Automating Risk of Bias Assessment for Clinical Trials

Iain J Marshall, Joël Kuiper, and Byron C Wallace

***Abstract*— *Systematic reviews*, which summarize the entirety of the evidence pertaining to a specific clinical question, have become critical for evidence-based decision making in healthcare. But such reviews have become increasingly onerous to produce due to the exponentially expanding biomedical literature base. This study proposes a step toward mitigating this problem by automating *risk of bias* assessment in systematic reviews, in which reviewers determine whether study results may be affected by biases (e.g., poor randomization or blinding). Conducting risk of bias assessment is an important but onerous task. We thus describe a machine learning approach to automate this assessment, using the standard *Cochrane Risk of Bias Tool* which assesses seven common types of bias. Training such a system would typically require a large labeled corpus, which would be prohibitively expensive to collect here. Instead, we use *distant supervision*, using data from *the Cochrane Database of Systematic Reviews* (a large repository of systematic reviews), to pseudoannotate a corpus of 2200 clinical trial reports in PDF format. We then develop a joint model which, using the full text of a clinical trial report as input, predicts the risks of bias while simultaneously extracting the text fragments supporting these assessments. This study represents a step toward automating or semiautomating extraction of data necessary for the synthesis of clinical trials.**

***Index Terms*—Evidence-based medicine, health informatics, machine learning, natural language processing.**

I. Introduction and Motivation

RANDOMIZED-controlled trials (RCTs) constitute the primary literature for evidence-based medicine (EBM). Systematic reviews of RCTs are considered the strongest form of evidence because they aim to provide an unbiased view that incorporates all relevant identified evidence [1]. Flaws in trial design, conduct, analysis or reporting in the individual studies comprising a systematic review can result in bias, thus resulting in treatment effects being over or underestimated [2], [3]. For example, *double-blinding*—where neither the participant nor the investigator are aware of which of the treatments are being administered—has been shown to reduce bias in trial results [4]. Assessing the risks of important biases in RCTs is therefore a critical step in interpreting and synthesizing trial reports.

Such bias assessments inform the analyses conducted in the systematic review. For example, trials judged to be at high risk of bias may be withheld in *sensitivity analyses*, allowing one to judge treatment efficacy from only the most robust evidence. Since a single systematic review may contain dozens of RCTs, and risk of bias assessment requires reading entire articles, performing such assessments is extremely time-consuming. Indeed, the time taken to conduct risk of bias assessments has been identified as a key factor preventing systematic reviews from being kept up-to-date [5].

As the number of articles describing clinical trials continues to grow exponentially (in 2010, more than 75 clinical trials were published daily, on average; the Cochrane Library [6] alone indexes 286 418 trials as having been conducted in the last decade [7]), the prospect of manually assessing the risk of bias for every publication becomes increasingly daunting. And the time required to complete each review means that they are often out-of-date [8]. Already the generation of primary evidence is outpacing our ability to synthesize it given pragmatic resource constraints [9], [10]. 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 [11], [12].

In this paper (a version of which was originally presented at ACM-BCB 2014 [13]), we present our automated system for determining risk of bias from clinical trials, describe the potential for clinical applications of this technology, and outline further developments needed to reach this goal. Our novel contributions in this study are summarized as follows:

1) We describe a machine learning approach to automatically judge the risk of bias across clinically important areas (see Fig. 1). Automating this quality assessment with reasonable fidelity may help with myriad EBM applications.
2) We demonstrate that existing systematic reviews may be used to *distantly supervise* [14] the annotation of a corpus of clinical trial reports, thus obviating the need for expensive manually annotated data.
3) We present a novel method for jointly judging the risk of bias associated with a given article *and* extracting the sentence that supports this judgment. This is in keeping with how humans perform risk of bias assessment. We demonstrate that this approach improves automated risk of bias assessment.

II. Related Work

Here we aim to facilitate semiautomated information extraction and summarization from articles describing clinical trials. It has previously been recognized that machine learning tools can assist abstract screening [15]–[18], data extraction [19]–[23], summarization [24], [25], and scoping [26].

Manuscript received December 14, 2014; revised April 17, 2015; accepted April 27, 2015. Date of publication May 8, 2015; date of current version July 23, 2015.

I. J Marshall is with the Department of Primary Care and Public Health Sciences, King's College London, London WC2R 2LS, U.K. (e-mail: iain.marshall@kcl.ac.uk).

J. Kuiper is with the Vortext Systems (e-mail: me@joelkuiper.eu).

B. C Wallace is with the School of Information, University of Texas at Austin, Austin, TX 78712 USA (e-mail: byron.wallace@utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2015.2431314

TABLE I
POSSIBLE SOURCES OF BIAS ASSESSED BY THE RISK OF BIAS TOOL

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

To our knowledge, however, there have been fewer efforts to automate data extraction from articles describing clinical trials compared to, e.g., work on methods to mine cancer-related and genetic literature [27], [28]. A particularly relevant example of the latter is due to Ling *et al.* [29], [30], in which they developed automated methods for generating gene summaries from biomedical literature. Such summaries may be viewed as *semistructured*, as they comprise free-text entries corresponding to several semantic aspects of interest. This is similar to the present effort, in which we aim to extract sentences that support judgements concerning the risks of bias across several domains. To train their gene summarization model, they also exploited existing resources (as we do here). Specifically, they generated training data from existing structured gene summaries in the "FlyBase" database, thus providing plentiful, if noisy, training data.

In a similar spirit to Ling and colleagues, we leverage previously curated data to provide indirect supervision to train our models, thus obviating the need for expensive manual supervision. In contrast, however, we explicitly have distant supervision for each domain (or aspect) of interest, whereas in their case they had to infer the text relevant to each facet [30]. We note that there is a wealth of work on models for the general task of information extraction from biomedical texts [31], but we do not attempt to survey them exhaustively here.

III. DATA USED FOR DISTANT SUPERVISION

A. Cochrane Database of Systematic Reviews (CDSR) 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 5800 systematic reviews[1], collectively published as *the*CDSR [6]. This database contains structured data manually extracted from the papers describing the included trials.

The Cochrane Collaboration has developed a tool for assessing bias in clinical trials. The tool has been adopted across all Cochrane systematic reviews since 2008 [2]. Additionally, the tool is now widely used outside of Cochrane [32]. The tool comprises seven domains by default (see Table I), 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

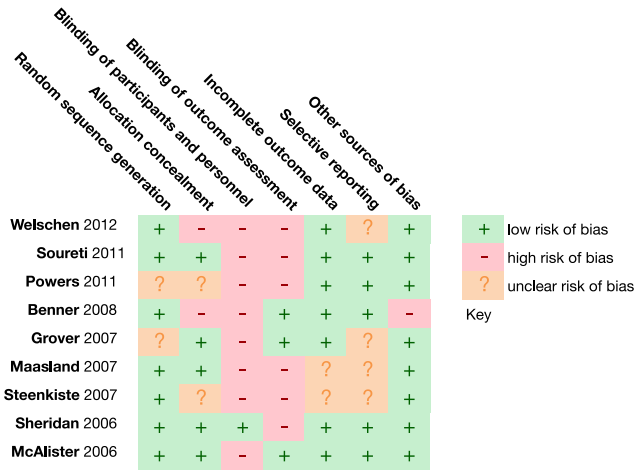


Fig. 1. Illustrative risk of bias output from the Cochrane tool. Each row represents a single study. In this study, we aim to automate the generation of such tables and extract snippets of text from articles to justify each judgement.

Review authors judge the risk of bias in each domain as *high*, *low*, or *unknown* (see Fig. 1). 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.

For many assessments, Cochrane reviewers justify their risk of bias assessments by quoting supporting text directly from the original study (see Fig. 2). This is desirable because it increases the transparency of the judgments. Here we exploit these manually extracted sentences as "distant" supervision with which to train our models. The benefit of this approach is that rather than acquiring expensive labels from domain experts, we are leveraging an existing corpus.

B. Data

In this study, we use descriptions of, and data about, clinical trials manually extracted by Cochrane reviewers for previously conducted systematic reviews (i.e., those in the CDSR). We use this structured data as a substitute for manual annotations. In this sense the strategy we take here is *distantly supervised* [14], [33].

1) Data Structure of Cochrane Reviews: The CDSR contains structured and semistructured data for the individual studies comprising each systematic review. Each review contains a

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"

Fig. 2. Review authors' justification for their score of an example study in the *allocation concealment* domain. Here the risk of bias was deemed low and the highlighted quote was extracted (into the CDSR) as support for this judgement. To train machine learning models that can automate this bias assessment (and supporting sentence extraction), we match entries in the CDSR to full-text articles that describe the corresponding clinical trials, and we identify the extracted sentences stored in the CDSR for said trials within these matched full-texts. This process will necessarily be *noisy*, i.e., introduce false positives and false negatives into the training data, but we show that it is precise enough to train reasonably accurate classifiers.

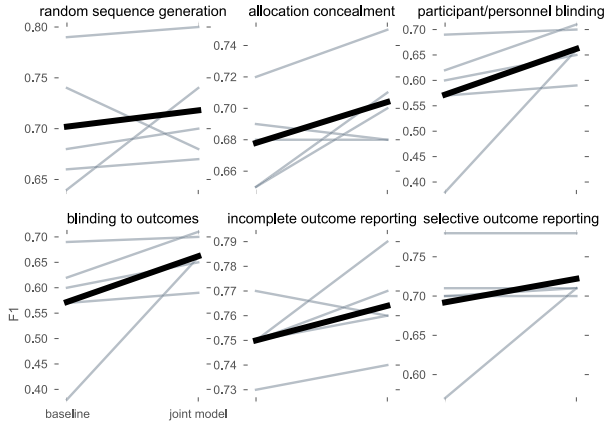


Fig. 3. Results from five-fold cross-validation across the six domains. The y-axis is F1 score. Lines connect results achieved on the same folds; the thick black lines are means (the gray lines correspond to individual fold results). The proposed joint model consistently outperforms the baseline approach.

wealth of (structured) data about the included clinical trials included in the review and there are usually multiple clinical trials described in each review. Cochrane reviews use basic clinical trial identifiers that 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 semistructured data (i.e., filtered snippets of text) that describe a specific 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) 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 handle transcription errors by Cochrane review authors, we used nonoverlapping 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 independent queries retrieved the same article. Using this strategy, we linked the semistructured descriptions of 52 454 clinical trials from Cochrane reviews to their unique PubMed identifiers, which allowed us to access citation information for these articles.

3) Justification for Risk of Bias Assessments: 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 judgment. 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 Fig. 2). We exploited this convention by searching for this pattern throughout the CDSR using a regular expression. With this approach we identified supporting quotes in at least one domain for a total of 3529 unique clinical trials. For 2200 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.

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 a matching string in the clinical trial report. For the sentence identification task, the clinical trial reports were word and sentence tokenized; sentences that matched a quote were labeled as "positives." All others were labeled as "negatives." For the document classification task, we labeled each full text trial report as being at *high*, *low*, or *unknown* risk of bias using the classification from the linked review (these labels are explicitly available in the CDSR). Approximately half of the trials included were judged to be at low risk of bias for each domain, whereas $<1\%$ of sentences were relevant to bias in any domain (see Table II).

IV. Machine Learning Methods

In this section, we introduce a baseline approach which independently learns risk of bias assessment and supporting sentence extraction. We then introduce a joint model that 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 predictions and thus result in improved predictive performance.

A. Overall Risk of Bias Prediction

We first consider the task of predicting the 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 (binary) output space $\mathcal{Y}$ comprises *low risk* and *unknown/high risk*. This dichotomization of the task is practical, since reviewers will typically conduct additional *sensitivity analyses* using only studies at low risk to investigate the robustness of their results.

We use the soft-margin support vector machine [34] as our classification model. We will denote each article by $\mathbf{x}_i$, its label for quality domain $q \in \mathcal{Q}$ (where $\mathcal{Q}$ is the set of quality domains

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

TABLE II
DESCRIPTION OF BASELINE LABEL FREQUENCIES IN THE TEST DATA

Domain	Documents at low risk of bias (%)	Sentences relevant to risk of bias (%)
Random sequence generation	1163/2088 (55.7%)	1396/565 134 (0.3%)
Allocation concealment	936/2182 (42.9%)	887/593 018 (0.2%)
Blinding of participants and personnel	981/2078 (47.2%)	1052/565 827 (0.2%)
Blinding of outcome assessment	363/714 (50.8%)	336/196 222 (0.2%)
Incomplete outcome data	1306/2081 (62.8%)	641/564 132 (0.1%)
Selective reporting	1105/1855 (59.6%)	83/500 006 ($<0.1\%$)

Denominators represent the number of documents or sentences which were able to be labeled (positively or negatively) for each domain.

enumerated in Table I) by y_i^q and a feature extracting function by ϕ. For the latter we use standard (unigram) bag-of-words 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.

B. Sentence Identification

We take a similar approach for identifying sentences as for the overall document-level judgments described above, though here labels indicate whether a given sentence was selected by a domain expert as supporting her judgment. 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

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

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

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

where the notation is similar to above (3) with the addition of m_i, which we used 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).

V. 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 independent models for classification and extraction since the sentences identified as describing bias ought to inform the overall risk of bias assessment. Intuitively, if the text describing random sequence generation contains words such as *computer* and *generated*, we would expect the document to be classified as being at *low* risk of bias for this domain (see Table III).

A. 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 sentence j in article i (sentence s_{ij}) supports the risk of bias judgment 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 (3) as follows:

$$y_i^q = \text{sign}\{\mathbf{w}_y^q \phi(\mathbf{x}_i) + \mathbf{w}_{y,s}^q \lambda_y(s_{i*}^q)\}. \quad (6)$$

Here, λ_y 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 presence in judgment-supporting sentences. We use $\mathbf{w}_{y,s}^q$ to denote the weight vector associated with the sentence interaction features for domain q. During training we minimize over $\mathbf{w}_y' = \mathbf{w}_y^q + \mathbf{w}_{y,s}^q$ (here $+$ denotes vector concatenation).

For unlabeled documents at test time, we will 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 it supports the judgment 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

TABLE III
DOCUMENT CLASSIFICATION RESULTS: BASELINE MODEL (SECTION IV-A) PERFORMANCE

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

Shown are averages over five-fold cross-validation (and ranges). We also include the four most informative features according the model for illustrative purposes.

TABLE IV
DOCUMENT CLASSIFICATION RESULTS: JOINT MODEL (SECTION V-A) PERFORMANCE

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

$-i$ represents the described "interaction" features, where the token occurs in a sentence deemed to be relevant to the bias domain.

quality assessment in a given article (because these predictions are made independently). We can write the whole predictive model out as follows:

$$y_i^q = \text{sign}\{\mathbf{w}_y^q \phi(\mathbf{x}_i) + \hat{l}_{i0}^q \mathbf{w}_{y,s}^q \boldsymbol{\lambda}_y(s_{i0}^q) + \ldots + \hat{l}_{im_i}^q \mathbf{w}_{y,s}^q \boldsymbol{\lambda}_y(s_{im_i}^q)\} \quad (7)$$

where the $\hat{l}_{ij}^q$ are predictions made via (4).

VI. EMPIRICAL RESULTS

We matched the full-texts of 2200 clinical trial reports to semistructured 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 III and IV and in Fig. 3 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 their harmonic mean. Performance for the sentence identification task is shown in Table V. As can be seen in Table IV, the joint model (where sentence predictions informed the overall document judgment) improved the predictions across all domains. And as can be seen in Table IV, interaction features comprised the majority 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.

TABLE V
RESULTS FOR AUTOMATING THE SENTENCE IDENTIFICATION TASK (USING A STANDARD BOW AND REGULARIZED LINEAR MODEL APPROACH)

Domain	Performance		
	F1	Precision	Recall
Random sequence generation	0.53	0.43	0.68
Allocation concealment	0.48	0.42	0.58
Blinding of participants and personnel	0.37	0.30	0.50
Blinding of outcome assessment	0.38	0.34	0.42
Incomplete outcome data	0.23	0.16	0.44
Selective reporting	0.06	0.11	0.04

VII. DISCUSSION

We 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 shown 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 and demonstrated that this novel approach improves document-level risk of bias assessment performance. Because 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.

But, assessing the risk of bias in a study is inherently subjective. A validation study of the Cochrane risk of bias tool found wide variations in judgments by different researchers in all domains, with the *selective reporting* domain showing the least agreement ($\kappa = 0.13$, 95% CI –0.05 to 0.31) [35]. 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 this 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 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. These issues imply that the results reported here may be pessimistic for this task. And, this offers an interesting perspective to the evaluation of our method: As the data are inherently noisy expert evaluation of our produced results might yield very different performance. We are currently conducting a "deployed" evaluation of these models which aims to shed light on such issues. To this end, we have recruited experienced systematic review authors, who will manually assess the quality of output from the tool. We plan to conduct a blinded comparison of model output versus human authored text and bias judgements relating to the same trials taken from published systematic reviews. This will allow us to empirically address whether automated methods provide accuracy comparable to human experts.

To further improve the performance of our system, related methods developed for sentiment analysis [36], [37] could be explored, as the task is conceptually similar. We are particularly interested in exploring probabilistic models that aim to jointly model sentiment and text fragments [38], [39]. We also note that a shortcoming of the proposed sentence extraction model is that the model for each domain (i.e., each weight vector $\mathbf{w}_s^q$) is fit independently of those for other domains. Going forward we hope to extend this model to take a *multitask* approach, i.e., jointly fit a single model over all domains [40], which could improve predictive performance further. Additional features, like journal impact factor, date of publication, or ontology terms derived from text, might also further improve performance.

Finally, the proposed distantly supervised method has the potential to be extended to extract other variables of interest from clinical trial reports. Specifically, the CDSR contains (semi)structured information on trial populations, interventions, outcomes, and results data. Other structured resources, e.g., SRDR[3] and ClinicalTrials.gov[4] contain highly-structured data on additional variables. Tools to automate these tasks could lead to a large reduction in the time required to produce systematic reviews.

In practice, we envision a hybrid computer–human system in which machine learning models guide the extraction process (thereby reducing manual labor). This necessitates the development of a tool to integrate the machine learning machinery described here with an intuitive graphical user interface. We have already built a prototype tool for risk of bias assessment [42].[5] We hope to pair this tool with our semiautomated abstract screening software [43]. This will further optimize reviewer workflow by enabling prioritized abstract screening coupled with semiautomated data extraction from relevant literature. Another option to optimize reviewer workflow would be to use the computer generated extractions for redundancy to improve data quality; i.e., rather than having two experts independently extract data, we might substitute the computer for one of them. However, as automated annotation becomes more wide-spread, questions about scalability and data provenance become relevant [44]. To address these questions we envision an integrated system that borrows from semantic web technology such as Open Annotations[6] and W3C Prov.[7] Such a system should ensure that one can trace the annotations and extractions back to their source documents, and to their authors, at any time.

We conclude that clinicians and health sciences researchers are overwhelmed with data. And, if we are to maintain the rigor and comprehensiveness of EBM products, new data mining methods are sorely needed to mitigate problems of information overload. This study is a step toward such larger aims.

[3]Systematic Review Data Repository [41] http://srdr.ahrq.gov/

[4]US regulatory database for clinical trial registrations https://clinicaltrials.gov/

[5]Available at: https://robot-reviewer.vortext.systems

[6]http://www.openannotation.org/spec/core/

[7]http://www.w3.org/TR/prov-overview/

REFERENCES

[1] C. D. Mulrow, D. J. Cook, and F. Davidoff, "Systematic reviews: Critical links in the great chain of evidence," *Ann. Internal Med.*, vol. 126, no. 5, pp. 389–391, Mar. 1997.

[2] 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*, vol. 343, p. d5928, 2011.

[3] A. R. Jadad, R. A. Moore, D. Carroll, C. Jenkinson, D. J. 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, Feb. 1996.

[4] J. Savović, H. Jones, D. Altman, R. Harris, P. Jni, J. Pildal, B. Als-Nielsen, E. Balk, C. Gluud, L. Gluud, J. Ioannidis, K. Schulz, R. Beynon, N. Welton, L. Wood, D. Moher, J. Deeks, and J. Sterne, "Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies," *Health Technol. Assessment (Winchester, England)*, vol. 16, no. 35, pp. 1–82, Sep. 2012.

[5] D. Tovey, R. Marshall, Bazian Ltd., S. Hopewell, and T. Rader, "Fit for purpose: Centralising updating support for high-priority Cochrane reviews," The Cochrane Collaboration, report available at http://editorial-unit.cochrane.org/fit-purpose-centralised-updating-support-high-priority-cochrane-reviews, *Nat. Inst. Health Res. Evaluation, Trials, Studies Coordinating Centre*, Tech. Rep., Jul. 2011.

[6] Cochrane Collaboration. (2014). The Cochrane database of systematic reviews [Online]. Available: http://www.thecochranelibrary.com

[7] 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 Med. Inform. Decision Making*, vol. 12, no. 1, art. no. 95, 2012.

[8] K. G. Shojania, M. Sampson, M. T. Ansari *et al.*, Updating Systematic Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US); 2007 Sep. (Technical Reviews, No. 16.) Available from: http://www.ncbi.nlm.nih.gov/books/NBK44099/

[9] H. Bastian, P. Glasziou, and I. Chalmers, "Seventy-five trials and eleven systematic reviews a day: How will we ever keep up?," *PLoS Med.*, vol. 7, no. 9, art. no. e1000326, 2010.

[10] B. Wallace, I. Dahabreh, C. Schmid, J. Lau, and T. Trikalinos, "Modernizing the systematic review process to inform comparative effectiveness: Tools and methods," *J. Comparative Effectiveness Res.*, vol. 2, no. 3, pp. 273–282, 2013.

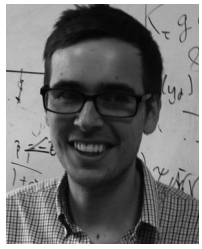
[11] I. Chalmers, M. B. Bracken, B. Djulbegovic, S. Garattini, J. Grant, A. Metin Glmezoglu, D. W. Howells, J. P A Ioannidis, and S. Oliver, "How to

increase value and reduce waste when research priorities are set," *Lancet*, vol. 383, no. 9912, pp. 156–165, Jan. 2014.
[12] J. Elliott, I. Sim, J. Thomas, N. Owens, G. Dooley, J. Riis, B. Wallace, J. Thomas, A. Noel-Storr, G. Rada, C. Struthers, T. Howe, H. MacLehose, L. Brandt, I. Kunnamo, and C. Mavergames, "Cochranetech: Technology and the future of systematic reviews," *Cochrane Database Syst. Rev.*, vol. 9, p. ED000091, 2014.
[13] I. Marshall, J. Kuiper, and B. Wallace, "Automating risk of bias assessment for clinical trials," in *Proc. ACM Conf. Bioinform. Comput. Biol. Biomed.*, 2014.
[14] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. Joint Conf. Annu. Meeting ACL Int. Joint Conf. Natural Lang. Process. AFNLP*, 2009, pp. 1003–1011.
[15] A. Cohen, W. Hersh, K. Peterson, and P.-Y. Yen, "Reducing workload in systematic review preparation using automated citation classification," *J. Amer. Med. Inform. Assoc.*, vol. 13, no. 2, pp. 206–219, Mar. 2006.
[16] T. Bekhuis and D. Demner-Fushman, "Towards automating the initial screening phase of a systematic review," *Studies Health Technol. Inform.*, vol. 160, pp. 146–150, 2010.
[17] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Active learning for biomedical citation screening," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 173–181, 2010.
[18] B. C. Wallace, K. Small, C. E. Brodley, J. Lau, and T. A. Trikalinos, "Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr," in *Proc. ACM SIGHIT Int. Health Inform. Symp.*, pp. 819–824, 2011.
[19] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings Bioinform.*, vol. 6, pp. 57–71, 2005.
[20] B. de Bruijn, S. Carini, S. Kiritchenko, J. Martin, and I. Sim, "Automated information extraction of key trial design elements from clinical trial publications," in *Proc. AMIA Annu. Symp.*, 2008, pp. 141–145.
[21] F. Boudin, J. Y. Nie, J. C. Bartlett, R. Grad, P. Pluye, and M. Dawes, "Combining classifiers for robust PICO element detection," *BMC Med. Inform. Decision Making*, vol. 10, art. no. 29, 2010.
[22] S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, and I. Sim, "ExaCT: Automatic extraction of clinical trial characteristics from journal publications," *BMC Med. Inform. Decision Making*, vol. 10, art. no. 56, 2010.
[23] S. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," *BMC Bioinform.*, Issue 12, (Supp. 2), art. no. S5, 2011.
[24] R. Summerscales, S. Argamon, S. Bai, J. Huperff, and A. Schwartzff, "Automatic summarization of results from clinical trials," in *Proc. Bioinform. Biomed.*, 2011, pp. 372–377.
[25] R. Summerscales, "Automatic summarization of clinical abstracts for evidence-based medicine," Ph.D. dissertation, The Department of Computer Science, Illinois Inst. Technol., Chicago, IL, USA, 2013.
[26] I. Shemilt, A. Simon, G. J. Hollands, T. M. Marteau, D. Ogilvie, A. OMara-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," *Res. Synthesis Methods*, vol. 5, no. 1, pp. 31–49, 2014.
[27] Y. Jin, R. T. McDonald, K. Lerman, M. A. Mandel, S. Carroll, M. Y. Liberman, F. C. Pereira, R. S. Winters, and P. S. White, "Automated recognition of malignancy mentions in biomedical literature," *BMC Bioinform.*, vol. 7, no. 1, p. 492, 2006.
[28] S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein, L. Ungar, S. Winters, and P. White, "Integrated annotation for biomedical information extraction," in *Proc. Human Lang. Technol./North Amer. Assoc. Comput. Linguistics*, 2004, pp. 61–68.
[29] X. Ling, J. Jiang, X. He, Q. Mei, C. Zhai, and B. Schatz, "Generating gene summaries from biomedical literature: A study of semi-structured summarization," *Inform. Proc. Manage.*, vol. 43, no. 6, pp. 1777–1791, 2007.
[30] X. Ling, Q. Mei, C. Zhai, and B. Schatz, "Mining multi-faceted overviews of arbitrary topics in a text collection," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 497–505.
[31] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings Bioinform.*, vol. 6, no. 1, pp. 57–71, 2005.
[32] 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, art. no. e003342, 2013.
[33] T. Nguyen and A. Moschitti, "End-to-end relation extraction using distant supervision from external semantic repositories," in *Proc. Annu. Meeting Assoc. Comput. Linguistics: Human Lang. Technol.*, 2011, pp. 277–282.
[34] V. Vapnik, *The Nature of Statistical Learning Theory*, New York, NY, USA: Springer, 2013.
[35] L. Hartling, M. Ospina, and Y. Liang, "Risk of bias versus quality assessment of randomised controlled trials: Cross sectional study," *BMJ*, vol. 339, art. no. b4012, 2009.
[36] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*. New York, NY, USA: Springer, 2012, pp. 415–463.
[37] B. Pang, B. Pang, L. Lee, and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inform. Retrieval*, vol. 2, pp. 1–135, 2008.
[38] I. Titov and R. McDonald, "A joint model of text and aspect ratings for sentiment summarization," in *Proc. Assoc. Comput. Linguistics*, 2008, p. 61801.
[39] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 804–812.
[40] H. Daumé III, A. Kumar, and A. Saha, "Frustratingly easy semi-supervised domain adaptation," in *Proc. Workshop Domain Adaptation Natural Lang. Process.*, 2010, pp. 53–59.
[41] S. Ip, N. Hadar, S. Keefe, C. Parkin, R. Iovin, E. M. Balk, and J. Lau, "A web-based archive of systematic review data," *Syst. Rev.*, vol. 1, no. 21, p. 15, 2012.
[42] 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*. New York, NY, USA: Springer, 2014, pp. 452–455.
[43] B. Wallace, K. Small, C. Brodley, J. Lau, and T. Trikalinos, "Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr," in *Proc. 2nd ACM SIGHIT Int. Health Inform. Symp.*, 2012, pp. 819–824.
[44] G. van Valkenhoef and J. Kuiper, "Crowdsourcing a comprehensive clinical trial repository," in *Proc. AAAI Workshop Modern Artif. Intell. Health Anal.*, 2014.

Iain J Marshall received his medical degree from the University of Aberdeen, Aberdeen, UK in 2004, and gained membership of the Royal College of General Practitioners in 2008. He is a family physician and researcher based at the Department of Primary Care and Public Health Sciences, King's College London, London, U.K., and an inner-city clinic at Tooting, London. His research interests are in systematic review methodology, cardiovascular disease prevention, and use of machine learning to automate data extraction from clinical trials.

Joël Kuiper received the Bachelor of Science degree in artificial intelligence in 2011 from the University of Groningen, Groningen, The Netherlands.

He is currently looking into the commercial applications of text mining, machine learning and information retrieval on PDF documents with his startup Vortext Systems.



Byron C Wallace received the Ph.D. degree in computer science in 2012 from Tufts University, Medford, MA, USA, under the supervision of C. Brodley.

He is an Assistant Professor at the University of Texas at Austin, Austin, TX, USA. He was previously Research Faculty at Brown University, where he was affiliated with the Center for Evidence-Based Medicine and the Brown Laboratory for Linguistic Processing. His research is in machine learning, data mining and natural language processing with an emphasis on applications in health informatics.