Automating Risk of Bias Assessment for Clinical Trials

Iain J. Marshall
King's College London
iain.marshall@kcl.ac.uk

Joël Kuiper
University of Groningen
joel.kuiper@rug.nl

Byron C. Wallace
University of Texas at Austin
byron.wallace@gmail.com

ABSTRACT

In medicine, the publication of clinical trials now far outpaces clinicians' ability to read them. *Systematic reviews*, which aim to summarize the entirety of the available evidence on a specific clinical question, have therefore become the linchpin of evidence-based decision making. A key task in systematic reviews is determining whether the results of included studies may be affected by biases, e.g., poor randomization or blinding. This is called *risk of bias* assessment and is now standard practice. Standardized tools are used to perform these assessments; a notable example being the *Cochrane risk of bias tool*, which covers seven different types of potential biases and involves researchers extracting sentences from articles to support their bias assessments. These assessments are crucial in interpretating published evidence, but due to the exponential growth of the biomedical literature base, manually assessing the risk of bias in clinical trials has grown burdensome for clinical researchers. Aiming to mitigate this workload, we explore automating risk of bias assessment. We demonstrate that systematic reviews may be used to *distantly supervise* text mining models, obviating the need for manually annotated clinical trial reports. Specifically, we leverage data from *the Cochrane Database of Systematic Reviews* (a large repository of systematic reviews), and link clinical trial reports to structured data from the same studies found in CDSR to produce a pseudo-annotated labeled corpus. We then develop a joint model which, using (the PDF of) a clinical trial report as input, predicts the risks of bias in each of the aforementioned seven areas while simultaneously extracting the text fragments supporting these assessments.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining

1. INTRODUCTION AND MOTIVATION

Reports describing randomized clinical trials constitute the primary literature for evidence-based medicine (EBM). When correctly designed and executed, randomized control trials provide reliable evidence regarding the effects of medical interventions (e.g., drugs): proper randomization allows one to assess the causal effect of treatments. But flaws in trial design, conduct, analysis or reporting can result in over or underestimating treatment effects [5]. Assessing the quality of clinical trials (i.e., the risk of bias in the reported results) is therefore a critical step for researchers weighing the evidence.

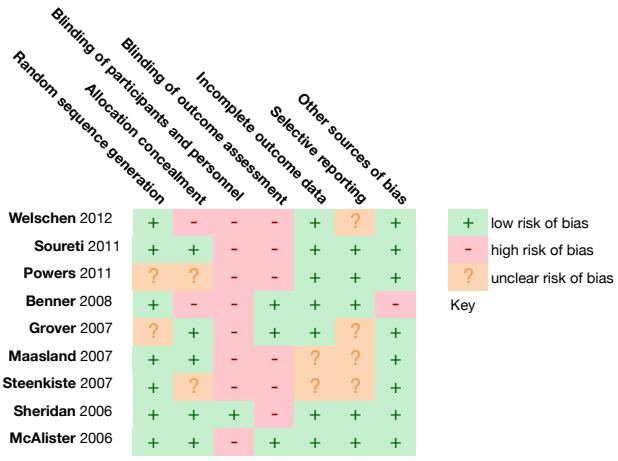

Figure 1: Illustrative risk of bias output from the Cochrane tool. Each row represents a single study. In this work, we aim to automate the generation of such tables (and to extract the sentences from articles supporting these judgements).

Assessing the risk of bias is especially important for those undertaking *systematic reviews*, which look to rigorously summarize the entirety of the evidence pertaining to a specific clinical question. Systematic reviews of clinical trials are the cornerstone of evidence-based medicine and are considered the strongest form of evidence, because they provide a summary effect estimate that incorporates all relevant published evidence.

Researchers performing such reviews must be careful to take into account potential biases in the literature. Failure to do so may result in inaccurate estimates of true treatment effects. In light of the role that systematic reviews play in shaping health policy guidelines and informing patient care,

such inaccuracies may ultimately be quite harmful. It is therefore imperative to carefully assess the risk of bias in clinical trial reports.

As the number of articles describing clinical trials continues to grow exponentially (in 2010, more than 75 clinical trials were published daily, and this number is increasing), the prospect of manually assessing risk of bias for every clinical trial publication becomes increasingly daunting (if not impossible). Already the generation of primary evidence is outpacing our ability to synthesize it given pragmatic resource constraints [1, 15]. If we are to keep systematic reviews and related evidence-based medicine products current, we need to optimize the steps involved in conducting EBM and evidence synthesis. Machine learning and data mining will play an important role modernizing the practice of EBM [12]. More generally, the overwhelming volume of published clinical literature requires the development of new data mining methods that can automatically process, analyze and otherwise make sense of clinical trial reports.

In this vein, we present a novel data mining method that automatically assesses the risk of bias across several quality domains using the full-text of published articles while simulteanously extracting sentences that support these judgements. Our contributions in this work are summarized as follows:

- So far as we are aware, this is the first attempt to automatically infer the risk of bias across clinically important dimensions (see Figure 1). Automating this quality assessment with reasonable fidelity may help with myriad evidence-based medicine applications.
- We demonstrate that systematic reviews may be used to distantly supervise [7] the training of text mining models, thus avoiding the need for expensive manually annotated data.
- We present a novel method for jointly judging the risk of bias associated with a given article *and* extracting the sentence that supports this judgement. This is in keeping with how humans perform risk of bias assessment. We demonstrate that this approach improves performance with respect to risk of bias predictions from free-text.

We show that automated risk of bias assesment is both feasible and we show (qualitatively) that it is potentially useful for systematic reviewers. Indeed, such an approach could have a major impact in the way reviews are conducted. More generally, the proposed method has the potential to facilitate rapid discovery of high-quality biomedical literature.

The remainder of this paper is structured as follows. In the next section we motivate this work by discussing the domain of evidence-based medicine. In Section 2.3, we then introduce a novel *distantly supervised* [7] method to construct a large corpus of pseudo-annotated articles describing clinical trials: for each article we find a corresponding risk-of-bias assessment across several domains and the specific sentences that support these assessments. It is this annotation – currently performed by clinical researchers at substantial cost – that we look to automate (using only the free-text of articles). In Section 3 we introduce our approach to doing so. Our method takes into account both document and sentence level information jointly. In Section 5, we evaluate this approach for both document and sentence level risk of bias assesments. In Section 6 we discuss methods for optimizing both the performance and usability of the proposed method, and how tools such as the one we introduce here can influence the way in which systematic reviews are conducted.

2. BACKGROUND

2.1 Systematic reviews

Randomized controlled trials (RCTs) provide the most reliable assessments of the efficacy and safety of medical interventions. As such, they should inform treatment decisions [3]: this is the idea behind evidence-based care. But achieving this aim is complicated by the massive numbers of trials that are conducted: for example, the Cochrane Library [10] alone indexes 286,418 trials as having been conducted in the last decade [13]. To practice in this environment of information overload, decision-makers often rely on systematic reviews to summarize the clinical trials that address a specific intervention for a particular disease (and for a particular population).

Systematic reviewing consists of specifying an inclusion criteria (i.e., the criteria studies must satisfy to be included in the review), searching the literature, screening the retrieved (potentially eligible) citations to identify eligible studies, extracting data from these eligible studies and, finally, summarizing the relevant evidence. The process of producing systematic reviews is thus extremely laborious, but of the utmost importance in terms of enabling informed clinical decisions. In the final step of statistical evidence synthesis, it is vital to take into account the perceived quality of the identified studies to be summarized. One would not want, for example, a poorly run study to greatly affect the summary results of a systematic review. Practically speaking, this means that one needs to carefully assess the potential risks of bias.

2.2 The Cochrane Database of Systematic Reviews and Risk of Bias Tool

The *Cochrane Collaboration* is a global network of researchers who work together to produce systematic reviews. At present, the group comprises over 30,000 researchers (mostly physicians and other health practitioners) who have produced upwards of 5,838 systematic reviews[1], collectively published as *the Cochrane Database of Systematic Reviews* (CDSR)[10]. This database contains structured data manually extracted from the papers describing the included trials.

The Cochrane Collaboration recently developed a tool for assesing bias in clinical trials, which has been adopted across all Cochrane systematic reviews since 2008 [5]. This tool aimed to unify and improve the myriad tools that were previously used. The new system comprises seven domains by default (see Table 1), but domains may be added or removed by authors based on the needs of their specific review.

[1] `http://www.cochrane.org/cochrane-reviews/cochrane-database-systematic-reviews-numbers`

Bias	Allocation concealment
Authors judgement	Low risk
Support for judgement	Quote: "The Family Practice Research Coordinator at the University of British Columbia held this sequence independently and remotely"

Figure 2: Review authors' justification for their score of an example study in domain 2, *Allocation Concealment*; we retrieved the highlighted part using a regular expression

Review authors judge the risk of bias in each domain as *high*, *low*, or *unknown* (see Figure 1). Most domains are assessed *per study* included in the review. However, some domains are assessed more than one time per study, i.e., once per *per outcome*. For example, it is regarded as good practice for a clinical trial to be *double-blinded* – that is, neither the participant nor the investigator should be aware of which of the treatments are being given to whom. In this paper we focus on the first six domains, which are used consistently across reviews; the seventh domain covers "other" risks, and therefore varies greatly according to the needs of individual studies.

The Cochrane tool uniquely does not score all studies which are not blinded as having a *high* risk of bias, but instead considers whether blinding is likely to have had an effect on the specific outcome being assessed. Measuring subjective outcomes in unblinded studies, such as participant satisfaction or pain, is more susceptible to bias than measuring more objective outcomes such as prolongment of life. Three of the seven domains are recommended to be scored *per outcome* for this reason. However, in practice review authors often give a single score for all outcomes, and this is the practice used in the majority of currently published reviews.

For many assessments, Cochrane reviewers justify their risk of bias assessments by quoting supporting text directly from the original study (see Figure 2). This is desirable because it increases the transparency of the judgements.

2.3 Data

Here we leverage descriptions of and data about clinical trials manually extracted for existing Cochrane systematic reviews. We use this structured data (which we glean from the CDSR) as a substitute for manual annotations. In this sense the strategy we take here is *distantly supervised* [7, 8].

2.3.1 Data structure of Cochrane reviews

The CDSR contains structured and semi-structured data for the individual studies comprising each systematic review. Internally, the Cochrane Collaboration stores working versions of their reviews as XML. Each review contains a wealth of (structured) data about the included clinical trials (i.e., those that met review inclusion criteria). There are usually multiple clinical trials described in a single review. Cochrane reviews use basic clinical trial identifiers which are unique per review (based on the first author surname and year of publication) throughout these files. It is therefore possible to extract structured data, and filtered snippets of full text

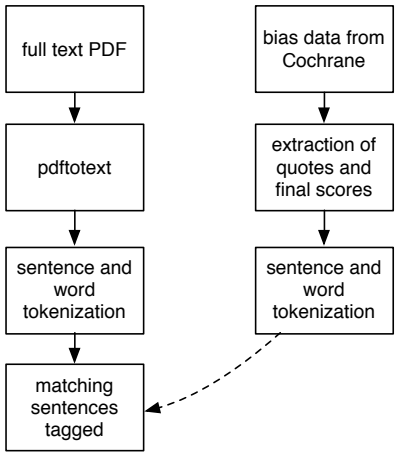

Figure 3: Corpus construction processing

data which describe a single clinical trial. Using these identifiers, we were able to obtain full structured citation data for the primary reference of all included studies across the entire CDSR.

2.3.2 Linking to full text studies

To facilitate retrieving the original trial reports, we linked the trials to PubMed, a popular portal to biomedical study citations. To cope with transcription errors by the Cochrane review authors, we used non-overlapping combinations of the citation elements to form multiple search queries. Each of the queries might be expected to uniquely retrieve the target paper; we assumed an accurate match in cases where two or more queries retrieved the same article. Using this strategy, we linked the semi-structured descriptions of 52,454 clinical trials from Cochrane reviews to their unique PubMed ID.

2.3.3 Justification for risk of bias decisions

The risk of bias classification (*high*, *low*, or *unknown*) is structured and retrievable per clinical trial for individual domains. The risk of bias tool allows much flexibility: review authors may remove core domains or add new domains depending on the needs of their review. For this reason, we restricted our task to the core default domains which have wide uptake.

The risk of bias tool requires review authors to record an explanation for each risk of bias judgement. This explanation is recorded as unstructured text, but is retrievable per study. It is permissible to use a quote from the original trial report to justify a decision, and many review authors have informally adopted a standardized way of recording this (see Figure 2). We exploited this convention by searching for the pattern throughout the CDSR using a regular expression. We identified quotes in one or more domains for a total of 3,529 clinical trials. For 2,200 of these trials, we were able to obtain full text original reports in PDF format. These PDFs linked with the structured and unstructured descriptions of the same trials from the CDSR formed our corpus.

Domain title	Explanation	Level of score
Random sequence generation	Was the method of randomisation scientifically valid	per study
Allocation concealment	Are researchers able to influence which groups participants are allocated to	per study
Blinding of participants and personnel	Were participants treatment groups concealed from them and study personnel	per outcome
Blinding of outcome assessment	Was the person assessing outcomes blinded to the participants' treatment group	per outcome
Incomplete outcome data	Might an imbalance in study withdrawals or dropouts lead to a bias in results	per outcome
Selective reporting	Have any outcomes studied not been published (usually by comparison with a protocol)	per study
Other sources of bias		per study

Table 1: Possible sources of bias assessed by the Risk of Bias tool

2.3.4 *Aligning Cochrane data with original trial reports*

PDFs of clinical trial reports were converted to plain text using the `pdftotext` utility from Xpdf.[2] We retrieved individual quotes from the Cochrane database, and sought the longest matching substring in the clinical trial report. For the sentence identification task, the clinical trial reports were word and sentence tokenized; sentences that overlapped a quote were labelled positively; all others negatively (see Figure 3). For the document classification task, we labelled each full text trial report as being at *high*, *low*, or *unknown* risk of bias using the classification from the linked review (i.e., these labels are explicitly available in the CDSR).

3. METHODS

In this section we first introduce the preliminary machinery that constitutes the baseline approaches we consider for the tasks of risk of bias assessment and supporting sentence extraction. We then introduce a model that jointly leverages both document level risk of bias assessments and the associated supporting quotes. The intuition here is that the identified sentences will inform the document level preductions and thus result in improved predictive performance.

3.1 Overall Risk of Bias Prediction

We first consider the task of predicting the overall (study-level) risk of bias from the full-text of articles. As an initial approach, we treat this as a standard binary classification task, where the output space $\mathcal{Y}$ comprises *low risk* and *unknown/high risk*. This dichotomization of the task is practical in that one typically wants to know whether or not an article exhibits some sort of bias. From this vantage, the distinction between *unknown* and *high risk* is unimportant; both would require further assessment by the domain expert. We note that a model that could reliably identify studies with low risk of bias across domains would be useful, e.g., in helping researchers rapidly discover high quality literature.

We leverage the soft-margin Support Vector Machine (SVM) [14] as our classification model. We will denote each instance (article) by $\mathbf{x}_i$, its label for quality domain $q \in \mathcal{Q}$ (where $\mathcal{Q}$ is the set of quality domains enumerated in Table 1) by y_i^q and a feature extracting function by ϕ. For the latter we use standard bag-of-words (BoW) text encoding. To map the problem into a binary task, we define a function $\mathcal{F}$ as follows:

$$\mathcal{F}(y_i^q) = \begin{cases} 1 & \text{if } y_i^q = \textit{low risk of bias} \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

Then, for each quality domain q we find a minimizing weight vector $\mathbf{w}_d^q$ (the d here is to distinguish this vector from those introduced for the sentence extraction task, below). We assume risk of bias labels assume the form:

$$y_i^q = \text{sign}\{\mathbf{w}_d^q \phi(\mathbf{x}_i)\} \quad (2)$$

And we find each $\mathbf{w}_d^q$ by solving the following objective:

$$\underset{\mathbf{w}_d^q}{\text{argmin}}\, \alpha \|\mathbf{w}_d^q\|^2 + \sum_{i=1}^{n^q} \mathcal{L}(\text{sign}\{\mathbf{w}_d^q \phi(\mathbf{x}_i)\}, \mathcal{F}(y_i^q)) \quad (3)$$

Where n^q denotes the number of labeled instances for the domain q and $\mathcal{L}$ is the usual hinge-loss function. The α parameter controls the degree of regularization: we tune this via grid-search over training data, maximizing for F1 score.

3.2 Sentence Prediction

We have just described an initial approach to overall (document level) risk of bias prediction. We now review our baseline approach to automatically extracting *sentences* supporting these quality judgements. A similar approach as for the overlall risk of bias prediction is taken, though here labels indicate whether or not a given sentence was selected by a domain expert as supporting her judgement regarding risk of bias for a specific quality domain. Denoting sentence j in document i by s_{ij} and its associated label (for target domain q) by l_{ij}^q, we posit the classification model:

$$l_{ij}^q = \text{sign}\{\mathbf{w}_s^q \phi(\mathbf{s}_{ij})\} \quad (4)$$

[2] `http://www.foolabs.com/xpdf/`

And we estimate the associated sentence extraction parameters $\mathbf{w}_s^q$ by optimizing the following (again for each domain q):

$$\underset{\mathbf{w}_s^q}{\operatorname{argmin}}\, \alpha\|\mathbf{w}_s^q\|^2 + \sum_{i=1}^{n^q}\sum_{j=1}^{m_i} \mathcal{L}(\text{sign}\{\mathbf{w}_s^q\phi(\mathbf{s}_{ij})\}, l_{ij}^q)) \quad (5)$$

where the notation is similar to above (Equation 3) with the addition of m_i, which we use to denote the number of sentences in document i. Note that we use the same feature extraction function ϕ as we did for the full-text predictions (here this extracts binary bag of words features).

4. A JOINT RISK OF BIAS AND SUPPORTING SENTENCE EXTRACTION MODEL

We now introduce a novel model that integrates the sentence extraction task with document level risk of bias prediction. A joint model is preferable to completely independent models for classification and extraction because the overall risk of bias assessment ought to inform the supporting sentence extraction. Intuitively, for example, if the (study-level) risk of bias due to poor random sequence generation is deemed to be *low*, then we would expect the supporting sentence to contain words such as *computer* and *generated* (Table 2).

4.1 Informing Overall Risk of Bias Prediction with Supporting Sentences

To realize a joint model, we introduce terms into the document level risk of bias prediction that interact n-gram indicator features with supporting sentence predictions. We will again denote the binary prediction regarding whether or not sentence j in article i (sentence s_{ij}) supports the risk of bias judgement for domain q by l_{ij}^q (we assume this is 0 or 1) and we will denote the corresponding predictions by $\hat{l}_{ij}^q$. Further, we denote the supporting sentence for domain q in document i by s_{i*}^q.

We then augment the baseline risk of bias model (Equation 3) as follows:

$$y_i^q = \text{sign}\{\mathbf{w}_d^q\phi(\mathbf{x}_i) + \mathbf{w}_{ds}^q\lambda_d(s_{i*}^q)\} \quad (6)$$

Here λ_d is a feature extraction function for supporting sentences: this can be viewed as adding terms that indicate tokens (unigrams) being present *in a supporting sentence* within a document. Put another way, these are interaction terms that cross bag-of-words features with their prescence in judgement-supporting sentences. We use $\mathbf{w}_{ds}^q$ to denote the (document-level) weight vector associated with the sentence interaction features for domain q. During training we minimize over $\mathbf{w}_d' = \mathbf{w}_d^q + \mathbf{w}_{ds}^q$ (here + denotes vector concatenation).

For unlabeled documents (i.e., at test time), we will of course not know which sentence supports quality assessment (i.e., which is s_{i*}^q). Instead, we rely on *predicted* sentence labels, $\hat{l}_{ij}^q$. In particular, for each quality domain q we predict for each sentence j in article i whether or not it supports the judgement for said domain. If the prediction is that it does, we add interaction terms accordingly. Note that at test time, we may therefore add interaction features from multiple sentences that are predicted as supporting quality assessment in a given article (because these predictions are made independently). We can write the whole predictive model out as follows:

$$\begin{aligned} y_i^q = \text{sign}\{\mathbf{w}_d^q\phi(\mathbf{x}_i) + \hat{l}_{i0}^q\mathbf{w}_{ds}^q\lambda_d(s_{i0}^q) + \\ ... + \hat{l}_{im_i}^q\mathbf{w}_{ds}^q\lambda_d(s_{im_i}^q)\} \end{aligned} \quad (7)$$

Where the $\hat{l}_{ij}^q$ are predictions made via Equation 4.

4.2 Sentence Extraction

We also consider a model that informs supporting sentence extraction with document level risk of bias information. In particular, we add terms to Equation 4 that interact sentence level features with the predicted article-level risk of bias assessments for the corresponding document. More specificially, we augment the representation of each sentence j comprising document i with terms that interact the document-level risk of bias assessment with sentence-level features. Formally, abbreviating low risk and high risk by lr and hr, respectively, we assume the following model:

$$\begin{aligned} l_{ij} = \text{sign}\{\mathbf{w}_s^q\phi(\mathbf{s}_{ij}) + \mathcal{I}_{lr}(y_i^q)\mathbf{w}_{s;lr}^q\lambda_{s;lr}(\mathbf{s}_{ij}) \\ +\mathcal{I}_{hr}(y_i^q)\mathbf{w}_{s;hr}^q\lambda_{s;hr}(\mathbf{s}_{ij})\} \end{aligned} \quad (8)$$

Where $\mathcal{I}_{lr}$ ($\mathcal{I}_{hr}$) is an indicator function that is 1 when the argument is low risk (high risk) of bias and 0 otherwise. Furthermore, we have introduced the sentence-level interaction feature extraction function λ_s, analogous to λ_y above, except that here we interact sentence feature indicators with the binary *low risk* (lr) and *high risk* (hr) document labels. We denote the weights associated with these features by $\mathbf{w}_{s;lr}^q$ $\mathbf{w}_{s;hr}^q$, respectively. We also introduce feature extraction functions that are parameterized by the document level labels ($\lambda_{s;lr}$ and $\lambda_{s;hr}$). These are convienence functions that generate unique 'interaction copies' of bag-of-words features for tokens in low and high risk sentences.

5. EMPIRICAL RESULTS

We matched the full-texts of 2,200 clinical trial reports to semi-structured descriptions of the same trials in the CDSR. We first consider the task of identifying studies with *low* risk of bias (or other). We show five-fold cross-validation results for this task in Tables 2 and 3, and Figure 5. We report precision, recall and F1 with respect to *low risk of bias* (or not). Precision is the fraction of studies classified as *low risk* that indeed were (as per the Cochrane reviewer's decision); recall is the total fraction of *low risk* studies correctly identified as such and F1 is the harmonic mean of these metrics.

As can be seen in Figure 5, the proposed joint model improved the predictions across all domains. And as can be seen in Table 3, interaction features comprised the majority

Domain	F1	precision	recall	most informative features
Random sequence generation	0.70 (0.64, 0.79)	0.67 (0.51, 0.82)	0.79 (0.52, 0.93)	computer, generated, random, randomization
Allocation concealment	0.68 (0.65, 0.72)	0.66 (0.60, 0.71)	0.72 (0.57, 0.82)	sealed, generated, envelopes, randomization
Blinding of participants and personnel	0.57 (0.38, 0.69)	0.66 (0.62, 0.69)	0.53 (0.26, 0.78)	blind, placebo, double, influence, summary
Blinding of outcome assessment	0.62 (0.54, 0.67)	0.52 (0.46, 0.56)	0.81 (0.69, 1.00)	blinded, secondary, nd, session, responsible
Incomplete outcome data	0.75 (0.73, 0.77)	0.63 (0.61, 0.70)	0.93 (0.82, 0.99)	immediately, aimed, id, compare, intravenous
Selective reporting	0.69 (0.57, 0.78)	0.62 (0.59, 0.71)	0.82 (0.48, 0.98)	march, finding, maintenance, institute, july

Table 2: Document classification results: baseline model (Section 3.1) performance. Shown are averages over five-fold cross-validation (and ranges). We also include the four most informative features according the model for illustrative purposes.

of the top-ranking (most informative) features. Thus the proposed strategy of incorporating features extracted from sentences deemed likely to support risk of bias assessments improves classification performance.

We define task 2 as identifying sentences reporting information about bias. We present results for this problem in Table 4. Unfortunately the proposed joint models (which included actual or predicted information from the document level) did not improve performance for this task. We note, however, that when the true (rather than predicted) document-level labels are used, we do see a consistent improvement in precision (though at a modest expense in recall). Furthermore, a caveat to these results is that the sentence labels we are using for evaluation are noisy (in contrast to the document level risk of bias labels); we discuss this point further below.

Randomisation method was computer generated and was not blinded.

(algorithm predicted sentence)

…women with early breast cancer (pT1-3a pN0-1 M0) at 17 centres in the UK were randomly assigned after primary surgery to receive…

(quote from the Cochrane review)

Figure 5: Example of where the algorithm (the hybrid model using the *predicted* document level label) has picked a better sentence justifying a risk of bias decision for *random sequence generation* than the quote in the training data.

6. DISCUSSION

In this paper we have demonstrated that systematic reviews may be used to distantly supervise the training of biomedical text extraction systems, thus obviating the need for expensive manual annotation. In particular we have demonstrated the feasibility of this approach for training models to perform risk of bias assessment for articles describing clinical trials. We have also described a joint model for this task that simultaneously identifies the text fragments justifying the assessment. We demonstrated that this novel approach improves document-level risk of bias assessment performance. We note that the Cochrane risk of bias tool requires authors to transparently describe the reasons for their decisions. An automated tool would therefore have to justify its decisions. The method presented here has the advantage of being able to provide the sentence from the trial report which led to the classification.

Assessing the risk of bias in a study is inherently subjective. Indeed, a validation study of the Cochrane risk of bias tool found wide variations in judgements by different researchers in all domains, with the *selective reporting* domain showing the least agreement (κ=0.13, 95% CI -0.05 to 0.31) [4]. The instructions for the risk of bias tool indicate that 'convincing text' from the original clinical trial reports is uncommon, and recommends consulting the trial protocol where possible. Our model was not able to predict sentences with any useful accuracy in this domain, though we do not think is surprising given the difficulty (as evidenced by the poor agreement between domain experts).

Concerning the sentence identification task, we used quotations from Cochrane as (distantly supervised) training and test data. But we note that when assessing the risk of bias, authors select what they deem to be the single best sentence as evidence. This means other, equally relevant supporting sentences, may not be marked by experts as such, thus resulting in false negatives. Ideally the test data would identify *all* relevant sentences as evidence. Indeed, Figure 5 shows that the proposed model sometimes produces arguably *better* (more pertinent) quotes than the ones actually reported in Cochrane, but here such quotes (sentence predictions) would still be counted as false positives, because of the limitations due to our distantly supervised approach. This implies that the results reported here are pessimistic for this task; it may also account for why the proposed joint model fails to improve performance for this task versus the baseline model. We plan to assemble a gold standard test corpus comprising a modest number of studies for which all acceptable supporting sentences for each quality domain have been recorded; this would provide a more accurate evaluation.

Similarly, implementing the system within an integrated screening pipeline would reduce the importance of low precision, as the predictions would serve as a reading guide to assist domain experts (reviewers), rather than as a replacement for their judgements altogether. We are currently working

Domain	F1	precision	recall	most informative features
Random sequence generation	0.72 (0.67, 0.80)	0.69 (0.52, 0.83)	0.78 (0.63, 0.94)	computer-i, computer, generated-i, random-i
Allocation concealment	0.70 (0.68, 0.75)	0.67 (0.55, 0.79)	0.77 (0.59, 0.88)	by-i, the-i, was-i, and-i, sealed, calculated
Blinding of participants and personnel	0.66 (0.59, 0.71)	0.65 (0.60, 0.73)	0.70 (0.50, 0.84)	blind, double, placebo, placebo-i, double-i, blind-i
Blinding of outcome assessment	0.67 (0.63, 0.69)	0.53 (0.46, 0.57)	0.92 (0.85, 1.00)	established, were-i, single, generated, blinded
Incomplete outcome data	0.76 (0.74, 0.79)	0.64 (0.61, 0.71)	0.94 (0.89, 1.00)	aimed, described, needed, wong, model, second
Selective reporting	0.72 (0.70, 0.78)	0.63 (0.59, 0.71)	0.87 (0.71, 0.98)	oral, issue, unrelated, march, maintenance

Table 3: Document classification results: joint model (Section 4.1) performance. -i represents 'interaction' features (where the word occured in a sentence predicted as supporting a quality assessment). Note the frequency of the interaction features amongst the more informative tokens (suggesting that these are indeed useful features).

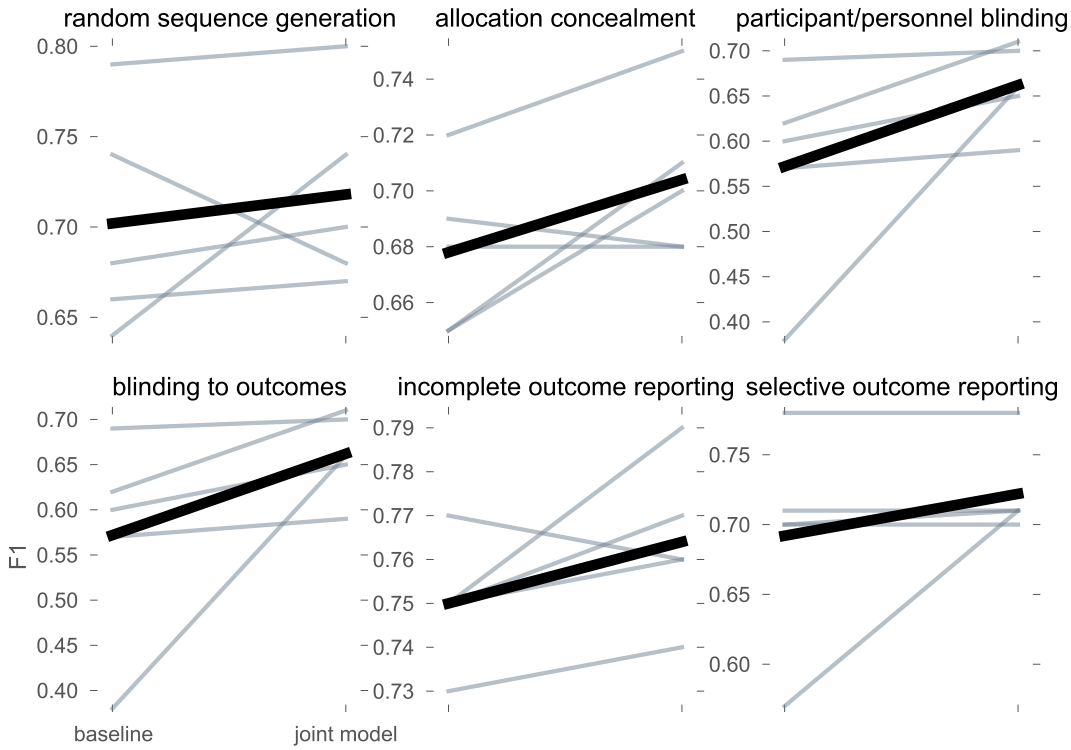

Figure 4: Results from five-fold cross-validation across the 6 domains. The y-axis is F1 score. Lines connect results achieved on the same folds; the thick black lines are means (the grey lines correspond to individual fold results). The proposed joint model consistently outperforms the baseline approach.

Domain	Baseline			Joint (true labels)			Joint (predicted labels)		
	F1	precision	recall	F1	precision	recall	F1	precision	recall
Random sequence generation	0.53	0.43	0.68	0.56	0.49	0.67	0.48	0.37	0.66
Allocation concealment	0.48	0.42	0.58	0.48	0.43	0.56	0.44	0.37	0.57
Blinding of participants and personnel	0.37	0.30	0.50	0.35	0.49	0.35	0.32	0.25	0.44
Blinding of outcome assessment	0.38	0.34	0.42	0.38	0.37	0.41	0.38	0.37	0.40
Incomplete outcome data	0.23	0.16	0.44	0.23	0.17	0.38	0.24	0.17	0.39
Selective reporting	0.06	0.11	0.04	0.06	0.05	0.07	0.03	0.02	0.06

Table 4: Results for the sentence identification task. The joint model does boost precision (at some cost in recall) when the *true* document-level labels are used, but the model does not seem to improve performance when the predicted labels are used in place.

on such a tool that leverages the model presented here.

To improve the performance of our system we plan to explore related methods developed for sentiment analysis [6, 9], as the task is conceptually similar. We are particularly interested in exploring probabalistic models that aim to jointly model sentiment and text fragments [11, 2], although these would need to be adapted to the supervised case. It might also be possible to leverage more of the existing background knowledge present in Cochrane.

Finally, we note that in future work we aim to extend these strategies to extract other variables of interest from clinical trial reports. Specifically, the CDSR contains structured and semi-structured information on trial populations, interventions, outcomes, and results data. Tools to automate these tasks could lead to a large reduction in the time required to produce systematic reivews.

Going forward, it will be crucial to assess the practical utility of automated extraction systems. In particular, we envision a hybrid computer-human system in which machine learning models guide the extraction process (thereby reducing manual labor). Another potential use of automated methods would be to use the computer generated extractions for redundancy to improve data quality; i.e., rather than having two experts independently perform 'double-extraction', we might substitute the computer for one of them. With these concrete applications of automated methods in mind, we then need to assess the level of predictive accuracy that models need to achieve in order to be useful in practice.

Clinicians and health sciences researchers are overwhelmed with data. If we are to maintain the rigor and comprehensiveness of evidence-based medicine products, new data mining methods are sorely needed to mitigate problems of information overload. This work is a step toward such larger aims.

7. REFERENCES

[1] H. Bastian, P. Glasziou, and I. Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9), 2010.

[2] S. Brody and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics, 2010.

[3] Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. 268(17):2420–2425, 1992.

[4] L. Hartling, M. Ospina, and Y. Liang. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ*, 339:b4012, 2009.

[5] J. Higgins, D. Altman, P. Gotzsche, P. Juni, D. Moher, A. Oxman, J. Savovic, K. Schulz, L. Weeks, and J. Sterne. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343:d5928, Oct. 2011.

[6] B. Liu and L. Zhang. A survey of Opinion Mining and Sentiment Analysis. In *Mining Text Data*, pages 415–463. 2012.

[7] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[8] T. Nguyen and A. Moschitti. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 277–282. Association for Computational Linguistics, 2011.

[9] B. Pang, B. Pang, L. Lee, and L. Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 2008.

[10] The Cochrane Collaboration. The Cochrane Database of Systematic Reviews, Jan. 2014.

[11] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Association for Computational Linguistics (ACL)*, volume 51, page 61801, 2008.

[12] G. Tsafnat, A. Dunn, P. Glasziou, and E. Coiera. The automation of systematic reviews. *BMJ: British Medical Journal*, 346, 2013.

[13] G. Valkenhoef, T. Tervonen, B. Brock, and H. Hillege. Deficiencies in the transfer and availability of clinical trials evidence: a review of existing systems and standards. *BMC Medical Informatics and Decision Making*, 12(1):95, 2012.

[14] V. Vapnik. The nature of statistical learning theory. *Data Mining and Knowledge Discovery*, pages 1–47, 6.

[15] B. Wallace, I. Dahabreh, C. Schmid, J. Lau, and T. Trikalinos. Modernizing the systematic review process to inform comparative effectiveness: tools and methods. *Journal of Comparative Effectiveness Research*, 2(3):273–282, 2013.