq at 3years version 3	3 yr 2 mnths
kgfolate	daily folate intake (microgrammes) from ffq at 3years version 3	3 yr 2 mnths
kgiodine	daily iodine intake (microgrammes) from ffq at 3years version 3	3 yr 2 mnths
kgiron	daily iron intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgmg	daily magnesium intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgmono	daily monounsaturated fat intake (g) from ffq at 3years version 3	3 yr 2 mnths
kgnceq	daily niacin equivalents intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgnmesugars	daily nme sugars intake (g) from ffq at 3years version 3	3 yr 2 mnths
kgnsp	daily nsp intake (g) from ffq at 3years version 3	3 yr 2 mnths
kgphosphorus	daily phosphorus intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgpoly	daily polyunsaturated fat intake (g) from ffq at 3years version 3	3 yr 2 mnths
kgpotassium	daily potassium intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgprotein	daily protein intake (g) from ffq at 3years version 3	3 yr 2 mnths
kgretinol	daily retinol intake (microgrammes) from ffq at 3years version 3	3 yr 2 mnths
kgribo	daily riboflavin intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgsfa	daily saturated fat intake (g) from ffq at 3years version 3	3 yr 2 mnths
kgselenium	daily selenium intake (microgrammes) from ffq at 3years version 3	3 yr 2 mnths
kgsodium	daily sodium intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgstarch	daily starch intake (g) from ffq at 3years version 3	3 yr 2 mnths
kgsugar	daily sugar intake (g) from ffq at 3years version 3	3 yr 2 mnths
kgthiamin	daily thiamin intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgvitc	daily vitamin c intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgvitb6	daily vitamin b6 intake (mg) from ffq at 3years version 3	3 yr 2 mnths
kgvitb12	daily vitamin b12 intake (microgrammes) from ffq at 3years version 3	3 yr 2 mnths

kgvitd	daily vitamin d intake (mg) from ffq at 3years version 3	3 yr 2 mnths	
kgvite	daily vitamin e intake (mg) from ffq at 3years version 3	3 yr 2 mnths	
kgzinc	daily zinc intake (mg) from ffq at 3years version 3	3 yr 2 mnths	
fd10cv_kcal	coefficient of variation for total energy intake (kcal), f10+	10 yrs	
kqpemotion	sdq emotional symptoms score (prorated)	6 yrs 9 mnths	
kqpconduct	sdq conduct problems score (prorated)	6 yrs 9 mnths	
kqphyper	sdq hyperactivity score (prorated)	6 yrs 9 mnths	
kqppeer	sdq peer problems score (prorated)	6 yrs 9 mnths	
kqpprosoc	sdq prosocial score (prorated)	6 yrs 9 mnths	
kqpebdtot	sdq total difficulties score (prorated)	6 yrs 9 mnths	
f8at065	Att Sky Search – Normative Score: F@8	8 yrs 6 mnths	
f8at146	Att Dual Task – Score: F@8	8 yrs 6 mnths	
f8at148	Att Dual Task – Normative Score: F@8	8 yrs 6 mnths	
f8at228	Att Opp Worlds Task – Normative score Same World: F@8	8 yrs 6 mnths	
f8at229	Att Opp Worlds Task – Normative score Opp World: F@8	8 yrs 6 mnths	
f8lc125	LoC – Locus of Control Score: F@8	8 yrs 6 mnths	
f8dv440	DANVA, All Faces – # Errors: F@8	8 yrs 6 mnths	YES
f8fs120	F&S Friends score: F@8	8 yrs 6 mnths	
f8aa150	Antisocial activities score: F@8	8 yrs 6 mnths	
f8gb041	Gender – CAI score: F@8	8 yrs 6 mnths	
f8bp026	Posting Behaviour – Irritability/Neg emotion Score: F@8	8 yrs 6 mnths	
f8bp036	Posting Behaviour – Impulsivity/Distractability Score: F@8	8 yrs 6 mnths	
f8bp046	Posting Behaviour – Approach Score: F@8	8 yrs 6 mnths	
f8bp056	Posting Behaviour – Sluggishness Score: F@8	8 yrs 6 mnths	
f8bp066	Posting Behaviour – Wariness Score: F@8	8 yrs 6 mnths	
f8se125	Self Esteem: Scholastic Competence Score: F@8	8 yrs 6 mnths	
f8se126	Self Esteem: Global Self Worth Score: F@8	8 yrs 6 mnths	
f8ws112	WISC – Total IQ: F@8	8 yrs 6 mnths	YES

f8ba026	Activities Behaviour – Irritability/Neg emotion Score: F@8	8 yrs 6 mnths	
f8ba036	Activities Behaviour – Impulsivity/Distractability Score: F@8	8 yrs 6 mnths	
f8ba046	Activities Behaviour – Approach Score: F@8	8 yrs 6 mnths	
f8ba056	Activities Behaviour – Sluggishness Score: F@8	8 yrs 6 mnths	
f8ba066	Activities Behaviour – Wariness Score: F@8	8 yrs 6 mnths	
f8sl040	S&L – WOLD comprehension – Raw Score: F@8	8 yrs 6 mnths	
f8lf110	SD score for FVC: LF, F@8	8 yrs 6 mnths	YES
se087b	DV: Activity symptoms score (prorated)	11 yrs 2 mnths (School year 6)	
se090b	DV: Attention symptoms score (prorated)	11 yrs 2 mnths (School year 6)	
se093b	DV: Attention/activity symptoms score (prorated)	11 yrs 2 mnths (School year 6)	
se098a	DV: Burden of attention/activity problems score	11 yrs 2 mnths (School year 6)	
se123b	DV: Awkward behaviours score (prorated)	11 yrs 2 mnths (School year 6)	
se126b	DV: Troublesome behaviours score (prorated)	11 yrs 2 mnths (School year 6)	
se129b	DV: Awkward/troublesome behaviours score (prorated)	11 yrs 2 mnths (School year 6)	
se134a	DV: Burden of awkward/troublesome behaviours score	11 yrs 2 mnths (School year 6)	
se161b	DV: SDQ prosocial score (prorated)	11 yrs 2 mnths (School year 6)	
se162b	DV: SDQ hyperactivity score (prorated)	11 yrs 2 mnths (School year 6)	

se163b	DV: SDQ emotional symptoms score (prorated)	11 yrs 2 mnths (School year 6)	
se164b	DV: SDQ conduct problems score (prorated)	11 yrs 2 mnths (School year 6)	
se165b	DV: SDQ peer problems score (prorated)	11 yrs 2 mnths (School year 6)	
se166b	DV: SDQ total difficulties score (prorated)	11 yrs 2 mnths (School year 6)	
sf573b	DV: CCEI anxiety subscale score (prorated)	11 yrs 2 mnths (School year 6)	
sf574b	DV: CCEI somatic subscale score (prorated)	11 yrs 2 mnths (School year 6)	
sf575b	DV: CCEI depression subscale score (prorated)	11 yrs 2 mnths (School year 6)	
sf576b	DV: CCEI total score (prorated)	11 yrs 2 mnths (School year 6)	
sf611b	DV: Bachman self esteem score (prorated)	11 yrs 2 mnths (School year 6)	
kd380a	HOME score	1 yr 6 mnths	YES
kk310	DV: Hygiene Score	4 yrs 6 mnths	
kk317	DV: Toilet Incontinence Score	4 yrs 6 mnths	
kk489	DV: CH Enjoyment of School Score	4 yrs 6 mnths	
kq316	DV: Sleep Worries Score	6 yrs 9 mnths	
kq378b	DV: Life events score since child's 5th birthday (prorated)	6 yrs 9 mnths	
kq425	DV: Locomotor Ability Score	6 yrs 9 mnths	
kq442	DV: Fine Motor Score	6 yrs 9 mnths	
kq462	DV: Cognitive Score	6 yrs 9 mnths	
kq475	DV: Playing & Sharing Score	6 yrs 9 mnths	

kq477	DV: Empathy Subscale Score	6 yrs 9 mnths
kq486	DV: Ball Skills Score	6 yrs 9 mnths
kq502	DV: Social Skills Score	6 yrs 9 mnths
kq517	DV: Communication Score	6 yrs 9 mnths
kq519	DV: Musical Subscale Score	6 yrs 9 mnths
kq525	DV: Speech Intelligibility Score	6 yrs 9 mnths
kq538	DV: Uncommunicative Score	6 yrs 9 mnths
kq558	DV: Developmental Worries Score	6 yrs 9 mnths
kq573	DV: Child Activity Score	6 yrs 9 mnths
kq597	DV: Female Parenting Score	6 yrs 9 mnths
kq622	DV: Male Parenting Score	6 yrs 9 mnths
kq653	DV: Sibling Interaction Score	6 yrs 9 mnths
kq680	DV: Feeding Difficulties Score	6 yrs 9 mnths
kr213b	DV: Separation anxiety symptoms score (prorated)	7 yrs 7 mnths
kr222a	DV: Burden of separation anxieties score	7 yrs 7 mnths
kr236b	DV: Particular fears score (prorated)	7 yrs 7 mnths
kr247a	DV: Burden of particular fears score	7 yrs 7 mnths
kr259b	DV: Social fears score (prorated)	7 yrs 7 mnths
kr275a	DV: Burden of social fears score	7 yrs 7 mnths
kr300b	DV: Stress reactions score (prorated)	7 yrs 7 mnths
kr309a	DV: Burden of stress reactions score	7 yrs 7 mnths
kr332b	DV: Compulsions score (prorated)	7 yrs 7 mnths
kr337b	DV: Compulsions/obsessions score (prorated)	7 yrs 7 mnths
kr351a	DV: Burden of compulsions/obsessions score	7 yrs 7 mnths
kr367b	DV: General anxieties score (prorated)	7 yrs 7 mnths
kr379b	DV: General anxiety symptoms score (prorated)	7 yrs 7 mnths
kr387a	DV: Burden of general anxieties score	7 yrs 7 mnths
kr429a	DV: Burden of moods score	7 yrs 7 mnths

kr447b	DV: Activity symptoms score (prorated)	7 yrs 7 mnths
kr459b	DV: Attention symptoms score (prorated)	7 yrs 7 mnths
kr462b	DV: Attention/activity symptoms score (prorated)	7 yrs 7 mnths
kr468b	DV: Teacher complaints score (prorated)	7 yrs 7 mnths
kr478a	DV: Burden of attention/activity problems score	7 yrs 7 mnths
kr492b	DV: Awkward behaviours score (prorated)	7 yrs 7 mnths
kr501a	DV: Burden of awkward behaviours score	7 yrs 7 mnths
kr519b	DV: Troublesome behaviours score (prorated)	7 yrs 7 mnths
kr554b	DV: Skuse social cognition score (prorated)	7 yrs 7 mnths
ku503b	DV: CCC – Intelligibility and fluency score (prorated)	9 yrs 7 mnths
ku504b	DV: CCC – Syntax score (prorated)	9 yrs 7 mnths
ku505b	DV: CCC – Inappropriate initiation score (prorated)	9 yrs 7 mnths
ku506b	DV: CCC – Coherence score (prorated)	9 yrs 7 mnths
ku507b	DV: CCC – Stereotyped conversation score (prorated)	9 yrs 7 mnths
ku508b	DV: CCC – Use of conversational context score (prorated)	9 yrs 7 mnths
ku509b	DV: CCC – Conversational rapport score (prorated)	9 yrs 7 mnths
ku510b	DV: CCC – Pragmatic aspects of communication score (prorated)	9 yrs 7 mnths
ku673b	DV: SMFQ depression score (prorated)	9 yrs 7 mnths
ku705b	DV: SDQ – Prosocial score (prorated)	9 yrs 7 mnths
ku706b	DV: SDQ – Hyperactivity score (prorated)	9 yrs 7 mnths
ku707b	DV: SDQ – Emotional symptoms score (prorated)	9 yrs 7 mnths
ku708b	DV: SDQ – Conduct problems score (prorated)	9 yrs 7 mnths
ku709b	DV: SDQ – Peer problems score (prorated)	9 yrs 7 mnths
ku710b	DV: SDQ – Total difficulties score (prorated)	9 yrs 7 mnths
f8at060L	Selective attention	8 yrs 6 mnths
f8at061L	Motor score	8 yrs 6 mnths
f8at147L	Dual task decrement score	8 yrs 6 mnths

Table B.3: List of descriptions of outcome variables included in our dataset, from the ALSPAC cohort Abbreviations: yr, year; mnths, months; DV, derived variable; standard deviation; LDL, low density lipoprotein; WISC, Wechsler Intelligence Scale for Children; DANVA, Diagnostic Analysis of Nonverbal Accuracy test; IL6, interleukin 6; SBP, systolic blood pressure; HDL, high density lipoprotein; VLDL, very low density lipoprotein; SDQ, Strengths and Difficulties Questionnaires; CCC, Childrens Communication Checklist; CCEI, Crown Crisp Experiental Index. For further information of ALSPAC variables See [146, 147] and the ALSPAC website: http://www.bristol.ac.uk/alspac/researchers/resources-available/ Variables included in the complete case dataset have prefix f8.

Rank	Outcome variable (original data with variable N)	SD change of inverse normal transformed outcome for a 1 SD change of BMI allele score				SD change of inverse normal transformed imputed outcome for a 1 SD change of BMI allele score (N=8,121)			
		Sample size	SD change	95% CI	P value (adjusted P value) [2]	Imputed rank	SD change	95% CI	P value (adjusted P value)[2]
1	leptin_9 *	4,249	0.138	0.11, 0.17	<0.001	1	0.122	0.09, 0.15	<0.001
2	crp_9 *	4,250	0.083	0.05, 0.11	<0.001	3	0.072	0.04, 0.10	<0.001
3	AGE_MENARCHE_YE *	2,946	-0.083	-0.12, -0.05	<0.001	2	-0.087	-0.12, -0.06	<0.001
4	hdl_9	4,250	-0.067	-0.10, -0.04	<0.001 (0.002)	13	-0.040	-0.08, 0.00	0.028 (1)
5	f7sa021 *	6,013	0.049	0.02, 0.07	<0.001	4	0.045	0.02, 0.07	<0.001
6	il6_9	4,240	0.053	0.02, 0.08	0.001 (0.091)	8	0.043	0.01, 0.08	0.011 (1)
7	kk489	5,807	0.041	0.02, 0.07	0.002 (0.255)	5	0.042	0.02, 0.07	0.001 (0.188)
8	f8se125	5,222	0.042	0.02, 0.07	0.002 (0.323)	6	0.039	0.01, 0.07	0.005 (0.787)
9	apob_9	4,250	0.043	0.01, 0.07	0.005 (0.788)	37	0.021	-0.01, 0.05	0.160 (1)
10	trig_9	4,250	0.042	0.01, 0.07	0.006 (0.970)	33	0.022	-0.01, 0.05	0.152 (1)
11	vldl_9	4,250	0.042	0.01, 0.07	0.006 (1)	34	0.022	-0.01, 0.05	0.157 (1)
12	apoai_9	4,250	-0.038	-0.07, -0.01	0.012 (1)	42	-0.022	-0.05, 0.01	0.198 (1)
13	insulini_15 *	2,859	0.047	0.01, 0.08	0.012	9	0.045	0.01, 0.08	0.016
14	se093b	4,541	0.037	0.01, 0.07	0.013 (1)	18	0.030	0.00, 0.06	0.041 (1)
15	kqpemotion	5,748	-0.030	-0.06, 0.00	0.022 (1)	14	-0.030	-0.06, 0.00	0.030 (1)
16	kk310	6,231	0.028	0.00, 0.05	0.024 (1)	20	0.027	0.00, 0.05	0.048 (1)
17	f8se126	5,214	0.031	0.00, 0.06	0.025 (1)	24	0.026	0.00, 0.05	0.076 (1)
18	f8lf110 *	5,276	0.030	0.00, 0.06	0.030 (1)	7	0.033	0.01, 0.06	0.008 (1)
19	kr351a	5,684	0.028	0.00, 0.05	0.031 (1)	15	0.026	0.00, 0.05	0.036 (1)
20	kr236b	5,734	-0.028	-0.05, 0.00	0.033 (1)	11	-0.030	-0.06, 0.00	0.019 (1)
21	glucosem_15 *	2,862	0.041	0.00, 0.08	0.038	16	0.041	0.00, 0.08	0.037
22	se129b	4,545	0.028	0.00, 0.06	0.051 (1)	10	0.030	0.00, 0.05	0.019 (1)

23	se123b	4,545	0.029	0.00, 0.06	0.052 (1)	12	0.030	0.00, 0.05	0.022 (1)
24	kr367b	5,703	-0.025	-0.05, 0.00	0.054 (1)	17	-0.027	-0.05, 0.00	0.040 (1)
25	se162b	4,546	0.028	0.00, 0.06	0.059 (1)	38	0.021	-0.01, 0.05	0.163 (1)
26	kq538	5,798	-0.025	-0.05, 0.00	0.062 (1)	22	-0.024	-0.05, 0.00	0.068 (1)
27	kr247a	5,299	-0.025	-0.05, 0.00	0.065 (1)	21	-0.025	-0.05, 0.00	0.061 (1)
28	kr379b	5,674	-0.023	-0.05, 0.00	0.072 (1)	19	-0.027	-0.05, 0.00	0.045 (1)
29	kq475	5,769	-0.023	-0.05, 0.00	0.078 (1)	23	-0.023	-0.05, 0.00	0.072 (1)
30	kr259b	5,692	-0.023	-0.05, 0.00	0.085 (1)	25	-0.023	-0.05, 0.00	0.076 (1)
31	ku709b	5,746	-0.022	-0.05, 0.00	0.089 (1)	30	-0.021	-0.05, 0.01	0.124 (1)
32	kgnsp	6,301	0.022	0.00, 0.05	0.090 (1)	29	0.019	0.00, 0.04	0.114 (1)
33	kgcarotene	6,301	0.021	0.00, 0.05	0.095 (1)	35	0.018	-0.01, 0.04	0.158 (1)
34	se090b	4,541	0.024	-0.01, 0.05	0.105 (1)	28	0.021	0.00, 0.05	0.107 (1)
35	adiponectin_9	4,247	-0.024	-0.05, 0.01	0.126 (1)	59	-0.016	-0.05, 0.02	0.326 (1)
36	se134a	4,514	0.022	-0.01, 0.05	0.129 (1)	41	0.021	-0.01, 0.05	0.196 (1)
37	ku705b	5,755	0.020	-0.01, 0.05	0.132 (1)	40	0.017	-0.01, 0.04	0.191 (1)
38	se165b	4,546	0.023	-0.01, 0.05	0.132 (1)	27	0.026	-0.01, 0.06	0.105 (1)
39	kq573	5,798	0.019	-0.01, 0.04	0.137 (1)	26	0.022	0.00, 0.05	0.097 (1)
40	kq622	5,635	0.020	-0.01, 0.05	0.141 (1)	32	0.021	-0.01, 0.05	0.129 (1)
41	f8at229	5,416	0.020	-0.01, 0.05	0.142 (1)	45	0.016	-0.01, 0.04	0.221 (1)
42	ldl_9	4,250	0.022	-0.01, 0.05	0.152 (1)	94	0.009	-0.02, 0.04	0.541 (1)
43	f8ba066	5,572	-0.019	-0.04, 0.01	0.158 (1)	36	-0.018	-0.04, 0.01	0.159 (1)
44	f8ba026	5,581	0.018	-0.01, 0.04	0.177 (1)	31	0.020	-0.01, 0.05	0.128 (1)
45	ku710b	5,732	-0.017	-0.04, 0.01	0.179 (1)	56	-0.014	-0.04, 0.01	0.298 (1)
46	kgvitb12	6,301	-0.017	-0.04, 0.01	0.181 (1)	43	-0.016	-0.04, 0.01	0.198 (1)
47	f8aa150	5,367	-0.018	-0.04, 0.01	0.184 (1)	51	-0.018	-0.05, 0.01	0.249 (1)
48	f8ba056	5,572	-0.018	-0.04, 0.01	0.184 (1)	52	-0.015	-0.04, 0.01	0.259 (1)
49	ku509b	5,693	0.017	-0.01, 0.04	0.196 (1)	54	0.015	-0.01, 0.04	0.294 (1)
50	f8bp046	5,559	0.017	-0.01, 0.04	0.199 (1)	84	0.009	-0.02, 0.04	0.493 (1)

51	sf574b	4,296	0.019	-0.01, 0.05	0.202 (1)	47	0.018	-0.01, 0.05	0.232 (1)
52	kq519	5,770	0.016	-0.01, 0.04	0.211 (1)	60	0.013	-0.01, 0.04	0.327 (1)
53	ku503b	5,793	0.016	-0.01, 0.04	0.219 (1)	63	0.012	-0.01, 0.04	0.335 (1)
54	kqpebdtot	5,724	-0.015	-0.04, 0.01	0.235 (1)	46	-0.015	-0.04, 0.01	0.227 (1)
55	se164b	4,545	0.017	-0.01, 0.05	0.241 (1)	70	0.013	-0.02, 0.04	0.386 (1)
56	kq597	5,748	0.015	-0.01, 0.04	0.258 (1)	44	0.016	-0.01, 0.04	0.212 (1)
57	kgretinol	6,301	-0.014	-0.04, 0.01	0.279 (1)	48	-0.014	-0.04, 0.01	0.235 (1)
58	ku707b	5,736	-0.014	-0.04, 0.01	0.290 (1)	57	-0.013	-0.04, 0.01	0.314 (1)
59	kr519b	5,664	0.013	-0.01, 0.04	0.308 (1)	83	0.010	-0.02, 0.04	0.473 (1)
60	kq525	5,765	-0.013	-0.04, 0.01	0.312 (1)	55	-0.014	-0.04, 0.01	0.296 (1)
61	f8sl040	5,560	-0.014	-0.04, 0.01	0.313 (1)	39	-0.017	-0.04, 0.01	0.186 (1)
62	se166b	4,546	0.015	-0.01, 0.04	0.327 (1)	74	0.012	-0.02, 0.04	0.403 (1)
63	sf611b	4,302	-0.014	-0.04, 0.01	0.327 (1)	71	-0.013	-0.04, 0.02	0.387 (1)
64	kgvitd	6,301	-0.013	-0.04, 0.01	0.328 (1)	62	-0.012	-0.04, 0.01	0.331 (1)
65	kq680	5,798	-0.013	-0.04, 0.01	0.336 (1)	49	-0.016	-0.04, 0.01	0.236 (1)
66	kgmg	6,301	0.012	-0.01, 0.04	0.345 (1)	61	0.011	-0.01, 0.03	0.329 (1)
67	kr492b	5,661	0.012	-0.01, 0.04	0.353 (1)	80	0.010	-0.02, 0.04	0.456 (1)
68	ku504b	5,775	-0.012	-0.04, 0.01	0.355 (1)	50	-0.016	-0.04, 0.01	0.240 (1)
69	f8dv440 *	5,108	-0.013	-0.04, 0.01	0.355	67	-0.012	-0.04, 0.01	0.361
70	kr447b	5,680	0.011	-0.01, 0.04	0.383 (1)	106	0.006	-0.02, 0.03	0.629 (1)
71	kr309a	5,665	-0.011	-0.04, 0.01	0.413 (1)	77	-0.011	-0.04, 0.02	0.430 (1)
72	f8ba036	5,567	-0.011	-0.04, 0.01	0.415 (1)	115	-0.006	-0.04, 0.02	0.690 (1)
73	f8at065	5,483	0.011	-0.02, 0.04	0.422 (1)	58	0.014	-0.01, 0.04	0.321 (1)
74	kgcalcium	6,301	-0.010	-0.03, 0.01	0.423 (1)	82	-0.009	-0.03, 0.01	0.458 (1)
75	kgiron	6,301	0.010	-0.01, 0.03	0.426 (1)	66	0.011	-0.01, 0.03	0.356 (1)
76	ku706b	5,756	-0.010	-0.04, 0.02	0.433 (1)	91	-0.008	-0.03, 0.02	0.537 (1)
77	kgvite	6,301	-0.010	-0.03, 0.02	0.442 (1)	81	-0.009	-0.03, 0.01	0.457 (1)
78	f8at061	5,427	-0.010	-0.04, 0.02	0.444 (1)	118	-0.005	-0.03, 0.02	0.694 (1)

79	kr429a	5,653	0.010	-0.02, 0.04	0.448 (1)	87	0.009	-0.02, 0.04	0.513 (1)
80	kgfolate	6,301	0.009	-0.02, 0.03	0.458 (1)	75	0.010	-0.01, 0.03	0.412 (1)
81	f8bp026	5,573	-0.010	-0.04, 0.02	0.461 (1)	116	-0.005	-0.03, 0.02	0.692 (1)
82	kqphyper	5,748	-0.010	-0.04, 0.02	0.464 (1)	73	-0.010	-0.03, 0.01	0.401 (1)
83	se161b	4,546	-0.011	-0.04, 0.02	0.472 (1)	72	-0.012	-0.04, 0.02	0.399 (1)
84	kc_und	6,853	-0.009	-0.03, 0.02	0.476 (1)	69	-0.011	-0.03, 0.01	0.372 (1)
85	chol_9	4,250	0.011	-0.02, 0.04	0.478 (1)	127	0.005	-0.03, 0.04	0.750 (1)
86	kgcholesterol	6,301	-0.009	-0.03, 0.02	0.488 (1)	85	-0.008	-0.03, 0.01	0.496 (1)
87	kr462b	5,689	0.009	-0.02, 0.03	0.488 (1)	103	0.006	-0.02, 0.03	0.618 (1)
88	ku505b	5,770	-0.009	-0.03, 0.02	0.491 (1)	53	-0.015	-0.04, 0.01	0.284 (1)
89	kgiodine	6,301	0.008	-0.02, 0.03	0.499 (1)	68	0.011	-0.01, 0.03	0.371 (1)
90	kgpoly	6,301	-0.008	-0.03, 0.02	0.507 (1)	88	-0.008	-0.03, 0.02	0.517 (1)
91	kq477	5,769	-0.008	-0.03, 0.02	0.527 (1)	79	-0.009	-0.03, 0.01	0.449 (1)
92	se087b	4,537	0.009	-0.02, 0.04	0.527 (1)	135	0.004	-0.02, 0.03	0.781 (1)
93	kgstarch	6,301	0.008	-0.02, 0.03	0.528 (1)	89	0.008	-0.02, 0.03	0.534 (1)
94	kq502	5,765	0.008	-0.02, 0.03	0.535 (1)	105	0.006	-0.02, 0.03	0.622 (1)
95	f8bp036	5,559	-0.008	-0.03, 0.02	0.544 (1)	132	-0.004	-0.03, 0.02	0.779 (1)
96	sf576b	4,318	0.009	-0.02, 0.04	0.546 (1)	122	0.006	-0.02, 0.04	0.707 (1)
97	kgdha	6,301	0.008	-0.02, 0.03	0.547 (1)	102	0.006	-0.02, 0.03	0.616 (1)
98	kgselenium	6,301	0.008	-0.02, 0.03	0.558 (1)	96	0.007	-0.02, 0.03	0.567 (1)
99	hb_f7	4,761	0.008	-0.02, 0.04	0.560 (1)	121	0.006	-0.02, 0.03	0.705 (1)
100	kq425	5,777	-0.007	-0.03, 0.02	0.576 (1)	76	-0.010	-0.03, 0.01	0.417 (1)
101	kd380a *	6,885	0.007	-0.02, 0.03	0.581	99	0.006	-0.02, 0.03	0.598
102	f8ba046	5,565	0.007	-0.02, 0.03	0.582 (1)	97	0.007	-0.02, 0.03	0.589 (1)
103	kgn3	6,301	0.007	-0.02, 0.03	0.588 (1)	112	0.005	-0.02, 0.03	0.679 (1)
104	kqppeer	5,752	-0.007	-0.03, 0.02	0.592 (1)	90	-0.008	-0.03, 0.02	0.535 (1)
105	ku708b	5,751	-0.007	-0.03, 0.02	0.593 (1)	158	-0.001	-0.03, 0.03	0.921 (1)
106	kqpconduct	5,755	0.007	-0.02, 0.03	0.595 (1)	104	0.006	-0.02, 0.03	0.621 (1)

107	kr332b	5,651	-0.007	-0.03, 0.02	0.595 (1)	78	-0.010	-0.04, 0.02	0.432 (1)
108	kgepa	6,301	0.007	-0.02, 0.03	0.595 (1)	111	0.005	-0.02, 0.03	0.675 (1)
109	f8gb041	5,301	-0.007	-0.03, 0.02	0.598 (1)	109	-0.006	-0.03, 0.02	0.664 (1)
110	kc_says	6,853	-0.006	-0.03, 0.02	0.610 (1)	92	-0.007	-0.03, 0.02	0.538 (1)
111	kq558	5,798	-0.007	-0.03, 0.02	0.614 (1)	120	-0.005	-0.03, 0.02	0.699 (1)
112	kr222a	5,683	-0.007	-0.03, 0.02	0.625 (1)	65	-0.013	-0.04, 0.01	0.345 (1)
113	f8at228	5,420	0.007	-0.02, 0.03	0.632 (1)	124	0.005	-0.02, 0.03	0.735 (1)
114	fd10cv_kcal	4,922	-0.007	-0.03, 0.02	0.643 (1)	64	-0.014	-0.04, 0.01	0.339 (1)
115	kr554b	5,666	0.006	-0.02, 0.03	0.648 (1)	152	0.002	-0.02, 0.03	0.879 (1)
116	kr478a	5,476	-0.006	-0.03, 0.02	0.659 (1)	163	0.001	-0.03, 0.03	0.943 (1)
117	f9sn702 *	5,286	-0.006	-0.03, 0.02	0.659	160	-0.001	-0.03, 0.02	0.930
118	se098a	4,420	0.006	-0.02, 0.04	0.662 (1)	128	0.004	-0.02, 0.03	0.751 (1)
119	kgfat	6,301	-0.005	-0.03, 0.02	0.679 (1)	110	-0.005	-0.03, 0.02	0.667 (1)
120	f8fs120	5,365	-0.005	-0.03, 0.02	0.687 (1)	129	-0.004	-0.03, 0.02	0.763 (1)
121	ku506b	5,775	0.005	-0.02, 0.03	0.688 (1)	153	0.002	-0.02, 0.03	0.883 (1)
122	ku508b	5,695	0.005	-0.02, 0.03	0.693 (1)	145	0.003	-0.02, 0.03	0.835 (1)
123	kr459b	5,673	0.005	-0.02, 0.03	0.695 (1)	101	0.007	-0.02, 0.03	0.616 (1)
124	kr213b	5,631	-0.005	-0.03, 0.02	0.704 (1)	93	-0.008	-0.03, 0.02	0.539 (1)
125	kgsfa	6,301	-0.005	-0.03, 0.02	0.705 (1)	117	-0.005	-0.03, 0.02	0.693 (1)
126	ku673b	5,749	-0.005	-0.03, 0.02	0.713 (1)	138	-0.003	-0.03, 0.02	0.800 (1)
127	kq442	5,761	-0.005	-0.03, 0.02	0.721 (1)	95	-0.007	-0.03, 0.02	0.552 (1)
128	kr300b	5,708	0.005	-0.02, 0.03	0.722 (1)	125	0.004	-0.02, 0.03	0.748 (1)
129	kq486	5,758	0.005	-0.02, 0.03	0.725 (1)	150	0.002	-0.02, 0.03	0.858 (1)
130	kgcarbohydrate	6,301	0.004	-0.02, 0.03	0.745 (1)	114	0.005	-0.02, 0.03	0.685 (1)
131	kgribo	6,301	-0.004	-0.03, 0.02	0.750 (1)	154	-0.002	-0.02, 0.02	0.888 (1)
132	kgnmesugars	6,301	0.004	-0.02, 0.03	0.751 (1)	100	0.006	-0.02, 0.03	0.599 (1)
133	kgprotein	6,301	-0.004	-0.03, 0.02	0.754 (1)	131	-0.003	-0.03, 0.02	0.769 (1)
134	f8at147	5,312	-0.004	-0.03, 0.02	0.765 (1)	161	-0.001	-0.03, 0.02	0.939 (1)

135	kr275a	5,718	-0.004	-0.03, 0.02	0.767 (1)	146	-0.003	-0.03, 0.02	0.836 (1)
136	f8ws112 *	5,516	-0.004	-0.03, 0.02	0.769	86	-0.008	-0.03, 0.02	0.513
137	kq316	5,780	0.004	-0.02, 0.03	0.782 (1)	123	0.005	-0.02, 0.03	0.717 (1)
138	f8at060	5,447	-0.004	-0.03, 0.02	0.791 (1)	139	-0.003	-0.03, 0.02	0.801 (1)
139	kgphosphorus	6,301	-0.003	-0.03, 0.02	0.798 (1)	130	-0.003	-0.03, 0.02	0.768 (1)
140	kgzinc	6,301	-0.003	-0.03, 0.02	0.798 (1)	136	-0.003	-0.03, 0.02	0.789 (1)
141	f9sn703 *	5,286	0.003	-0.02, 0.03	0.800	98	0.007	-0.02, 0.03	0.589
142	ku510b	5,668	-0.003	-0.03, 0.02	0.805 (1)	113	-0.006	-0.03, 0.02	0.683 (1)
143	sf575b	4,318	0.004	-0.03, 0.03	0.817 (1)	164	0.001	-0.03, 0.03	0.945 (1)
144	f8lc125	4,793	-0.003	-0.03, 0.03	0.824 (1)	134	-0.004	-0.03, 0.02	0.781 (1)
145	se163b	4,546	-0.003	-0.03, 0.03	0.830 (1)	157	-0.002	-0.03, 0.03	0.917 (1)
146	kq378b	5,775	0.003	-0.02, 0.03	0.833 (1)	147	0.003	-0.02, 0.03	0.843 (1)
147	kgmono	6,301	-0.002	-0.03, 0.02	0.842 (1)	155	-0.001	-0.02, 0.02	0.896 (1)
148	f8bp066	5,566	-0.003	-0.03, 0.02	0.849 (1)	165	0.001	-0.03, 0.03	0.949 (1)
149	kqpprosoc	5,754	-0.002	-0.03, 0.02	0.858 (1)	143	-0.003	-0.03, 0.02	0.820 (1)
150	kr387a	5,673	0.002	-0.02, 0.03	0.859 (1)	142	-0.003	-0.03, 0.02	0.820 (1)
151	kr501a	5,700	0.002	-0.02, 0.03	0.879 (1)	149	0.002	-0.02, 0.03	0.858 (1)
152	kgsodium	6,301	0.002	-0.02, 0.03	0.883 (1)	151	0.002	-0.02, 0.02	0.873 (1)
153	kr468b	5,688	0.002	-0.02, 0.03	0.886 (1)	156	0.002	-0.03, 0.03	0.904 (1)
154	f8at146	5,340	-0.002	-0.03, 0.02	0.889 (1)	172	0.000	-0.03, 0.03	0.999 (1)
155	kgvitb6	6,301	-0.002	-0.03, 0.02	0.896 (1)	140	-0.003	-0.03, 0.02	0.802 (1)
156	kk317	6,209	0.002	-0.02, 0.03	0.898 (1)	137	0.003	-0.02, 0.03	0.795 (1)
157	kq653	5,216	0.002	-0.03, 0.03	0.902 (1)	108	0.006	-0.02, 0.03	0.661 (1)
158	kgalcohol	6,301	-0.002	-0.03, 0.02	0.906 (1)	148	-0.002	-0.03, 0.02	0.854 (1)
159	kq517	5,771	-0.001	-0.03, 0.02	0.908 (1)	162	-0.001	-0.03, 0.03	0.941 (1)
160	sf573b	4,321	-0.002	-0.03, 0.03	0.912 (1)	119	-0.006	-0.04, 0.02	0.697 (1)
161	kgpotassium	6,301	0.001	-0.02, 0.03	0.917 (1)	144	0.002	-0.02, 0.03	0.830 (1)
162	kq462	5,774	0.001	-0.02, 0.03	0.919 (1)	126	0.004	-0.02, 0.03	0.749 (1)

163	kr337b	5,651	-0.001	-0.03, 0.02	0.922 (1)	133	-0.004	-0.03, 0.02	0.779 (1)
164	kgsugar	6,301	-0.001	-0.03, 0.02	0.939 (1)	169	0.000	-0.02, 0.02	0.983 (1)
165	kgvitc	6,301	0.001	-0.02, 0.03	0.945 (1)	159	-0.001	-0.03, 0.02	0.924 (1)
166	kgthiamin	6,301	0.001	-0.02, 0.03	0.951 (1)	171	0.000	-0.02, 0.02	0.993 (1)
167	f8bp056	5,567	0.001	-0.03, 0.03	0.968 (1)	107	0.006	-0.02, 0.03	0.631 (1)
168	f8at148	5,315	0.001	-0.03, 0.03	0.968 (1)	170	0.000	-0.02, 0.02	0.988 (1)
169	kgnceq	6,301	0.000	-0.02, 0.03	0.971 (1)	166	0.000	-0.02, 0.02	0.968 (1)
170	se126b	4,545	0.000	-0.03, 0.03	0.975 (1)	168	0.001	-0.03, 0.03	0.970 (1)
171	kgenergy	6,301	0.000	-0.02, 0.02	0.990 (1)	167	0.000	-0.02, 0.02	0.969 (1)
172	ku507b	5,751	0.000	-0.03, 0.03	1.000 (1)	141	-0.004	-0.03, 0.03	0.812 (1)

Table B.4: Ranking by association strength (P value) of the stage one tests: Outcome associations with BMI allele score for original and imputed datasets Full names of variables are given in Table B.3. All outcomes are transformed to normal distributions using a rank-based inverse normal transformation. Exposure and outcome variables are standardised. Outcome as dependent variable, BMI allele score as independent variable.

[1] Using Stata IVregress command and robust option. BMI allele score is the instrumental variable for log BMI age 8. First stage predicting log bmi at age 8. The second stage performs an unadjusted association of these log bmi age 8 predictions with the outcome.

[2] Adjusted P values are adjusted for the 160 tests performed using the Bonferroni correction: pcorrected = poriginal160. Adjusted P values greater than 1 are rounded to 1.

* Variables in validation set.

Abbreviated variable name	Variable
Leptin, 9	leptin_9
CRP, 9	crp_9
Age menarche	AGE_MENARCHE_YE
HDL, 9	hdl_9
SBP, 7	f7sa021
IL6, 9	il6_9
Enjoyment of School Score, 4	kk489
Self Esteem: Scholastic Competence, 8	f8se125
Apolipoprotein B, 9	apob_9
Triglycerides, 9	trig_9
VLDL age 9	vldl_9
Apolipoprotein al, 9	apoai_9
Insulin, 15	insulini_15
Attention/activity symptoms score, 11	se093b
SDQ emotional symptoms score, 6	kqpemotion
Hygiene Score, 4	kk310
Self Esteem: Global Self Worth Score, 8	f8se126
FVC: LF, 8	f8lf110
Burden of compulsions/obsessions score, 7	kr351a
Particular fears score, 7	kr236b
Glucose, 15	glucosem_15

Table B.5: List of the variables associated with the BMI allele score, with abbreviated and full variable names. Abbreviations: BMI, body mass index; CI, confidence interval; SD, standard deviation; IV, instrumental variable; CRP, c-reactive protein; LDL, low density lipoprotein; IL6, interleukin 6; SBP, systolic blood pressure; HDL, high density lipoprotein; VLDL, very low density lipoprotein; SDQ, Strengths and Difficulties Questionnaires; LF, lung function; FVC, forced vital capacity.

Outcome variable	Data transformation
Leptin 9	Logarithm
CRP, 9	Logarithm
HDL, 9	Logarithm
IL6, 9	Logarithm
VLDL, 9	Logarithm
Triglycerides, 9	Logarithm
Insulin, 15	Logarithm
Enjoyment of School Score, 4	Binary <19, $\geqslant 19$
Self Esteem: Scholastic Competence, 8	Binary <18, $\geqslant 18$
Attention/activity symptoms score, 11	Binary 0, >0
SDQ emotional symptoms score, 6	Binary 0, >0
Self Esteem: Global Self Worth Score, 8	Binary <20, $\geqslant 20$
Burden of compulsions/obsessions score, 7	Binary 0, >0
Particular fears score, 7	Binary <4, $\geqslant 4$

Table B.6: Data transformations of outcome variables used in stage 2 analysis. Logarithm tranformations use base 10.

Bibliography

[1] R. Peto, S. Darby, H. Deo, P. Silcocks, E. Whitley, and R. Doll, "Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies," *BMJ*, vol. 321, no. 7257, pp. 323–329, 2000.

[2] A. S. Team *et al.*, "ALSPAC–the Avon Longitudinal Study of Parents and Children," *Paediatric and perinatal epidemiology*, vol. 15, no. 1, pp. 74–87, 2001.

[3] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[4] J. P. T. Higgins and S. Green, eds., *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Ltd, 2008.

[5] H. Bastian, P. Glasziou, and I. Chalmers, "Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?," *PLoS medicine*, vol. 7, no. 9, p. e1000326, 2010.

[6] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani, and E. Coiera, "Systematic review automation technologies," *Syst Rev*, vol. 3, p. 74, 2014.

[7] S. Ananiadou, B. Rea, N. Okazaki, R. Procter, and J. Thomas, "Supporting systematic reviews using text mining," *Social Science Computer Review*, vol. 27, no. 4, pp. 509–523, 2009.

[8] J. Thomas, J. McNaught, and S. Ananiadou, "Applications of text mining within systematic reviews," *Research Synthesis Methods*, vol. 2, no. 1, pp. 1–14, 2011.

[9] A. M. Cohen, C. E. Adams, J. M. Davis, C. Yu, P. S. Yu, W. Meng, L. Duggan, M. McDonagh, and N. R. Smalheiser, "Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools," in *Proceedings of the 1st ACM International Health Informatics Symposium*, pp. 376–380, ACM, 2010.

[10] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, S. Ananiadou, *et al.*, "Using text mining for study identification in systematic reviews: a systematic review of current approaches," *Systematic reviews*, vol. 4, no. 1, p. 5, 2015.

[11] 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*, vol. 17, no. 4, pp. 446–453, 2010.

[12] 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*, vol. 5, no. 1, pp. 31–49, 2014.

[13] A. M. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen, "Reducing workload in systematic review preparation using automated citation classification," *Journal of the American Medical Informatics Association*, vol. 13, no. 2, pp. 206–219, 2006.

[14] S. Jonnalagadda and D. Petitti, "A new iterative method to reduce workload in systematic review process," *International journal of computational biology and drug design*, vol. 6, no. 1, pp. 5–17, 2013.

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

[16] 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*, vol. 11, no. 1, p. 55, 2010.

[17] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Active learning for biomedical citation screening," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 173–182, ACM, 2010.

[18] O. Frunza, D. Inkpen, S. Matwin, W. Klement, and P. Oblenis, "Exploiting the systematic review protocol for classification of medical abstracts," *Artificial intelligence in medicine*, vol. 51, no. 1, pp. 17–25, 2011.

[19] S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, and I. Sim, "ExaCT: automatic extraction of clinical trial characteristics from journal publications," *BMC Medical Informatics and Decision Making*, vol. 10, p. 56, Sept. 2010.

[20] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," *BMC bioinformatics*, vol. 12, no. Suppl 2, p. S5, 2011.

[21] J. Kuiper, I. Marshall, B. Wallace, and M. Swertz, "Spá: A web-based viewer for text mining in evidence based medicine," in *Machine Learning and Knowledge Discovery in Databases*, pp. 452–455, Springer, 2014.

[22] I. Marshall, J. Kuiper, and B. Wallace, "RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials," *Journal of the American Medical Informatics Association : JAMIA*, 2015.

[23] I. J. Marshall, J. Kuiper, and B. C. Wallace, "Automating risk of bias assessment for clinical trials," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 88–95, ACM, 2014.

[24] H. M. Colhoun, P. M. McKeigue, and G. Davey Smith, "Problems of reporting genetic associations with complex outcomes," *The Lancet*, vol. 361, no. 9360, pp. 865–872, 2003.

[25] L. R. Cardon and J. I. Bell, "Association study designs for complex diseases," *Nature Reviews Genetics*, vol. 2, no. 2, pp. 91–99, 2001.

[26] L. A. C. Millard, N. Davies, N. Timpson, K. Tilling, P. A. Flach, and G. Davey Smith, "MR-PheWAS: hypothesis-prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization," *Scientific reports*, vol. 5, 2015.

[27] L. A. C. Millard, P. A. Flach, and J. P. T. Higgins, "Rate-constrained ranking and the rate-weighted AUC," in *Machine Learning and Knowledge Discovery in Databases*, pp. 386–403, Springer, 2014.

[28] L. A. C. Millard, M. Kull, and P. A. Flach, "Rate-oriented point-wise confidence bounds for ROC curves," in *Machine Learning and Knowledge Discovery in Databases*, pp. 404–421, Springer, 2014.

[29] 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*, 2015. doi:10.1093/ije/dyv306.

[30] G. Davey Smith, D. A. Lawlor, R. Harbord, N. Timpson, I. Day, and S. Ebrahim, "Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology," *PLoS Medicine*, vol. 4, no. 12, p. e352, 2007.

[31] D. A. Lawlor, G. Davey Smith, K. R. Bruckdorfer, D. Kundu, and S. Ebrahim, "Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence?," *The Lancet*, vol. 363, no. 9422, pp. 1724–1727, 2004.

[32] C. D. Mulrow, "Systematic reviews: rationale for systematic reviews," *BMJ*, vol. 309, no. 6954, pp. 597–599, 1994.

[33] J. Lau, E. M. Antman, J. Jimenez-Silva, B. Kupelnick, F. Mosteller, and T. C. Chalmers, "Cumulative meta-analysis of therapeutic trials for myocardial infarction," *New England Journal of Medicine*, vol. 327, no. 4, pp. 248–254, 1992.

[34] R. Ganann, D. Ciliska, and H. Thomas, "Expediting systematic reviews: methods and implications of rapid reviews," *Implementation Science*, vol. 5, no. 1, p. 56, 2010.

[35] 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 Technology Assessment*, vol. 7, no. 1, pp. 1–76, 2003.

[36] J. P. T. Higgins, D. G. Altman, P. C. Gøtzsche, P. Jüni, D. Moher, A. D. Oxman, J. Savović, K. F. Schulz, L. Weeks, and J. A. Sterne, "The Cochrane Collaborations tool for assessing risk of bias in randomised trials," *BMJ*, vol. 343, 2011.

[37] A. W. Chan and D. G. Altman, "Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors," *BMJ*, vol. 330, no. 7494, p. 753, 2005.

[38] K. F. Schulz, I. Chalmers, R. J. Hayes, and D. G. Altman, "Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials," *Jama*, vol. 273, no. 5, pp. 408–412, 1995.

[39] A. R. Jadad, R. A. Moore, D. Carroll, C. Jenkinson, D. J. M. Reynolds, D. J. Gavaghan, and H. J. McQuay, "Assessing the quality of reports of randomized clinical trials: is blinding necessary?," *Controlled clinical trials*, vol. 17, no. 1, pp. 1–12, 1996.

[40] L. Wood, M. Egger, L. L. Gluud, K. F. Schulz, P. Jüni, D. G. Altman, C. Gluud, R. M. Martin, A. J. Wood, and J. A. Sterne, "Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study," *BMJ*, vol. 336, no. 7644, pp. 601–605, 2008.

[41] J. Savović, H. E. Jones, D. G. Altman, R. J. Harris, P. Jüni, J. Pildal, B. Als-Nielsen, E. M. Balk, C. Gluud, L. L. Gluud, *et al.*, "Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials," *Annals of internal medicine*, vol. 157, no. 6, pp. 429–438, 2012.

[42] P. Jüni, D. G. Altman, and M. Egger, "Assessing the quality of controlled clinical trials," *BMJ*, vol. 323, no. 7303, pp. 42–46, 2001.

[43] K. F. Schulz and D. A. Grimes, "Generation of allocation sequences in randomised trials: chance, not choice," *The Lancet*, vol. 359, no. 9305, pp. 515–519, 2002.

[44] K. F. Schulz and D. A. Grimes, "Allocation concealment in randomised trials: defending against deciphering," *The Lancet*, vol. 359, no. 9306, pp. 614–618, 2002.

[45] S. J. Day, D. G. Altman, *et al.*, "Blinding in clinical trials and other studies," *BMJ*, vol. 321, no. 7259, p. 504, 2000.

[46] S. V. Katikireddi, M. Egan, and M. Petticrew, "How do systematic reviews incorporate risk of bias assessments into the synthesis of evidence? A methodological study," *Journal of epidemiology and community health*, vol. 69, no. 2, pp. 189–195, 2015.

[47] C. Begg, M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, K. F. Schulz, D. Simel, *et al.*, "Improving the quality of reporting of randomized controlled trials: the CONSORT statement," *Jama*, vol. 276, no. 8, pp. 637–639, 1996.

[48] S. Hopewell, S. Dutton, L.-M. Yu, A.-W. Chan, and D. G. Altman, "The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed," *BMJ*, vol. 340, 2010.

[49] A. C. Plint, D. Moher, A. Morrison, K. Schulz, D. G. Altman, C. Hill, and I. Gaboury, "Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review," *Medical Journal of Australia*, vol. 185, no. 5, p. 263, 2006.

[50] J. Pildal, A.-W. Chan, A. Hróbjartsson, E. Forfang, D. G. Altman, and P. C. Gøtzsche, "Comparison of descriptions of allocation concealment in trial protocols and the published reports: cohort study," *BMJ*, vol. 330, no. 7499, p. 1049, 2005.

[51] C. L. Vale, J. F. Tierney, S. Burdett, *et al.*, "Can trial quality be reliably assessed from published reports of cancer trials: evaluation of risk of bias assessments in systematic reviews," *BMJ*, vol. 346, p. f1798, 2013.

[52] H. P. Soares, S. Daniels, A. Kumar, M. Clarke, C. Scott, S. Swann, B. Djulbegovic, *et al.*, "Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the radiation therapy oncology group," *BMJ*, vol. 328, no. 7430, pp. 22–24, 2004.

[53] K. Huwiler-Müntener, P. Jüni, C. Junker, and M. Egger, "Quality of reporting of randomized trials as a measure of methodologic quality," *Jama*, vol. 287, no. 21, pp. 2801–2804, 2002.

[54] R. Mhaskar, B. Djulbegovic, A. Magazin, H. P. Soares, and A. Kumar, "Published methodological quality of randomized controlled trials does not reflect the actual quality assessed in protocols," *Journal of clinical epidemiology*, vol. 65, no. 6, pp. 602–609, 2012.

[55] A. R. Jadad, D. J. Cook, A. Jones, T. P. Klassen, P. Tugwell, M. Moher, and D. Moher, "Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals," *Jama*, vol. 280, no. 3, pp. 278–280, 1998.

[56] A. W. Jørgensen, J. Hilden, and P. C. Gøtzsche, "Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review," *BMJ*, vol. 333, no. 7572, p. 782, 2006.

[57] O. Olsen, P. Middleton, J. Ezzo, P. C. Gøtzsche, V. Hadhazy, A. Herxheimer, J. Kleijnen, and H. McIntosh, "Quality of Cochrane reviews: assessment of sample from 1998," *BMJ*, vol. 323, no. 7317, pp. 829–832, 2001.

[58] S. Hopewell, I. Boutron, D. G. Altman, and P. Ravaud, "Incorporation of assessments of risk of bias of primary studies in systematic reviews of randomised trials: a cross-sectional study," *BMJ open*, vol. 3, no. 8, p. e003342, 2013.

[59] A. M. Moseley, M. R. Elkins, R. D. Herbert, C. G. Maher, and C. Sherrington, "Cochrane reviews used more rigorous methods than non-Cochrane reviews: sur-

vey of systematic reviews in physiotherapy," *Journal of clinical epidemiology*, vol. 62, no. 10, pp. 1021–1030, 2009.

[60] J. Savović, L. Weeks, J. A. Sterne, L. Turner, D. G. Altman, D. Moher, and J. P. T. Higgins, "Evaluation of the Cochrane collaborations tool for assessing the risk of bias in randomized trials: focus groups, online survey, proposed recommendations and their implementation," *Syst Rev*, vol. 3, no. 37, pp. 10–1186, 2014.

[61] L. Hartling, M. P. Hamm, A. Milne, B. Vandermeer, P. L. Santaguida, M. Ansari, A. Tsertsvadze, S. Hempel, P. Shekelle, and D. M. Dryden, "Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs," *Journal of Clinical Epidemiology*, vol. 66, no. 9, pp. 973–981, 2013.

[62] L. Hartling, M. Ospina, Y. Liang, D. M. Dryden, N. Hooton, J. Krebs Seida, and T. P. Klassen, "Risk of bias versus quality assessment of randomised controlled trials: cross sectional study," *BMJ*, vol. 339, 2009.

[63] L. Hartling, K. Bond, B. Vandermeer, J. Seida, D. M. Dryden, and B. H. Rowe, "Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma," *PLoS One*, vol. 6, no. 2, p. e17242, 2011.

[64] S. Lensen, C. Farquhar, and V. Jordon, "Risk of bias: are judgements consistent between reviews?," *Cochrane Database of Systematic Reviews supplement*, vol. 1, no. 150, p. 30, 2014.

[65] S. Armijo-Olivo, M. Ospina, B. R. da Costa, M. Egger, H. Saltaji, J. Fuentes, C. Ha, and G. G. Cummings, "Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane risk of bias tool in physical therapy trials," *PloS One*, vol. 9, no. 5, p. e96920, 2014.

[66] J. Thomas, M. Newman, and S. Oliver, "Rapid evidence assessments of research to inform social policy: taking stock and moving forward," *Evidence & Policy: A Journal of Research, Debate and Practice*, vol. 9, no. 1, pp. 5–27, 2013.

[67] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, and T. A. Trikalinos, "Modeling annotation time to reduce workload in comparative effectiveness reviews," in *Proceedings of the 1st ACM International Health Informatics Symposium*, pp. 28–35, ACM, 2010.

[68] 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*, vol. 18, no. 1, pp. 104–104, 2011.

[69] A. M. Cohen, "Optimizing feature representation for automated systematic review work prioritization," in *AMIA annual symposium proceedings*, vol. 2008, p. 121, American Medical Informatics Association, 2008.

[70] M. Khabsa, A. Elmagarmid, I. Ilyas, H. Hammady, and M. Ouzzani, "Learning to identify relevant studies for systematic reviews using random forest and external information," *Machine Learning*, pp. 1–18, 2015.

[71] A. M. Cohen, N. R. Smalheiser, M. S. McDonagh, C. Yu, C. E. Adams, J. M. Davis, and S. Y. Philip, "Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine," *Journal of the American Medical Informatics Association*, p. ocu025, 2015.

[72] N. Cristianini and M. W. Hahn, *Introduction to computational genomics: a case studies approach*. Cambridge University Press, 2006.

[73] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[74] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine learning*, vol. 77, no. 1, pp. 103–123, 2009.

[75] P. A. Flach, "The geometry of ROC space: understanding machine learning metrics through ROC isometrics," in *Proceedings of the 20th International Conference on Machine Learning*, ICML 2003, pp. 194–201, 2003.

[76] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[77] J. P. T. Higgins and D. G. Altman, "Assessing risk of bias in included studies," *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series*, pp. 187–241, 2008.

[78] L. E. Dodd and M. S. Pepe, "Partial AUC estimation and regression," *Biometrics*, vol. 59, no. 3, pp. 614–623, 2003.

[79] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests.," *Radiology*, vol. 201, no. 3, pp. 745–750, 1996.

[80] D. K. McClish, "Analyzing a portion of the ROC curve," *Medical Decision Making*, vol. 9, no. 3, pp. 190–195, 1989.

[81] A. P. Bradley, "Half-AUC for the evaluation of sensitive or specific classifiers," *Pattern Recognition Letters*, vol. 38, pp. 93–98, 2014.

[82] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–48, ACM, 2000.

[83] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[84] R. P. Sheridan, S. B. Singh, E. M. Fluder, and S. K. Kearsley, "Protocols for bridging the peptide to nonpeptide gap in topological similarity searches," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 5, pp. 1395–1406, 2001.

[85] J.-F. Truchon and C. I. Bayly, "Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem," *Journal of Chemical Information and Modeling*, vol. 47, no. 2, pp. 488–508, 2007.

[86] S. J. Swamidass, C.-A. Azencott, K. Daily, and P. Baldi, "A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval," *Bioinformatics*, vol. 26, no. 10, pp. 1348–1356, 2010.

[87] W. Zhao, K. E. Hevener, S. W. White, R. E. Lee, and J. M. Boyett, "A statistical framework to evaluate virtual screening," *BMC Bioinformatics*, vol. 10, no. 1, p. 225, 2009.

[88] J. Albert, "LearnBayes: Functions for learning Bayesian inference," *R package version 2.12*, 2008.

[89] F. J. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *ICML*, vol. 98, pp. 445–453, 1998.

[90] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Machine Learning*, vol. 31, pp. 1–38, 2004.

[91] S. Macskassy, F. Provost, and S. Rosset, "Pointwise ROC confidence bounds: An empirical evaluation," Proceedings of the Workshop on ROC Analysis in Machine Learning, 2005.

[92] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A first course in order statistics*, vol. 54. Siam, 1992.

[93] G. Campbell, "Advances in statistical methodology for the evaluation of diagnostic and laboratory tests," *Statistics in Medicine*, vol. 13, no. 5-7, pp. 499–508, 1994.

[94] S. A. Macskassy, F. Provost, and S. Rosset, "ROC confidence bands: an empirical evaluation," in *Proceedings of the 22nd International Conference on Machine Learning*, ICML 2005, (New York, NY, USA), pp. 537–544, 2005.

[95] J. B. Tilbury, W. Van Eetvelt, J. M. Garibaldi, J. Curnsw, and E. C. Ifeachor, "Receiver operating characteristic analysis for intelligent medical systems-a new approach for finding confidence intervals," *Biomedical Engineering, IEEE Transactions on*, vol. 47, no. 7, pp. 952–963, 2000.

[96] P. Hall, R. J. Hyndman, and Y. Fan, "Nonparametric confidence intervals for receiver operating characteristic curves," *Biometrika*, vol. 91, no. 3, pp. 743–750, 2004.

[97] M. M. Leeflang, J. J. Deeks, C. Gatsonis, and P. M. Bossuyt, "Systematic reviews of diagnostic test accuracy," *Annals of internal medicine*, vol. 149, no. 12, pp. 889–897, 2008.

[98] S. Macskassy and F. Provost, "Confidence bands for ROC curves: Methods and an empirical study," Proceedings of the First Workshop on ROC Analysis in AI, 2004.

[99] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, ACM, 2005.

[100] P. A. Flach, "ROC analysis," in *Encyclopedia of Machine Learning*, pp. 869–875, Springer, 2010.

[101] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[102] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[103] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers," in *ICML*, vol. 1, pp. 609–616, Citeseer, 2001.

[104] P. Flach, *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

[105] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nature Reviews Genetics*, vol. 9, no. 5, pp. 356–369, 2008.

[106] J. P. Ioannidis, "Why most published research findings are false," *PLoS medicine*, vol. 2, no. 8, p. e124, 2005.

[107] D. J. Hunter, "Lessons from genome-wide association studies for epidemiology," *Epidemiology*, vol. 23, no. 3, pp. 363–367, 2012.

[108] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9362–9367, 2009.

[109] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[110] A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, *et al.*, "Genetic studies of body mass index yield new insights for obesity biology," *Nature*, vol. 518, no. 7538, pp. 197–206, 2015.

[111] H. H. Maes, M. C. Neale, and L. J. Eaves, "Genetic and environmental factors in relative body weight and human adiposity," *Behavior genetics*, vol. 27, no. 4, pp. 325–351, 1997.

[112] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of GWAS discovery," *The American Journal of Human Genetics*, vol. 90, no. 1, pp. 7–24, 2012.

[113] C. J. Patel, M. R. Cullen, J. P. Ioannidis, and A. J. Butte, "Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels," *International journal of epidemiology*, vol. 41, no. 3, pp. 828–843, 2012.

[114] C. J. Patel, R. Chen, and A. J. Butte, "Data-driven integration of epidemiological and toxicological data to select candidate interacting genes and environmental factors in association with disease," *Bioinformatics*, vol. 28, no. 12, pp. i121–i126, 2012.

[115] I. Tzoulaki, C. J. Patel, T. Okamura, Q. Chan, I. J. Brown, K. Miura, H. Ueshima, L. Zhao, L. Van Horn, M. L. Daviglus, *et al.*, "A nutrient-wide association study on blood pressure," *Circulation*, pp. CIRCULATIONAHA–112, 2012.

[116] S. Pendergrass and M. D. Ritchie, "Phenome-wide association studies: Leveraging comprehensive phenotypic and genotypic data for discovery," *Current Genetic Medicine Reports*, vol. 3, no. 2, pp. 92–100.

[117] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford, "PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations," *Bioinformatics*, vol. 26, no. 9, pp. 1205–1210, 2010.

[118] M. A. Hall, A. Verma, K. D. Brown-Gentry, R. Goodloe, J. Boston, S. Wilson, B. McClellan, C. Sutcliffe, H. H. Dilks, N. B. Gillani, *et al.*, "Detection of pleiotropy through a phenome-wide association study (PheWAS) of epidemiologic data as part of the environmental architecture for genes linked to environment (EAGLE) study," *PLoS genetics*, vol. 10, no. 12, p. e1004678, 2014.

[119] G. Davey Smith and S. Ebrahim, "Data dredging, bias, or confounding: they can all get you into the BMJ and the friday papers," *BMJ: British Medical Journal*, vol. 325, no. 7378, p. 1437, 2002.

[120] G. Davey Smith and S. Ebrahim, "Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease?," *International journal of epidemiology*, vol. 32, no. 1, pp. 1–22, 2003.

[121] G. Davey Smith and G. Hemani, "Mendelian randomization: genetic anchors for causal inference in epidemiological studies," *Human molecular genetics*, vol. 23, no. R1, pp. R89–R98, 2014.

[122] D. A. Lawlor, R. M. Harbord, J. A. Sterne, N. Timpson, and G. Davey Smith, "Mendelian randomization: using genes as instruments for making causal inferences in epidemiology," *Statistics in medicine*, vol. 27, no. 8, pp. 1133–1163, 2008.

[123] L. R. Cardon and L. J. Palmer, "Population stratification and spurious allelic association," *The Lancet*, vol. 361, no. 9357, pp. 598–604, 2003.

[124] A. L. Tyler, F. W. Asselbergs, S. M. Williams, and J. H. Moore, "Shadows of complexity: what biological networks reveal about epistasis and pleiotropy," *Bioessays*, vol. 31, no. 2, pp. 220–227, 2009.

[125] V. Didelez and N. Sheehan, "Mendelian randomization as an instrumental variable approach to causal inference," *Statistical methods in medical research*, vol. 16, no. 4, pp. 309–330, 2007.

[126] T. J. VanderWeele, E. J. T. Tchetgen, M. Cornelis, and P. Kraft, "Methodological challenges in Mendelian randomization," *Epidemiology*, vol. 25, no. 3, pp. 427–435, 2014.

[127] M. A. Hernán and J. M. Robins, "Instruments for causal inference: an epidemiologist's dream?," *Epidemiology*, vol. 17, no. 4, pp. 360–372, 2006.

[128] G. Vazquez, S. Duval, D. R. Jacobs, and K. Silventoinen, "Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis," *Epidemiologic reviews*, vol. 29, no. 1, pp. 115–128, 2007.

[129] C. Friedemann, C. Heneghan, K. Mahtani, M. Thompson, R. Perera, A. M. Ward, *et al.*, "Cardiovascular disease risk in healthy children and its association with body mass index: systematic review and meta-analysis," *BMJ*, vol. 345, p. e4759, 2012.

[130] H. Oude Luttikhuis, L. Baur, H. Jansen, V. A. Shrewsbury, C. O'Malley, R. P. Stolk, and C. D. Summerbell, "Interventions for treating obesity in children," *The Cochrane Library*, 2009.

[131] K. Shaw, H. Gennat, P. ORourke, and C. Del Mar, "Exercise for overweight or obesity," *Cochrane Database Syst Rev*, vol. 4, no. 4, 2006.

[132] N. J. Timpson, R. Harbord, G. Davey Smith, J. Zacho, A. Tybjærg-Hansen, and B. G. Nordestgaard, "Does greater adiposity increase blood pressure and hyper-

tension risk? Mendelian randomization using the FTO/MC4R genotype," *Hypertension*, vol. 54, no. 1, pp. 84–90, 2009.

[133] H. S. Mumby, C. E. Elks, S. Li, S. J. Sharp, K.-T. Khaw, R. N. Luben, N. J. Wareham, R. J. Loos, and K. K. Ong, "Mendelian randomisation study of childhood BMI and early menarche," *Journal of obesity*, vol. 2011, 2011.

[134] P. Brennan, J. McKay, L. Moore, D. Zaridze, A. Mukeria, N. Szeszenia-Dabrowska, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, *et al.*, "Obesity and cancer: Mendelian randomization approach utilizing the FTO genotype," *International journal of epidemiology*, vol. 38, no. 4, pp. 971–975, 2009.

[135] M. Kivimäki, G. Davey Smith, N. J. Timpson, D. A. Lawlor, G. D. Batty, M. Kähönen, M. Juonala, T. Rönnemaa, J. S. Viikari, T. Lehtimäki, *et al.*, "Lifetime body mass index and later atherosclerosis risk in young adults: examining causal links using Mendelian randomization in the cardiovascular risk in Young Finns study," *European heart journal*, vol. 29, no. 20, pp. 2552–2560, 2008.

[136] P. Welsh, E. Polisecki, M. Robertson, S. Jahn, B. M. Buckley, A. J. de Craen, I. Ford, J. W. Jukema, P. W. Macfarlane, C. J. Packard, *et al.*, "Unraveling the directional link between adiposity and inflammation: a bidirectional Mendelian randomization approach," *The Journal of Clinical Endocrinology & Metabolism*, vol. 95, no. 1, pp. 93–99, 2010.

[137] J. A. Hubacek, O. Viklicky, D. Dlouha, S. Bloudickova, R. Kubinova, A. Peasey, H. Pikhart, V. Adamkova, I. Brabcova, E. Pokorna, *et al.*, "The FTO gene polymorphism is associated with end-stage renal disease: two large independent case–control studies in a general population," *Nephrology Dialysis Transplantation*, vol. 27, no. 3, pp. 1030–1035, 2012.

[138] R. M. Freathy, N. J. Timpson, D. A. Lawlor, A. Pouta, Y. Ben-Shlomo, A. Ruokonen, S. Ebrahim, B. Shields, E. Zeggini, M. N. Weedon, *et al.*, "Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI," *Diabetes*, vol. 57, no. 5, pp. 1419–1426, 2008.

[139] T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. Perry, K. S. Elliott, H. Lango, N. W. Rayner, *et al.*, "A common

variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity," *Science*, vol. 316, no. 5826, pp. 889–894, 2007.

[140] B. G. Nordestgaard, T. M. Palmer, M. Benn, J. Zacho, A. Tybjærg-Hansen, G. Davey Smith, and N. J. Timpson, "The effect of elevated body mass index on ischemic heart disease risk: causal estimates from a Mendelian randomisation approach," *PLoS medicine*, vol. 9, no. 5, p. e1001212, 2012.

[141] N. Timpson, B. Nordestgaard, R. Harbord, J. Zacho, T. Frayling, A. Tybjærg-Hansen, and G. Davey Smith, "C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization," *International Journal of Obesity*, vol. 35, no. 2, pp. 300–308, 2011.

[142] T. Lyngdoh, P. Vuistiner, P. Marques-Vidal, V. Rousson, G. Waeber, P. Vollenweider, and M. Bochud, "Serum uric acid and adiposity: deciphering causality using a bidirectional Mendelian randomization approach," *PloS one*, vol. 7, no. 6, p. e39321, 2012.

[143] T. M. Palmer, B. G. Nordestgaard, M. Benn, A. Tybjærg-Hansen, G. Davey Smith, D. A. Lawlor, N. J. Timpson, *et al.*, "Association of plasma uric acid with ischaemic heart disease and blood pressure: Mendelian randomisation analysis of two large cohorts," *BMJ*, vol. 347, p. f4262, 2013.

[144] K. S. Vimaleswaran, D. J. Berry, C. Lu, E. Tikkanen, S. Pilz, L. T. Hiraki, J. D. Cooper, Z. Dastani, R. Li, D. K. Houston, *et al.*, "Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts," *PLoS medicine*, vol. 10, no. 2, p. e1001383, 2013.

[145] A. Thakkinstian, L. Chailurkit, D. Warodomwichit, W. Ratanachaiwong, S. Yamwong, S. Chanprasertyothin, J. Attia, P. Sritara, and B. Ongphiphadhanakul, "Causal relationship between body mass index and fetuin-A level in the asian population: a bidirectional Mendelian randomization study," *Clinical endocrinology*, vol. 81, no. 2, pp. 197–203, 2014.

[146] J. R. Golding J, Pembrey M and the ALSPAC Study Team, "ALSPAC–the Avon Longitudinal Study of Parents and Children," *Paediatric and perinatal epidemiology*, vol. 15, no. 1, pp. 74–87, 2001.

[147] A. Fraser, C. Macdonald-Wallis, K. Tilling, A. Boyd, J. Golding, G. Davey Smith, J. Henderson, J. Macleod, L. Molloy, A. Ness, *et al.*, "Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort," *International journal of epidemiology*, vol. 42, no. 1, pp. 97–110, 2013.

[148] A. Boyd, J. Golding, J. Macleod, D. A. Lawlor, A. Fraser, J. Henderson, L. Molloy, A. Ness, S. Ring, and G. Davey Smith, "Cohort profile: the 'children of the 90s' – the index offspring of the Avon Longitudinal Study of Parents and Children," *International journal of epidemiology*, p. dys064, 2012.

[149] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," *Nature protocols*, vol. 5, no. 9, pp. 1564–1573, 2010.

[150] E. K. Speliotes, C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, A. U. Jackson, H. L. Allen, C. M. Lindgren, J. Luan, R. Mägi, *et al.*, "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index," *Nature genetics*, vol. 42, no. 11, pp. 937–948, 2010.

[151] S. Burgess and S. G. Thompson, "Use of allele scores as instrumental variables for Mendelian randomization," *International journal of epidemiology*, vol. 42, no. 4, pp. 1134–1144, 2013.

[152] J. Booth, P. Tomporowski, J. Boyle, A. Ness, C. Joinson, S. Leary, and J. Reilly, "Obesity impairs academic attainment in adolescence: findings from ALSPAC, a UK cohort," *International Journal of Obesity*, vol. 38, no. 10, pp. 1335–1342, 2014.

[153] R. J. Hancox, R. Poulton, J. M. Greene, C. R. McLachlan, M. S. Pearce, and M. R. Sears, "Associations between birth weight, early childhood weight gain and adult lung function," *Thorax*, vol. 64, no. 3, pp. 228–232, 2009.

[154] B. A. Dennison, T. A. Erb, and P. L. Jenkins, "Television viewing and television in bedroom associated with overweight risk among low-income preschool children," *Pediatrics*, vol. 109, no. 6, pp. 1028–1035, 2002.

[155] L. StataCorp, "Stata version 11.0," *College Station, TX: StataCorp LP*, 2009.

[156] J. H. Stock, J. H. Wright, and M. Yogo, "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business & Economic Statistics*, vol. 20, no. 4, 2002.

[157] P. S. Clarke and F. Windmeijer, "Identification of causal effects on binary outcomes using structural mean models," *Biostatistics*, vol. 11, no. 4, pp. 756–770, 2010.

[158] P. S. Clarke and F. Windmeijer, "Instrumental variable estimators for binary outcomes," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1638–1652, 2012.

[159] T. M. Palmer, D. A. Lawlor, R. M. Harbord, N. A. Sheehan, J. H. Tobias, N. J. Timpson, G. Davey Smith, and J. A. Sterne, "Using multiple genetic variants as instrumental variables for modifiable risk factors," *Statistical methods in medical research*, vol. 21, no. 3, pp. 223–242, 2012.

[160] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice," *Statistics in medicine*, vol. 30, no. 4, pp. 377–399, 2011.

[161] P. Royston, "Multiple imputation of missing values: update," *Stata Journal*, vol. 5, no. 2, p. 188, 2005.

[162] U. Sovio, D. O. Mook-Kanamori, N. M. Warrington, R. Lawrence, L. Briollais, C. N. Palmer, J. Cecil, J. K. Sandling, A.-C. Syvänen, M. Kaakinen, *et al.*, "Association between common variation at the FTO locus and changes in body mass index from infancy to late childhood: the complex nature of genetic association through growth and development," *PLoS genetics*, vol. 7, no. 2, p. e1001307, 2011.

[163] J. P. Ioannidis, "How to make more published research true," *PLoS medicine*, vol. 11, no. 10, p. e1001747, 2014.

[164] A. Heini, C. Lara-Castro, K. Kirk, R. Considine, J. Caro, and R. Weinsier, "Association of leptin and hunger-satiety ratings in obese women," *International journal of obesity*, vol. 22, no. 11, pp. 1084–1087, 1998.

[165] T. Fall, S. Hägg, R. Mägi, A. Ploner, K. Fischer, M. Horikoshi, A.-P. Sarin, G. Thorleifsson, C. Ladenvall, M. Kals, *et al.*, "The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis," *PLoS medicine*, vol. 10, no. 6, p. e1001474, 2013.

[166] M. V. Holmes, L. A. Lange, T. Palmer, M. B. Lanktree, K. E. North, B. Almoguera, S. Buxbaum, H. R. Chandrupatla, C. C. Elbers, Y. Guo, *et al.*, "Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis," *The American Journal of Human Genetics*, vol. 94, no. 2, pp. 198–208, 2014.

[167] P. Würtz, Q. Wang, A. J. Kangas, R. C. Richmond, J. Skarp, M. Tiainen, T. Tynkkynen, P. Soininen, A. S. Havulinna, M. Kaakinen, *et al.*, "Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change," *PLoS medicine*, vol. 11, no. 12, p. e1001765, 2014.

[168] K. K. Ong, P. Emmett, K. Northstone, J. Golding, I. Rogers, A. R. Ness, J. C. Wells, and D. B. Dunger, "Infancy weight gain predicts childhood body fat and age at menarche in girls," *The Journal of Clinical Endocrinology & Metabolism*, vol. 94, no. 5, pp. 1527–1532, 2009.

[169] L. D. Howe, K. Tilling, B. Galobardes, G. Davey Smith, A. R. Ness, and D. A. Lawlor, "Socioeconomic disparities in trajectories of adiposity across childhood," *International Journal of Pediatric Obesity*, vol. 6, no. 2Part2, pp. e144–e153, 2011.

[170] A. Matijasevich, C. G. Victora, J. Golding, F. C. Barros, A. M. Menezes, C. L. Araujo, and G. Davey Smith, "Socioeconomic position and overweight among adolescents: data from birth cohort studies in Brazil and the UK," *BMC Public Health*, vol. 9, no. 1, p. 105, 2009.

[171] A. R. Ness, S. Leary, J. Reilly, J. Wells, J. Tobias, E. Clark, and G. Davey Smith, "The social patterning of fat and lean mass in a contemporary cohort of children," *International Journal of Pediatric Obesity*, vol. 1, no. 1, pp. 59–61, 2006.

[172] J. P. Ioannidis, "Why most discovered true associations are inflated," *Epidemiology*, vol. 19, no. 5, pp. 640–648, 2008.

[173] K. J. Rothman, "No adjustments are needed for multiple comparisons.," *Epidemiology*, vol. 1, no. 1, pp. 43–46, 1990.

[174] Z. Fewell, G. Davey Smith, and J. A. Sterne, "The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study," *American Journal of Epidemiology*, vol. 166, no. 6, pp. 646–655, 2007.

[175] M. V. Holmes, F. W. Asselbergs, T. M. Palmer, F. Drenos, M. B. Lanktree, C. P. Nelson, C. E. Dale, S. Padmanabhan, C. Finan, D. I. Swerdlow, *et al.*, "Mendelian randomization of blood lipids for coronary heart disease," *European heart journal*, p. eht571, 2014.

[176] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, pp. 333–359, 2011.

[177] L. Hindorff, J. MacArthur, J. Morales, H. Junkins, P. Hall, A. Klemm, and T. Manolio, "A catalog of published genome-wide association studies." Available at: www.genome.gov/gwastudies.