A distantly supervised dataset for automated data extraction from diagnostic studies

Norman, C.; Leeflang, M.; Spijker, R.; Kanoulas, E.; Névéol, A.

DOI
10.18653/v1/W19-5012

Publication date
2019

Document Version
Final published version

Published in
SIGBioMed Workshop on Biomedical Natural Language Processing

A distantly supervised dataset for automated data extraction from diagnostic studies

Christopher Norman,[1,2] **Mariska Leeflang,**[2] **René Spijker,**[3,4]
Evangelos Kanoulas,[5] and **Aurélie Névéol**[1]
[1] LIMSI, CNRS, Université Paris-Saclay
[2] KEBB, Amsterdam Public Health, Amsterdam UMC, University of Amsterdam
[3] Medical Library, Amsterdam Public Health, Amsterdam UMC, University of Amsterdam
[4] Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, UMCU, Utrecht University
[5] Informatics Institute and Amsterdam Business School, University of Amsterdam
norman@limsi.fr, m.m.leeflang@amc.uva.nl,
R.Spijker-2@umcutrecht.nl, E.Kanoulas@uva.nl,
neveol@limsi.fr

Abstract

Systematic reviews are important in evidence based medicine, but are expensive to produce. Automating or semi-automating the data extraction of index test, target condition, and reference standard from articles has the potential to decrease the cost of conducting systematic reviews of diagnostic test accuracy, but relevant training data is not available. We create a distantly supervised dataset of approximately 90,000 sentences, and let two experts manually annotate a small subset of around 1,000 sentences for evaluation. We evaluate the performance of BioBERT and logistic regression for ranking the sentences, and compare the performance for distant and direct supervision. Our results suggest that distant supervision can work as well as, or better than direct supervision on this problem, and that distantly trained models can perform as well as, or better than human annotators.

1 Background

Evidence based medicine is founded on systematic reviews, which synthesize all published evidence addressing a given research question. By examining multiple studies, a systematic review can examine the variation between different studies, the discrepancies between them, as well as look at the quality of evidence across studies in a way that is difficult in a single trial. Since a systematic review needs to consider the entire body of published literature, producing a systematic review is expensive and labor-intensive process, often requiring months of manual work (O'Mara-Eves et al., 2015).

To ensure that the results of a systematic review are as comprehensive and unbiased as possible, their production follows a strict and systematic procedure. To catch and resolve disagreements, all steps of the process are performed in duplicate by at least two reviewers. There have recently been examples of systematic reviews using automation in a limited capacity (Bannach-Brown et al., 2019; Przybyła et al., 2018; Lerner et al., 2019), but the impact of automation on the reliability of systematic reviews is not yet fully understood. Automation is not part of accepted practice in current guidelines (De Vet et al., 2008).

After a set of potentially included studies have been identified, systematic reviewers complete a so-called *data extraction form* for each study. These forms comprise a semi-structured summary of the studies, identifying and extracting a consistent, pre-specified set of data items from abstracts or full-text articles in a coherent format (see the left part of Table 1 for sample exerpts). The coherent format allows the data from the studies to be synthesized qualitatively or quantitatively to address the research question of the review.

In this study we will focus on systematic reviews of diagnostic test accuracy (DTA), which examine the accuracy of tests and procedures for diagnosing medical conditions, and which have seen little attention in previous literature on automated data extraction. To compare and synthesize results across studies, reviewers extract diagnostic accuracy from each study, but also determine the *index test* (the specific diagnostic test or procedure that is being tested), what *target condition* the test seeks to diagnose, and the *reference standard* (the diagnostic test or procedure that is being used as the gold standard) (see Fig 1 for an example). These data must be determined for each study to know if the diagnostic accuracy in different studies can be compared.

Proceedings of the BioNLP 2019 workshop, pages 105–114
Florence, Italy, August 1, 2019.

Original		Cleaned	
Review: CD008892, study: Dutta 2006			
Index tests:	TUBEX Typhidot	Index test: Index test:	TUBEX Typhidot
Target condition and reference standard(s):	Target condition Salmonella Typhi Reference standard: peripheral blood culture	Target condition: Target condition: Reference standard:	Salmonella Typhi Typhoid fever Peripheral blood culture
Note: These are the data items corresponding to the example text in Fig. 1			
Review: CD010502, study: Schwartz 1997b			
Index tests:	Throat swab: not reported Commercial name of the RADT: QuickVue In-Line Strep A (Quidel) Type of RADT: EIA	Index test: Index test: Index test:	QuickVue In-Line Strep A EIA ELISA Immunoassays
Target condition and reference standard(s):	See Schwartz 1997a	Target condition: Target condition: Reference standard: Reference standard:	Group A streptococcus Group A streptococcal infection Microbial culture Bacterial culture
Note: Neither the target condition nor the reference standard were mentioned in the table for Schwartz 1997a, but assumed the same for all studies included in this systematic review (they were presumably considered obvious by the authors).			

Table 1: Examples of raw data from three data extractions forms in unstructured format (left) and a structured summary of the data intended for distant supervision by pattern matching (right).

Although typhoid fever is confirmed by culture of Salmonella enterica serotype Typhi , rapid and simple diagnostic serologic tests would be useful in developing countries. We examined the performance of Widal test in a community field site and compared it with Typhidot and Tubex tests for diagnosis of typhoid fever . [...] Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the 3 serologic tests were calculated using culture-confirmed typhoid fever cases as "true positives" and paratyphoid fever and malaria cases as "true negatives". [...] The sensitivity, specificity, PPV, and NPV of Typhidot and Tubex were not better than Widal test. There is a need for more efficient rapid diagnostic test for typhoid fever especially during the acute stage of the disease. Until then, culture remains the method of choice.

Legend: Target condition Index Test Reference standard

Figure 1: Examples of data items highlighted in text, with supporting context underlined. Based on the manual annotation by one expert (ML) on a study by Dutta et al. (2006).

1.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model that is unsupervisedly pretrained on a large general language corpus, then supervisedly fine-tuned on natural language processing tasks (Devlin et al., 2018). Despite being a general approach, with almost no task-specific modifications, BERT achieves state-of-the-art performance across a number of natural language processing tasks, including text classification, question answering, inference, and named entity recognition.

Pretrained models like BERT can be used directly for screening automation or automated data extraction. However, by default BERT is trained on a general language corpus, which differs radically in word choice and grammar from the special language found in biomedicine and related fields (Sager et al., 1980). Pretraining on biomedical corpora, rather than general corpora, has been demonstrated to improve performance on several biomedical natural language processing tasks (Lee et al., 2019; Beltagy et al., 2019; Si et al., 2019).

1.2 Objectives

In this study we seek to:

1. Construct a dataset for training machine learning models to identify and extract data from full-text articles on diagnostic test accuracy. We focus on the target condition, index test, and reference standard.

2. Train models to identify specific data items in full-text articles on diagnostic test accuracy

One of the main aims of our study is to determine how such a dataset should be constructed to allow for training well performing models. In particular, do we need directly supervised data, or can we build reliable models with distantly supervised data? If we do need directly supervised data, how much is necessary?

2 Related Work

There have been attempts to extract several types of data relevant to systematic reviews, most notably extracting PICO[1] statements from article text (Wallace et al., 2016; Kiritchenko et al., 2010; Kim et al., 2011; Nye et al., 2018). Other data items include background and study design (Kim et al., 2011), as well as automatically performing risk of bias assessments (Marshall et al., 2014). There is also a recent TAC track for data extraction in systematic reviews of environmental agents.[2] Similarly, previous work by Kiritchenko et al. (2010) aimed to extract 21 different kinds of data from articles, including treatment name, sample size, as well as the primary and secondary outcome from article text. Furthermore, the key criterion for extraction in a systematic review is not the actual data, but the context it appears in. For instance, both intervention studies and a diagnostic studies have target conditions, but these refer to different things: the intervention study seek to *treat* the condition while the diagnostic study seeks to *diagnose* it. As a consequence, in an intervention study the inclusion criterion often mentions the disease, while in a diagnostic study inclusion criteria may mention symptoms rather than the actual disease. This means that a data extraction system trained on interventions may not work as well (or at all) for systematic reviews of diagnostic test accuracy, even though it may seem that the same data is extracted in both. Furthermore, unlike the data required in diagnostic reviews, many previously considered data items are mentioned once in articles, often using formulaic expressions (e.g. sex, blinding, randomization).

Conventional methods for automated data extraction split articles into sentences and classify these individually using conventional machine learning methods (e.g. SVM, Naive Bayes) (Jonnalagadda et al., 2015), or label spans in the text and classify these using sequence tagging (e.g. CRF, LSTM) (Nye et al., 2018).

Despite the body of previous work on automation, many data items relevant to systematic reviews have been overlooked. A 2015 systematic review of data extraction found 26 articles describing the attempted extraction of 52 different data items, but almost all focused on interventions (Jonnalagadda et al., 2015). No study considered any data item specific to diagnostic studies, except for general data items common to both interventions and diagnostic studies, such as age, sex, blinding, or the generation of random allocation sequences. The likely reason for this is that traditional data extraction systems require bespoke training data for each particular data item to extract, which is generally only available through expensive, manual annotation by experts.

A cheaper way to construct datasets for data extraction is to use distant supervision, where the dataset is annotated per article or per review, rather than per sentence or per text span. Supervised methods are then trained on fuzzy annotations derived heuristically for each sentence. For instance, Wallace et al. (2016) used supervised distant supervision to learn to identify PICO statements in full text, and Marshall et al. (2014) used supervised distant learning with SVMs to identify risk of bias assessments.

There is likely a trade-off between quality and data size. All else being equal, direct supervision is generally better than distant supervision (distantly supervised training data adds a source of noise not present for direct supervision). At the same time, it may not be feasible for experts to annotate large amounts of data. Crowd-sourcing is sometimes used as an alternative to a group of known experts, but if a high degree of expertise is necessary to annotate, crowd-sourcing may not give sufficient guarantees about the expertise of the annotators.

3 Material

We used data from a previous dataset, the LIMSI-Cochrane dataset (Norman et al., 2018),[3] to identify references included in previous systematic reviews of diagnostic test accuracy. The LIMSI-Cochrane dataset comprises 1,738 references to DTA studies from 63 DTA systematic reviews. The dataset includes the data extraction forms for each

[1] Population, intervention, control group, and outcome.

[2] https://tac.nist.gov/2018/SRIE/index.html

[3] DOI: 10.5281/zenodo.1303259

Target Condition

	pos	neg	total
Distant train	11,336	63,204	74,540
test	2,884	13,572	16,456
total	14,220	77,776	90,996
Annotated by ML	92	889	981
Annotated by RS	48	983	1,031

Index Test

	pos	neg	total
Distant train	14,280	63,343	77,623
test	2,675	13,992	16,667
total	16,955	77,335	94,290
Annotated by ML	93	888	981
Annotated by RS	87	944	1,031

Reference Standard

	pos	neg	total
Distant train	7,006	56,638	63,644
test	1,258	14,602	15,860
total	8,264	71,240	79,504
Annotated by ML	26	955	981
Annotated by RS	26	1,005	1,031

Table 2: The number of sentences in our dataset, broken into distantly annotated training and test sets, as well as a manually annotated subset. Distant annotations for each data type were not available for all studies, and the total number of labelled sentences are therefore different for each data type.

study completed by the systematic review authors.

The dataset itself does not contain abstracts or full-texts, but include identifiers in the form of PubMed IDs and DOIs which can be used to retrieve abstracts or full-texts.

We used the reference identifiers (PMID and/or DOI) taken from the LIMSI-Cochrane dataset to construct a collection of PDF articles. We used EndNote's 'find full text' feature, which retrieves PDF articles from a range of publishers.[4] The PDF articles were then converted into XML format using Grobid (Lopez, 2009).

We randomly split the dataset into dedicated training and evaluation sets, where we used 48 of the systematic reviews as the training set, and we kept the remaining 15 systematic reviews for evaluation. For each of the 15 systematic reviews in the evaluation set, we randomly selected one article to be annotated manually. The remaining articles in the evaluation set were not used for training, since training and testing on the same systematic review is known to overestimate classification performance (Cohen, 2008). The goal of this work is to learn the semantics of the context, rather than the semantics of particular terms, and these contexts should be consistent across reviews.

3.0.1 Distant annotation

The data forms from the systematic reviews were intended to be read by and be useful to the human systematic review authors. The contents are therefore usually semi-structured rather than structured, and will include different kinds of data depending on what is relevant to the systematic review (see Table 1).

We create a dataset of distant annotations from the LIMSI-Cochrane dataset by manually converting the semi-structured data into structured data items, and by ensuring that these items can be found in the corresponding article using pattern matching (see Table 1).

We split each of the XML documents into sentences using the nltk sentence splitter.[5] The sentences are then divided into positive and negative depending on whether the relevant data items occur as a partial match in the sentence. Partial matches were calculated using *tf·idf* cosine similarity between the data item and the sentence, where we took the 20 top ranking sentences for each pair of data item and article, with a similarity score of 0.1 or higher. We chose 20 as a target number of sentences since we felt this was a reasonable upper limit on the number of relevant sentences in a single article. We added an absolute threshold of 0.1 to keep the system from annotating obviously non-relevant sentences (scores close to zero) when no matches could be found in the article. For articles that have multiple data items we used the concatenation of all data items. For example, in Table 1, the data items for 'Schwartz 1997b' would be: target condition: 'Group A streptococcus; Group A streptococcal infection', index test: 'QuickVue In-Line Strep A; EIA; ELISA Immunoassays', and reference standard 'Microbial culture; Bacterial culture'.

We excluded all articles where the data items were not provided in the data form (because the reviewers did not extract this data), or where data forms were missing from the systematic review. Since we do not know which sentences were relevant or not in these articles we did not use these

[4] https://endnote.com/

[5] https://www.nltk.org/

articles as either positive or negative data. As a consequence the total amount of sentences differ for the target condition, index test and reference standard.

We repeated the matching precedure for the target condition, the index test and the reference standard, resulting in three distinct datasets.

3.0.2 Expert annotation

We randomly split the evaluation set into three sets of five systematic reviews. Two experts (ML and RS) on systematic reviews of diagnostic test accuracy manually annotated the 15 articles by highlighting all sentences in the text that 1) mentions the target condition, index test, and reference standard 2) makes it clear that these are the target condition, index test and reference standard, and 3) do not simply mention these same items in an unrelated context. The annotation instructions were written and adjusted twice to remove ambiguity, and the reasons for disagreement were discussed and resolved after two rounds of annotation. As a compromise between getting more data and being able to use the agreement between the experts as baseline for the performance, one expert annotated the first five studies, the second expert annotated the next five studies, and both annotated the last five studies.

4 Method

We construct three pipelines, one for each of the target condition, index test, and reference standard, and we train and evaluate these separately.

We varied our experiments in three dimensions: We tried A) two machine learning algorithms, B) two levels of preprocessing, and C) distantly supervised training data versus directly supervised training data. The directly and distantly supervised models were evaluated on the same data.

4.0.1 A1: BioBERT

We here used a pointwise learning-to-rank approach, where we trained a sentence ranking model by using BioBERT, a version of BERT pretrained on PubMed and PMC (Lee et al., 2019), and fine-tuned the model by training it to regress probability scores. This model was thus trained to map sentences to relevance scores.

To train and evaluate, we used the default BERT setup for the GLUE datasets,[6] modified to output a relevance score rather than a binary value. We used default parameters.

4.0.2 A2: Logistic Regression

We here used a pairwise learning-to-rank approach, where we trained a logistic regression model using stochastic gradient descent (sklearn). As features we used 1) lowercased, *tf·idf* weighted word *n*-grams, 2) lowercased, binary word *n*-grams, 3) lowercased, *tf·idf* weighted, stemmed word *n*-grams, 4) lowercased, stemmed, binary word *n*-grams, as well as *i*) lowercased, *tf·idf* weighted character *n*-grams, and *ii*) *non*-lowercased, *tf·idf* weighted character *n*-grams. We used word *n*-grams up to length 3, and character *n*-grams up to length 6. The first set of features is intended to capture contextual information ('for the diagnosis of ...'); the second set of features is intended to capture medical technical terms, which are often distinctive at the morpheme level (e.g. 'ischemia', 'anemia'). We deliberately did not use stop-words, since doing so would discard almost all the contextual information. This results in a sparse feature matrix consisting of approximately 1.8 million features for the distantly supervised experiments, and approximately 300,000 features for the directly supervised experiments.

We handled class imbalance by setting the weight for the positive class to 80. This was previously determined to be a reasonable weight in experiments on screening automation in diagnostic test accuracy systematic reviews, a problem with similar class imbalance.

4.0.3 B1: Raw Sentences

Here we used the sentences as they appear in the articles.

4.0.4 B2: Sentences with UMLS Concepts

In this setup we used the *Unified Medical Language System*, a large ontology of medical concepts maintained by the National Library of Medicine (Bodenreider, 2004; Lindberg et al., 1993). We used MetaMap[7] to locate concept mentions in the sentences, and to replace these with their corresponding UMLS semantic types. For instance the sentence '*Typhoid fever is a febrile and often serious systemic illness caused by Salmonella enterica serotype Typhi*' was transformed into '*DSYN is a FNDG and TMCO serious DSYN caused by BACT enterica BACT*'.

[6] https://github.com/google-research/bert

[7] https://metamap.nlm.nih.gov/

Target condition

	Auto	ML	RS
Auto	1.00	0.07	0.04
ML	0.90	1.00	0.38
RS	1.00	0.62	1.00

Index test

	Auto	ML	RS
Auto	1.00	0.09	0.07
ML	1.00	1.00	0.61
RS	0.93	0.70	1.00

Reference standard

	Auto	ML	RS
Auto	1.00	0.01	0.03
ML	1.00	1.00	0.86
RS	1.00	0.40	1.00

Table 3: Agreement in terms of recall where columns are considered ground truth, e.g. annotator RS chose 62% of ML's annotations for the target condition.

4.0.5 C1: Directly Supervised Training

We here trained and evaluated on the articles manually annotated by our two experts (ML and RS), using leave-one-out cross-validation. In other words, to evaluate on each of the ten articles annotated by each annotator we used the remaining 9 articles annotated by the same expert as training data. This was done separately for each expert, and the annotations from the other expert was not used.

4.0.6 C2: Distantly Supervised Training

We here trained on the distant annotations from the 48 systematic reviews in the training set, and evaluated on the 15 manually annotated articles in the evaluation set, where each annotator provided annotation data for 10 articles (with a 5 article overlap). The articles used for evaluation were the same as in C1.

4.1 Evaluation

Since our model output ranked sentences, rather than a binary classification, we evaluated all experiments in terms of average precision.

As a comparison, we also evaluated the average precision using the ranking given by the other annotator. In plain language, we tried to evaluate how useful it would have been for the expert to highlight sentences for each other. The expert annotations were binary (Yes/No), rather than a ranking score, so we calculated the average precision by interpolating ties in the ranking.

5 Results

Out of the 1,738 references in the LIMSI-Cochrane dataset, 1152 had either a PMID or DOI assigned. EndNote was able to retrieve PDF articles for 666 of these references. A total of 90,996 sentences were distantly labeled for target condition, 94,290 sentences were distantly labeled for index test, and 79,504 sentences were distantly labeled for reference standard. The first annotator (ML) annotated 981 sentences and the second annotator (RS) annotated 1,031 sentences (Table 2).

We present the results of our algorithm evaluated on the annotations by ML in Table 4, and evaluated on the annotations by RS in Table 5.

The ranking performance exhibited large variations. Neither BioBERT or logistic regression were consistently better than the other, neither distant supervision or direct supervision were consistently better than the other, and neither raw sentence nor sentences augmented with UMLS concepts were consistently better than the other. For the target condition, the best performance was achieved by logistic regression on raw sentences using either distant or direct supervision, with a maximum at 0.412 compared to human performance at 0.376 and 0.386 respectively. For the index test, the performance fell within the range 0.344–0.468 compared to human performance at 0.525 and 0.516 respectively. For the reference standard, BioBERT exhibited substantially inferior results on the reference standard compared to logistic regression, while logistic regression performance fell within the range 0.345–0.467, compared to human performance at 0.267 and 0.381 respectively.

The performance also varied between systematic reviews, with consistently close to perfect performance on a few reviews (CD007394 and CD0008782), and consistently very low performance on a few (CD009647 and CD010339). These also correspond to the articles with the highest and lowest inter-annotator agreement. The consensus of the two experts is that CD010339 is not a diagnostic test accuracy study.

6 Discussion

Raw sentences worked consistently better for logistic regression on the target condition (8/8), and worked better than UMLS concepts as a general trend (20/24). While general concepts could theoretically improve performance by help-

	n pos	BioBERT				Logistic Regression				As ranked by the other expert (RS)
		Distant		Supervised		Distant		Supervised		
		Raw	UMLS	Raw	UMLS	Raw	UMLS	Raw	UMLS	
Target condition										
CD007394	1	1.000	0.500	0.143	0.250	1.000	0.500	1.000	0.500	0.500
CD007427	14	0.228	0.267	0.500	0.588	0.423	0.573	0.462	0.509	—
CD008054	10	0.197	0.353	0.060	0.182	0.167	0.118	0.170	0.148	—
CD008782	2	1.000	1.000	0.283	0.567	0.500	0.417	0.500	0.583	0.700
CD008892	29	0.182	0.274	0.384	0.247	0.368	0.439	0.290	0.333	0.338
CD009372	29	0.110	0.117	0.461	0.543	0.328	0.250	0.378	0.276	—
CD010339	16	0.192	0.179	0.642	0.513	0.537	0.432	0.482	0.495	0.154
CD010653	2	0.053	0.035	0.023	0.015	0.107	0.112	0.062	0.086	—
CD011420	6	0.070	0.074	0.239	0.175	0.189	0.138	0.254	0.157	0.190
mean:		0.336	0.311	0.304	0.342	0.402	0.331	0.400	0.343	0.376
Index test										
CD007394	2	1.000	1.000	0.643	0.361	0.750	0.500	0.583	0.583	1.000
CD007427	17	0.354	0.225	0.580	0.568	0.551	0.526	0.534	0.484	—
CD008054	10	0.388	0.305	0.449	0.281	0.170	0.161	0.195	0.218	—
CD008782	2	0.833	1.000	0.079	0.523	0.750	0.750	0.750	0.750	0.700
CD008892	34	0.342	0.473	0.458	0.391	0.471	0.484	0.496	0.529	0.524
CD009372	8	0.269	0.351	0.194	0.225	0.261	0.270	0.303	0.390	—
CD010339	1	0.167	0.050	0.067	0.067	0.071	0.100	0.013	0.017	0.010
CD011420	19	0.251	0.342	0.284	0.218	0.288	0.266	0.280	0.256	0.391
mean:		0.450	0.468	0.344	0.329	0.414	0.382	0.394	0.403	0.525
Reference standard										
CD007394	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
CD007427	2	0.145	0.032	0.081	0.034	0.052	0.037	0.035	0.041	—
CD008054	6	0.215	0.108	0.239	0.076	0.635	0.619	0.525	0.515	—
CD008892	13	0.112	0.097	0.152	0.154	0.408	0.351	0.264	0.255	0.201
CD009372	3	0.052	0.095	0.253	0.414	0.681	0.692	0.679	0.729	—
CD010653	1	0.020	0.016	0.020	0.059	0.029	0.034	0.067	0.067	—
CD011420	1	0.034	0.100	1.000	0.014	1.000	1.000	0.500	0.500	0.333
mean:		0.097	0.075	0.291	0.125	0.467	0.455	0.345	0.351	0.267

Table 4: Average precision results for the 8 different machine learning models on the data annotated by the first annotator (ML), compared to the performance of an independent human expert (annotator RS). The 'Raw' columns denote results for models trained and evaluated on raw sentences. The 'UMLS' columns denote results for models trained and evaluated on sentences where the concept mentions have been replaced with their corresponding UMLS semantic types. The '*n* pos' column denotes the number of positive sentences labeled by ML for each article. Rows were omitted for which no sentences were labeled positive. In the baseline results, cells are marked '—' if the article was not annotated by the other expert (RS).

ing the models generalize, this may also remove important semantic information from the sentences, keeping the models from ranking accurately. We also note that BioBERT already encodes a language model (similar to word embeddings), and concepts may therefore be unhelpful for the model.

BioBERT performed consistently better than logistic regression on the index test when using distant supervision (4/4), but not when using direct supervision (0/4). Logistic regression performed consistently better than BioBERT on both the target condition and the reference standard (16/16). On the reference standard the difference in performance is substantial, with BioBERT scoring very poorly, and logistic regression performing much better than human performance. The reason for BioBERT's poor performance on the reference standard may be due to the relative sparsity of the annotations for this subtask (see Table 2).

Distant supervision was consistently on par with or better than direct supervision. The top performing models also outperformed the human annotators on the target condition and the reference standard, and came comparatively close on the index test (0.468 versus 0.525 and 0.444 versus 0.516).

6.1 Limitations

We only manually annotated a small sample of the dataset. The small size is further compounded by problems with converting PDF to text, which may also bias the training and evaluation in favor of articles where the conversion works better (mainly articles from big publishers).

The dataset was constructed from articles included in previous systematic reviews of diagnostic test accuracy. These include articles that con-

	n pos	BioBERT				Logistic Regression				As ranked by the other expert (ML)
		Distant		Supervised		Distant		Supervised		
		Raw	UMLS	Raw	UMLS	Raw	UMLS	Raw	UMLS	
Target condition										
CD007394	2	0.750	0.500	0.667	0.040	0.833	0.500	1.000	0.833	0.667
CD008081	8	0.136	0.198	0.213	0.371	0.504	0.380	0.394	0.388	—
CD008760	5	0.200	0.144	0.283	0.163	0.252	0.300	0.481	0.300	—
CD008782	1	1.000	1.000	0.500	1.000	0.500	0.333	1.000	0.500	0.500
CD008892	15	0.170	0.270	0.088	0.342	0.440	0.505	0.667	0.542	0.564
CD009647	2	0.036	0.021	0.021	0.047	0.020	0.026	0.012	0.023	—
CD010339	2	0.061	0.040	0.066	0.062	0.044	0.029	0.063	0.023	0.019
CD010360	2	0.089	0.080	0.093	0.261	0.181	0.083	0.244	0.064	—
CD010705	7	0.189	0.269	0.127	0.341	0.382	0.359	0.254	0.402	—
CD010420	4	0.036	0.044	0.209	0.097	0.210	0.214	0.302	0.132	0.178
mean:		0.267	0.257	0.227	0.273	0.337	0.273	0.412	0.321	0.386
Index test										
CD007394	2	1.000	1.000	0.417	0.393	0.750	0.500	0.700	0.750	1.000
CD008081	11	0.464	0.229	0.463	0.454	0.431	0.412	0.394	0.447	—
CD008760	9	0.357	0.411	0.512	0.475	0.457	0.470	0.481	0.476	—
CD008782	1	1.000	1.000	1.000	0.500	1.000	1.000	1.000	1.000	0.500
CD008892	27	0.499	0.539	0.717	0.758	0.740	0.666	0.667	0.474	0.692
CD009647	1	0.053	0.015	0.020	0.006	0.006	0.009	0.012	0.040	—
CD010339	6	0.085	0.054	0.040	0.047	0.053	0.041	0.063	0.047	0.058
CD010360	8	0.154	0.119	0.233	0.278	0.222	0.202	0.244	0.242	—
CD010705	14	0.599	0.533	0.292	0.270	0.352	0.327	0.254	0.327	—
CD010420	8	0.234	0.296	0.280	0.251	0.259	0.235	0.302	0.257	0.328
mean:		0.444	0.420	0.397	0.343	0.427	0.386	0.412	0.406	0.516
Reference standard										
CD008081	3	0.254	0.132	0.134	0.177	0.867	0.698	1.000	1.000	—
CD008760	2	0.101	0.553	0.529	0.013	0.667	0.833	0.667	0.833	—
CD008892	11	0.110	0.212	0.283	0.108	0.356	0.286	0.334	0.225	0.417
CD010339	1	0.012	0.010	0.029	0.009	0.224	0.031	0.071	0.028	n/a
CD010360	1	0.200	0.037	0.111	0.038	0.810	0.023	0.167	0.143	—
CD010705	5	0.150	0.152	0.194	0.086	0.224	0.122	0.172	0.125	—
CD010420	3	0.167	0.347	0.358	0.019	0.810	0.806	0.692	0.694	0.345
mean:		0.142	0.206	0.234	0.064	0.428	0.400	0.443	0.435	0.381

Table 5: Average precision results for the 8 different machine learning models on the data annotated by the second annotator (RS), compared to the performance of an independent human expert (annotator ML). Abbreviations are the same as in Table 4. In the baseline results, cells are marked '—' if the article was not annotated by the other expert (ML).

tain diagnostic results, while not being diagnostic test accuracy studies. Arguably, these should be excluded from training or evaluation, and possibly even from the dataset.

7 Conclusions

Our results suggest that distant supervision is sufficient to train models to identify target condition, index test, and reference standard in diagnostic articles. Our results also suggest that such models can perform on par with human annotators.

We constructed a dataset of full-text articles of diagnostic test accuracy studies, with distant annotations for target condition, index test and reference standard, that can be used to train machine learning models. We also provide a subset of the data manually annotated by experts for evaluation. Our dataset cannot be publicly distributed due to copyright restrictions, but will be available upon request. We also plan to distribute the code for the distant annotations and data preprocessing, as well as the cleaned data extraction forms.

7.1 Future Work

The dataset is being updated, and we plan to increase the amount of manually annotated data to improve the statistical reliability of the experiments. We also plan to let all experts annotate the same articles to simplify the comparisons.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew SC Rice, Sophia Ananiadou, Jing Liao, and Malcolm Robert Macleod. 2019. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic reviews*, 8(1):23.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. SciBERT: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Aaron M Cohen. 2008. Optimizing feature representation for automated systematic review work prioritization. *AMIA Annual Symposium proceedings*, pages 121–5.

HCW De Vet, A Eisinga, II Riphagen, B Aertgeerts, D Pewsner, and R Mitchell. 2008. Chapter 7: searching for studies. *Cochrane handbook for systematic reviews of diagnostic test accuracy version 0.4 [updated September 2008]. The Cochrane Collaboration*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shanta Dutta, Dipika Sur, Byomkesh Manna, Bhaswati Sen, Alok Kumar Deb, Jacqueline L Deen, John Wain, Lorenz Von Seidlein, Leon Ochiai, John D Clemens, et al. 2006. Evaluation of new-generation serologic tests for the diagnosis of typhoid fever: data from a community-based surveillance in calcutta, india. *Diagnostic microbiology and infectious disease*, 56(4):359–365.

Siddhartha R Jonnalagadda, Pawan Goyal, and Mark D Huffman. 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic reviews*, 4(1):78.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, page S5. BioMed Central.

Svetlana Kiritchenko, Berry de Bruijn, Simona Carini, Joel Martin, and Ida Sim. 2010. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10:56.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Ivan Lerner, Perrine Créquit, Philippe Ravaud, and Ignacio Atal. 2019. Automatic screening using word embeddings achieved high sensitivity and workload reduction for updating living network meta-analyses. *Journal of clinical epidemiology*, 108:86–94.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of Medical Informatics*, 2(01):41–51.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer.

Iain J. Marshall, Joël Kuiper, and Byron C. Wallace. 2014. Automating risk of bias assessment for clinical trials. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '14*, pages 88–95.

Christopher Norman, Mariska Leeflang, and Aurélie Névéol. 2018. Data extraction and synthesis in systematic reviews of diagnostic test accuracy: A corpus for automating and evaluating the process. In *AMIA Annual Symposium Proceedings*, volume 2018, page 817. American Medical Informatics Association.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access.

Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):5.

Piotr Przybyła, Austin J Brockmeier, Georgios Kontonatsios, Marie-Annick Le Pogam, John McNaught, Erik von Elm, Kay Nolan, and Sophia Ananiadou. 2018. Prioritising references for systematic reviews with robotanalyst: A user study. *Research synthesis methods*, 9(3):470–488.

Juan C Sager, David Dungworth, Peter F McDonald, et al. 1980. *English special languages: principles and practice in science and technology*. John Benjamins Pub Co.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embedding. *arXiv preprint arXiv:1902.08691*.

Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Brian Zhu, and Iain J Marshall. 2016. Extracting PICO sentences from clinical trial reports

using supervised distant supervision. *Journal of Machine Learning Research*, 17(132):1–25.