CLEF 2018 Technologically Assisted Reviews in Empirical Medicine Overview

Evangelos Kanoulas[1], Dan Li[1], Leif Azzopardi[2], and Rene Spijker[3]

[1] Informatics Institute, University of Amsterdam, Netherlands, E.Kanoulas@uva.nl, D.Li@uva.nl
[2] Computer and Information Sciences, University of Strathclyde, Glasgow, UK, leif.azzopardi@strath.ac.uk
[3] Cochrane Netherlands and UMC Utrecht, Julius Center for Health Sciences and Primary Care, Netherlands, R.Spijker-2@umcutrecht.nl

Abstract. Conducting a systematic review is a widely used method to obtain an overview over the current scientific consensus on a topic of interest, by bringing together multiple studies in a reliable, transparent way. The large and growing number of published studies, and their increasing rate of publication, makes the task of identifying all relevant studies in an unbiased way both complex and time consuming to the extent that jeopardizes the validity of their findings and the ability to inform policy and practice in a timely manner. The CLEF 2018 e-Health *Technology Assisted Reviews in Empirical Medicine* task aims at evaluating search algorithms that seek to identify all studies relevant for conducting a systematic review in empirical medicine. The task had a focus on Diagnostic Test Accuracy (DTA) reviews, and consisted of two subtasks: 1) given a number of relevance criteria as described in a systematic review protocol, search a large medical database of article abstracts (PubMed) to find the studies to be included in the review, and 2) given the article abstracts retrieved by a carefully designed Boolean Query, prioritize them to reduce the effort required by experts to screen the abstracts for inclusion in the review. Seven teams participated in the task, with a total of 12 runs submitted for subtask 1 and 19 runs for subtask 2. This paper reports both the methodology used to construct the benchmark collection, and the results of the evaluation.

Keywords: Systematic Reviews, Technology Assisted Reviews, TAR, Diagnostic Test Accuracy, DTA, PubMed, Cochrane, e-Health, Information Retrieval, Text Classification, Evaluation, Test Collection, Benchmarking, High Recall, Active Learning, Relevance Feedback

1 Introduction

Evidence-based medicine has become an important pillar in health care and policy making. In order to practice evidence-based medicine, it is important to have a clear overview over the current scientific consensus. These overviews are

provided in systematic review articles, that summarize all available evidence regarding a certain topic (e.g., a treatment or a diagnostic test). To write a systematic review, researchers have to conduct a search that will retrieve all the studies that are relevant to the topic. The large and growing number of published studies, and their increasing rate of publication, makes the task of identifying relevant studies in an unbiased way both complex and time consuming to the extent that jeopardizes the validity of their findings and the ability to inform policy and practice in a timely manner. Hence, the need for automation in this process becomes of utmost importance. Finding all relevant studies in a corpus is a difficult task, known in the Information Retrieval (IR) domain as the "total recall" problem [7].

To this date, the retrieval of studies that contain the necessary evidence to inform systematic reviews is being conducted in multiple stages:

1. Identification: At the first stage a systematic review protocol, which describes the rationale, hypothesis, and planned methods of the review, is prepared. The protocol is used as a guide to carry out the review. Beyond other information, it provides the criteria that need to be met for a study to be included in the review. Further, a Boolean query that attempts to express these criteria is constructed by an information specialist. The query is then submitted to a medical database containing titles, abstracts, and indexing terms of a controlled vocabulary of medical studies. The result is a set, A, of potentially relevant studies.
2. Screening: At a second stage experts are screening the titles and abstracts of the returned set and decide which one of those hold potential value for their systematic review, a set D. If screening an abstract has a cost C_a, screening all |A| abstracts has a cost of $C_a * |A|$.
3. Eligibility: At a third stage experts are downloading the full text of the potentially relevant abstracts, D, identified in the previous phase and examine the content to decide whether indeed these studies are relevant or not. Examining a document has typically a larger cost than the cost of examining an abstract, $C_d > C_a$. The result of the second screening is the set of studies to be included in the systematic review.

Unfortunately, the precision of the Boolean query is typically low, hence reviewers often need to manually examine many thousands of irrelevant titles and abstracts in order to identify a small number of relevant ones. Further, there is no guarantee that the Boolean query will retrieve all relevant studies, jeopardizing the validity of the reviews. To overcome some of the limitations of the Boolean search, researchers have been testing the effectiveness of machine learning and information retrieval methods. O'Mara-Eves et al. [15] provide a systematic review of the use of text mining techniques for study identification in systematic reviews.

The focus of the CLEF 2018 e-Health *Technology Assisted Reviews in Empirical Medicine* (TAR), similar to last year [10], lies on Diagnostic Test Accuracy (DTA) reviews. Search in this area is generally considered the hardest, and a breakthrough in this field would likely be applicable to other areas as well [11].

The goal of the lab is to bring together academic, commercial, and government researchers that will conduct experiments and share results on automatic methods to retrieve relevant studies with high precision and high recall, and release a reusable test collection that can be used as a reference for comparing different retrieval and mining approaches in the field of medical systematic reviews.

This paper is organized as follows: Section 2 describes the two subtasks of the lab in detail, Section 3 describes the constructed benchmark collection, and Section 4 the evaluation measures used; in Section 5 we briefly describe the participating systems, and in Section 6 we discuss the results of the evaluation. Section 7 concludes the article.

2 Task Description

In this section we describe the two subtasks of the TAR lab, the input provided to participants for each one of the subtasks and the expected participant's output submitted to the lab for evaluation.

2.1 Subtask 1: No Boolean Search

Prior to constructing a Boolean Query researchers have to design and write a systematic review protocol that in detail defines what constitutes a relevant study for their review. In this experimental task of the TAR lab, participants are provided with the relevant pieces of a protocol, in an attempt to complete search effectively and efficiently by-passing the construction of the Boolean query.

In particular, for each systematic review that needs to be conducted (also referred to as *topic* in the IR terminology), participants are provided with the following input data:

1. topic ID;
2. the title of the review written by Cochrane experts;
3. parts of the protocol, which includes the *Objective*, the *Type of Study*, the *Participants*, the *Index Tests*, the *Target Conditions*, and the *Reference Standards*;
4. the PubMed database, provided by the National Center for Biotechnology Information (NCBI), part of the U.S. National Library of Medicine (NLM).

Participants are provided with 30 topics on Diagnostic Test Accuracy (DTA) reviews. For each one of these topics participants are asked to submit: (a) a ranked linked of PubMed articles, and (b) a threshold over this ranked list. Participant can submit up to 3 submissions ("runs"). A run is the output of the participants' algorithm for all the topics, in the form of a text file, with each line of the file following the format:

```
TOPIC-ID    THRESHOLD    PMID    RANK    SCORE    RUN-ID
```

Each line represents a PubMed article in the ranked list for a given topic, with RANK indicating the index of this article in the ranked list. TOPIC-ID is the id of the topic for which the document has been retrieved, and THRESHOLD is either 0 or 1, with 1 indicating that the given rank is the rank of the threshold. PMID is the PubMed Document Identifier of the article ranked at that position, SCORE is the score the algorithm gives to the article, and RUN-ID is an identifier for the submitted run. Participants are allowed to submit a maximum of 5,000 ranked PMIDs per topic, i.e. a total maximum of 150,000 lines per run.

2.2 Subtask 2: Title and Abstract Screening

Given the results of the Boolean Search from the first stage of the systematic review process as the starting point, participants are asked to rank the set of abstracts. The task has two goals: (i) to produce an the efficient ordering of the documents, such that all of the relevant abstracts are retrieved as early as possible, and (ii) to identify a subset which contains all or as many of the relevant abstracts for the least effort (i.e. total number of abstracts to be assessed).

In particular, for each systematic review that needs to be conducted (also refereed to as *topic* in the IR terminology), participants are provided with the following input data:

1. topic ID
2. the title of the review written by Cochrane experts;
3. the Boolean query manually constructed by Cochrane experts;
4. the set of PubMed Document Identifiers (PMID's) returned by running the query in MEDLINE.

Participants are provided with 30 topics on Diagnostic Test Accuracy (DTA) reviews, which are the same topics as those provided in subtask 1. As in subtask 1 participants are asked to submit: (a) a ranked linked of the PubMed articles in the given set, and (b) a threshold over this ranked list. Participant can submit up to 3 submissions, and the format of each submission follows the format of subtask 1 submissions. Further, given that subtask 2 was the main task of the CLEF 2017 e-Health *Technology Assisted Reviews in Empirical Medicine* [10], participants were allowed, if not encouraged, to also submit any of their 2017 system over the new 30 topics outputs.

3 Benchmark Collection

In what follows we describe the collection of articles used in the task, the topics released to participants, and how they were developed, as well as the relevance labels used in the evaluation.

3.1 Articles

The collection used in the lab is PubMed Baseline Repository last updated on November 28, 2017, and available on the NCBI FTP site under the `ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline` directories. PubMed comprises more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites. NLM produces a baseline set of MEDLINE/PubMed citation records in XML format for download on an annual basis. The annual baseline is released in December of each year. The complete baseline consists of files pubmed18n0001 through pubmed18n0928.

3.2 Topics

To construct the benchmark collection, the organizers of the task considered 30 systematic reviews on Diagnostic Test Accuracy already conducted by Cochrane researchers. These reviews are publicly available and can be found in the Cochrane Library[4]; they can be identified by setting the topic filter in the library to "Diagnosis" and "Diagnostic Test Accuracy" and the stage filter to "Review".

At the time of the topic construction 88 such systematic reviews were available; 50 of them were used in the 2017 task [10,6], and out of the remaining 38, 30 were chosen to constitute the 2018 topic set. The 30 systematic reviews considered can be found in Tables 9 and 10. The tables provide the topic id, a substring of DOI of the document (e.g. the DOI for the topic ID CD008122 is `10.1002/14651858.CD008122.pub2`), the title of the systematic review that corresponds to the topic, and the publication date.

Participants were provided with two sets of topics: (a) a development set, and (b) a test set. The development set consisted of 42 topics out of the 50 topics provided in the 2017 version of the lab. [5] The 50 topics released in 2017 were re-examined by our Cochrane information specialist, and co-author of this paper, Rene Spijker, and 8 of them were found not reliable for training or testing purposes, and hence removed from the development set. In particular, the search strategies used within these reviews had a different objective than the objective of the lab. For this task we set-out to use searches that are sensitive in nature to inform a specific question for one review. Some of the reviews we removed were part of an overarching project where one search query was used to inform multiple reviews. We believe that including these would not reflect our intended practice and would misinform the algorithms and strategies developed. Other reviews had a different approach where a local registry was built on a broad topic (dementia) which would inform the review and the MEDLINE search was only intended as a highly specific top-up search, again not the intended approach for this task. So the reviews themselves were reliable but the methods used deviated from this task making them unsuitable. The IDs of the 8 topics are the following:

[4] http://www.cochranelibrary.com/

[5] For subtask 1, two topics, CD011548 and CD011984, were not provided to participants, resulting in 40 training topics.

CD007431 (10), CD010772 (41), CD010775 (2), CD010896 (39), CD010771 (45), CD011145 (42), CD010783 (56), CD010860 (57), where in parenthesis is the filename of the topic in the 2017 release of the data.

Topic Description for Subtask 1: In subtask 1 each topic file was generated through the following procedure: First, the topic ID was extracted from the DOI of the systematic review. Then, the title of the systematic review was considered. Last, for each systematic review, the corresponding protocol was identified, and the objective of the review as described in the protocol was also considered. These three elements, topic ID, title and objective constitute the topic provided to participants. An example can be seen below:

```
Topic: CD008122

Title: Rapid diagnostic tests for diagnosing uncomplicated P. falciparum
       malaria in endemic countries

Objectives: To assess the diagnostic accuracy of RDTs for detecting
            clinical P. falciparum malaria (symptoms suggestive of
            malaria plus P. falciparum parasitaemia detectable by
            microscopy) in persons living in malaria endemic areas
            who present to ambulatory healthcare facilities with
            symptoms of malaria, and to identify which types and
            brands of commercial test best detect clinical P. falciparum
            malaria.
```

Cochrane DTA review titles follow a particular structure [9] with a few alternatives. For instance, in the example above the title follows the structure: "Index test(s) for [target condition(s) in [participant description]". The objective of a DTA systematic review can be: (a) to make comparisons between tests concerning their global accuracy, (b) to estimate the accuracy of a test operating at a particular threshold, or (c) to understand why results of studies vary. In the example above the objective is to estimate accuracy. Furthermore, participants were provided with other relevant parts of the protocol and in particular, the secondary objectives, if any, the type of study, the participants, the index tests, the target conditions, the comparator tests, and the reference standards.

The description of these relevant parts of the protocol as described in the Cochrane Handbook for DTA reviews [8] is can be found in the gray box below.

Types of studies: Identifiable design features of eligible studies must be stated. Review authors should describe the design as well as using a design name, as there is no universal terminology for diagnostic study designs. Key aspects include whether only prospective or both prospective and retrospective studies are to be included, to describe how and where participants were recruited (e.g. as a consecutive series of new presentations in primary care), and whether the study was cross-sectional or included longitudinal assessment for the reference standard. Authors should always state whether they included or excluded diagnostic case-control studies or the strategy used to make this decision. Any restrictions based on a minimal quality standard, minimal sample sizes, or numbers of diseased cases should be stated, but there is no clear guidance on how these limitations should be determined. In reviews that include comparisons between tests, alternative study designs which make within-study comparisons of tests may be sought, notably studies where all individuals receive all tests, and those where all individuals receive the reference standard but are randomized to receive different index tests. These latter studies should be described as randomized trials of test accuracy. Some reviews which compare tests may restrict study inclusion only to studies of these designs which make within-study comparisons, but others may include studies that evaluate one or other of the tests individually (particularly where few such published studies exist). Any such restrictions should be stated. Randomized trials of patient outcomes are rarely eligible for inclusion. They can only be included if individuals received both the index test and a reference standard – occasionally this information is available.

Participants: Review authors should specify the participants for whom the test would be applicable, including any restrictions on diagnoses, age groups and settings. Planned subgroup analyses related to participant characteristics should not be listed here – they should be listed under the sources of heterogeneity in the secondary objectives.

Index tests: Review authors should specify the test(s) to be evaluated in the review. If multiple tests are being reviewed and compared with each other details for each test should be given. In the first Cochrane DTA protocols and reviews tests were separated into new index tests or existing comparator tests. However it is often difficult to distinguish index from comparator tests and tests are no longer divided into these two categories. However, where it is clear that some tests are new experimental tests and others are existing standard comparative tests this should be noted.

Target conditions: The target condition is a particular disease or disease stage that the index test is intended to identify. Some reviews may evaluate the ability of tests to differentiate between several target conditions – if this is the case, the multiple target conditions should all be listed here.

Reference standards: Describe the clinical reference standards required to establish the presence or absence of the target condition in the tested population. If any reference standards are commonly used but considered inadequate this should be stated here as an exclusion criteria. If the review covers multiple target conditions, the reference standard for each should be stated.

Topic Description for Subtask 2: In subtask 2 each topic file was generated through the following procedure: For each systematic review, we reviewed the search strategy from the corresponding study in Cochrane Library. A search strategy, among other things, consists of the exact Boolean query developed and submitted to a medical database, at the time the review was conducted, and typically can be found in the Appendix of the study. Rene Spijker, a co-author of this work and a Cochrane information specialist examined the grammatical correctness of the search query and specified the date range which dictated the valid dates for the articles to be included in this systematic review. The date range was necessary because a study published after the systematic review should not be included even though it might be relevant, since that would require manually examining its content to quantify its relevance.

A number of medical databases, and search interfaces to these databases is available for search, and for each one information specialists construct a different variation of their query that better fits the data and meta-data of the database. For this task, we only considered the Boolean query constructed for the MEDLINE database, using the Wolters Kluwer Ovid interface. Then we submitted the constructed Boolean query to the OVID system at `http://demo.ovid.com/demo/ovidsptools/launcher.htm` and collected all the returned PubMed document identification numbers (PMID's) which satisfied the date range constraint. This step was automated by a Python script we put together and through an interface available to the University of Amsterdam.

The topic file is in a text format and contains four sections, Topic, Title, Query, and PMID's. PMID's are the PubMed document IDs returned by the Boolean query. The PMIDs can be used to access the corresponding document through the National Center for Biotechnology Information (NCBI)[6]. An example of a topic file can be viewed below.

```
Topic: CD008122

Title: Rapid diagnostic tests for diagnosing uncomplicated
       P. falciparum malaria in endemic countries

Query:
1.  Exp Malaria/
2.  Exp Plasmodium/
3.  Malaria.ti,ab
4.  1 or 2 or 3
5.  Exp Reagent kits, diagnostic/
6.  rapid diagnos* test*.ti,ab
7.  RDT.ti,ab
8.  Dipstick*.ti,ab
9.  Rapid diagnos* device*.ti,ab
10. MRDD.ti,ab
11. OptiMal.ti,ab
12. Binax NOW.ti,ab
```

[6] https://www.ncbi.nlm.nih.gov/books/NBK25497/

```
13. ParaSight.ti,ab
14. Immunochromatograph*.ti,ab
15. Antigen detection method*.ti,ab
16. Rapid malaria antigen test*.ti,ab
17. Combo card test*.ti,ab
18. Immunoassay Immunoassay/
19. Chromatography Chromatography/
20. Enzyme-linked immunosorbent assay/
21. Rapid test*.ti,ab
22. Card test*.ti,ab
23. Rapid AND (detection* or diagnos*).ti,ab
24. 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14
    or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23
25. 4 and 24
26. Limit 25 to Humans
27. limit 26 to ed=19400101-20100114

Pids:
    19164769
    9557953
    7688346
    18509532
    ...
```

3.3 Relevance Labels

The original systematic reviews written by Cochrane experts included a reference section that listed Included, Excluded, and Additional references to medical studies. Included are the studies that are relevant to the systematic review. Excluded are the studies that in the abstract and title screening stage were considered relevant, but at the article screening phase were considered irrelevant to the study and hence excluded from it. Additional are the studies that do not impact the outcome of the review, and hence irrelevant to it. The union of Included and Excluded references are the studies that were screened at a Title and Abstract level and were considered for further examination at a full content level. These constituted the relevant documents at the abstract level, while the Included references constituted the relevant documents at the full content level.

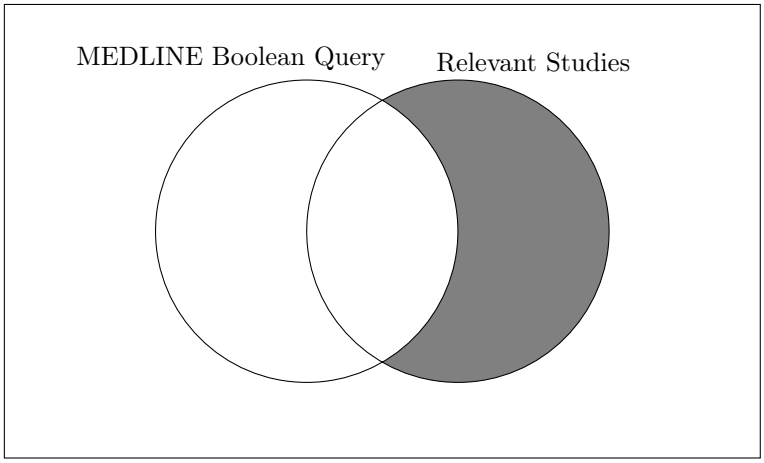
The majority of the references included their corresponding PMID, but not all of them. For those references missing the PMID, the title was extracted from the reference, and it was used as a query to Google Search Engine over the domain `https://www.ncbi.nlm.nih.gov/pubmed/`. The top-scored document returned by Google was selected, and the title of the study contained in landing page, as identified in the metadata extracted. The title was compared then with the title of the study used as search query. If the Edit Distance between the two titles was up to 3 (just to account for spaces, parentheses, etc.) then the study reference was replaced by the PMID also extracted from the metadata of the landing page. If (a) the title had an edit distance greater than 3 but less

than 20, or (b) the study was an included study, or (c) no title was contained in the Google result metadata, or (d) no Google results were returned, then the query was submitted at `https://www.ncbi.nlm.nih.gov/pubmed/` and the results were manually examined. All other studies were discarded under the assumption that they are not contained in PubMed. The format of the qrels followed the standard TREC format:

Topic Iteration Document Relevance

where Topic is the topic ID of the systematic review, Iteration in our case is a dummy field always zero and not used, Document is the PMID, and Relevancy is a binary code of 0 for not relevant and 1 for relevant studies. The order of documents in the qrel files is not indicative of relevance. Studies that were returned by the Boolean query but were not relevant based on the above process, were considered irrelevant. Those are studies that were excluded at the abstract and title screening phase. All other documents in MEDLINE were also assumed to be irrelevant, given that they were not judged by the human assessor.

Note that, as mentioned earlier, the references of a systematic review were produced after a number of Boolean queries were submitted to a number of medical databases, and their titles and abstracts were screened. The PMID's provided however were only those that came out of the MEDLINE query. Therefore, there was a number of abstract-level relevant studies (the gray area in the Venn diagram below) that were not part of the result set of the Boolean query provided to the participants. Studies that were cited in the systematic review but did not appear in the results of the Boolean query were excluded from the label set for Subtask 2, but included for Subtask 1. Hence, the total number of relevant abstracts in the test set for Subtask 1 is 4,656, while in Subtask 2 it is 3,964; further the total number of relevant studies in Subtask 1 is 759, while for Subtask 2 it is 678.



Topic	# total PMIDs	# abs rel	# doc rel	% abs rel	% doc rel
Development Set					
CD010438	3250	39	3	1.20	0.09
CD007427	1521	123	17	8.09	1.12
CD009593	14922	78	24	0.52	0.16
CD011549	12705	2	1	0.02	0.01
CD011134	1953	215	49	11.01	2.51
CD008686	3966	7	5	0.18	0.13
CD011975	8201	619	60	7.55	0.73
CD009323	3881	122	9	3.14	0.23
CD009020	1584	162	12	10.23	0.76
CD011548	12708	113	5	0.89	0.04
CD011984	8192	454	28	5.54	0.34
CD010409	43363	76	41	0.18	0.09
CD008054	3217	274	41	8.52	1.27
CD009591	7991	144	41	1.80	0.51
CD008691	1316	73	20	5.55	1.52
CD010632	1504	32	14	2.13	0.93
CD007394	2545	95	47	3.73	1.85
CD008643	15083	11	4	0.07	0.03
CD009944	1181	117	64	9.91	5.42
CD008803	5220	99	99	1.90	1.90
CD008782	10507	45	34	0.43	0.32
CD009647	2785	56	17	2.01	0.61
CD009135	791	77	19	9.73	2.40
CD008760	64	12	9	18.75	14.06
CD009519	5971	104	46	1.74	0.77
CD009372	2248	25	10	1.11	0.44
CD010276	5495	54	24	0.98	0.44
CD009551	1911	46	16	2.41	0.84
CD012019	10317	3	1	0.03	0.01
CD008081	970	26	10	2.68	1.03
CD009185	1615	92	23	5.70	1.42
CD010339	12807	114	9	0.89	0.07
CD010653	8002	45	0	0.56	0.00
CD010542	348	20	8	5.75	2.30
CD010023	981	52	14	5.30	1.43
CD010705	114	23	18	20.18	15.79
CD010633	1573	4	3	0.25	0.19
CD010173	5495	23	10	0.42	0.18
CD009786	2065	10	6	0.48	0.29
CD010386	626	2	1	0.32	0.16
CD009579	6455	138	79	2.14	1.22
CD009925	6531	460	55	7.04	0.84

42个topic

Table 1. Statistics of topics in the development set. The total PMIDs are the ones retrieved by the Boolean Query, with the percentage of relevant articles also computed over this retrieved set.

Topic	# total PMIDs	# abs rel	# doc rel	% abs rel	% doc rel
Test Set					
CD008122	1911	272	57	0.142	0.030
CD012599	8048	575	19	0.071	0.002
CD009175	5644	65	7	0.012	0.001
CD009694	161	16	9	0.099	0.056
CD009263	79786	124	10	0.002	0.000
CD010502	2985	229	71	0.077	0.024
CD010680	8405	26	0	0.003	0.000
CD010864	2505	44	3	0.018	0.001
CD011431	1182	297	26	0.251	0.022
CD011602	6157	8	1	0.001	0.000
CD011420	251	42	5	0.167	0.020
CD011686	9443	55	2	0.006	0.000
CD012179	9832	304	117	0.031	0.012
CD012281	9876	23	9	0.002	0.001
CD011053	2235	12	7	0.005	0.003
CD011515	7244	127	1	0.018	0.000
CD008587	9158	79	35	0.009	0.004
CD011926	4050	40	29	0.010	0.007
CD012165	10222	308	47	0.030	0.005
CD012083	322	11	5	0.034	0.016
CD008892	1499	69	30	0.046	0.020
CD011126	6000	13	9	0.002	0.002
CD010657	1859	139	35	0.075	0.019
CD008759	932	60	42	0.064	0.045
CD010296	4602	53	38	0.012	0.008
CD010213	15198	599	33	0.039	0.002
CD012009	536	37	4	0.069	0.007
CD011912	1406	36	18	0.026	0.013
CD012010	6830	290	8	0.042	0.001
CD012216	217	11	1	0.051	0.005

Table 2. Statistics of topics in the test set. The total PMIDs are the ones retrieved by the Boolean Query, with the percentage of relevant articles also computed over this retrieved set.

Table 1 and Table 2 show the distribution of the relevant documents at abstract or document level for all the topics in the development set and the test set. The total number of unique PMID's released for the training set was 241,669 (an average of 5754 per topic) and for the test set 218,496 (an average of 7283 per topic). The average percentage of relevant documents at Abstract level in the training set is 3.8% of the total number of PMID's released, and in the test set 4.7%, while at the content level the average percentage is 1.5% in the training set, and 1% in the test set. In [17], a test collection was developed based on a random selection of 93 Cochrane systematic reviews (not just DTAs), and reported a slightly higher rate of relevance ($\frac{14}{1159} = 1.2\%$). However, compared with the TREC campaign, the rate of relevant documents is 5.45% and 2.78% for the Adhoc track of TREC 8, and the Web track of TREC 2002, respectively. Overall, the number of relevant documents is not very high in this lab, making locating them quite a difficult task.

4 Evaluation

Evaluation within the context of using technology to assist in the reviewing process is very much dependent on how the users interact with the system, and on the goal of the technology assistance. For example, if the goal of the assistance is to autonomously predict which studies should be assessed by the end-user at a document level, then the problem can be viewed as a classification problem; the system screens all abstracts and returns a subset of them as relevant. If the goal of the assistance is to identify all the relevant documents as quick as possible but let the human decide when to stop screening, then the problem can be viewed as a ranking problem. There are, of course, many other possible variations. For the purposes of the 2018 lab, we consider the problem as a ranking problem - that is, to rank the set of documents associated with the topic in decreasing order of relevance.

Furthermore, the two subtasks although very similar in terms of evaluation, i.e. in both subtasks participants' runs are rankings of article, with a designated threshold, they also differ: in subtask 2 the set of articles to be prioritized contains all the relevant articles, while in subtask 1 the relevant articles need to be found within the entire PubMed database, and hence there is no guarantee that all relevant articles will appear in the top 5000. Further, in subtask 1, the length of the ranked lists vary significantly across different topics.

For the evaluation of runs employ a number of standard IR measures, along with measures that have been developed for the particular task of technology assisted reviews [4,2]. A list of the used measures can be seen below:

- Subtask 1
  1. Average Precision
  2. Number of Relevant Found
  3. Precision @ last relevant found
  4. Recall @ rank k, with k in [50, 100, 200, 500, 1000, 2000, 5000]

5. Recall @ threshold

– Subtask 2
1. Average Precision
2. Recall @ k % of top ranked abstracts, with k in [5, 10, 20, 30]
3. Work Saved over Sampling at recall r, $WSS@r = (TN + FN)/N(1 - r)$ [2]
4. Reliability = $loss_r + loss_e$ [4], with $loss_r = (1 - r)^2$, where r is the recall at the threshold, and $loss_e = (n/(R + 100) * 100/N)^2$, where n is the number of returned documents by the system up to the threshold, N is the size of the collection, and R the number of relevant documents.
5. Recall @ threshold

The lab organizers developed an evaluation software similar to trec_eval for the easy evaluation of the submitted runs, also provided to participants. The code of the tar_eval software is available at `https://github.com/CLEF-TAR/tar`.

5 Participants

The 2018 task received submissions from 7 teams, including one team from Canada (UWA), one team from the USA (UIC/OHSU), one team from the UK (Sheffield), one team from China (ECNU), one team from Greece (AUTH), one team from Italy (UNIPD), one team from France (Limsi-CNRS). The participating teams are:

1. Aristotle University of Thessaloniki, Greece (AUTH)
2. Centre Nationnal de la Recherche Scientifique, France & Amsterdam Medical Center, The Netherlands (CNRS)
3. East China Normal University, China (ECNU)
4. University of Illinois College of Medicine, Chicago, Illinois, USA and Oregon Health & Science University, Portland, Oregon, USA (UIC/OHSU)
5. University of Padua, Italy (UNIPD)
6. University of Sheffield, United Kingdom (Sheffield)
7. University of Waterloo, Canada (UWA)

For the subtask 1, we received 12 runs from 4 teams. For the subtask 2, we received 19 runs from 7 teams.

The 7 teams used a variety of learning methods including batch supervised learning, continuous active learning, a variety of learning algorithms including logistic regression, support vector machines, and neural networks, as well as unsupervised retrieval methods, such as TT-IDF, BM25, with or without traditional relevance feedback methods, such as the Rocchio's Algorithm, and a variety of text representation methods including simple count-based methods and neural embeddings.

Tables 3 and 4 categorize the participating runs in the two subtasks along five dimensions: (a) automatic vs manual runs; (b) use of the development set; (c) use of supervised and semi-supervised learning algorithms, and (d) use of

Subtask 1: No Boolean Search					
Run	Automatic	Development	Supervision	Feedback	Threshold
auth_run1	✓	✓	✓	content	fixed
auth_run2	✓	✓	✓	content	fixed
auth_run3	✓	✓	✓	content	fixed
ECNU_RUN1	✓	x	x	x	x
ECNU_RUN2	✓	✓	✓	x	x
ECNU_RUN3	✓	✓	✓	x	x
shef-bm25	✓	x	x	x	x
shef-tfidf	✓	x	x	x	x
shef-bool	✓	x	x	x	x
UWA	✓	x	✓	abs	auto
UWG	x	x	✓	manual	auto
UWX	x	x	✓	manual & abs	auto

Table 3. Categorization of participants' runs in subtask 1 along four dimensions.

Subtask 2: Title and Abstract Screening					
Run	Automatic	Development	Supervision	Feedback	Threshold
auth_run1	✓	✓	✓	content	fixed
auth_run2	✓	✓	✓	content	fixed
auth_run3	✓	✓	✓	content	fixed
cnrs_RF_uni	✓	x	✓	abs & content	x
cnrs_RF_bi	✓	x	✓	abs & content	x
cnrs_comb	✓	✓	✓	abs & content	x
ECNU_RUN1	✓	x	x	x	x
ECNU_RUN2	✓	✓	✓	x	x
ECNU_RUN3	✓	✓	✓	x	x
unipd_t500	✓	x	✓	abs	fixed
unipd_t1000	✓	x	✓	abs	fixed
unipd_t1500	✓	x	✓	abs	fixed
shef-feed	✓	x	x	abs	x
shef-general	✓	x	x	x	x
shef-query	✓	x	x	x	x
uic_model7	✓	x	✓	x	x
uic_model8	✓	x	✓	x	x
UWA	✓	x	✓	abs	auto
UWB	✓	x	✓	abs & content	auto

Table 4. Categorization of participants' runs in subtask 2 along four dimensions.

relevance feedback, and (e) the type of thresholding used. The categorization has been performed by the lab coordinators – not by the participants – based on the submitted participants description of their algorithms. Hence, there is always a chance of mis-classifying some run. In subtask 1 participants employed both supervised and unsupervised methods for ranking articles. A total of 5 runs were trained over the provided development set, and their generalization was tested against the test topics, while 7 made no explicit use of it; it may be the case that participants tried different models and algorithms over the development set, and selected to submit the best performing ones, hence there may be a flavor of model selection, however we did not consider this as use of the development set. Participants represented the textual data in a variety of ways, including document-topic features, bag-of-words, topic model distributions, embeddings, metadata. Out of the 19 runs submitted for subtask 2, 6 trained over the development set, 12 used the relevance feedback provided per topic, either at an abstract or content level, while 6 runs used a fixed threshold, 2 an automatic thresholding method, and the rest did not threshold the ranking at all. Below we provide a short description of the submitted runs for both subtask 1 and 2.

AUTH took a learning-to-rank approach, using both batch and active learning. Their model consists of two parts: an inter-topic model which utilizes XGBoost and is trained over the entire development corpus (for subtask 1 it is 2500 articles returned by PubMed search, and for subtask 2 the articles provided by the organizers) and an intra-topic model, an iteratively-built SVM, trained over relevance feedback provided partially in the test topics. For the inter-topic model a total of 48 for subtask 1 and 31 for subtask 2 topic-document (or solely topic) features were computed over the title and the abstract of the articles and the query. For the intra-topic model a TF-IDF vectorization of the articles was used [12].

CNRS trained a logistic regression model on a large number ($>$ 500,000) of features over the development set. The logistic regression model is intended to capture features that are related to DTA studies independent of the topic. They further used an active learning approach which continuously learn to find relevant articles within each topic. A model that combines the two using a feed-forward neural network was also used [13].

ECNU used the BM25 algorithm for subtask 1 to acquire a baseline. Furthermore, query expansion based on MeSH terms and pseudo relevance feedback (PRF) was used to improve the results. In sub-task 2, they employed Paragraph2Vector to represent query and documents for similarity calculation [18].

UIC/OHSU first applied a clustering algorithm over a large number of PubMed articles to identify 6 publication types, including DTA studies, but also Randomized Controlled Trials, Cross-sectional Studies, Cross-over Studies, Cohort Studies, and Case-Control Studies. The clusters were then represented by a feature vector of the centroid with each article in the cluster represented by 300 weighted terms most associated to the words in the article. Then, each article in the provided dataset was compared to the 6 clusters and a number of

similarity measures were computed. These were used as features to be used by an SVM to classify articles against the 6 clusters. [3]

UNIPD used a two-dimensional probabilistic version of BM25 to rank articles, using relevance feedback up to a certain number of articles shown to the user, and switched to a Naive Bayes classifier for the remaining of the articles until a fixed threshold point [14].

Sheffield used RAKE [16] keyword extraction algorithm for subtask 1 to interpret protocols, extract keywords and form them into queries designed to retrieve relevant documents, while Apache Lucene was used as the IR engine. Their approach to subtask 2 was to enrich queries with terms designed to identify diagnostic test accuracy studies and also by making use of relevance feedback [1].

UWA applied the Baseline Model Implementation (BMI) from the TREC Total Recall Track (2015-2016) and the CLEF 2017 eHealth. They further applied their "knee-method" stopping criterion to BMI to determine how many abstracts should be examined for each topic. The difference between different submissions came from the selection of feedback to be used to retrain the model with the options being abstract-level, content-level, or manual feedback provided by the participants themselves [5].

6 Results

In this section we provide the results of the evaluation for both subtasks.

6.1 Subtask 1: No Boolean Search

Tables 5, and 6 provide the results of the evaluation for subtask 1 for a subset of the evaluation measures. All participats' runs are evaluated both against the document and the abstract level relevance labels. What is impressive in these results is that without putting any manual effort to construct a Boolean Query – a rather time-consuming and error-prone process – the best system achieves a 96.7% recall, missing only 25 Included studies out of all 759.

Figure 1 shows the box plots for Average Precision against the document level labels for each one of the participant's runs in Subtask 1, with the Mean Average Precision denoted by a blue dashed line in the box plot.

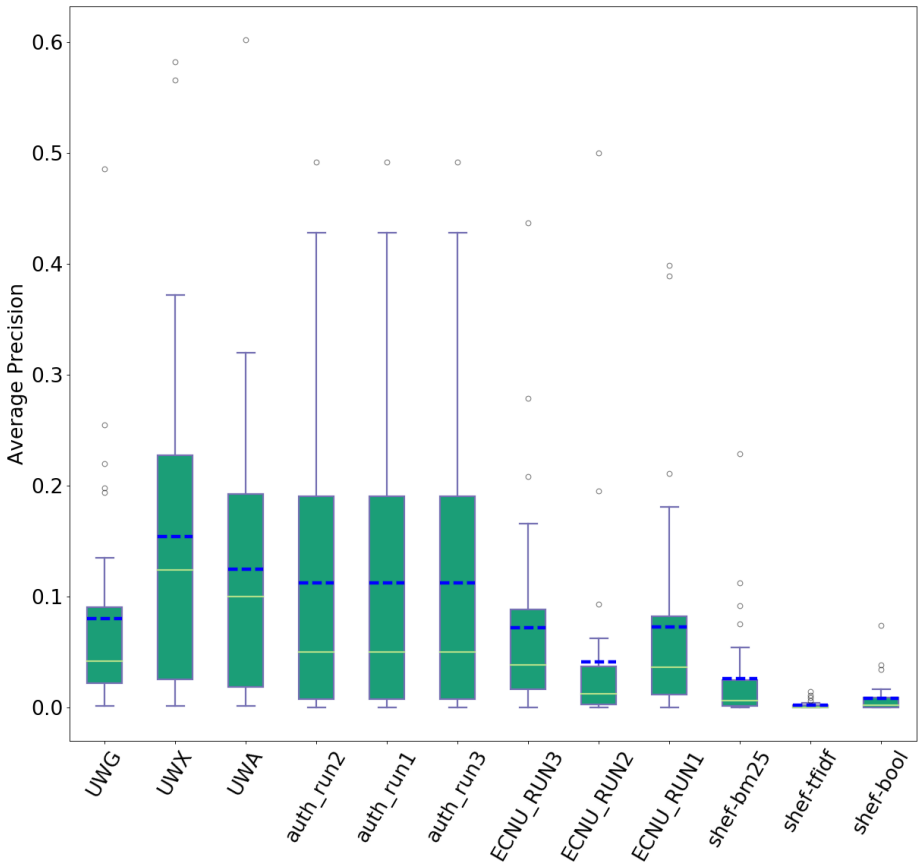


Fig. 1. Average precision using the document level relevance judgments.

Table 5. Average scores for the submitted runs in Subtask 1; relevance is considered at the document level, i.e. only Included studies are considered relevant. In total there are 759 studies that are Included in the 30 systematic reviews conducted.

Run	Total Rel	Rel Found	P@ Last Rel	MAP	R@50	R@100	R@200	R@300	R@400	R@500	R@1000	R@2000	R@5000	R@k	k
auth_run1	759	619	0.217	0.113	0.188	0.341	0.510	0.610	0.660	0.693	0.787	0.802	0.816	0.816	5000
auth_run2	759	619	0.217	0.113	0.188	0.341	0.510	0.610	0.660	0.693	0.787	0.802	0.816	0.809	2500
auth_run3	759	619	0.217	0.113	0.188	0.341	0.510	0.610	0.660	0.693	0.787	0.802	0.816	0.787	1000
ECNU_RUN1	759	426	0.118	0.072	0.170	0.242	0.339	0.393	0.431	0.472	0.561	0.561	0.561	0.472	500
ECNU_RUN2	759	310	0.080	0.041	0.076	0.145	0.216	0.281	0.340	0.378	0.408	0.408	0.408	0.378	500
ECNU_RUN3	759	426	0.109	0.072	0.173	0.246	0.341	0.411	0.452	0.485	0.561	0.561	0.561	0.485	500
shef-bm25	759	323	0.443	0.026	0.045	0.063	0.108	0.149	0.169	0.187	0.261	0.315	0.426	0.426	5000
shef-tfidf	759	202	0.523	0.002	0.005	0.005	0.017	0.029	0.042	0.057	0.086	0.126	0.266	0.266	5000
shef-bool	759	227	0.467	0.008	0.022	0.049	0.069	0.097	0.111	0.124	0.170	0.221	0.299	0.299	5000
UWA	759	727	0.225	0.124	0.256	0.428	0.592	0.693	0.771	0.806	0.912	0.947	0.958	0.951	3559
UWG	759	734	0.239	0.080	0.121	0.273	0.462	0.590	0.675	0.729	0.883	0.959	0.967	0.962	3611
UWX	759	727	0.221	0.154	0.254	0.386	0.564	0.673	0.743	0.784	0.884	0.950	0.958	0.951	3613

Table 6. Average scores for the submitted runs in Subtask 1; relevance is considered at the abstract level, i.e. both Included and Excluded studies are considered relevant. In total there are 4656 studies that are identified as potential relevant during the title and abstract screening in the 30 systematic reviews conducted.

Run	Total Rel	Rel Found	P@ Last Rel	MAP	R@50	R@100	R@200	R@300	R@400	R@500	R@1000	R@2000	R@5000	R@k	k
auth_run1	4656	2879	0.707	0.149	0.064	0.129	0.214	0.276	0.331	0.368	0.486	0.548	0.618	0.618	5000
auth_run2	4656	2879	0.707	0.149	0.064	0.129	0.214	0.276	0.331	0.368	0.486	0.548	0.618	0.568	2500
auth_run3	4656	2879	0.707	0.149	0.064	0.129	0.214	0.276	0.331	0.368	0.486	0.548	0.618	0.486	1000
ECNU_RUN1	4656	1626	0.167	0.110	0.069	0.109	0.166	0.209	0.238	0.265	0.349	0.349	0.349	0.265	500
ECNU_RUN2	4656	1232	0.117	0.046	0.041	0.078	0.128	0.167	0.202	0.226	0.265	0.265	0.265	0.226	500
ECNU_RUN3	4656	1626	0.164	0.109	0.070	0.113	0.173	0.217	0.248	0.274	0.349	0.349	0.349	0.274	500
shef-bm25	4656	1606	0.758	0.035	0.030	0.047	0.075	0.096	0.110	0.123	0.177	0.239	0.345	0.345	5000
shef-tfidf	4656	989	0.681	0.005	0.004	0.008	0.015	0.023	0.029	0.035	0.061	0.105	0.212	0.212	5000
shef-bool	4656	1007	0.739	0.017	0.015	0.027	0.044	0.061	0.072	0.082	0.114	0.157	0.216	0.216	5000
UWA	4656	4354	0.632	0.274	0.110	0.208	0.351	0.460	0.537	0.591	0.746	0.853	0.935	0.909	3559
UWG	4656	4352	0.638	0.202	0.050	0.117	0.265	0.380	0.466	0.527	0.713	0.848	0.935	0.909	3612
UWX	4656	4346	0.670	0.291	0.111	0.203	0.345	0.450	0.534	0.588	0.739	0.853	0.933	0.909	3613

Figure 2 shows the recall-effort curves for the participats' runs, that is the recall value at different percentage of documents shown to the user. The green curve with the square marker corresponds to the Oracle run, which achieves an optimal recall at the different effort levels.

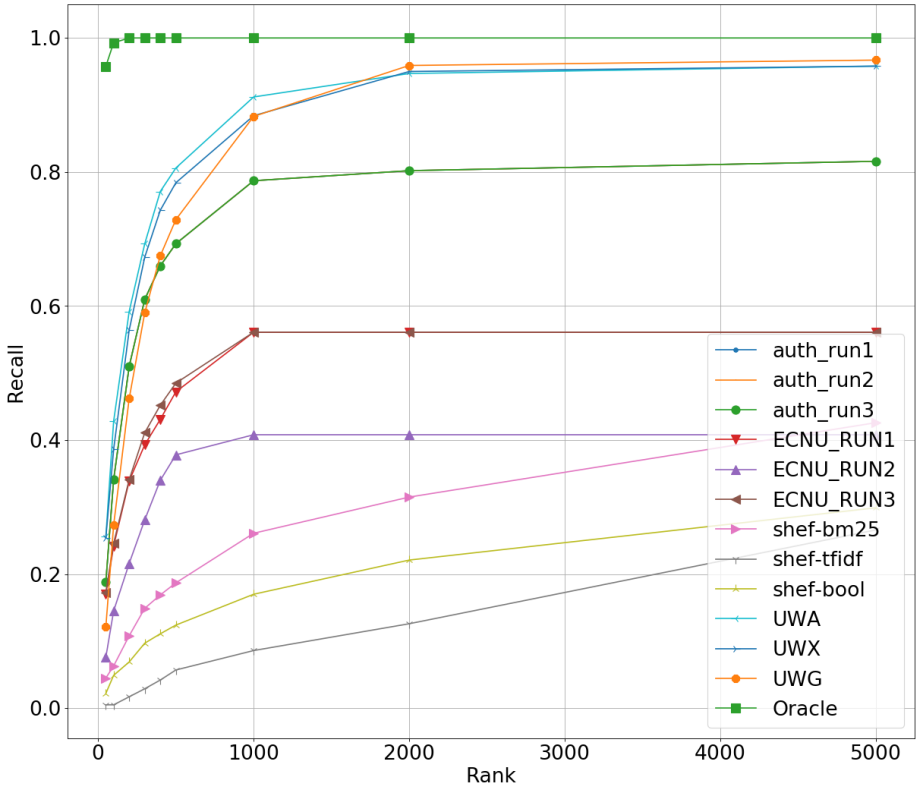


Fig. 2. Recall at different top-k percentages of shown abstracts. Recall is computed using the abstract level relevance labels.

Figure 3 presents the recall obtained by the participants' runs at the point of the threshold as a function of the number of documents presented to the user. As expected the more documents presented to the user (the lower the threshold) the higher the achieved recall. Nevertheless, there are still algorithms that dominate others. The figure present the Pareto frontier.

Figure 4 demonstrates the bar plot of average precision values per topic; the dashed blue line in the box plots designates the average Average Precision (AAP) for each topic, a measure that can be seen as a proxy for topic difficulty.

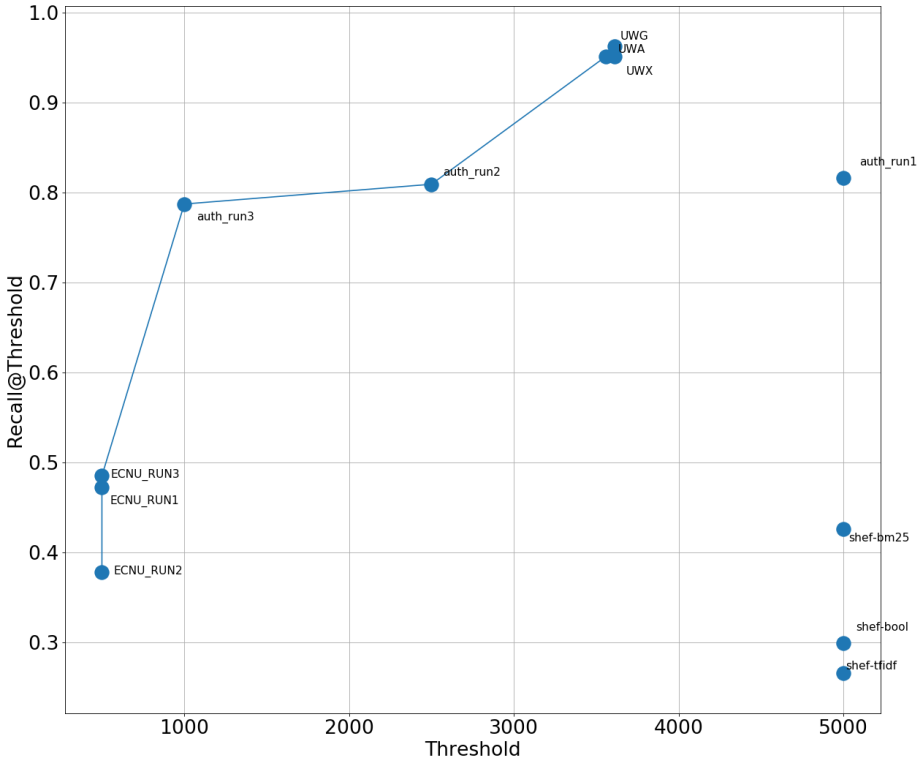


Fig. 3. Recall at the threshold rank as a function of the number of documents shown to the user.

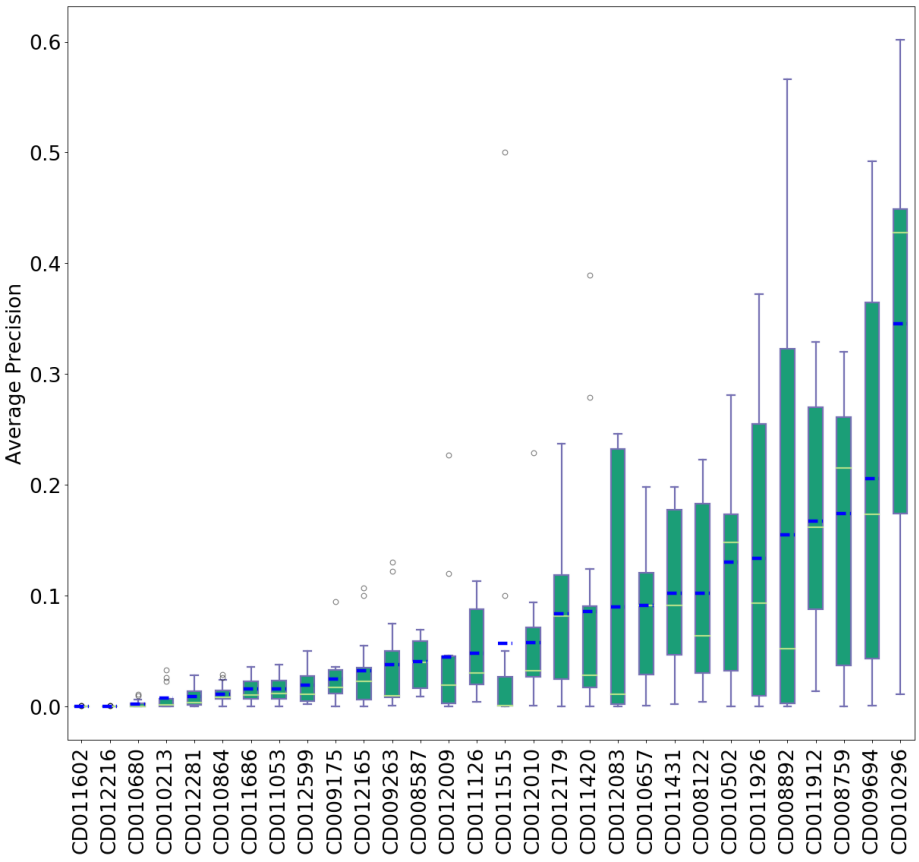


Fig. 4. Average Average Precision at document level relevance labels.

6.2 Subtask 2: Title and Abstract Screening

Tables 7, and 8 provide the results of the evaluation for subtask 2 for a subset of the evaluation measures. All participats' runs are evaluated both against the document and the abstract level relevance labels, respectively. As one can observe, the best run can achieve a recall of 99.4% by reviewing 30% of the abstracts, i.e. missing 4 out of 678 Included studies, while by only reviewing 10% of the abstracts the best run can still achieve 90.6% recall.

Figure 5 shows the box plots for Average Precision against the abstract level labels for each one of the participants' runs in Subtask 2, with the Mean Average Precision denoted by a blue dashed line in the box plot.

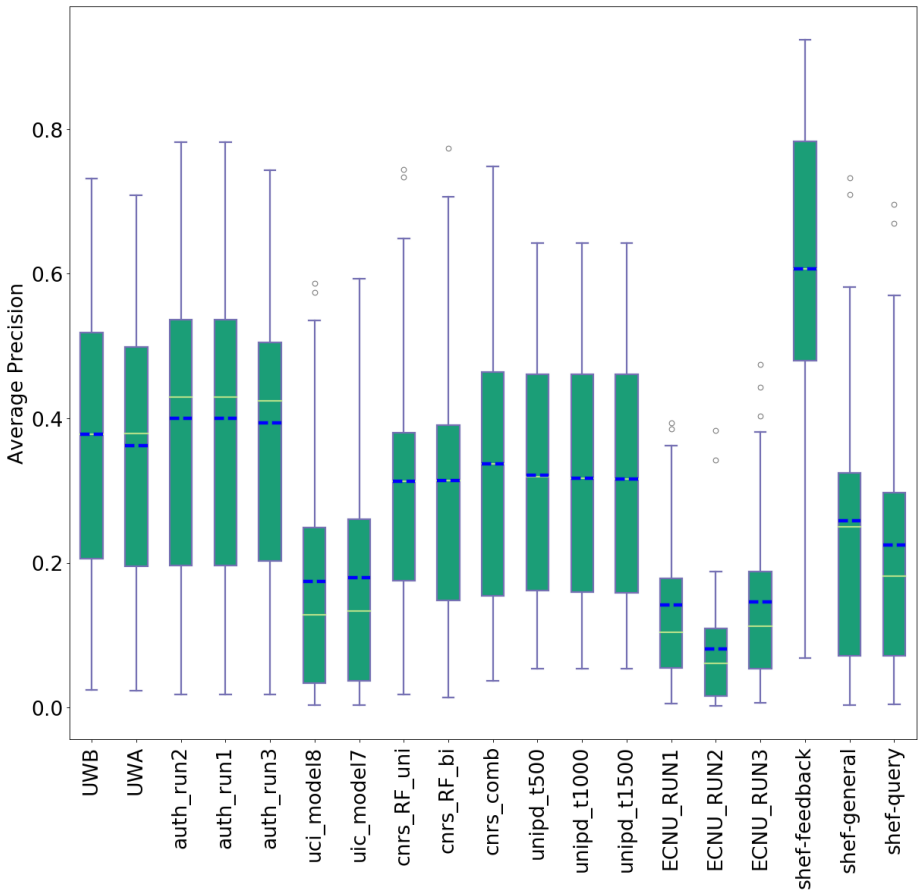


Fig. 5. Average precision using the abstract level relevance judgments.

Table 7. Average scores for the submitted runs in Subtask 2; relevance is considered at the document level, i.e. only Included studies are considered relevant. In total there are 759 studies that are Included in the 30 systematic reviews conducted.

Run	Total Rel	Avg Last Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Reliability	R@k	k
auth_run1	678	571	0.196	0.684	0.898	0.981	0.994	0.848	0.860	0.711	1.000	7245
auth_run2	678	571	0.196	0.684	0.898	0.981	0.994	0.848	0.860	0.245	0.981	877
auth_run3	678	584	0.194	0.689	0.906	0.978	0.993	0.836	0.858	0.246	0.980	877
cnrs_RF_uni	678	3349	0.169	0.665	0.873	0.940	0.954	0.741	0.640	0.720	1.000	7245
cnrs_RF_bi	678	1305	0.176	0.665	0.882	0.945	0.971	0.815	0.762	0.720	1.000	7245
cnrs_comb	678	1225	0.203	0.72	0.903	0.954	0.976	0.824	0.779	0.720	1.000	7245
ECNU_RUN1	678	6355	0.053	0.245	0.353	0.509	0.606	0.123	0.142	0.444	0.587	464
ECNU_RUN2	678	2926	0.032	0.184	0.329	0.494	0.587	0.115	0.119	0.515	0.432	465
ECNU_RUN3	678	6341	0.057	0.291	0.422	0.571	0.646	0.147	0.178	0.446	0.579	464
unipd_t500	678	1190	0.184	0.596	0.801	0.935	0.962	0.792	0.781	0.263	0.922	870
unipd_t1000	678	1390	0.184	0.596	0.789	0.920	0.956	0.765	0.744	0.360	0.962	1587
unipd_t1500	678	1341	0.184	0.597	0.788	0.920	0.947	0.754	0.735	0.421	0.974	2161
shef-feed	678	3740	0.291	0.622	0.780	0.906	0.953	0.759	0.753	0.720	1.000	7245
shef-general	678	3799	0.132	0.420	0.602	0.779	0.872	0.681	0.664	0.720	1.000	7245
shef-query	678	4366	0.103	0.373	0.532	0.724	0.829	0.594	0.588	0.720	1.000	7245
uic_model7	678	4342	0.109	0.348	0.504	0.643	0.706	0.486	0.456	0.235	0.704	2142
uic_model8	678	4382	0.108	0.342	0.494	0.628	0.689	0.467	0.439	0.257	0.636	1778
UWB	678	511	0.174	0.664	0.895	0.988	0.994	0.841	0.860	0.358	0.963	1535
UWA	678	528	0.149	0.653	0.889	0.981	0.993	0.833	0.845	0.429	0.999	2738

Table 8. Average scores for the submitted runs in Subtask 1; relevance is considered at the abstract level, i.e. both Included and Excluded studies are considered relevant. In total there are 3964 studies that are identified as potential relevant during the title and abstract screening in the 30 systematic reviews conducted.

Run	Total Rel	Avg Last Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Reliability	R@k	k
auth_run1	3964	3405	0.400	0.515	0.731	0.891	0.945	0.749	0.611	0.394	1.000	7283
auth_run2	3964	3405	0.400	0.515	0.731	0.891	0.945	0.749	0.611	0.171	0.944	881
auth_run3	3964	4295	0.393	0.519	0.729	0.881	0.932	0.734	0.563	0.172	0.943	881
cnrs_RF_uni	3964	5708	0.313	0.410	0.603	0.771	0.836	0.513	0.349	0.398	1.000	7283
cnrs_RF_bi	3964	5173	0.314	0.408	0.609	0.789	0.874	0.617	0.460	0.398	1.000	7283
cnrs_comb	3964	4379	0.337	0.412	0.609	0.785	0.881	0.657	0.510	0.398	1.000	7283
ECNU_RUN1	3964	7173	0.142	0.205	0.311	0.467	0.585	0.027	0.026	0.427	0.520	465
ECNU_RUN2	3964	4726	0.081	0.157	0.266	0.429	0.550	0.019	0.000	0.502	0.371	466
ECNU_RUN3	3964	7172	0.146	0.233	0.355	0.517	0.619	0.029	0.025	0.409	0.534	465
unipd_t500	3964	3936	0.321	0.412	0.602	0.800	0.886	0.616	0.475	0.209	0.856	874
unipd_t1000	3964	4102	0.317	0.411	0.587	0.771	0.866	0.572	0.410	0.241	0.920	1601
unipd_t1500	3964	4259	0.316	0.412	0.589	0.767	0.847	0.543	0.396	0.270	0.945	2188
shef-feed	3964	5171	0.607	0.470	0.623	0.783	0.859	0.635	0.444	0.398	1.000	7283
shef-general	3964	5519	0.258	0.272	0.429	0.648	0.781	0.552	0.431	0.398	1.000	7283
shef-query	3964	5737	0.224	0.240	0.388	0.604	0.742	0.506	0.377	0.398	1.000	7283
uic_model8	3964	6386	0.174	0.205	0.326	0.477	0.575	0.255	0.154	0.326	0.513	1753
uic_model7	3964	6186	0.180	0.206	0.332	0.491	0.588	0.264	0.164	0.276	0.576	2121
UWA	3964	2546	0.362	0.519	0.724	0.888	0.947	0.751	0.608	0.289	0.990	2926
UWB	3964	2655	0.378	0.525	0.730	0.894	0.946	0.756	0.610	0.287	0.927	1764

Figure 6 shows the recall-effort curves for the participats' runs, that is the recall value at different percentage of abstracts shown to the user.

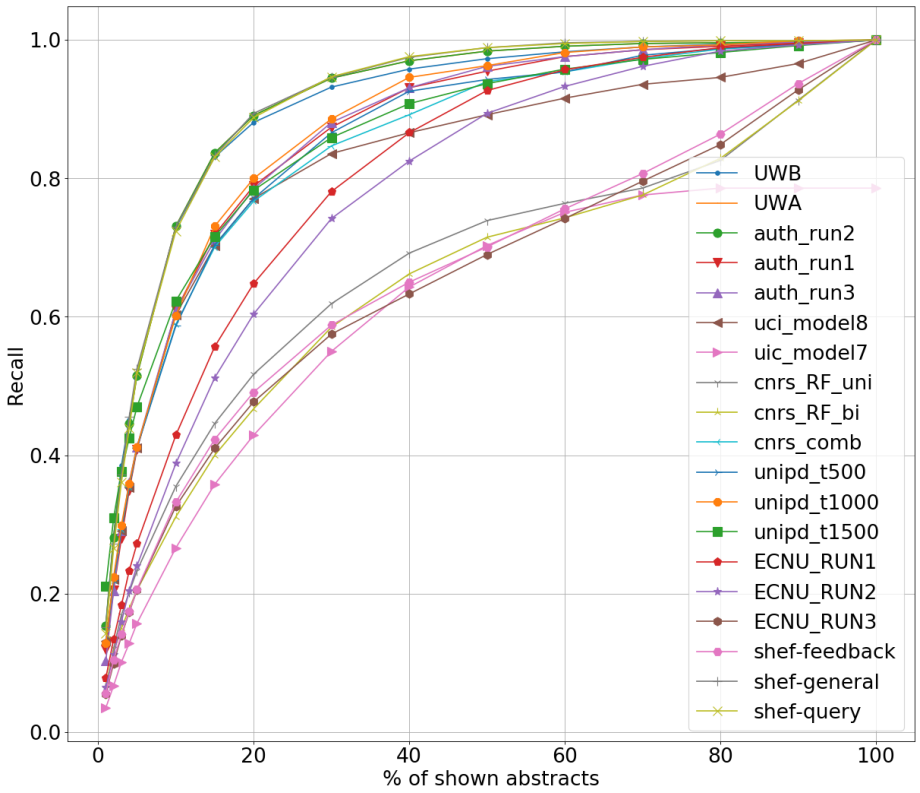


Fig. 6. Recall at different ranks.

Figure 7 presents the recall obtained by the participants' runs at the point of the threshold as a function of the number of abstracts presented to the user. As expected the more abstract presented to the user (the lower the threshold) the higher the achieved recall. Nevertheless, there are still algorithms that dominate others. The figure present the Pareto frontier.

Figure 8 demonstrates the bar plot of average precision values per topic; the dashed blue line in the box plots designates the average Average Precision (AAP) for each topic, a measure that can be seen as a proxy for topic difficulty.

7 Conclusions

The CLEF 2018 e-Health Lab Task 2 constructed a benchmark collection of 30 Diagnostic Test Accuracy systematic reviews to study the effectiveness and effi-

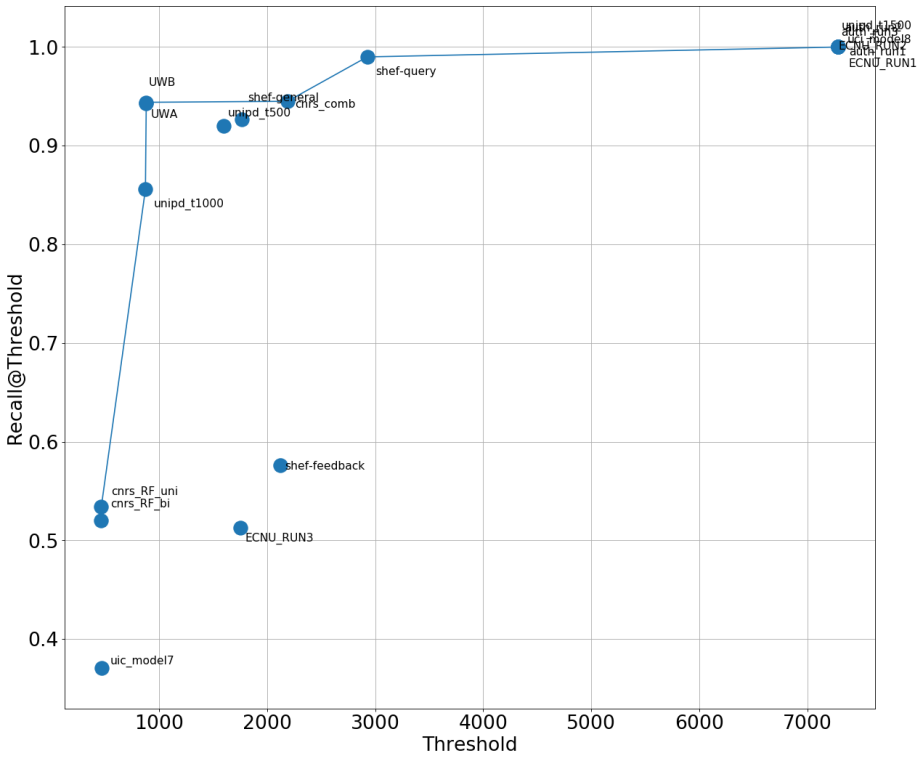


Fig. 7. Recall at the threshold rank as a function of the number of abstracts shown to the user.

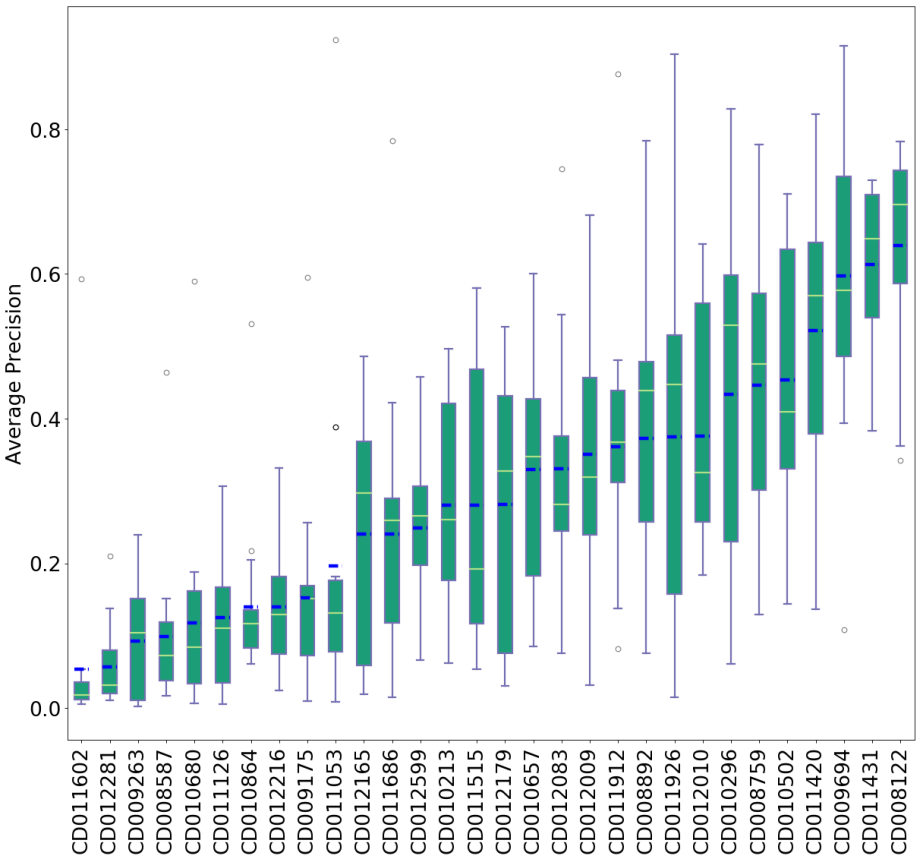


Fig. 8. Average Average Precision at abstract level relevance labels.

ciency of information retrieval and machine learning algorithms both in finding relevant articles in a large medical databse without explicitely constructing a Boolean query and in prioritizing the studies to be screened at the abstract and title screening stage, and providing a stopping criterion over the ranked list. The results demonstrate that automatic methods can be trusted for finding most, if not all, relevant studies in a fraction of the time manual screening can do the same. Further, many of the algorithms retrieved articles that were not in the results of the Boolean query, hence raising even concerns for the validity of the current practice in conducting systematic reviews. Given that across different runs many parameters change simultaneously it is not easy to come to certain conclusions about the relative performance of automatic methods.

Regarding the benchmark collection itself, there is a number of limitations to be considered: (a) Pivoting on the results of the the OVID MEDLINE Boolean query limits our ability to identify all relevant studies, i.e. relevant studies that are outputted by Boolean queries over different databases, and relevant studies that are actually not found by these Boolean queries. The former can be overcome by considering all the different queries submitted; for the latter extra manual judgments would be required. (b) Pivoting on abstract and title only we miss the opportunity to study the effect of automatic methods when applied to the full text of the studies, that would present an opportunity to completely overcome the multi-stage process of systematic reviews. However, most of the full text articles are protected under copyright laws that do not give all participants access to those. (c) The evaluation setup of ranking does not allows us to consider the cost of the process, since given a ranking a researcher would have to still go over all studies ranked. A more realistic setup, e.g. a double-screening setup, could be considered. (d) In the construction of relevant judgments we considered the included and excluded references of the systematic reviews under study, which prevented us to study the noise and disagreement between reviewers. (e) In our effort to allow iterative algorithms, e.g. active learning algorithms, to be submitted, we handed the test sets' relevant judgments directly to the participants, which is rather unusual for this type of evaluation exercises. An alternative would be the setup used by the TREC Total Recall, where participants submitted their running algorithms to the organizers. (f) When it comes to evaluation measures there is a large variety of those, all of which take a different often useful view point on the effectiveness of algorithm, but which makes it difficult to decide upon a single golden measure to rank participants' runs.

References

1. Alharbi, A., Briggs, W., Stevenson, M.: Retrieving and ranking studies for systematic reviews: University of sheffieldâĂŹs approach to clef ehealth 2018 task 2. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, CEUR-WS.org (2018)

2. Cohen, A.M., Hersh, W.R., Peterson, K., Yen, P.Y.: Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association 13(2), 206–219 (2006)
3. Cohen, A.M., Smalheiser, N.R.: Ohsu clef 2018 task 2 diagnostic test accuracy ranking using publication type cluster similarity measures. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, CEUR-WS.org (2018)
4. Cormack, G.V., Grossman, M.R.: Engineering quality and reliability in technology-assisted review. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 75–84. SIGIR '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2911451.2911510
5. Cormack, G.V., Grossman, M.R.: Technology-assisted review in empirical medicine: Waterloo participation in clef ehealth 2018. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, CEUR-WS.org (2018)
6. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J.R.M., Zuccon, G.: CLEF 2017 ehealth evaluation lab overview. In: Jones, G.J.F., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings. Lecture Notes in Computer Science, vol. 10456, pp. 291–303. Springer (2017), https://doi.org/10.1007/978-3-319-65813-1_26
7. Grossman, M.R., Cormack, G.V., Roegiest, A.: TREC 2016 total recall track overview. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016. vol. Special Publication 500-321. National Institute of Standards and Technology (NIST) (2016), http://trec.nist.gov/pubs/trec25/papers/Overview-TR.pdf
8. Group, C.D.T.A.W., et al.: Handbook for dta reviews (2009)
9. Higgins, J.P., Green, S.: Cochrane handbook for systematic reviews of interventions, vol. 4. John Wiley & Sons (2011)
10. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2017 technologically assisted reviews in empirical medicine overview. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017), http://ceur-ws.org/Vol-1866/invited_paper_12.pdf
11. Leeflang, M.M., Deeks, J.J., Takwoingi, Y., Macaskill, P.: Cochrane diagnostic test accuracy reviews. Systematic reviews 2(1), 82 (2013)
12. Minas, A., Lagopoulos, A., Tsoumakas, G.: Aristotle university's approach to the technologically assisted reviews in empirical medicine task of the 2018 clef ehealth lab. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, CEUR-WS.org (2018)
13. Norman, C., Leeflang, M., Neveol, A.: Limsi@clef ehealth 2018 task 2: Technology assisted reviews by stacking active and static learning. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, CEUR-WS.org (2018)

14. Nunzio, G.M.D., Ciuffreda, G., Vezzani, F.: Interactive sampling for systematic reviews. ims unipd at clef 2018 ehealth task 2. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, CEUR-WS.org (2018)
15. O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., Ananiadou, S.: Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic Reviews 4(1), 5 (2015)
16. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text Mining: Applications and Theory pp. 1–20 (2010)
17. Scells, H., Zuccon, G., Koopman, B., Deacon, A., Geva, S., Azzopardi, L.: A test collection for evaluating retrieval of studies for inclusion in systematic reviews. In: To appear in Proceedings of the 40th international ACM SIGIR conference on Research and development in Information Retrieval. ACM (2017)
18. Wu, H., Wang, T., Chen, J., Chen, S., Hu, Q., He, L.: Ecnu at 2018 ehealth task 2: Technologically assisted reviews in empirical medicine. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, CEUR-WS.org (2018)

Topic ID	Topic Title	Publication Date
CD008122	Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries	2010/01/14
CD012599	First and second trimester serum tests with and without first trimester ultrasound tests for Down's syndrome screening	2011/08/25
CD009175	Clinical symptoms and signs for the diagnosis of Mycoplasma pneumoniae in children and adolescents with community-acquired pneumonia	2012/06/26
CD009694	Computed tomography (CT) angiography for confirmation of the clinical diagnosis of brain death	2012/08/31
CD009263	123I-MIBG scintigraphy and 18F-FDG-PET imaging for diagnosing neuroblastoma	2012/09/21
CD010502	Rapid antigen detection test for group A streptococcus in children with pharyngitis	2013/02/01
CD010680	Ankle brachial index for the diagnosis of lower limb peripheral arterial disease	2013/02/01
CD010864	D-dimer test for excluding the diagnosis of pulmonary embolism	2013/12/12
CD011431	Rapid diagnostic tests for diagnosing uncomplicated non-falciparum or Plasmodium vivax malaria in endemic countries	2013/12/31
CD011602	Ultrasonography for diagnosis of alcoholic cirrhosis in people with alcoholic liver disease	2015/01/31
CD011420	Lateral flow urine lipoarabinomannan assay for detecting active tuberculosis in HIV-positive adults	2015/02/28
CD011686	Triage tools for detecting cervical spine injury in pediatric trauma patients	2015/02/28
CD012179	Blood biomarkers for the non-invasive diagnosis of endometriosis	2015/05/01
CD012281	Combination of the non-invasive tests for the diagnosis of endometriosis	2015/05/31
CD011053	Imaging for the exclusion of pulmonary embolism in pregnancy	2015/07/28

Table 9. The provided to participants set of testing topics (PART I).

Topic ID	Topic Title	Publication Date
CD011515	Diagnostic accuracy of different imaging modalities following computed tomography (CT) scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer	2015/11/05
CD008587	Cytology versus HPV testing for cervical cancer screening in the general population	2015/11/30
CD011926	Molecular assays for the diagnosis of sepsis in neonates	2016/01/19
CD012165	Endometrial biomarkers for the non-invasive diagnosis of endometriosis	2016/02/16
CD012083	Ultrasonography for confirmation of gastric tube placement	2016/02/28
CD008892	Rapid diagnostic tests for typhoid and paratyphoid (enteric) fever	2016/03/01
CD011126	Three-dimensional saline infusion sonography compared to two-dimensional saline infusion sonography for the diagnosis of focal intracavitary lesions	2016/03/30
CD010657	Dimercaptosuccinic acid scan or ultrasound in screening for vesicoureteral reflux among children with urinary tract infections	2016/03/31
CD008759	Platelet count, spleen length, and platelet count-to-spleen length ratio for the diagnosis of oesophageal varices in people with chronic liver disease or portal vein thrombosis	2016/06/30
CD010296	Ultrasonography for endoleak detection after endoluminal abdominal aortic aneurysm repair	2016/07/01
CD010213	Imaging modalities for characterising focal pancreatic lesions	2016/07/19
CD012009	Amylase in drain fluid for the diagnosis of pancreatic leak in post-pancreatic resection	2017/02/28
CD011912	Pulse oximetry screening for critical congenital heart defects	2017/03/30
CD012010	Serum amylase and lipase and urinary trypsinogen and amylase for diagnosis of acute pancreatitis	2017/03/30
CD012216	18F PET with florbetapir for the early diagnosis of AlzheimerâĂŹs disease dementia and other dementias in people with mild cognitive impairment (MCI)	2017/05/01

Table 10. The provided to participants set of testing topics (PART II).