# Data Extraction and Synthesis in Systematic Reviews of Diagnostic Test Accuracy: A Corpus for Automating and Evaluating the Process

**Christopher Norman, M Sc,[1,2] Mariska Leeflang, PhD,[2] Aurélie Névéol, PhD[1]**
[1]**LIMSI, CNRS, Université Paris Saclay, F-91405 Orsay**
[2] **Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands**

**Abstract**

*Background: Systematic reviews are critical for obtaining accurate estimates of diagnostic test accuracy, yet these require extracting information buried in free text articles, which is often laborious. Objective: We create a dataset describing the data extraction and synthesis processes in 63 DTA systematic reviews, and demonstrate its utility by using it to replicate the data synthesis in the original reviews. Method: We construct our dataset using a custom automated extraction pipeline complemented with manual extraction, verification, and post-editing. We evaluate using manual assessment by two annotators and by comparing against data extracted from source files. Results: The constructed dataset contains 5,848 test results for 1,354 diagnostic tests from 1,738 diagnostic studies. We observe an extraction error rate of 0.06–0.3%. Conclusions: This constitutes the first dataset describing the later stages of the DTA systematic review process, and is intended to be useful for automating or evaluating the process.*

## 1 Introduction

Accurate estimates of diagnostic test accuracy (DTA)[1] are critical for deciding what tests to recommend or use, as well as for interpreting test results, and is therefore important to clinicians and policy makers, as well as to individual patients. Diagnostic test accuracy results are usually reported independently in several small studies. In order to achieve accurate and generalizable estimates, we typically need to perform a systematic review, i.e., identify all studies evaluating the diagnostic test of interest, and pool the results of these together.

However, the process of producing systematic reviews is almost entirely manual and therefore costly. Reducing the cost would not only serve to reduce public spending on research, but might also make systematic reviews feasible that would today require too much time or resources.[2] This is particularly true for systematic reviews of diagnostic test accuracy, in which the workload may be considerably higher than for other kinds of systematic reviews.[3]

In this work we present a novel dataset describing the work done by systematic review authors during the data extraction and synthesis stages in past systematic reviews. In the long term, we hope that this dataset will prove to be useful in automating the process, or in evaluating such automated systems. In the short term however, the data can also be used to evaluate the work done by human systematic review authors, and we demonstrate this by independently replicating the pooled results reported in the systematic reviews in the dataset.

### 1.1 Automated Data Extraction and Synthesis

Systematic reviews are conducted in a multistep process as illustrated in Figure 1, each step following a systematic and highly controlled procedure to ensure close to perfect recall, and a minimum of mistakes. The data extracted from the identified studies is pooled and synthesized, and the conclusions of the review are based on this synthesis. In a DTA systematic review, these results typically come in the form of a summary score, the mean sensitivity and specificity* estimated from the synthesized data. The high recall requirements, as well the high stakes associated with errors means that all of the stages are manual in practice.

Most steps in the process are costly and could benefit from assistance through technical means,[4] but previous work has focused on reducing the workload mainly in the article selection process,[2] i.e. in stage 2 in Fig. 1. Less work has been done towards reducing the workload in the article retrieval, article screening, data extraction, data synthesis, and the analysis steps (3–7 in figure 1), even though these too are laborious and still entirely manual processes.

*Defined as the number of true positives divided by the total number of positives and the number of true negatives divided by the total number of negatives respectively.

Consequently, datasets exist describing the abstract screening stage in DTA systematic review, or describing the data extraction in other domains, and these have been used for work towards automating these processes. We are aware of no datasets describing stages 3–7 for DTA systematic reviews. In this paper we aim to partially fill this gap, by presenting a corpus describing the processes performed by human authors in the data extraction and synthesis stages in DTA systematic reviews. This is intended to be useful for eventually automating the process, but also for reasoning about the work done by human systematic review authors.

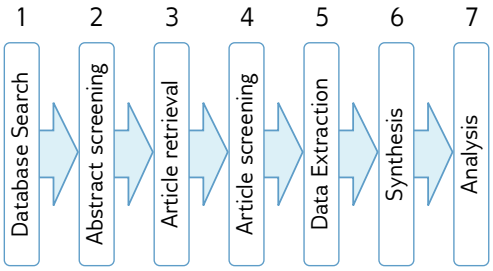



**Figure 1:** Overview of the systematic review process. Simplified from Tsafnat et al..[4]

### 1.2 Systematic Review Reproducibility

Multiple levels of reproducibility of research have been proposed.[5,6] Exact definition may differ, but generally we *reproduce* research by redoing experiments in the same setting, we *replicate* research by redoing the analysis on the reported data, and we *repeat* research by retracing exactly the steps of published results.

Research reproducibility has been receiving increasing amount of attention by the research community in the last few decades.[7,8] In practice however, it is often difficult to even replicate the results of a paper, that is to say, to use the data presented to redo the analysis. In the setting of a systematic review on diagnostic test accuracy, provided the data used to calculate the summary scores are reported, we should be able to redo the calculations to yield the same results. To demonstrate the usefulness of our newly created dataset, we will test to what extent this is possible for the systematic reviews in the Cochrane Library, by replicating the calculation of the key summary scores reported in each systematic review.

### 1.3 Related Work

#### 1.3.1 Datasets Describing the Systematic Review Process

Datasets have been published describing the database search and abstract screening steps in the systematic review process, both for DTA systematic reviews and for other topics. For instance, the included studies as well as the database search queries from 50 of the systematic reviews on DTA in the Cochrane Library have previously been published in one of the CLEF eHealth shared tasks,[9,10] thus addressing stages 1 and 2 in Figure 1. Similarly, Cohen has previous published a dataset describing the included and excluded studies in 15 systematic reviews on drug class efficacy,[11] thus addressing stage 2 in Figure 1, albeit in a different domain.

Datasets addressing the data extraction stage do not exist for diagnostic test accuracy, but exist for other domains, such as the PIBOSO corpus.[12] Work has also been done on automatically extracting PICO* statements,[13,14] as well as other clinical trial information from article full text.[14]

In order to extract data from DTA studies automatically using supervised machine learning, we need labeled gold standard datasets describing what data was extracted from each primary study, i.e. the data extraction forms in each systematic review. Such a dataset targeting systematic reviews of diagnostic test accuracy should include data extraction forms for the data necessary to perform the systematic review analysis, such as the index test, reference standard, target condition, and the $2 \times 2$ tables,† preferable with an emphasis on those items most difficult to extract manually. We are aware of no such datasets in current literature.

*Population, intervention, control group, and outcome.

†The true/false positives and the true/false negatives for the test results, roughly equivalent to a confusion matrix in computer science.

### 1.3.2 Systematic Review Replication

Replication in science has been the focus of an increasing amount of discussion recently.[7] However, we are not aware of work on replicating systematic reviews.

## 2 Objectives

We extract and reconstruct the reported data from each open-access or free systematic review in the diagnostic test accuracy section in the Cochrane Library,* with the following goals in mind:

1. **Data extraction form corpus for DTA systematic reviews**: We create a dataset by collecting the data extraction forms, the summary scores, and the included and excluded articles from each systematic review in the Cochrane Library. The dataset is intended to describe the work being done by human screeners in the data extraction and synthesis stages of a DTA systematic review, by documenting the input and output of these stages in past reviews.
2. **Summary score replication**: We demonstrate the usefulness of our corpus by replicating the summary scores reported for the main tests in each systematic review. Our aim is to identify obstacles to replication, and to measure the discrepancies between our calculated summary scores and those reported.

## 3 Material and Methods

Our raw data consists of the systematic reviews on diagnostic test accuracy published in the Cochrane Library.† The Cochrane Library is the repository for the systematic reviews conducted under the auspices of Cochrane, one of the leading organizations for systematic reviews worldwide, which imposes more rigorous standards on systematic reviews than do paper-based journals.[15]

We downloaded a snapshot of the systematic reviews on DTA in October 2017 to keep the data consistent during processing. For four of the reviews, we were able to obtain the XML source files used to compile the published systematic reviews from the authors, and we use these to evaluate the extraction quality. We do not have access to the source files for the other 59 systematic reviews, and for these we need to extract the data from the published articles.

### 3.1 The Cochrane DTA Data Form Corpus

We construct our dataset from the published review articles, which come in a mixture of free text, formatted as HTML, and data tables, formatted as PNG images. Our contribution consists not only of extracting these elements piecemeal, but also in linking the elements together. We use automated extraction methods where possible, and complement these with manual extraction, verification and post-editing.

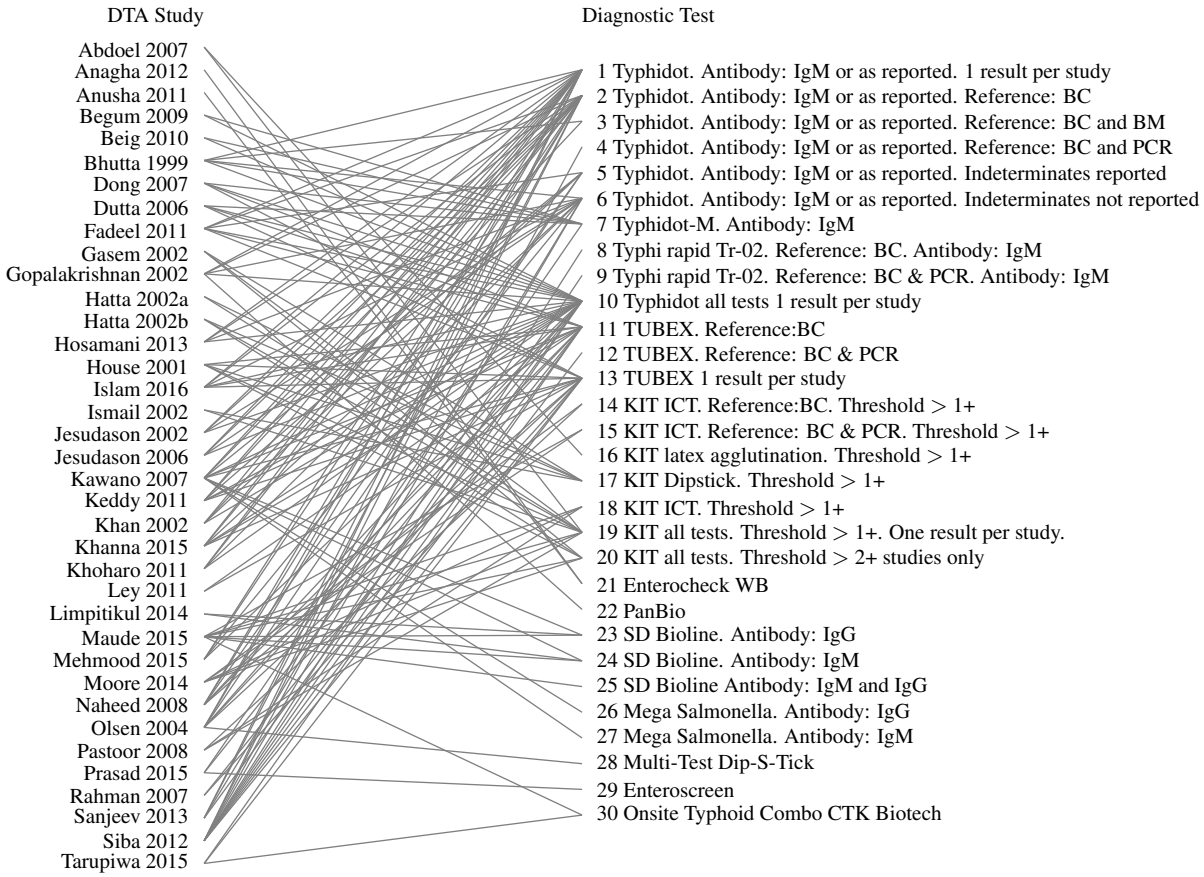
Figure 2a presents the results of text extraction for a sample systematic review:[16] the list of included studies (left column), the list of diagnostic tests (right column) and which study evaluated which tests (links). Figure 2b presents a sample $2 \times 2$ image table from the same review. Each row in the table describe the results on the same test (test 5) performed independently in six studies, and so each row corresponds to a link in Figure 2a.

Two annotators (CN and AN) manually post-edited the data from one systematic review sampled randomly. Based on this evaluation, we decided which parts of the automated extraction requires manual verification and post-editing, and which parts can be extracted automatically.

To assess the output quality, we also compare the post-edited data against the source files where available.

**HTML processing** The HTML contents are processed using the LXML Python package.‡ Our system parses the HTML articles, and locates each section in the article using HTML ID and class attributes, as well as headers, and extracts the text contents.

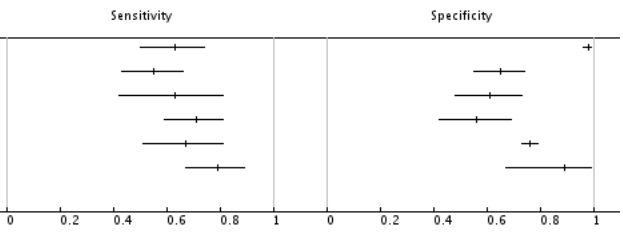
*http://www.cochranelibrary.com/topic/Diagnosis/Diagnostic%20test%20accuracy/
†www.cochranelibrary.com
‡https://pypi.python.org/pypi/lxml




**(a)** The relations between the diagnostic tests and the studies included in the systematic review. Most of the tests are evaluated by several studies, and so are connected by several edges in the graph.

Review: Rapid diagnostic tests for typhoid and paratyphoid (enteric) fever
Test: 5 Typhidot. Antibody: IgM or as reported. Indeterminates reported

| Study | TP | FP | FN | TN | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Fadeel 2011 | 42 | 5 | 25 | 309 | 0.63 [ 0.50, 0.74 ] | 0.98 [ 0.96, 0.99 ] |
| Kawano 2007 | 41 | 36 | 34 | 66 | 0.55 [ 0.43, 0.66 ] | 0.65 [ 0.55, 0.74 ] |
| Keddy 2011 | 17 | 25 | 10 | 39 | 0.63 [ 0.42, 0.81 ] | 0.61 [ 0.48, 0.73 ] |
| Khan 2002 | 49 | 26 | 20 | 33 | 0.71 [ 0.59, 0.81 ] | 0.56 [ 0.42, 0.69 ] |
| Naheed 2008 | 29 | 197 | 14 | 627 | 0.67 [ 0.51, 0.81 ] | 0.76 [ 0.73, 0.79 ] |
| Olsen 2004 | 46 | 2 | 12 | 17 | 0.79 [ 0.67, 0.89 ] | 0.89 [ 0.67, 0.99 ] |




**(b)** The diagnostic test results, as reported in the systematic review, for test 5 in the graph above. Each row describes the test results reported in one study corresponds to a single edge in the graph above.

**Figure 2:** Example of parts of the data in a systematic review on diagnostic test accuracy of Salmonella infection,[16] showing the relations between the data elements (a), and the source data from which the relations were extracted (b).

**Image processing** The diagnostic test results are only presented in images, and therefore requires optical character recognition (OCR) for extraction. In our extraction system, the images are first passed through a preprocessing stage where the images are scaled to roughly double the original size and antialised. The data is then extracted using Tesseract.* We also use domain knowledge to correct mistakes, and flag possible errors for manual verification.

*https://github.com/tesseract-ocr/tesseract

For each systematic review, we collect references to the *included and excluded studies*, i.e. the output of stage 4 in the systematic review process (Fig. 1). We collect these automatically using HTML processing.

For each primary study we collect the *data extraction forms* reported in the systematic review, i.e. the data extracted from each included primary study. We also collect the *data tables*, documenting the test results for each test, i.e. the $2 \times 2$ tables, sensitivity and specificity along with their 95% confidence intervals (see Fig. 2b). Together, these constitute the output of stage 5 in the systematic review process (Fig. 1) for diagnostic test accuracy. We do this using HTML processing to locate the data tables, and image processing to extract the table contents.

For each systematic review we collect the reported *summary scores* for each diagnostic test, i.e. the means estimated from the pooled test results. This constitutes the output of stage 6 in the systematic review process (Fig. 1) for diagnostic test accuracy. We do this manually by reading the summary of findings and locating the relevant data table matching the description in the text.

We also keep track of which test was performed by which study, which studies were included in which systematic review and which test results were used in which summary score calculation. These relations are not simple one-to-one relations, and in practice systematic reviews include sets of studies which may overlap with the included or excluded studies in other systematic reviews. The diagnostic tests performed by the studies within a systematic review generally overlap* (see Fig. 2a and 2b). A summary score pair should normally be connected to a single diagnostic test (but several test results), or two tests if it measures the relative performance of contrasted pairs of tests, but the summary scores are usually not reported for all diagnostic tests.

### 3.2 Summary Score Replication

We attempt to replicate the summary scores reported in each systematic review. We take note of which summary scores were reported with sufficient clarity that it would be possible to exactly repeat the original summary score calculations. However, in this work we do not attempt exact repetition, only replication using equivalent methods following Cochrane guidelines.

Summary scores should be calculated to account for the inter- and intrastudy variance.[1] There are multiple software packages available to perform these calculations, such as the SAS `NLMixed` procedure,† the Stata `xtmelogit` or `meqrlogit` routines,‡ or the `reitsma` function from the `mada` R package.[17] Cochrane guidelines give no recommendation as to which software package to use,[1] and the choice consequently differs from review to review, as does the exact software version. Any of these choices is however considered valid, and should produce equivalent, albeit not necessarily identical results.

While the choice of software package and version is often reported in systematic reviews, it would be infeasible to repeat all systematic reviews using the exact same software package and version. Our intent is not however exactly repeating the original calculations, but replicating them using equivalent software, and we therefore use the same software package (`mada`) for all our trials.

## 4 Results

From the 63 systematic reviews we extracted 5,848 test results together evaluating 1,354 unique diagnostic tests in 1,738 DTA studies (Table 1). We also extracted 589 reported summary score pairs, of which we replicate 103 and compare with the values reported in the systematic reviews.

### 4.1 Dataset Construction

Table 1 presents statistics of the dataset contents, which are publicly available.§

*Pooling the results of diagnostic tests performed by multiple primary studies is one of the primary objectives of a DTA systematic review and so this overlap is expected in a successful systematic review.

†`https://support.sas.com/documentation/onlinedoc/stat/141/nlmixed.pdf`

‡`https://www.stata.com/help.cgi?xtmelogit`

§DOI: `10.5281/zenodo.1303259`

| Data | Extracted | Auto-corrected | perc. | Evaluated | | | |
|---|---|---|---|---|---|---|---|
| | | | | Manually | perc. | Against source | perc. |
| Systematic reviews | 63 | - | - | 1 | 1.6% | 4 | 6.3% |
| Included studies | 1,738 | - | - | 49 | 2.8% | 203 | 11.7% |
| Data forms | 1,356 | - | - | 49 | 3.6% | 145 | 10.7% |
| Text entries | 29,201 | - | - | 1,176 | 4.0% | 3,281 | 11.2% |
| Excluded studies | 6,699 | - | - | 132 | 2.0% | 337 | 5.0% |
| Diagnostic tests | 1,354 | 796 | 58.8% | 43 | 3.2% | 28 | 2.1% |
| Test results | 5,848 | 1,981 | 33.9% | 144 | 2.5% | 330 | 5.6% |
| Study IDs | 5,848 | 1,706 | 29.2% | 144 | 2.5% | 330 | 5.6% |
| Numerical | 58,480 | 1,018 | 1.7% | 1,440 | 2.5% | 3,300 | 5.6% |
| Summary scores | 589 | - | - | - | - | - | - |

**Table 1:** The nature and amount of extracted data of the each type, along with the portions of auto-corrected and evaluated elements. We consider test results to be auto-corrected if they contain at least one auto-corrected value. We consider diagnostic tests to be auto-corrected if they contain at least one auto-corrected test result.

**Evaluation by manual annotation by two annotators** Two independent annotators (AN and CN) manually verified and post-edited the included and excluded studies, the data tables, and the data forms automatically extracted from one systematic review. During the annotation, we highlighted the data elements flagged by the automatic validation. We observed a 100% inter-annotator agreement on the corrected data. No errors were found in the extraction of the included and excluded studies, the data forms, or the numerical values from the tables. We did however find 3 errors among the study IDs extracted from the data tables. As these errors were all flagged by the automatic validation process, we decided that validation in subsequent reviews would be performed by one annotator (CN) focusing on the flagged data.

**Evaluation by comparing with the source files** We compared our extracted data after post-editing against ground truth data from four source files.

We observed 4 errors out of 330 study IDs (1.2%), of which 3 could be spotted by either checking whether the study IDs were in the list of included references, or by checking that the study IDs for each table appeared in alphabetical order. We thus observed 1 error out of 330 (0.3%) after sanity checking.

We observed 2 errors out of 3,300 numerical values (0.06%), both of which could be spotted by sanity checking that the 2 by 2 table matches the sensitivity and specificity for each table row.

### 4.2 Summary Score Replication

Figure 3 presents the flow of reviews in the dataset according to replicability status. For the 103 of the 589 summary score pairs that could be meaningfully replicated, we plot the distribution of discrepancies in Figure 4, and we list the larger discrepancies in Table 2 (top section).

Twenty-five of the summary score pairs occurred in reviews with no data section. A further 76 of the summary scores descriptions did not clearly match any of the data tables in the data section. For 49 of the summary scores the number of studies or participants differed between the data tables and what was reported in the summary score description.

We excluded 60 summary scores from our replication attempts because they used measures other than the mean, such as the median, or the range. We also excluded 276 scores because they were based on less than three primary studies, and therefore can not be used to calculate reliable estimates.[1]

We replicated the remaining 103 summary scores pairs and plotted the difference in Fig 4. Of the 603 scalar values in the 103 summary scores,* we observed a 5 point difference in 79 / 603 (13%) of the scalars, and a 10 point difference in 25 / 603 (4%) scalars. In Table 2 (top section), we list the summary scores with at least one 10 point difference, roughly corresponding to the outliers in Fig. 4. We also list all the replicated summary scores for the summary scores where the number of studies or participants differ and had at least a 10 point difference in Table 2 (bottom section).

*A summary score is composed of 3 or 6 scalar values depending on whether both sensitivity and specificity are reported.

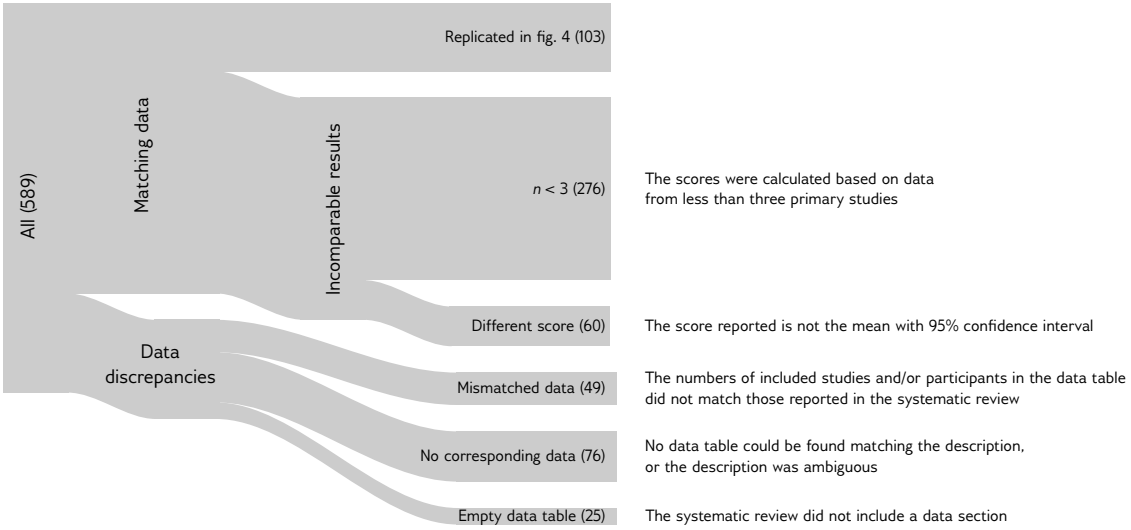



**Figure 3:** Flow of the summary scores in the replication, describing how many summary scores (primary studies) were excluded for each reason in our replication attempt.

| Test | | Replicated summary scores with 95% CI | | | | | | Reported summary scores with 95% CI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Description | $n$ | Sensitivity (%) | | | Specificity (%) | | $n$ | Sensitivity (%) | | | Specificity (%) | |
| Reported scores with matching number of studies and participants, at least one 10 point difference | | | | | | | | | | | | | |
| CD010705 37 | Direct; SLID; cul... | 8 | 71.6 | [35.6, | 92.0] | 97.6 | [90.0, 99.4] | 8 | 87.0 | [38.1, | 98.6] | 99.5 | [93.6, 100] |
| CD012179 26 | Anti-endometrial ... | 4 | 80.0 | [64.2, | 90.0] | 80.5 | [60.0, 91.9] | 4 | 81.0 | [76.0, | 87.0] | 75.0 | [46.0, 100] |
| CD012179 82 | 42.5. CA-19.9 (ca... | 3 | 36.5 | [29.7, | 44.0] | 90.2 | [61.7, 98.1] | 3 | 36.0 | [26.0, | 45.0] | 87.0 | [75.0, 99.0] |
| CD009591 21 | Pelvic MRI | 7 | 76.6 | [64.0, | 85.8] | 69.2 | [53.9, 81.2] | 7 | 79.0 | [70.0, | 88.0] | 72.0 | [51.0, 92.0] |
| CD009591 7 | RVS TVUS | 10 | 66.9 | [40.3, | 85.8] | 97.4 | [93.3, 99.0] | 10 | 88.0 | [82.0, | 94.0] | 100 | [98.0, 100] |
| CD009591 30 | RVS MRI | 3 | 76.7 | [52.7, | 90.7] | 91.5 | [56.9, 98.9] | 3 | 81.0 | [70.0, | 93.0] | 86.0 | [78.0, 95.0] |
| CD009591 9 | Vaginal TVUS | 6 | 56.5 | [31.6, | 78.5] | 97.3 | [92.8, 99.0] | 6 | 57.0 | [21.0, | 94.0] | 99.0 | [96.0, 100] |
| CD009591 31 | Vaginal MRI | 4 | 74.3 | [61.1, | 84.2] | 93.2 | [81.9, 97.6] | 4 | 77.0 | [67.0, | 88.0] | 97.0 | [92.0, 100] |
| CD009591 33 | POD MRI | 5 | 86.2 | [74.2, | 93.1] | 89.8 | [70.5, 97.0] | 5 | 90.0 | [76.0, | 100] | 98.0 | [89.0, 100] |
| CD009591 19 | Rectosigmoid TRUS | 4 | 89.5 | [82.9, | 93.8] | 91.4 | [78.5, 96.9] | 4 | 91.0 | [85.0, | 98.0] | 96.0 | [91.0, 100] |
| CD009591 38 | Rectosigmoid MDCT-e | 3 | 95.6 | [80.3, | 99.1] | 97.9 | [92.7, 99.4] | 3 | 98.0 | [94.0, | 100] | 99.0 | [97.0, 100] |
| CD010023 1 | CT | 4 | 72.7 | [59.4, | 82.9] | 97.8 | [94.3, 99.2] | 4 | 72.0 | [36.0, | 92.0] | 99.0 | [71.0, 100] |
| CD010023 2 | MRI | 5 | 78.4 | [61.8, | 89.1] | 96.2 | [87.3, 98.9] | 5 | 88.0 | [64.0, | 97.0] | 100 | [38.0, 100] |
| CD010023 3 | BS | 6 | 95.3 | [88.1, | 98.2] | 84.7 | [69.8, 93.0] | 6 | 99.0 | [69.0, | 100] | 86.0 | [73.0, 94.0] |
| CD009579 8 | CCA POC haematobium | 4 | 39.9 | [12.3, | 75.9] | 77.5 | [43.7, 93.9] | 4 | 39.0 | [6.0, | 73.0] | 78.0 | [55.0, 100] |
| CD010653 2 | Diagnosis of schi... | 16 | 57.9 | [50.1, | 65.3] | 73.8 | [64.0, 81.7] | 16 | 58.0 | [50.3, | 65.3] | 74.7 | [85.2, 82.3] |
| CD010079 7 | IQCODE cut-off 3.6 | 3 | 74.4 | [64.2, | 82.4] | 91.3 | [83.9, 95.5] | 3 | 78.0 | [68.0, | 86.0] | 87.0 | [71.0, 95.0] |
| CD008054 14 | Triage of LSIL wi... | 4 | 95.9 | [90.6, | 98.3] | 22.4 | [16.7, 29.5] | 4 | 97.5 | [69.6, | 99.8] | 24.8 | [7.3, 58.1] |
| CD008054 18 | Triage of LSIL wi... | 4 | 80.4 | [63.5, | 90.6] | 48.9 | [22.8, 75.7] | 4 | 84.6 | [48.6, | 97.0] | 44.4 | [16.0, 76.9] |
| Reported scores with mismatched number of studies and participants, at least one 10 point difference | | | | | | | | | | | | | |
| CD012165 26 | PGP 9.5 (protein ... | 8 | 88.7 | [67.4, | 96.7] | 76.9 | [69.2, 83.1] | 7 | 96.0 | [91.0, | 100] | 86.0 | [70.0, 100] |
| CD009591 28 | USL MRI | 4 | 85.4 | [78.2, | 90.5] | 81.6 | [51.7, 94.8] | 4 | 86.0 | [80.0, | 92.0] | 84.0 | [68.0, 100] |
| CD007394 2 | Children | 7 | 44.2 | [15.1, | 77.8] | 96.4 | [92.7, 98.3] | 6 | 84.0 | [66.0, | 93.0] | 88.0 | [60.0, 97.0] |
| CD011975 5 | Total hCG | 2 | 7.4 | [2.7, | 18.3] | 95.0 | [94.1, 95.8] | 3 | 19.0 | [4.0, | 58.0] | Assumed 95 % | |
| CD008803 40 | OCT: ONH horizont... | 12 | 53.7 | [40.8, | 66.1] | 88.9 | [84.0, 92.4] | 6 | 41.0 | [26.0, | 58.0] | 94.0 | [90.0, 96.0] |
| CD008803 34 | OCT: ONH Cup area | 15 | 52.6 | [38.9, | 65.9] | 88.3 | [84.0, 91.5] | 9 | 45.0 | [26.0, | 67.0] | 92.0 | [87.0, 95.0] |
| CD008803 38 | OCT: ONH Cup volume | 17 | 43.6 | [29.8, | 58.3] | 89.3 | [85.3, 92.3] | 9 | 30.0 | [16.0, | 49.0] | 94.0 | [92.0, 96.0] |
| CD008803 37 | OCT: ONH Nerve he... | 8 | 52.8 | [40.8, | 64.5] | 88.3 | [81.9, 92.6] | 4 | 44.0 | [28.0, | 62.0] | 93.0 | [87.0, 96.0] |
| CD008803 36 | OCT: ONH Rim volume | 11 | 56.6 | [45.7, | 66.9] | 89.7 | [84.7, 93.2] | 6 | 49.0 | [35.0, | 62.0] | 95.0 | [92.0, 96.0] |
| CD008081 2 | OCT for detection... | 3 | 80.3 | [72.3, | 86.4] | 82.6 | [71.9, 89.8] | 3 | 74.0 | [68.0, | 86.0] | 92.0 | [87.0, 97.0] |

Point difference scale.

**Table 2:** Replicated vs reported summary score pairs differing from the reported summary scores by at least 10 point on one of the six scalar values (one cell per row in the table). Larger differences are highlighted.

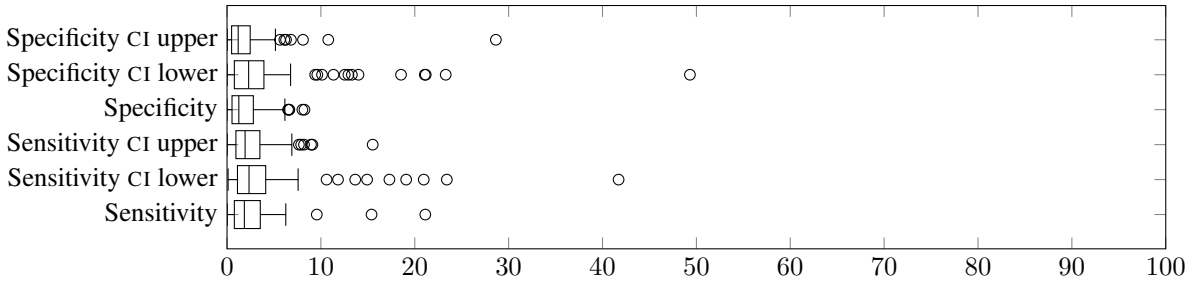



**Figure 4:** Distribution of differences between reported and replicated summary scores for each of the six scalar values in the summary scores. All differences are in absolute point difference.

## 5 Discussion

In this section, we discuss the data extraction and provide insight on the growth and possible uses of the dataset. We then discuss the findings and implications of our replication study in Section 5.2

### 5.1 Dataset Construction

Our manual extraction by two annotators on the automatically extracted parts of the data had a 100% inter-annotator agreement. Furthermore, we only observe errors for data extracted from the data tables, using OCR, and all of the errors were flagged for inspection by the automated extraction. In light of this, we content ourselves with letting a single annotator check and post-edit the extracted data tables for the remainder of the dataset, and do not verify the extraction of the lists of included and excluded studies or the data extraction forms.

Manual post-editing still lets the occasional error through, as we can see from the results in Section 4.1, but these can generally be spotted by automatic sanity checking.

The amount of manual effort required for to process a single review varies, from a few minutes to several hours for a single review. The effort required for verification and post-editing chiefly depends on the success rate of the automated extraction of the data tables, and the effort required to manually extract the summary scores chiefly depends the clarity of presentation in the systematic review. The amount of data in the review also plays a role, but only in the presence of automatic extraction failures, and unclear presentation.

### 5.2 Summary Score Replication

As we can see in Figure 4, our replicated summary scores are generally close to the summary scores reported in the original systematic reviews, but we also observe a large number of discrepancies. The discrepancies tend to be larger and more common for the lower bound of the confidence intervals (Figure 4).

When there is a mismatch in terms of number of studies or participants between the summary scores and the data tables, this is typically deliberate, and the authors often state the reason why some of the studies or participants were excluded from the calculations. In order to replicate such summary scores, it may be necessary to modify the data tables manually. Simply using the original unmodified tables can obviously give different summary scores, although this varies (Table 2, bottom section).

In some cases the reason for the mismatch is not stated, but may be due to different definitions of number of studies or participants in different parts of the systematic review.*

The mismatch for ‘CD011975 Total hCG’ in Figure 2 (bottom section) appears to be due to a copy-paste error however, and the results presented is identical to the results for ‘CD011975 Inhibin.’ The summary scores are apparently calculated from data table 11 rather than from data table 5. This summary score is not mentioned in the systematic review body however, and so does not appear to have influenced the findings of the review.

*A paper can contain multiple studies, and for instance a diagnostic test for glaucoma may count individual eyes as participants.

A large part of the manual work required to connect summary scores to data tables were due to their order often being different in the two sections. We therefore recommend that data tables and summary scores presentation be aligned in future systematic reviews to make it easier to verify the results, as well as catch errors. We also note that this would go far towards automating the synthesis step in the systematic review process for diagnostic test accuracy.

The frequency of genuine errors in the systematic review is low (1 out of 150). However, the one error we did find could be spotted simply by verifying the numbers of studies and participants. Such verification could potentially be performed automatically, provided the summary of findings and the data tables are organized consistently, and a standard definition of number of included studies and participants are used throughout the systematic review.

### 5.3 Dataset Applications and Future Work

This data is intended to be used to 1) train and evaluate methods for automating the data extraction and data synthesis stages in DTA systematic reviews, 2) perform replication studies, like the one we describe here, and 3) perform robustness studies by e.g. evaluating how the results of the analysis would differ on different subsets of the data (subset analysis or ablation studies).

In particular, this dataset contains all the information that needs to be extracted from the included studies in a systematic review on diagnostic test accuracy, and can therefore be used towards training supervised machine learning models to automate this process in future reviews.

## 6 Conclusion

In this paper, we presented a dataset describing the input and output of the data extraction and synthesis stages in systematic reviews on diagnostic test accuracy. The data extraction was manually validated and found successful with an error rate of 6 in 10,000 for the numerical values, 0.3% for the study ids, and no observed errors on the other data types. This is the first dataset to provide material for addressing automation of the later stages of the DTA systematic review process, including the data extraction and synthesis.

We demonstrate the value of this dataset by conducting a replication study of 103 summary scores from the data synthesis in 27 of the systematic review. Overall, we are able to replicate the results reported in the systematic reviews with less than 5 point difference for 87% of the values. We did not try to replicate the interpretation of the results to test whether these differences would have affected the general conclusions reached in the reviews. Our findings mirror insights gained in other fields: it is often not straightforward to replicate reported results, even when these are reported clearly.

We believe that the availability of material presented here (data and tools) can be helpful for the community to leverage the information contained in systematic reviews to a fuller extent, for instance to make it easier to replicate or update the data synthesis and analysis in systematic reviews.

### Acknowledgments

Clive Adams from the Cochrane Schizophrenia Group kindly provided us with the source RevMan file for one of the systematic reviews we used for evaluation.

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

1. Petra Macaskill, Constantine Gatsonis, Jonathan Deeks, Roger Harbord, and Yemisi Takwoingi. Cochrane handbook for systematic reviews of diagnostic test accuracy. *Version 0.9. 0. London: The Cochrane Collaboration*, 2010.
2. Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):5, 2015.

3. Henry Petersen, Josiah Poon, Simon K Poon, and Clement Loy. Increased workload for systematic review literature searches of diagnostic tests compared with treatments: Challenges and opportunities. *JMIR medical informatics*, 2(1):e11, 2014.

4. Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. Systematic review automation technologies. *Systematic reviews*, 3(1):74, 2014.

5. Steven N Goodman, Daniele Fanelli, and John PA Ioannidis. What does research reproducibility mean? *Science translational medicine*, 8(341):341ps12–341ps12, 2016.

6. K Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany J Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E Hunter. Three dimensions of reproducibility in natural language processing. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2018, page 156. NIH Public Access, 2018.

7. Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452, 2016.

8. Christian Collberg and Todd A Proebsting. Repeatability in computer systems research. *Communications of the ACM*, 59(3):62–69, 2016.

9. Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. Overview of the CLEF technologically assisted reviews in empirical medicine. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017.*, CEUR Workshop Proceedings. CEUR-WS.org, 2017.

10. Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. Overview of the CLEF technologically assisted reviews in empirical medicine 2018. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*, CEUR Workshop Proceedings, 2018.

11. A M Cohen, W R Hersh, K Peterson, and P Yen. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. pages 206–219, 2006.

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

13. Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Brian Zhu, and Iain J Marshall. Extracting pico sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research*, 17(132):1–25, 2016.

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

15. Alejandro R Jadad, Deborah J Cook, Alison Jones, Terry P Klassen, Peter Tugwell, Michael Moher, and David Moher. Methodology and reports of systematic reviews and meta-analyses: a comparison of cochrane reviews with articles published in paper-based journals. *Jama*, 280(3):278–280, 1998.

16. Lalith Wijedoru, Sue Mallett, and Christopher M Parry. Rapid diagnostic tests for typhoid and paratyphoid (enteric) fever. *The Cochrane Library*, 2017.

17. Philipp Doebler and Heinz Holling. Meta-analysis of diagnostic accuracy with mada. *Reterieved at: https://cran. rproject. org/web/packages/mada/vignettes/mada. pdf*, 2015.