Citation for published version (APA):
Marshall, I. J., Kuiper, J., Banner, E., & Wallace, B. C. (2017). Automating Biomedical Evidence Synthesis: RobotReviewer. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.
Automating Biomedical Evidence Synthesis: RobotReviewer

Iain J. Marshall,¹ Joël Kuiper,² Edward Banner³ and Byron C. Wallace³
¹Department of Primary Care and Public Health Sciences, Kings College London
²Doctor Evidence, §College of Computer and Information Science, Northeastern University
iain.marshall@kcl.ac.uk, jkuiper@doctorevidence.com
banner.ed@husky.neu.edu, byron@ccs.neu.edu

Abstract

We present RobotReviewer, an open-source web-based system that uses machine learning and NLP to semi-automate biomedical evidence synthesis, to aid the practice of Evidence-Based Medicine. RobotReviewer processes full-text journal articles (PDFs) describing randomized controlled trials (RCTs). It appraises the reliability of RCTs and extracts text describing key trial characteristics (e.g., descriptions of the population) using novel NLP methods. RobotReviewer then automatically generates a report synthesising this information. Our goal is for RobotReviewer to automatically extract and synthesise the full-range of structured data needed to inform evidence-based practice.

1 Introduction and Motivation

Decisions regarding patient healthcare should be informed by all available evidence; this is the philosophy underpinning Evidence-based Medicine (EBM) (Sackett, 1997). But realizing this aim is difficult, in part because clinical trial results are primarily disseminated as free-text journal articles. Moreover, the biomedical literature base is growing exponentially (Bastian et al., 2010). It is now impossible for a practicing clinician to keep up to date by reading primary research articles, even in a narrow specialty (Moss and Marcus, 2017). Thus healthcare decisions today are often made without full consideration of the existing evidence.

Systematic reviews (SRs) are an important tool for enabling the practice of EBM despite this data deluge. SRs are reports that exhaustively identify and synthesise all published evidence pertinent to a specific clinical question. SRs include an assessment of research biases, and often a statistical meta-analysis of trial results. SRs inform all levels of healthcare, from national policies and guidelines to bedside decisions. But the expanding primary research base has made producing and maintaining SRs increasingly onerous (Bastian et al., 2010; Wallace et al., 2013). Identifying, extracting, and combining evidence from free-text articles describing RCTs is difficult, time-consuming, and laborious. One estimate suggests that a single SR requires thousands of person hours (Allen and Olkin, 1999); and a recent analysis suggests it takes an average of nearly 70 weeks to publish a review (Borah et al., 2017). This incurs huge financial cost, particularly because reviews are performed by highly-trained persons.

To keep SRs current with the literature then we must develop new methods to expedite evidence synthesis. Specifically, we need tools that can help identify, extract, assess and summarize evidence relevant to specific clinical questions from free-text articles describing RCTs. Toward this end, this paper describes RobotReviewer (RR; Figure 1), an open-source system that automates aspects...
of the data-extraction and synthesis steps of a systematic review using novel NLP models.

2 Overview of RobotReviewer (RR)
RR is a web-based tool which processes journal article PDFs (uploaded by end-users) describing the conduct and results of related RCTs to be synthesised. Using several machine learning (ML) data-extraction models, RR generates a report summarizing key information from the RCTs, including, e.g., details concerning trial participants, interventions, and reliability. Our ultimate goal is to automate the extraction of the full range of variables necessary to perform evidence synthesis. We list the current functionality of RR and future extraction targets in Table 1.

RR comprises several novel ML/NLP components that target different sub-tasks in the evidence synthesis process, which we describe briefly in the following section. RR provides access to these models both via a web-based prototype graphical interface and a REST API service. The latter provides a mechanism for integrating our models with existing software platforms that process biomedical texts generally and that facilitate reviews specifically (e.g., Covidence2). We provide a schematic of the system architecture in Figure 2. We have released the entire system as open source via the GPL v 3.0 license. A live demonstration version with examples, a video, and the source code is available at our project website.3

3 Tasks and Models
We now briefly describe the tasks RR currently automates and the ML/NLP models that we have developed and integrated into RR to achieve this.

3.1 Risks of Bias (RoB)
Critically appraising the conduct of RCTs (from the text of their reports) is a key step in evidence synthesis. If a trial does not rigorously adhere to a well-designed protocol, there is a risk that the results exhibit bias. Appraising such risks has been formalized into the Cochrane4 Risk of Bias (RoB) tool (Higgins et al., 2011). This defines several ‘domains’ with respect to which the risk of bias is to be assessed, e.g., whether trial participants were adequately blinded.

EBM aims to make evidence synthesis transparent. Therefore, it is imperative to provide support for one’s otherwise somewhat subjective appraisals of risks of bias. In practice, this entails extracting quotes from articles supporting judgments, i.e. rationales (Zaidan et al., 2007). An automated system needs to do the same. We have therefore developed models that jointly (1) categorize articles as describing RCTs at ‘low’ or ‘high/unknown’ risk of bias across domains, and, (2) extract rationales supporting these categorizations (Marshall et al., 2014; Marshall et al., 2016; Zhang et al., 2016).

We have developed two model variants for automatic RoB assessment. The first is a multi-task (across domains) linear model (Marshall et al., 2014). The model induces sentence rankings (w.r.t. to how likely they are to support assessment for a given domain) which directly inform the overall RoB prediction through ‘interaction’ features (interaction of n-gram features with whether identified as rationale [yes/no]).

To assess the quality of extracted sentences, we conducted a blinded evaluation by expert systematic reviewers, in which they assessed the quality of manually and automatically extracted sentences. Sentences extracted using our model were scored comparably to those extracted by human reviewers (Marshall et al., 2016). However, the accuracy of the overall classification of articles as describing high/unclear or low risk RCTs achieved by our model remained 5-10 points lower than that achieved in published (human authored) SRs (estimated using articles that had been independently assessed in multiple SRs).

We have recently improved overall document classification performance using a novel variant of Convolutional Neural Networks (CNNs) adapted for text classification (Kim, 2014; Zhang and Wallace, 2015). Our model, the ‘rationale-augmented CNN’ (RA-CNN), explicitly identifies and up-weights sentences likely to be rationales. RA-CNN induces a document vector by taking a weighted sum over sentence vectors (output from a sentence-level CNN), where weights are set to reflect the predicted probability of sentences being rationales. The composite document vector is fed through a softmax layer for overall article classification. This model achieved gains of 1-2% abso-
3.2 PICO

The Population, Interventions/Comparators and Outcomes (PICO) together define the clinical question addressed by a trial. Characterising and representing these is therefore an important aim for automating evidence synthesis.

3.2.1 Extracting PICO sentences

Past work has investigated identifying PICO elements in biomedical texts (Demner-Fushman and Lin, 2007; Boudin et al., 2010). But these efforts have largely considered only article abstracts, limiting their utility: not all clinically salient data is always available in abstracts. One exception to this is a system called ExaCT (Kiritchenko et al., 2010), which does operate on full-texts, although...
Table 1: Typical variables required for an evidence synthesis (Centre for Reviews and Dissemination, 2009), and current RR functionality. Text: extracted text snippets describing the variable (e.g. ‘The randomization schedule was produced using a statistical computer package’). Structured: translation to e.g., standard bias scores or medical ontology concepts.

| General                              | Text | Structured | Extraction type | Text | Structured |
|--------------------------------------|------|------------|-----------------|------|------------|
| Record number                        | ✓    |            | Intervention and setting | Setting | ✓          |
| Author                               | ✓    |            |                  |       |            |
| Article title                        | ✓    |            |                  |       |            |
| Citation                             | ✓    |            |                  |       |            |
| Type of Publication                  | ✓    |            |                  |       |            |
| Country of origin                    | ✓    |            |                  |       |            |
| Source of funding                    | ✓    |            |                  |       |            |
| Study characteristics                | ✓    | ✓          |                  |       |            |
| Aims/objectives                      | ✓    |            |                  |       |            |
| Study design                         | ✓    | ✓          |                  |       |            |
| Inclusion criteria                   | ✓    |            |                  |       |            |
| Randomization/blinding               | ✓    | ✓          |                  |       |            |
| Unit of allocation                   | ✓    |            |                  |       |            |
| Participants                         |      |            |                  |       |            |
| Age                                  | ✓    |            |                  |       |            |
| Gender                               | ✓    |            |                  |       |            |
| Ethnicity                            | ✓    |            |                  |       |            |
| Socio-economic status                | ✓    |            |                  |       |            |
| Disease characteristics              | ✓    | ✓          |                  |       |            |
| Co-morbidities                       | ✓    |            |                  |       |            |

3.2.2 PICO embeddings

We have begun to explore learning dense, low-dimensional embeddings of biomedical abstracts specific to each PICO dimension. In contrast to monolithic document embedding approaches, such as doc2vec (Le and Mikolov, 2014), PICO embeddings are an example of disentangled representations.

Briefly, we have developed a neural approach which assumes access to manually generated free-text aspect summaries (here, one per PICO element) with corresponding documents (abstracts). The objective is to induce vector representations (via an encoder model) of abstracts and aspect summaries that satisfy two desiderata. (1) The embedding for a given abstract/aspect should be close to its matched aspect summary; (2) but far from the embeddings of aspect summaries for other abstracts, specifically those which differ with respect to the aspect in question.

To train this model, we used data recorded for previously conducted SRs to train our embedding model. Specifically we collected...
30,000+ abstract/aspect summary pairs stored in the Cochrane Database of Systematic Reviews (CDSR). We have demonstrated that the induced aspect representations improve performance an information retrieval task for EBM: ranking RCTs relevant to a given systematic review.\(^5\)

![Figure 5: PICO embeddings. Here, a mouse-over event has occurred on the point corresponding to Humiston et al. in the intervention embedding space, triggering the display of the three uni-/bi-grams that most excited the encoder model.](image)

For RR, we incorporate these models to induce abstract representations and then project these down to two dimensions using a PCA model pretrained on the CDSR. We then present a visualisation of study positions in this reduced space, thus revealing relative similarities and allowing one, e.g., to spot apparently outlying RCTs. To facilitate interpretation, we display the uni and bi-grams most activated for each study by filters in the learned encoder model on mouse-over. Figure 5 shows such an example. We are actively working to refine our approach to further improve the interpretability of these embeddings.

### 3.3 Study design

RCTs are regarded as the gold standard for providing evidence on the effectiveness of health interventions (Chalmers et al., 1993) Yet these articles form a small minority of the available medical literature. We employ an ensemble classifier, combining multiple CNN models, Support Vector Machines (SVMs), and which takes account of meta-data obtained from PubMed. Our evaluation on an independent dataset has found this approach achieves very high accuracy (area under the Receiver Operating Characteristics curve $= 0.987$), outperforming previous ML approaches and manually created boolean filters.\(^6\)

### 4 Discussion

We have presented RobotReviewer, an open-source tool that uses state-of-the-art ML and NLP to semi-automate biomedical evidence synthesis. RR incorporates the underlying trained models with a prototype web-based user interface, and a REST API that may be used to access the models. We aim to continue adding functionality to RR, automating the extraction and synthesis of additional fields: particularly structured PICO data, outcome statistics, and trial participant flow. These additional data points would (if extracted with sufficient accuracy) provide the information required for statistical synthesis.

For example, for assessing bias, RR is competitive with, but modestly inferior to the accuracy of a conventional manually produced systematic review (Marshall et al., 2016) We therefore recommended that RR be used as a time-saving tool for manual data extraction, or that one of two humans in the conventional data-extraction process be replaced by the automated process.

However, there is an increasing need for methods that trade a small amount of accuracy for increased speed (Tricco et al., 2015). The opportunity cost of maintaining current rigor in SRs is vast: reviews do not exist for most clinical questions (Smith, 2013), and most reviews are out of date soon after publication (Shojania et al., 2007).

RR used in a fully automatic workflow (without manual checks) might improve upon relying on the source articles alone, particularly given those in clinical practice are unlikely to have time to read the full texts. To explore how automation should be used in practice, we plan to experimentally evaluate RR in real-world use: in terms of time saved, user experience, and the resultant review quality.

### Acknowledgments

RobotReviewer is supported by the National Library of Medicine (NLM) of the National Institutes of Health (NIH) under award R01LM012086. The content is solely the

\(^{5}\)Under review; preprint available at [http://www.byronwallace.com/static/articles/PICO-vectors-preprint.pdf](http://www.byronwallace.com/static/articles/PICO-vectors-preprint.pdf).

\(^{6}\)Under review; pre-print available at [https://kclpure.kcl.ac.uk/portal/iain.marshall.html](https://kclpure.kcl.ac.uk/portal/iain.marshall.html).
responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. IJM is supported by the Medical Research Council (UK), through its Skills Development Fellowship program (MR/N015185/1).

References

IE Allen and I Olkin. 1999. Estimating time to conduct a meta-analysis from number of citations retrieved. The Journal of the American Medical Association (JAMA), 282(7):634–635.

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

R Borah, AW Brown, PL Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the Prospero registry. BMJ open, 7(2):e012545.

F Boudin, J-Y Nie, and M Dawes. 2010. Positional language models for clinical information retrieval. In EMNLP, pages 108–115.

Centre for Reviews and Dissemination. 2009. Systematic reviews: CRD’s guidance for undertaking reviews in health care. University of York, York.

I Chalmers, M Enkin, and MJNC Keirse. 1993. Preparing and updating systematic reviews of randomized controlled trials of health care. Milbank Q., 71(3):411.

J de Borda. 1784. A paper on elections by ballot. Sommerlad F, McLean I (1989, eds) The political theory of Condorcet, pages 122–129.

D Demner-Fushman and J Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics, 33(1):63–103.

JPT Higgins, DG Altman, PC Gøtzsche, P Jüni, D Moher, AD Oxman, J Savović, KF Schulz, L Weeks, and JAC Sterne. 2011. The Cochrane Collaborations tool for assessing risk of bias in randomised trials. BMJ, 343:d5928.

Y Kim. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

S Kiritchenko, B de Bruijn, S Carini, J Martin, and I Sim. 2010. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC medical informatics and decision making, 10(1):56.

J Kuiper, IJ Marshall, BC Wallace, and MA Swertz. 2014. Spá: A web-based viewer for text mining in evidence based medicine. In ECML-PKDD, pages 452–455. Springer.

QV Le and T Mikolov. 2014. Distributed representations of sentences and documents. In ICML, volume 14, pages 1188–1196.

IJ Marshall, J Kuiper, and BC Wallace. 2014. Automating risk of bias assessment for clinical trials. In ACM-BCR, pages 88–95.

IJ Marshall, J Kuiper, and BC Wallace. 2016. RobotReviewer: Evaluation of a System for Automatically Assessing Bias in Clinical Trials. Journal of the American Medical Informatics Association (JAMIA), 23(1):193–201.

M Mintz, S Bills, R Snow, and D Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In IJCNLP, pages 1003–1011.

AJ Moss and FI Marcus. 2017. Changing times in cardiovascular publications: A commentary. Am. J. Med., 130(1):11–13, January.

DL Sackett. 1997. Evidence-based medicine: how to practice and teach EBM. WB Saunders Company.

K G Shojiangia, M Sampson, M T Ansari, J Ji, C Garrity, T Rader, and D Moher. 2007. Updating Systematic Reviews. Technical Review No. 16. Agency for Healthcare Research and Quality (US), 1 September.

Richard Smith. 2013. The Cochrane collaboration at 20. BMJ, 347:f7383, 18 December.

Andrea C Tricco, Jesmin Antony, Wasifa Zarin, Lisa Strifler, Marco Ghassemi, John Ivory, Laure Perrier, Brian Hutton, David Moher, and Sharon E Straus. 2015. A scoping review of rapid review methods. BMC Med., 13:224, 16 September.

BC Wallace, IJ Dahabreh, CH Schmid, J Lau, and TA Trikalinos. 2013. Modernizing the systematic review process to inform comparative effectiveness: tools and methods. Journal of Comparative Effectiveness Research (JCER), 2(3):273–282.

BC Wallace, J Kuiper, A Sharma, M Zhu, and IJ Marshall. 2016. Extracting PICO Sentences from Clinical Trial Reports using Supervised Distant Supervision. Journal of Machine Learning Research, 17(132):1–25.

O Zaidan, J Eisner, and CD Piatko. 2007. Using “annotator rationales” to improve machine learning for text categorization. In NAACL, pages 260–267.

Y Zhang and B Wallace. 2015. A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

Y Zhang, IJ Marshall, and BC Wallace. 2016. Rationale-Augmented Convolutional Neural Networks for Text Classification. In EMNLP, pages 795–804.