RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials

RECEIVED 3 November 2014
REVISED 16 April 2015
ACCEPTED 18 April 2015
PUBLISHED ONLINE FIRST 23 June 2015

Iain J Marshall[1], Joël Kuiper[2] and Byron C Wallace[3]

ABSTRACT

Objective To develop and evaluate RobotReviewer, a machine learning (ML) system that automatically assesses bias in clinical trials. From a (PDF-formatted) trial report, the system should determine risks of bias for the domains defined by the Cochrane Risk of Bias (RoB) tool, and extract supporting text for these judgments.

Methods We algorithmically annotated 12,808 trial PDFs using data from the Cochrane Database of Systematic Reviews (CDSR). Trials were labeled as being at low or high/unclear risk of bias for each domain, and sentences were labeled as being informative or not. This dataset was used to train a multi-task ML model. We estimated the accuracy of ML judgments versus humans by comparing trials with two or more independent RoB assessments in the CDSR. Twenty blinded experienced reviewers rated the relevance of supporting text, comparing ML output with equivalent (human-extracted) text from the CDSR.

Results By retrieving the top 3 candidate sentences per document (top3 recall), the best ML text was rated more relevant than text from the CDSR, but not significantly (60.4% ML text rated 'highly relevant' *v* 56.5% of text from reviews; difference +3.9%, [−3.2% to +10.9%]). Model RoB judgments were less accurate than those from published reviews, though the difference was <10% (overall accuracy 71.0% with ML *v* 78.3% with CDSR).

Conclusion Risk of bias assessment may be automated with reasonable accuracy. Automatically identified text supporting bias assessment is of equal quality to the manually identified text in the CDSR. This technology could substantially reduce reviewer workload and expedite evidence syntheses.

Keywords: systematic review, data mining, natural language processing, randomized controlled trials as topic, bias

RESEARCH AND APPLICATIONS

BACKGROUND AND SIGNIFICANCE

Assessing bias is a core part of systematic review methodology. Reviews typically use standardized checklists or tools to assess trial quality.[1] The Cochrane Risk of Bias (RoB) tool is one such tool.[2] It has been adopted across the Cochrane Library, and increasingly in systematic reviews published elsewhere. The tool comprises six core domains (Box 1), which reviewers score as being at *high*, *low*, or *unclear* risk of bias.

Bias assessment is time-consuming, taking experienced reviewers around 20 minutes for every study included in a systematic review.[3] The requirement to assess bias has been identified as an important factor preventing Cochrane reviews from being kept up to date.[4] Bias assessment is also subjective: individual reviewers have low rates of agreement,[3] though this improves somewhat when review specific guidance is provided.[5]

Technology to assist reviewers in assessing bias has the potential to substantially reduce workload. An accurate system could make bias assessment quicker and more reliable, freeing up researcher time to concentrate on thoughtful evidence synthesis, and ultimately help keep systematic reviews up to date.[6]

Our preliminary work demonstrated the feasibility of automated risk of bias assessment.[7] However, in this prior work we evaluated our method with respect to an imperfect reference standard, thus demonstrating the *internal validity* of our approach, but not whether the technology is mature enough to be used in practice. Specifically, since RoB assessment is subjective, we need to compare the quality of the ML approach with that of a human assessment.

In this paper, we introduce a novel machine-learning (ML) approach, which 1) models risks of bias simultaneously across all domains while 2) identifying text supporting these judgments. We substantially extend our previous model, namely by introducing a *multi-task* approach that exploits correlations between different bias types. We then evaluate this approach, benchmarking our predictive accuracy against an estimate of how well humans perform for this task. We obtain these human benchmarks by exploiting the fact that many trials are included in multiple systematic reviews (and therefore have multiple, independently conducted risk of bias assessments available). To evaluate the quality of supporting text, we use a blinded expert panel, who were asked to rate the quality of algorithm output versus the supporting text chosen by the original review authors.

OBJECTIVES

We describe the development and evaluation of RobotReviewer, (RobotReviewer refers to the user interface coupled with our machine learning technologies. Eventually, it will do more than automatic RoB assessment), a system to automate the assessment of bias of randomized controlled trials using the Cochrane RoB tool. For each domain in the RoB tool, the system should reliably perform two tasks: 1)

Correspondence to Iain J Marshall, Department of Primary Care and Public Health Sciences, King's College London, 7th Floor, Capital House, 42 Weston Street, LONDON, SE1 3QD UK; iain.marshall@kcl.ac.uk; +44 (0) 207 848 8675

© The Author 2015. Published by Oxford University Press on behalf of the American Medical Informatics Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

For numbered affiliations see end of article.

Box 1: Items from the Cochrane Risk of Bias

- Random sequence generation
- Allocation concealment
- Blinding of participants and personnel
- Blinding of outcome assessment
- Incomplete outcome data
- Selective outcome reporting

Box 2: Example of the risk of bias data stored in a Cochrane review for the domain allocation concealment, from Higgins et al.[16]

Domain:	Allocation concealment
risk of bias:	High
justification:	Quote: " . . . using a table of random numbers." Comment: Probably not done.

determine whether a trial is at low risk of bias (document classification of low *v* high or unclear risk of bias), and 2) identify text from the trial report that supports these bias judgments (sentence classification of relevant *v* irrelevant). In the evaluation, we aim to compare model performance with the consistency of RoB assessments in published systematic reviews, to help judge to what extent bias assessments could be automated in real review production.

METHODS

From a machine-learning vantage point, 1) is a *classification* task and 2) is a *data extraction* problem. For both we can use *supervised machine learning*, in which model parameters are learned from manually annotated documents.[8] Unfortunately, collecting human annotations with which to train such systems is time-consuming and therefore expensive. This is especially true for biomedical text mining tasks, as these require costly expert annotators. Training corpora used in previous efforts for related tasks have been relatively small, comprising 100–350 documents.[8,9]

Here we take a different approach: to obtain a large corpus of labeled data we use *distant supervision*, a machine learning methodology that exploits structured data in existing databases in place of direct human supervision.[10,11] Distant supervision involves automatically deriving labels for unlabeled data from existing resources, typically using heuristics that cover the majority of cases but that are imperfect (e.g., string matching). This produces *noisy* labels (having a higher rate of errors than manual annotation). However, because these are 'free' labels, we can build and exploit larger training datasets than would be otherwise feasible. Larger training datasets, in turn, have been shown to improve model performance.[12]

We derive distant supervision from the Cochrane Database of Systematic Reviews (CDSR). The CDSR comprises more than 5,400 systematic reviews on health produced by members of the Cochrane Collaboration, an international non-profit organization. This dataset includes expert risk of bias assessments for clinical trials (see Box 2). Crucially, in many of these assessments, the review authors include direct quotes from the original trial report to justify their judgments. Therefore, in addition to the *article-level* risk of bias labels, we can also derive *sentence-level* annotations within full texts that indicate whether sentences were used in assessing the risk of bias for a particular domain. This derivation is accomplished by string matching. Figure 1 illustrates this schematically.

Automating the labeling of the clinical trials corpus

Our method for corpus construction is outlined in Figure 2.

Trial Linkage

First, we sought full-text PDFs of the primary citation for clinical trials that were included in systematic reviews in the CDSR. The CDSR contains semi-structured reference data, but not unique identifiers. We therefore used the following high-precision strategy. For each trial included in a systematic review in the CDSR we conducted multiple searches of PubMed. Each search used non-overlapping subsets of citation information, any of which might be expected to uniquely retrieve the trial (e.g., search 1: articles matching full title; search 2: articles with exact author combination with matching publication year; search 3: articles with matching journal name, volume, issue, and page number). We considered a positive match where two or more searches retrieved the same article. We linked 52,454 of 67,894 studies included in reviews in the CDSR to a unique publication using this method, and obtained 12,808 of these publications in PDF format.

Pre-processing of CDSR Data

Although the Risk of Bias tool assesses 6 core biases, Cochrane review authors are free to assess other (often idiosyncratic) biases if they feel they are relevant. The CDSR contains >1400 unique strings identifying bias domains. Most of these referred to one of the core domains listed in Box 1. We manually mapped alternative descriptions to the domain labels, and excluded domains unique to individual reviews. For each linked study, we extracted the types of bias assessed, the bias judgments, and the justifications for the judgments.

Labeling PDFs Using Distant Supervision

Plain text was extracted from the PDFs using the pdftotext utility from xPDF.[13] The extracted text was tokenized into sentences and words.

For task 1 (document annotation), we algorithmically labeled each document as being at 'low' or 'high/unclear' risk of bias, using the judgment from the linked Cochrane review. We dichotomized this outcome consistent with typical practice in systematic reviews: reviewers often conservatively assume that 'unclear' studies have a 'high' risk of bias and conduct sensitivity analyses including only studies at low risk of bias.

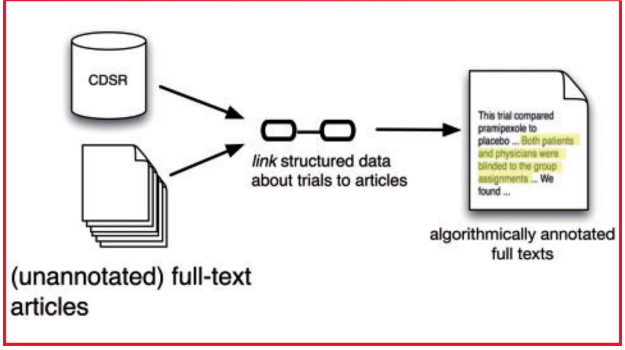
For task 2 (sentence annotation) where quote data was available in the linked Cochrane review, sentences containing exactly matching text were labeled as relevant to the risk of bias domain. All other sentences were labeled as irrelevant. This strategy produces incomplete labels and specifically would be expected to have high precision and low recall. Cochrane review authors are likely to quote one or two sentences that they feel best justify a risk of bias judgment. Ideally, all text relevant to the bias decision would be labeled.

Machine-Learning Approach

We use a novel *multi-task* variant of the soft-margin Support Vector Machine (SVM) [14] that maps articles to risk of bias assessments (low or high/unclear) (task 1) and simultaneously extracts sentences supporting these judgments (task 2). Multi-task learning refers to scenarios where the aim is to induce classifiers for multiple, related classification problems or 'tasks'.[15] Here, bias assessment for the respective domains constitute our related tasks.[16]

Our approach includes two novel components. First, we explicitly incorporate features derived from sentences that support risk of bias assessment into the document-level model that predicts the RoB. Second, we jointly model RoB across all domains, for both sentence and document-level predictions.

Figure 1:. Algorithmic annotation of clinical trial PDFs using data from the CDSR.


CDSR
link structured data about trials to articles
This trial compared pramipexole to placebo ... Both patients and physicians were blinded to the group assignments ... We found ...
algorithmically annotated full texts
(unannotated) full-text articles

The document-level feature space comprises the uni- and bi-grams of the full-text documents as a foundation. To incorporate supporting sentence information into this basic document-level model, we append interaction features for each sentence, which cross the sentence relevance to the domain of interest (relevant *v* not relevant) with the textual features from that sentence (uni- and bi-grams). We have presented details of this method elsewhere.[7] Intuitively, the word 'computer' appearing in a sentence concerning randomization may strongly indicate a low risk of bias, but the same word elsewhere in the document may have little predictive power. The interaction features we introduce aim to capture such information. To create these interaction features when training the model, we use data on sentence relevance taken directly from the CDSR. At test time, however, we do not know which sentences support risk of bias assessments for the respective domains, and we therefore use the sentences predicted as relevant by our sentence classifier.

Here we extend this model to borrow strength across risk of bias domains, using a multi-task approach for both sentence and document-level classifications. Specifically, we introduce 'interaction features' that represent the intersection of domains and token (word) indicators. Denoting the number of domains by k, we insert k copies of each feature vector $\mathbf{x}$ (one per domain) for each instance, in addition to a shared copy of $\mathbf{x}$ common to all domains (see Figure 3 for a schematic of this approach). For example, there will be a feature corresponding to the presence of the word 'computer' *and* the target

Figure 2: Schematic of corpus construction, and outline of the distant supervision process.

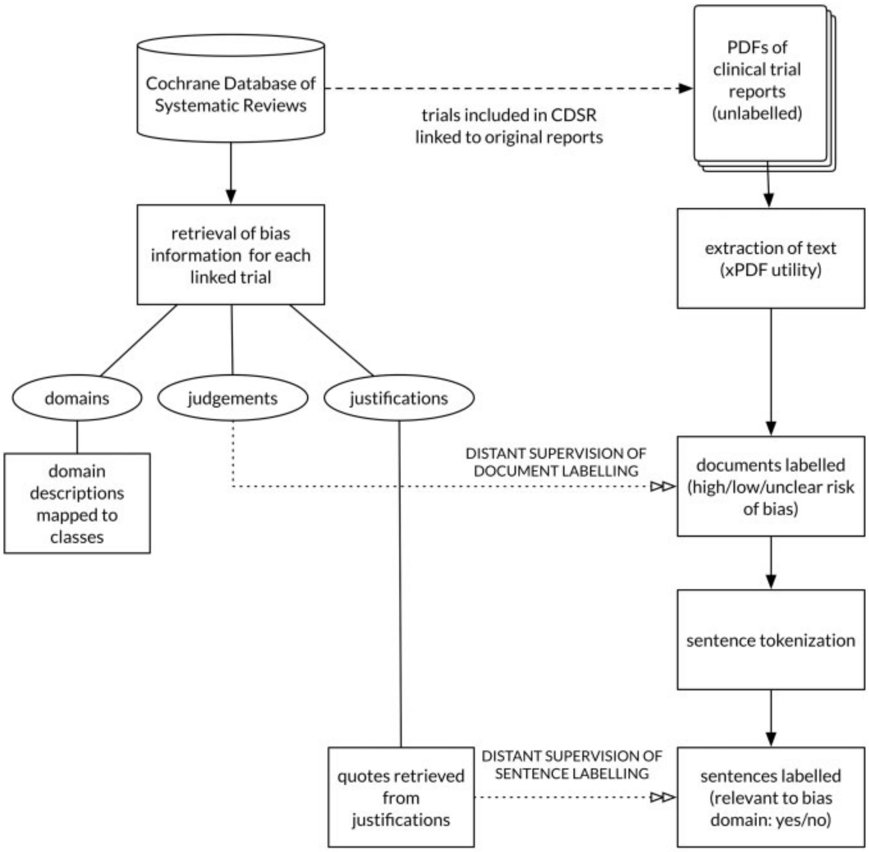
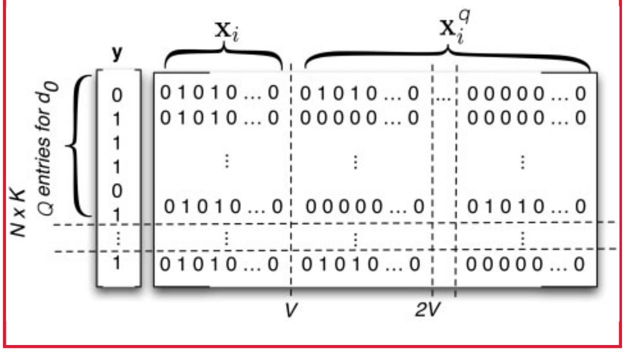

Cochrane Database of Systematic Reviews
trials included in CDSR linked to original reports
PDFs of clinical trial reports (unlabelled)
retrieval of bias information for each linked trial
extraction of text (xPDF utility)
domains
judgements
justifications
domain descriptions mapped to classes
DISTANT SUPERVISION OF DOCUMENT LABELLING
documents labelled (high/low/unclear risk of bias)
sentence tokenization
quotes retrieved from justifications
DISTANT SUPERVISION OF SENTENCE LABELLING
sentences labelled (relevant to bias domain: yes/no)

Figure 3: A schematic depiction of our multi-task learning approach. We define a joint classification model across domains. To achieve this, we include k representations of each instance (e.g., document) d in the design matrix, one per risk of bias domain. We construct the target vector y with the corresponding per-domain labels for d (derived from the CDSR). Each of the k entries for d comprises k+1 concatenated vectors of length equal to the vocabulary size (V), where for the document model this vocabulary is the union of unique uni- and bi-grams appearing in at least two articles, and for the sentence model V is the union of uni- and bi-grams appearing in at least two supporting sentences. The first copy of V in each row is shared across all entries representing the corresponding article; the remaining are domain specific copies. Thus in any given row, all but two of these sub-vectors will be zero vectors. Specifically, each row i will contain two non-zero vectors that are copies of the bag-of-words representation (binary indicators for V) for instance i (x_i): one shared and the other domain specific. The shared component allows borrowing of strength across domains, while the domain-specific components enable the model to learn words that signal low risk of bias (or the likelihood of supporting RoB assessment, for the sentence prediction task) only in specific domains.



domain *randomization*. This will be non-zero only in columns that comprise the copy of **x** specific to *randomization*. Note that this can be viewed as an instantiation of Daumé's *frustratingly easy* domain adaptation approach.[17] Our sentence model is thus trained jointly across all risk of bias domains; the shared component enables information sharing between them. We adopt this approach for both the sentence and the document-level models.

For a new article (at test time), we then use the multi-task sentence model to generate predictions for each sentence regarding whether it is likely to support assessments for the respective domains. Before a document-level RoB prediction is made, indicators corresponding to the tokens comprising sentences predicted to be relevant are inserted into the vectors representing documents. This is done for each domain. These document representations include a shared component across domains (again enabling borrowing of strength). Therefore, sentence-level predictions (made via a multi-task sentence-level model) directly inform our multi-task document-level model. This realizes a joint approach to predicting sentence-level relevance and document-level assessments across related tasks.

For both the sentence- and document-level model, we adopt a linear classification model defined by a weight vector **w**, such that $y = L(\mathbf{w}\cdot\mathbf{x})$, where L maps the continuous score to a categorical label of 0 (when $\mathbf{w}\cdot\mathbf{x} < 0$) or 1 ($\mathbf{w}\cdot\mathbf{x} \geq 0$). We use the hinge-loss function, which imposes no loss when the model prediction is correct and a loss proportional to the magnitude of $\mathbf{w}\cdot\mathbf{x}$ when the prediction is incorrect. We combine this with a squared L2 penalty term on the model parameters to form our objective, which describes a linear-kernel soft-margin SVM.[14] This objective is shown in Equation 1. For the sentence-level model, we sum the loss over the sentences comprising each distantly labeled document. The joint, multi-task modeling we have proposed is effectively realized by augmenting the feature space; for example, by inserting into document or sentence vectors shared and domain-specific copies of token indicators. Note that in our multi-task model, there is only a single, shared **w** for document classification (and another for sentence classification), whereas in our previous approach we fit separate weight vectors for each domain for both the document and sentence models.

The objective function we minimise is shown in equation 1. The general objective function we minimize. This is the form we use across all domains for both the document and sentence-level models, however for the latter we need to sum over the sentences comprising documents. Note that the innovations we have proposed (multi-task learning and a joint model incorporating sentence labels or predictions into the document-level model) are realized by manipulation of the feature spaces, hence the form the objective is unchanged.

$$w = arg\ min_w \sum_i \{hinge - loss(w \cdot \mathrm{x}_i,\ \mathrm{y}_i)\} + \alpha||\mathbf{w}||^2 \qquad (1)$$

We fit this model (estimate **w**) by minimizing our objective via Stochastic Gradient Descent (a standard optimization procedure in which one optimizes parameters to minimize an objective function by following the gradient as approximated by iterative evaluation on individual instances). The objective includes a hyper-parameter α, which trades regularization strength (model simplicity) against empirical loss. We tune α via line search ranging over values from 10^{-4} to 10^{-1}, equidistant in log-space, and select the value that performs best on average according to nested cross-fold validation.

All models we consider leverage token-based features encoded in a binary 'bag-of-words' representation. For sentence prediction, our vocabulary *V* comprises all unique uni- and bi-grams present in at least two supporting sentences. For document prediction, *V* comprises all uni- and bi-grams that appear in at least two articles. We preprocessed the text by removing English 'stop words' (uninformative words like "the" and "as") and converting to lowercase. Because using full-text and interaction features produces a very large feature space, we use the 'feature hashing' trick to keep the model tractable. The hashing trick maps strings to vector indices via a hashing function. This has been shown to work well for large-scale multi-task text classification.[18]

To summarize, we use a novel model for risk of bias prediction that 1) jointly makes article- and sentence-level predictions, and 2) borrows strength across the related risk of bias assessment tasks via multi-task learning, for both article and sentence predictions.

We have made the code used for the entire distant supervision pipeline and evaluation available at https://www.github.com/ijmarshall (under cochrane-nlp and cochrane-nlp-experiments). Systematic reviewers might instead use our prototype web-based tool which generates and presents RoB assessments for articles uploaded by users (Figure 5).[19]

Evaluation

Task 1: Document Prediction

For task 1 (document judgments) we exploited the fact that many trials (ranging from 239–1148 trials for each bias domain) are described in

Table 1: Results from the document evaluation task: Baseline=accuracy achieved by labeling all test documents with majority class for that domain; Model 1=separate bag-of-words model for each domain; Model 2=multi-task model jointly modeling all domains and incorporating information about sentence relevance as features; Model 3=same multi-task model excluding domains 5 and 6; Cochrane=estimate of human accuracy obtained by comparing a second risk of bias assessment (of the same trials) from another systematic review

Domain	Trials (n)	baseline	model 1	model 2	model 3	cochrane	P (model 2 versus cochrane)
Overall	6610	56.4%	69.3%	71.0%	-	78.3%	$P < 0.001$
1. Random sequence generation	1225	59.3%	72.5%	73.9%	75.8%	84.8%	$P < 0.001$
2. Allocation concealment	2089	53.7%	72.4%	74.0%	73.3%	80.0%	$P < 0.001$
3. Blinding of participants and personnel	1051	50.4%	72.6%	73.0%	73.7%	78.1%	$P = 0.003$
4. Blinding of outcome assessment	250	57.7%	64.0%	61.5%	67.4%	83.2%	$P < 0.001$
5. Incomplete reporting of outcomes	1149	60.9%	63.9%	65.1%	-	71.3%	$P < 0.001$
6. Selective reporting	846	59.9%	61.8%	67.6%	-	73.0%	$P = 0.010$

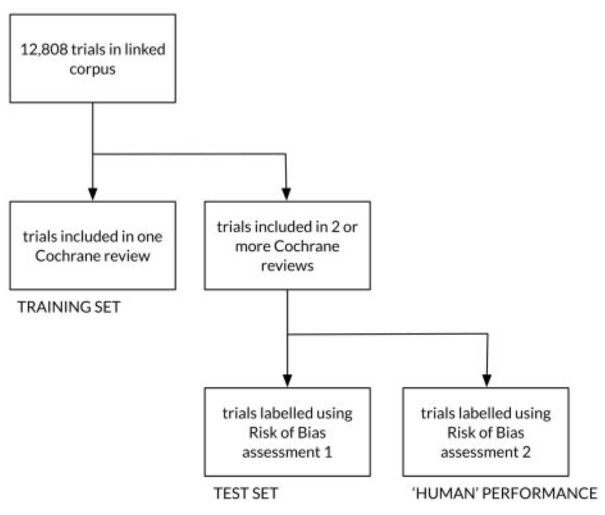
more than one Cochrane review. We could therefore access two independently conducted RoB assessments for each of these trials. We held out this set of trials to use as test data.

We were therefore able to evaluate both the accuracy of the automated RoB assignment and agreement among human reviewers on the same set of trials. This is important because agreement between human reviewers is imperfect (see Table 1). This establishes an upper bound on the performance we might hope to achieve using an automated approach. Figure 4 depicts our evaluation setup schematically.

For document prediction, we consider three models. Model 1 is a standard uni- and bi-gram model using the entire document text; each RoB domain is modeled independently in this approach.

Models 2 and 3 are the multi-task models that we have proposed in this work, which incorporate supporting text (output from task 2) and also share features across domains for document-level prediction. Model 2 uses the multi-task approach to jointly model all 6 domains. Model 3 is identical to model 2, except that it excludes *incomplete reporting of outcomes* and *selective reporting*. We excluded these domains post hoc under the assumption that predictions in these noisy domains were adversely affecting performance in other domains in the case of our multi-task model. Specifically, *incomplete reporting of outcomes* would typically involve calculation of withdrawal and dropout rates (which is not possible using bag-of-words modeling), and assessing *selective reporting* would usually require reference to a trial protocol. Finally, for comparison, we report a *baseline* result, where all documents are labeled with the majority class for the domain.

Figure 4: Test/training set partition for the document-level judgments, and use of the second Risk of Bias assessment as a surrogate for human performance.


12,808 trials in linked corpus
trials included in one Cochrane review
TRAINING SET
trials included in 2 or more Cochrane reviews
trials labelled using Risk of Bias assessment 1
TEST SET
trials labelled using Risk of Bias assessment 2
'HUMAN' PERFORMANCE

Task 2: Sentence Prediction

For task 2, we automatically labeled sentences comprising held-out documents using the trained model. We considered the evaluation criteria: top1, where the top scoring sentence per document was identified, and top3, where the top three scoring sentences were identified. The latter is particularly relevant to *semi*-automating RoB assessment; if relevant sentences are amongst the top three, then we can rapidly guide reviewers to these sentences, thus expediting assessment.

We used two control strategies for comparison: *cochrane*, where the text justifying a bias decision was taken directly from the published review in the CDSR describing the trial (to estimate human performance; mean 1.3 sentences per trial), and *baseline* where a sentence was drawn randomly from the document. We aim to assess the comparative relevance of sentences selected by our automated approach and the two control strategies. We cannot derive this information from the CDSR, however, and thus need to rely on human expertise to manually assess relevance. To this end, we recruited a panel of 20 systematic review authors (median published reviews per author: 19, IQR 7.5 to 51.5); all members had substantial experience of the Cochrane RoB tool.

Sentences identified by each strategy were presented to the panel members, who were blinded to the text source (i.e., whether sentences were selected by human, at random, or by our algorithm). The assessors were asked to assess sentence relevance to a particular domain in the RoB tool using a Likert-like scale (3=highly relevant, 2=somewhat relevant, and 1=not relevant). The assessors were provided with additional definitions for each of these categories. Two assessors piloted the evaluation and had substantial agreement (Kappa=0.79). We calculated based on pilot data that at least 350

Table 2: Proportion of studies for which *highly relevant* text was identified using four methods: baseline, one random sentence chosen per document; top1, the one most informative sentence according to the algorithm; top3, the top three most informative sentences according to the algorithm; and *cochrane*, being text quoted in published Cochrane reviews to justify bias decisions (mean 1.3 sentences per document). Where more than one sentence was identified, the one highest rated sentence contributes to the score.

Domain	Trials (n)	baseline	top1	top3	cochrane	top1 v cochrane	top3 v cochrane
Overall	378	0.5%	45.0%	60.4%	56.5%	−11.6% (−18.5% to −4.4%); $P < 0.001$	+3.9%, (−3.2% to +10.9%); $P = 0.141$
1. Random sequence generation	81	0.0%	55.6%	65.4%	60.5%		
2. Allocation concealment	75	0.0%	44.0%	60.0%	60.0%		
3. Blinding of participants and personnel	76	0.0%	55.3%	72.4%	68.4%		
4. Blinding of outcome assessment	56	0.0%	39.3%	62.5%	57.1%		
5. Incomplete reporting of outcomes	67	3.0%	40.9%	57.6%	50.8%		
6. Selective reporting	23	0.0%	0.0%	4.6%	4.6%		

Table 3: Proportion of studies for which *highly* or *somewhat relevant* text was identified using four methods; see caption of Table 2 for details of models.

Domain	Trials (n)	baseline	top1	top3	cochrane	top1 v cochrane	top3 v cochrane
Overall	378	3.7%	69.7%	84.9%	83.8%	-14.1% (−20.0% to −8.1%); $P < 0.001$	+1.0% (−4.2% to +6.2%); $P = 0.35$
1. Random sequence generation	81	4.9%	88.9%	92.6%	88.9%		
2. Allocation concealment	75	0.0%	72.0%	88.0%	89.3%		
3. Blinding of participants and personnel	75	4.0%	68.5%	84.2%	81.6%		
4. Blinding of outcome assessment	56	3.6%	58.9%	83.9%	82.1%		
5. Incomplete reporting of outcomes	67	7.5%	71.2%	90.9%	88.1%		
6. Selective reporting	23	0.0%	18.2%	31.9%	45.5%		

trials would be needed to detect a 10% difference in model output quality with 80% power with significance of $P < 0.05$. We collected a total of 1731 judgments from 20 experts from 371 trials.

RESULTS

Document-Level Results

Document-level results are presented in Table 1. Model 2 judgments were less accurate than those from published reviews, though the difference was <10% (overall accuracy 71.0% with ML *v* 78.3% with CDSR; $P < 0.001$). Model 1 (which does not include supporting sentences and models each domain separately) achieved substantially greater accuracy than baseline. Model 2 (which jointly models all domains, and incorporates information about whether sentences are judged relevant) improved performance compared with Model 1 in all but one domain (*blinding of outcome assessment*). Model 3, which ignores the noisy *selective reporting* and *incomplete reporting of outcomes* domains, resulted in uniform improvement compared to Model 1 across all the domains it included. Our study is not powered to assess the significance of differences between the three models; a larger dataset would be needed to determine whether these apparent differences are statistically significant.

Sentence-Level Results

The results from task 2 (identifying sentence with information about RoB) are presented in Tables 2 and 3. The *top1* model (which retrieves one sentence per study) performed substantially better than baseline, but produced text judged less relevant than that in the CDSR (10% fewer documents with highly relevant output, and 14% fewer documents with highly, or somewhat relevant output). The best text from the *top3* model was rated as *more* relevant than text from the CDSR overall, and in individual domains, but differences were not statistically significant.

DISCUSSION

We report the development and evaluation of RobotReviewer, a system for automating RoB assessment. Our system determines whether a trial is at low risk of bias for each domain in the Cochrane RoB tool, and identifies text that supports these judgments. We demonstrated strong performance on these tasks. Automatic document judgments were of reasonable accuracy, lagging our estimate of human reviewer accuracy by ∼7%. Our automated approach identified text supporting RoB judgments of similar quality to that found in published systematic reviews.

While our algorithm is not ready to replace manual RoB assessment altogether, we envisage several ways it could reduce author workload in practice. Since justifications are provided, most errors should be easy to identify, meaning reviewers need consult the full paper only where judgments are not adequately justified (see Box 3). This should mitigate a common concern about automation technologies: that they act as *black boxes*.[21] Elsewhere, we have described a prototype tool which presents the model predictions to the user directly within the original PDF document.[19] Figure 5 shows the system in use. This has the additional advantage of preserving the link between published reviews and their source data, a key omission in current practice.[20] Alternatively, this system could also be used to draw reviewers' attention to sentences that are likely to be relevant, leaving them to make the final judgment: this would expedite RoB assessment. Finally, we note that current practice is for two reviewers to assess the RoB of trials independently, then reach a consensus. An alternative workflow would be to replace one of these two reviewers with the automated approach, thus still having a second independent assessment.

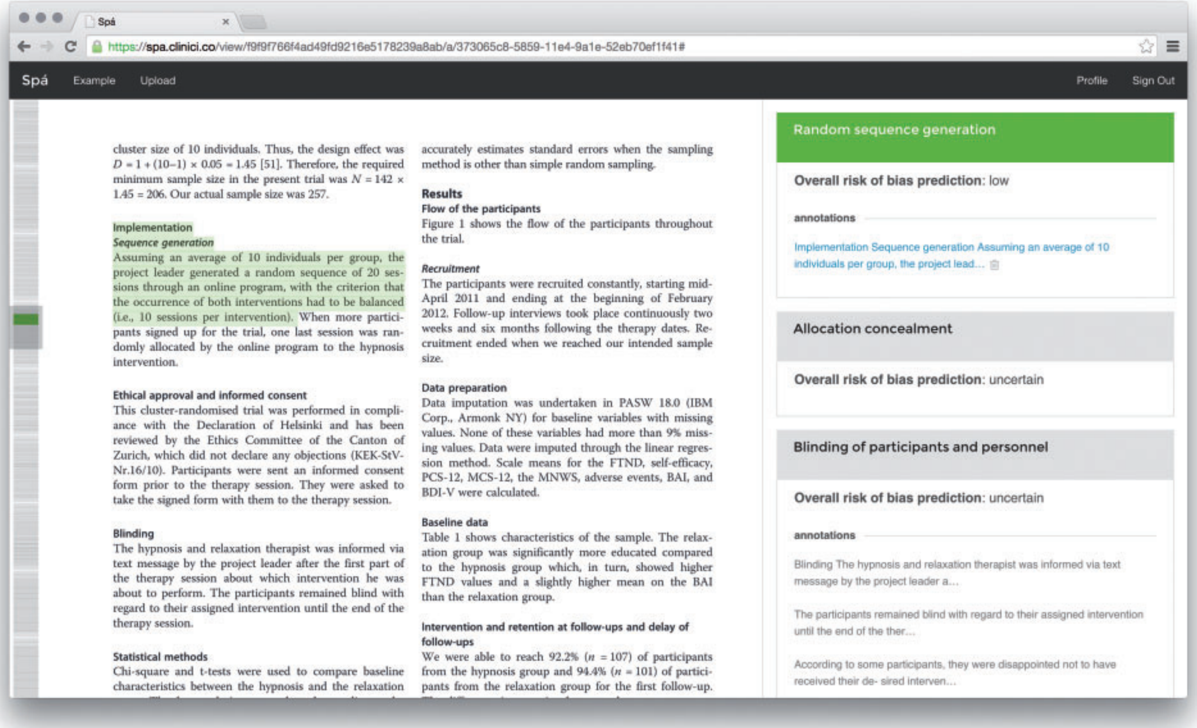
Box 3: Example of model output where a reviewer could verify the judgment by reference to the justifying text, without reference to the original paper.

Domain	Random sequence generation
Risk of bias	Low
Text justifying judgment	*Sequence generation* *Assuming an average of 10 individuals per group, the project leader generated a random sequence of 20 sessions through an online program, with the criterion that the occurrence of both interventions had to be balanced (i.e., 10 sessions per intervention)*

One potential weakness is that our corpus comprises a fraction of the studies reported in the Cochrane Library. The 12,808 PDFs were a convenience sample comprising PDFs available via university library subscriptions. Studies with unobtainable PDFs might have increased RoB, particularly those in lesser-known journals, those reported as conference abstracts, and older study reports. Post-hoc, we found these were 6% less likely to be at low RoB. However, our system is designed for PDF use, and our corpus should generalize to the types of PDFs that researchers use.

In this work we sought to estimate the accuracy of manual RoB assessment by making use of trials with 2 or more RoB assessments in the CDSR. Our approach makes the assumption that discrepancies between RoB assessments represent errors or disagreements. In

Figure 5. Example of our prototype system showing the bias assessment for random sequence generation and a supporting sentence.


Spá
https://spa.clinici.co/view/f9f9f766f4ad49fd9216e5178239a8ab/a/373065c8-5859-11e4-9a1e-52eb70ef1f41#
Spá
Example
Upload
Profile
Sign Out

cluster size of 10 individuals. Thus, the design effect was $D = 1 + (10-1) \times 0.05 = 1.45$ [51]. Therefore, the required minimum sample size in the present trial was $N = 142 \times 1.45 = 206$. Our actual sample size was 257.

Implementation

Sequence generation

Assuming an average of 10 individuals per group, the project leader generated a random sequence of 20 sessions through an online program, with the criterion that the occurrence of both interventions had to be balanced (i.e., 10 sessions per intervention). When more participants signed up for the trial, one last session was randomly allocated by the online program to the hypnosis intervention.

Ethical approval and informed consent

This cluster-randomised trial was performed in compliance with the Declaration of Helsinki and has been reviewed by the Ethics Committee of the Canton of Zurich, which did not declare any objections (KEK-StV-Nr.16/10). Participants were sent an informed consent form prior to the therapy session. They were asked to take the signed form with them to the therapy session.

Blinding

The hypnosis and relaxation therapist was informed via text message by the project leader after the first part of the therapy session about which intervention he was about to perform. The participants remained blind with regard to their assigned intervention until the end of the therapy session.

Statistical methods

Chi-square and t-tests were used to compare baseline characteristics between the hypnosis and the relaxation

accurately estimates standard errors when the sampling method is other than simple random sampling.

Results

Flow of the participants

Figure 1 shows the flow of the participants throughout the trial.

Recruitment

The participants were recruited constantly, starting mid-April 2011 and ending at the beginning of February 2012. Follow-up interviews took place continuously two weeks and six months following the therapy dates. Recruitment ended when we reached our intended sample size.

Data preparation

Data imputation was undertaken in PASW 18.0 (IBM Corp., Armonk NY) for baseline variables with missing values. None of these variables had more than 9% missing values. Data were imputed through the linear regression method. Scale means for the FTND, self-efficacy, PCS-12, MCS-12, the MNWS, adverse events, BAI, and BDI-V were calculated.

Baseline data

Table 1 shows characteristics of the sample. The relaxation group was significantly more educated compared to the hypnosis group which, in turn, showed higher FTND values and a slightly higher mean on the BAI than the relaxation group.

Intervention and retention at follow-ups and delay of follow-ups

We were able to reach 92.2% ($n = 107$) of participants from the hypnosis group and 94.4% ($n = 101$) of participants from the relaxation group for the first follow-up.

Random sequence generation

Overall risk of bias prediction: low

annotations

Implementation Sequence generation Assuming an average of 10 individuals per group, the project lead...

Allocation concealment

Overall risk of bias prediction: uncertain

Blinding of participants and personnel

Overall risk of bias prediction: uncertain

annotations

Blinding The hypnosis and relaxation therapist was informed via text message by the project leader a...

The participants remained blind with regard to their assigned intervention until the end of the ther...

According to some participants, they were disappointed not to have received their de- sired interven...

practice, RoB judgments may be influenced by individual review question, or by outcome. However, our estimate of human performance shows substantially greater agreement for this task than previous estimates. Where multiple Cochrane reviews contain the same trial, it is likely that they were produced by the same review group, and will share editors and/or author teams who may reach similar bias decisions more often than independent groups.

There are several promising routes for improving performance. First, our model was trained on trials from any clinical specialty. Since reviewer agreement increases when review specific guidance is given,[5] training a model on trials relevant to a review of interest may improve performance.

Second, the Cochrane Handbook recommends that some domains of bias should be assessed per *outcome*, rather than per study.[16] While our tool should identify text that is relevant to bias for all outcomes assessed, it produces one overall judgment per paper. This may partly explain relatively poor performance in the domain *blinding of outcome assessment*, which seems likely to vary substantially for different outcomes in the same trial. Most Cochrane reviews at present (and hence our training data) still assess all RoB domains per study, but we aspire to judge bias for each outcome in future.

Third, the domain *selective outcome reporting* requires reference to a trial protocol, to find if any pre-specified outcomes were not reported in the final paper. That our model performed better than baseline for this domain is probably explained by correlations between selective reporting bias and other biases that are more readily determined from the trial publication. Compulsory registration of clinical trials means that trial protocols are now easily obtainable: the WHO collates protocols from international registries in machine-readable format, including lists of outcomes. Linking with this dataset may yield better performance.

Finally, by incorporating reviewer corrections into the model, we could make use of *online* learning. This strategy may improve model performance with use. This would make it possible to stay up-to-date with changes in research methodology or bias assessment practice over time, and would learn to assess bias in a way which is tailored to the needs of a particular clinical area, review group, or individual authors.

CONCLUSION

We have outlined a method for automating RoB assessment, including the tasks of categorizing articles as at *low* or *high/unclear* risk and extraction of sentences supporting these categorizations, across several domains. While model performance lags human accuracy, it can identify text in trial PDFs with similar relevance to that in published reviews. Future methodological improvements are likely to close the gap between automated and human performance, and may eventually be able to replace manual RoB assessment altogether.

FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

All authors contributed to the study concept, data linkage, corpus construction, model development, validation study design, recruitment of the expert panel, and data analysis. All authors jointly drafted the manuscript, and all have approved the final version of the manuscript. IM is the corresponding author.

ACKNOWLEDGEMENTS

We would like to thank the following people who participated in our expert panel: Tom Trikalinos, Director of the Centre of Evidence Based Medicine (CEBM), Brown University; Rachel Marshall, Editor, Cochrane Editorial Unit; Emma Welsh, Managing Editor, Cochrane Airways Group; Kayleigh Kew, Systematic reviewer, Cochrane Airways Group; Ethan Balk, Assistant professor, CEBM, Brown University; James Thomas, Professor, Institute of Education, University of London; Michelle Richardson, Researcher, Institute of Education, University of London; Mark Newman, Reader in evidence-based policy, Institute of Education, University of London; Oluwaseun Akinyede, Research assistant, ECRI institute; Jonathan Treadwell, Associate director, ECRI institute; Joann Fontanarosa, Research assistant, ECRI institute; Karla Soares-Weiser, Managing Director, Enhance Reviews; Hanna Bergman, Senior researcher, Enhance Reviews; Mei Chung, Professor of Evidence Based Medicine, Tufts University; Issa Dahabreh, Assistant professor, CEBM, Brown University; Matthew Page, Systematic reviewer, Monash University; Miranda Cumpston, Editor, Cochrane Public Health Group; Ian Shemilt, Senior Research Associate, University of Cambridge; Jani Ruotsalainen, Managing Editor, Cochrane Occupational Safety and Health Review Group; Therese Dalsbø, Senior Adviser, Norwegian Knowledge Centre for the Health Services. We thank Chris Mavergames at the Cochrane Collaboration for his help in recruiting the panel. The research was supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London, and the NIHR Collaboration for Leadership in Applied Health Research and Care South London at King's College Hospital NHS Foundation Trust. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

REFERENCES

1. Centre for Reviews and Dissemination. 1.3.4. Assessing quality. In: *Systematic reviews: CRD's guidance for undertaking reviews in healthcare. CRD*, University of York 2009.
2. Higgins JPT, Altman DG, Gotzsche PC, *et al.* The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011; 343: d5928.
3. Hartling L, Ospina M, Liang Y. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ* 2009; 339: b4012.
4. Tovey D, Marshall R, Hopewell S, *et al.* Fit for purpose: centralising updating support for high-priority Cochrane Reviews. National Institute for Health Research Evaluation, Trials, and Studies Coordinating Centre 2011. [Available from http://editorial-unit.cochrane.org/fit-purpose-centralised-updating-support-high-priority-cochrane-reviews].
5. Hartling L, Bond K, Vandermeer B, *et al.* Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PLoS One* 2011; 6(2):e17242.
6. Tsafnat G, Glasziou P, Choong MK, *et al.* Systematic review automation technologies. *Systematic reviews journal* 2014; 3(1):74.
7. Marshall I, Kuiper J, Wallace B. Automating Risk of Bias Assessment for Clinical Trials. In: *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine.* New York, 2014; 88–95 [Available from: https://kclpure.kcl.ac.uk/portal/files/28954780/acmbcb2014_final_2.pdf].
8. Summerscales RL. Automatic Summarization of Clinical Abstracts for Evidence-Based Medicine [PhD thesis]. Illinois Institute of Technology 2013. [Available from: http://www.andrews.edu/~summersc/summerscales_phdthesis2013.pdf].
9. Kiritchenko S, Bruijn B de, Carini S, *et al.* ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Medical Informatics and Decision Making* 2010; 10(1): 56.

10. Mintz M, Bills S, Snow R, *et al*. 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. 2009;1003-11.
11. Ling, X, Jiang, J, He, X, *et al*. Automatically generating gene summaries from biomedical literature. *Pacific Symposium on Biocomputing* 2006;(11) 40–51.
12. Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 2001: 26–33. [Available from: http://dl.acm.org/citation.cfm?id=1073017].
13. xPDF, Open Source PDF viewer [website], http://www.foolabs.com/xpdf/ Accessed January 2015.
14. Boser, BE, Guyon, IM, Vapnik, VN. A training algorithm for optimal margin classifiers. In: *Proceedings of the Annual Workshop on Computational Learning Theory*. ACM 2010. 144-152.
15. Caruana, R. Multitask learning. Springer US, 1998.
16. Higgins J, Altman D, Sterna J. Chapter 8: Assessing risk of bias in included studies. In: *The Cochrane handbook for systematic reviews of interventions*. 2011. [Available from: www.cochrane-handbook.org].
17. Daumé H. Frustratingly easy domain adaptation. In: *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. 2010; 53–59. [Available from: http://arxiv.org/abs/0907.1815].
18. Weinberger K, Dasgupta A, Langford J, *et al*. Feature hashing for large scale multitask learning. In: *Proceedings of the 26th Annual International Conference on Machine Learning. ACM*; 2009; 1113–1120.
19. Kuiper J, Marshall I, Wallace B, *et al*. Spá: A web-based viewer for text mining in evidence based medicine. In: *Machine Learning and Knowledge Discovery in Databases. Springer,* Berlin and Heidelberg 2014; 452–455.
20. Adams CE, Polzmacher S, Wolff A. Systematic reviews: work that needs to be done and not to be done. *Journal of Evidence-Based Medicine* 2013; 6(4):232–235.
21. Thomas J. Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation? *OA Evidence-Based Medicine* 2013; 1(2):1–6.

AUTHOR AFFILIATIONS

[1]Department of Primary Care and Public Health Sciences, King's College London, UK

[2]University Medical Center, University of Groningen, Groningen, The Netherlands

[3]School of Information, University of Texas at Austin, Austin, Texas, USA