CLEF 2019 Technology 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. Systematic reviews are a widely used method to provide an overview over the current scientific consensus, by bringing together multiple studies in a systematic, reliable, and 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 2019 e-Health TAR Lab accommodated two tasks. Task 1 focused on retrieving relevant studies from PubMed without the use of a Boolean query, while Task 2 focused on the efficient and effective ranking of studies during the abstract and title screening phase of conducting a systematic review. In the 2019 lab we also expanded upon the type of systematics reviews considered. Hence, beyond Diagnostic Test Accuracy reviews, we also included Intervention, Prognosis, and Qualitative systematic reviews. We constructed a benchmark collection of 31 reviews published by Cochrane, and the corresponding relevant and irrelevant articles found by the original Boolean query. Three teams participated in Task 2, submitting automatic and semi-automatic runs, using information retrieval and machine learning algorithms over a variety of text representations, in a batch and iterative manner. This paper reports both the methodology used to construct the benchmark collection, and the results of the evaluation.

Keywords: Evaluation, Information Retrieval, Systematic Reviews, TAR, Text Classification, Active Learning

1 Introduction

Evidence-based medicine has become an important pillar in current health care and policy making. In order to practice evidence-based medicine, it is important

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

CEUR-WS.org/Vol-2380/paper_250.pdf

to have a clear overview of the current scientific consensus. These overviews are preferably provided in systematic reviews, that appraise, summarize, and synthesize 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 studies that are relevant to a 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 an 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 the 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 [**?**].

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, by doing so prospectively one tries to minimize risk of bias during conduct of a systematic 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 bibliographic 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 meet the inclusion criteria 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. [**?**] provide a systematic review of the use of text mining techniques for study identification in systematic reviews.

The focus of the CLEF 2017 and 2018 e-Health *Technology Assisted Reviews in Empirical Medicine* (TAR) [**?**,**?**], lied on Diagnostic Test Accuracy (DTA) re-

views. Identifying DTA studies has additional difficulties over the more common intervention studies caused by poorer reporting and indexing of these studies with a lot of heterogeneity in terminology, a breakthrough in this field would likely be applicable to other areas as well [?]. During the past two years search and classification algorithms were developed demonstrating good retrieval performance over the DTA studies. In 2019 we extended our focus to Intervention, Prognosis, and Qualitative systematic reviews.

The goal of the lab, as part of the CLEF e-Health 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 3 describes the two subtasks of the lab in detail, Section 2 describes the constructed benchmark collection, and Section 4 the evaluation measures used; in Section 5 we discuss the results of the evaluation. Section 6 concludes the article.

2 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.

2.1 Articles

The collection used in the lab is PubMed Baseline Repository last updated on 11/12/2018, 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 pubmed19n0001 through pubmed19n0972.

2.2 Topics

To construct the benchmark collection, the organizers of the task used 8 Diagnostic Test Accuracy, 20 Intervention, 1 Prognosis, and 2 Qualitative systematic reviews already conducted by Cochrane researchers. These reviews can be found in the Cochrane Library[4]. 72 DTA systematic reviews used in the 2017 and 2018 versions of the lab [?,?,?,?], as well as 20 different Intervention

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

reviews were also collected and made available to the participants as a development set. The 123 systematic review in both the development and test can be found in Tables 1, 2, 3, and 4. The tables provide the topic id, which is a substring of the DOI of the document (e.g. the DOI for the topic ID CD008122 is `10.1002/14651858.CD008122.pub2`), and the title of the systematic review that corresponds to the topic.

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.
```

Furthermore, participants were provided with other relevant parts of the protocol, which varies per type of review. The protocol for DTA reviews includes the type of study, the participants, the index tests, the target conditions, the comparator tests, and the reference standards. The protocol for Intervention reviews includes the types of studies, the type of participants, the types of interventions, and the type of outcome measures. The protocol for Prognosis reviews includes the types of studies, the types of participants, and the types of outcome measures. The protocol for Qualitative reviews includes types of studies and types of participants.

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 bibliographic 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. Although the date ranges reflect the time of the review a complete mirror image of the database as it was at the time is impossible as records get added and removed retrospectively so using the date range gives us the best approximation of the content at the moment of the review.

A number of medical databases, and search interfaces to these databases is available for searching, 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)[5]. 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
13. ParaSight.ti,ab
14. Immunochromatograph*.ti,ab
15. Antigen detection method*.ti,ab
16. Rapid malaria antigen test*.ti,ab
```

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

```
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
    ...
```

2.3 Relevance Labels

The original systematic reviews written by Cochrane researchers included a reference section that listed Included, Excluded, and Additional references to 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 full text 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 both Subtask 1 and Subtask 2 (while in 2018 they were included for Subtask 1).

The average percentage of relevant abstract in the training set is 6.5% of the total number of PMID's released, and in the test set 8.9%, while at the content level the average percentage is 2.6% in the training set, and 3.9% in the test set. Table 5, Table 6, Table 7, and Table 8 show the distribution of the relevant documents at abstract and document level for all the topics in the test set. A break down of the average percentage of relevant abstracts/documents are: DTA 12.9%/5.3%, Intervention 7.6%/3.4%, Prognosis 15.7%/9.4%, Qualitative 2.6%/1.0%.

3 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.

3.1 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;
4. the PubMed database, provided by the National Center for Biotechnology Information (NCBI), part of the U.S. National Library of Medicine (NLM).

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 an unlimitted number of 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.

3.2 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.

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 an unlimitted number of runs, and the format of each submission follows the format of subtask 1 submissions.

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.

For the evaluation of the two runs we employ a number of standard IR measures, along with measures that have been developed for the particular task of technology assisted reviews [**?**,**?**]. 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, $WSS@r = (TN + FN)/N(1 - r)$ [**?**]
  4. Reliability = $loss_r + loss_e$ [**?**], 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 Results

The 2019 task received submissions from 3 teams, all from Europe, including one team from The Netherlands (UvA), one team from the UK (Sheffield), and one team from Italy (UNIPD). For Subtask 1, we received no runs. For Subtask 2, we received 36 runs from the three teams. The three teams used a variety of ranking methods including traditional BM25, interactive BM25, continuous active learning, relevance feedback, as well as a variety of stopping criteria to provide a threshold on the ranking. The results on a selected subset of metrics on DTA, Intervention, Prognosis, and Qualitative studies, on abstract-level relevance, are shown in Tables 9, 7, 11, 12, respectively. Figures 1, 2, 3, and 4 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. Figures 9, 10, 11, 12 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 figures present the Pareto frontier.

6 Conclusions

The CLEF e-Health TAR has now constructed a benchmark collection of 80 Diagnostic Test Accuracy, 40 Intervention, 1 Prognosis, and 2 Qualitative systematic reviews to study the effectiveness and efficiency of information retrieval and machine learning algorithms in retrieving relevant studies from medical databses, and prioritizing the studies to be screened at the abstract and title screening stage, while 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. 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.

7 Appendix: Tables and Figures

Topic ID	Topic Title
CD012567	Positron emission tomography (PET) and magnetic resonance imaging (MRI) for assessing tumour resectability in advanced epithelial ovarian/fallopian tube/primary peritoneal cancer
CD012669	Point-of-care ultrasonography for diagnosing thoracoabdominal injuries in patients with blunt trauma
CD012233	Transabdominal ultrasound and endoscopic ultrasound for diagnosis of gallbladder polyps
CD008874	Airway physical examination tests for detection of difficult airway management in apparently normal adult patients
CD012768	Xpert MTB/RIF assay for extrapulmonary tuberculosis and rifampicin resistance
CD012080	Non-invasive diagnostic tests for Helicobacter pylori infection
CD011686	Triage tools for detecting cervical spine injury in pediatric trauma patients
CD009044	Diagnostic tests for autism spectrum disorder (ASD) in preschool children

Table 1. The provided to participants set of testing DTA topics.

Topic ID	Topic Title
CD010239	Lower versus higher oxygen concentrations titrated to target oxygen saturations during resuscitation of preterm infants at birth
CD012551	Non-pharmacological interventions for treating chronic prostatitis/chronic pelvic pain syndrome
CD011571	Antistreptococcal interventions for guttate and chronic plaque psoriasis
CD011140	Implantable miniature telescope (IMT) for vision loss due to end-stage age-related macular degeneration
CD012455	Melatonin for the promotion of sleep in adults in the intensive care unit
CD009642	Continuous intravenous perioperative lidocaine infusion for postoperative pain and recovery in adults
CD007867	Prescribed hypocaloric nutrition support for critically-ill adults
CD011768	Educational interventions for improving primary caregiver complementary feeding practices for children aged 24 months and under
CD011977	Blue-light filtering intraocular lenses (IOLs) for protecting macular health
CD012164	Subfascial endoscopic perforator surgery (SEPS) for treating venous leg ulcers
CD010038	Face-to-face interventions for informing or educating parents about early childhood vaccination
CD009069	Prophylactic vaccination against human papillomaviruses to prevent cervical cancer and its precursors
CD001261	Vaccines for preventing typhoid fever
CD010753	Antidepressants for insomnia in adults
CD006468	Anticoagulation for people with cancer and central venous catheters
CD010558	Psychological therapies for treatment-resistant depression in adults
CD000996	Inhaled corticosteroids for bronchiectasis
CD012069	Methylphenidate for attention deficit hyperactivity disorder (ADHD) in children and adolescents âĂŞ assessment of adverse events in non-randomised studies
CD004414	Interventions for preventing occupational irritant hand dermatitis
CD012342	Comparison of a therapeutic-only versus prophylactic platelet transfusion policy for people with congenital or acquired bone marrow failure disorders

Table 2. The provided to participants set of testing Intervention topics.

Topic ID	Topic Title
CD012661	Development of type 2 diabetes mellitus in people with intermediate hyperglycaemia

Table 3. The provided to participants set of testing Prognosis topics.

Topic ID	Topic Title
CD011787	Parents' and informal caregivers' views and experiences of communication about routine childhood vaccination: a synthesis of qualitative evidence
CD011558	Factors that influence the provision of intrapartum and postnatal care by skilled birth attendants in low- and middle-income countries: a qualitative evidence synthesis

Table 4. The provided to participants set of testing Qualitative topics.

Table 5. Statistics of topics in the test set of the DTA studies.

Topic	# total PMIDs	# abs rel	# doc rel	% abs rel	% doc rel
Diagnostic Test Accuracy					
CD008874	2382	130	121	0.055	0.051
CD009044	3169	47	8	0.015	0.003
CD011686	9729	74	3	0.008	0.000
CD012080	6643	85	85	0.013	0.013
CD012233	472	54	10	0.114	0.021
CD012567	6735	12	5	0.002	0.001
CD012669	1260	82	31	0.065	0.025
CD012768	131	100	41	0.763	0.313

Table 6. Statistics of topics in the test set of the Intervention studies.

Topic	# total PMIDs	# abs rel	# doc rel	% abs rel	% doc rel
Intervention					
CD000996	281	10	6	0.036	0.021
CD001261	571	85	26	0.149	0.046
CD004414	336	32	13	0.095	0.039
CD006468	3874	91	15	0.023	0.004
CD007867	943	31	15	0.033	0.016
CD009069	1757	94	6	0.054	0.003
CD009642	1922	90	72	0.047	0.037
CD010038	8867	36	12	0.004	0.001
CD010239	224	23	12	0.103	0.054
CD010558	2815	75	16	0.027	0.006
CD010753	2539	35	21	0.014	0.008
CD011140	289	4	0	0.014	0.000
CD011571	146	21	6	0.144	0.041
CD011768	9160	81	31	0.009	0.003
CD011977	195	65	38	0.333	0.195
CD012069	3479	425	327	0.122	0.094
CD012164	61	10	3	0.164	0.049
CD012342	2353	9	0	0.004	0.000
CD012455	1593	12	5	0.008	0.003
CD012551	591	86	34	0.146	0.058

Table 7. Statistics of topics in the test set of the Prognosis studies.

Topic	# total PMIDs	# abs rel	# doc rel	% abs rel	% doc rel
Prognosis					
CD012661	3367	527	317	0.157	0.094

Table 8. Statistics of topics in the test set of the Qualitative studies.

Topic	# total PMIDs	# abs rel	# doc rel	% abs rel	% doc rel
Qualitative					
CD011558	2168	51	27	0.024	0.012
CD011787	4369	125	34	0.029	0.008

Table 9. DTA studies with abstract-level QRELs

Run	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Rely	R@k	k
ILPS/DTA/abs-hh-ratio-ilps@uva.out	2420	0.493	0.589	0.682	0.789	0.834	0.406	0.304	0.189	0.815	1132
ILPS/DTA/abs-th-ratio-ilps@uva.out	2676	0.399	0.418	0.536	0.661	0.734	0.312	0.253	0.273	0.744	1558
Padua/DTA/2018_stem_original_p10_t400.out	1190	0.229	0.448	0.634	0.818	0.895	0.662	0.512	0.136	0.963	605
Padua/DTA/distributed_effort_p10_t1500.out	1111	0.229	0.445	0.63	0.814	0.895	0.652	0.513	0.204	0.963	2453
Padua/DTA/2018_stem_original_p10_t1000.out	1141	0.229	0.445	0.63	0.814	0.893	0.658	0.509	0.19	0.986	1195
Padua/DTA/2018_stem_original_p10_t200.out	1282	0.229	0.445	0.634	0.823	0.891	0.66	0.507	0.115	0.877	336
Padua/DTA/2018_stem_original_p10_t500.out	1200	0.229	0.445	0.634	0.818	0.893	0.662	0.509	0.147	0.97	719
Padua/DTA/2018_stem_original_p10_t300.out	1280	0.229	0.452	0.627	0.816	0.893	0.66	0.5	0.113	0.936	477
Padua/DTA/2018_stem_original_p10_t1500.out	1126	0.229	0.445	0.63	0.814	0.895	0.657	0.514	0.228	0.995	1524
Padua/DTA/distributed_effort_p10_t1000.out	1109	0.229	0.445	0.63	0.814	0.895	0.649	0.514	0.129	0.93	1776
Padua/DTA/2018_stem_original_p10_t100.out	2024	0.221	0.418	0.609	0.791	0.868	0.525	0.399	0.291	0.604	180
Padua/DTA/baseline_bm25_t500.out	2470	0.119	0.236	0.402	0.548	0.65	0.39	0.252	0.342	0.638	451
Padua/DTA/distributed_effort_p10_t300.out	1111	0.232	0.445	0.63	0.814	0.886	0.649	0.528	0.117	0.818	802
Padua/DTA/2018_stem_original_p50_t1000.out	1127	0.229	0.445	0.63	0.811	0.893	0.652	0.528	0.235	0.995	1473
Padua/DTA/distributed_effort_p10_t100.out	1271	0.204	0.439	0.614	0.77	0.839	0.61	0.468	0.308	0.572	284
Padua/DTA/2018_stem_original_p50_t200.out	1291	0.229	0.445	0.634	0.82	0.898	0.66	0.499	0.141	0.89	364
Padua/DTA/baseline_bm25_t1000.out	2395	0.119	0.236	0.389	0.543	0.659	0.396	0.26	0.274	0.761	826
Padua/DTA/distributed_effort_p10_t500.out	1116	0.229	0.445	0.63	0.814	0.891	0.634	0.521	0.096	0.874	1083
Padua/DTA/baseline_bm25_t300.out	2493	0.119	0.239	0.405	0.541	0.652	0.391	0.244	0.415	0.499	280
Padua/DTA/baseline_bm25_t100.out	2130	0.12	0.239	0.414	0.564	0.659	0.394	0.295	0.683	0.241	101
Padua/DTA/2018_stem_original_p50_t400.out	1189	0.229	0.448	0.634	0.816	0.891	0.654	0.527	0.154	0.965	672
Padua/DTA/2018_stem_original_p50_t300.out	1272	0.229	0.452	0.627	0.814	0.893	0.656	0.518	0.146	0.945	522
Padua/DTA/2018_stem_original_p50_t100.out	2027	0.222	0.418	0.609	0.786	0.868	0.549	0.394	0.308	0.618	189
Padua/DTA/distributed_effort_p10_t200.out	1194	0.225	0.445	0.632	0.811	0.877	0.663	0.509	0.17	0.735	566
Padua/DTA/baseline_bm25_t400.out	2492	0.119	0.239	0.405	0.539	0.65	0.386	0.246	0.355	0.596	367
Padua/DTA/2018_stem_original_p50_t1500.out	1056	0.229	0.445	0.63	0.814	0.898	0.651	0.537	0.31	1.0	2018
Padua/DTA/2018_stem_original_p50_t500.out	1200	0.229	0.445	0.634	0.809	0.889	0.649	0.524	0.169	0.97	820
Padua/DTA/baseline_bm25_t1500.out	2476	0.119	0.236	0.389	0.541	0.652	0.364	0.254	0.256	0.853	1171
Padua/DTA/baseline_bm25_t200.out	2253	0.12	0.234	0.405	0.55	0.652	0.409	0.278	0.504	0.407	192
Padua/DTA/distributed_effort_p10_t400.out	1116	0.231	0.445	0.63	0.814	0.886	0.634	0.528	0.1	0.856	942
Sheffield/DTA/DTA_sheffield-Chi-Squared.out	1964	0.222	0.305	0.45	0.641	0.73	0.475	0.375	0.479	1.0	3815
Sheffield/DTA/DTA_sheffield-baseline.out	2250	0.175	0.22	0.336	0.525	0.675	0.451	0.338	0.479	1.0	3815
Sheffield/DTA/DTA_sheffield-Odds_Ratio.out	2184	0.248	0.382	0.561	0.707	0.805	0.49	0.347	0.479	1.0	3815
Sheffield/DTA/DTA_sheffield-Log_Likelihood.out	1972	0.234	0.35	0.527	0.668	0.759	0.487	0.381	0.479	1.0	3815

Table 10. Intervention studies with abstract-level QRELs

Run	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Rely	R@k	k
ILPS/Int/abs-hh-ratio-ilps@uva.out	958	0.567	0.518	0.628	0.736	0.813	0.526	0.48	0.213	0.915	773
ILPS/Int/abs-th-ratio-ilps@uva.out	986	0.556	0.478	0.576	0.692	0.774	0.535	0.45	0.197	0.868	555
Padua/Int/2018_stem_original_p10_t400.out	985	0.28	0.307	0.502	0.663	0.744	0.632	0.511	0.334	0.941	487
Padua/Int/distributed_effort_p10_t1500.out	981	0.28	0.306	0.499	0.664	0.745	0.633	0.517	0.247	0.968	1349
Padua/Int/2018_stem_original_p10_t1000.out	977	0.28	0.306	0.499	0.664	0.745	0.63	0.51	0.415	0.973	870
Padua/Int/2018_stem_original_p10_t200.out	1180	0.28	0.312	0.501	0.671	0.775	0.617	0.488	0.267	0.901	301
Padua/Int/2018_stem_original_p10_t500.out	975	0.28	0.306	0.502	0.662	0.742	0.63	0.514	0.353	0.946	560
Padua/Int/2018_stem_original_p10_t300.out	1141	0.28	0.313	0.496	0.665	0.771	0.617	0.494	0.322	0.922	405
Padua/Int/2018_stem_original_p10_t1500.out	952	0.28	0.306	0.499	0.664	0.745	0.63	0.522	0.474	0.984	1117
Padua/Int/distributed_effort_p10_t1000.out	992	0.279	0.306	0.499	0.664	0.745	0.62	0.492	0.157	0.921	975
Padua/Int/2018_stem_original_p10_t100.out	1153	0.274	0.306	0.483	0.639	0.737	0.54	0.474	0.292	0.711	164
Padua/Int/baseline_bm25_t500.out	1233	0.222	0.191	0.282	0.41	0.515	0.435	0.394	0.481	0.741	402
Padua/Int/distributed_effort_p10_t300.out	974	0.276	0.306	0.499	0.664	0.733	0.592	0.481	0.122	0.794	441
Padua/Int/2018_stem_original_p50_t1000.out	836	0.29	0.306	0.498	0.688	0.795	0.643	0.542	0.493	0.988	1139
Padua/Int/distributed_effort_p10_t100.out	1114	0.248	0.315	0.444	0.604	0.704	0.458	0.372	0.402	0.45	156
Padua/Int/2018_stem_original_p50_t200.out	1185	0.29	0.312	0.499	0.693	0.792	0.63	0.481	0.331	0.911	334
Padua/Int/baseline_bm25_t1000.out	1241	0.222	0.191	0.282	0.408	0.524	0.446	0.392	0.471	0.827	682
Padua/Int/distributed_effort_p10_t500.out	991	0.278	0.306	0.499	0.664	0.743	0.606	0.483	0.115	0.842	594
Padua/Int/baseline_bm25_t300.out	1262	0.222	0.187	0.286	0.41	0.523	0.44	0.398	0.506	0.664	270
Padua/Int/baseline_bm25_t100.out	1397	0.223	0.186	0.291	0.429	0.557	0.414	0.368	0.485	0.507	99
Padua/Int/2018_stem_original_p50_t400.out	985	0.29	0.307	0.501	0.685	0.767	0.646	0.514	0.374	0.949	572
Padua/Int/2018_stem_original_p50_t300.out	1144	0.29	0.313	0.495	0.682	0.788	0.639	0.497	0.355	0.933	462
Padua/Int/2018_stem_original_p50_t100.out	1150	0.284	0.306	0.483	0.653	0.752	0.556	0.481	0.362	0.728	188
Padua/Int/distributed_effort_p10_t200.out	965	0.271	0.306	0.482	0.651	0.752	0.56	0.445	0.165	0.714	312
Padua/Int/baseline_bm25_t400.out	1242	0.222	0.191	0.286	0.412	0.523	0.434	0.393	0.485	0.713	337
Padua/Int/2018_stem_original_p50_t1500.out	796	0.29	0.306	0.498	0.688	0.785	0.642	0.553	0.541	0.999	1425
Padua/Int/2018_stem_original_p50_t500.out	1001	0.29	0.306	0.501	0.691	0.779	0.65	0.505	0.395	0.961	677
Padua/Int/baseline_bm25_t1500.out	1203	0.222	0.191	0.282	0.411	0.533	0.453	0.399	0.461	0.933	932
Padua/Int/baseline_bm25_t200.out	1263	0.222	0.189	0.284	0.417	0.535	0.438	0.396	0.466	0.624	191
Padua/Int/distributed_effort_p10_t400.out	981	0.277	0.306	0.499	0.663	0.734	0.595	0.483	0.116	0.822	518
Sheffield/Int/Int_sheffield-Log_likelihood.out	1132	0.293	0.258	0.378	0.583	0.695	0.458	0.381	0.599	1	2100
Sheffield/Int/Int_sheffield-Odds_Ratio.out	1070	0.261	0.267	0.404	0.569	0.7	0.462	0.384	0.599	1	2100
Sheffield/Int/Int_sheffield-baseline.out	1276	0.245	0.22	0.334	0.507	0.653	0.47	0.386	0.599	1	2100
Sheffield/Int/Int_sheffield-Chi_Squared.out	1149	0.262	0.238	0.36	0.537	0.687	0.469	0.415	0.599	1	2100

Table 11. Prognosis studies with abstract-level QRELs

Run	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Rely	R@k	k
ILPS/Pro/abs/abs-hh-ratio-ilps@uva	2885	0.673	0.562	0.714	0.875	0.911	0.591	0.143	0.018	0.948	1221
ILPS/Pro/abs/abs-th-ratio-ilps@uva	2537	0.628	0.521	0.682	0.818	0.927	0.566	0.247	0.014	0.922	867
Padua/Pro/abs/2018_stem_original_p10_t400	2967	0.235	0.214	0.484	0.812	0.901	0.567	0.119	0.035	0.828	735
Padua/Pro/abs/distributed_effort_p10_t1500	2594	0.235	0.214	0.484	0.812	0.896	0.554	0.23	0.049	0.99	2165
Padua/Pro/abs/2018_stem_original_p10_t1000	2644	0.235	0.214	0.484	0.812	0.896	0.554	0.215	0.022	0.943	1332
Padua/Pro/abs/2018_stem_original_p10_t200	2911	0.242	0.214	0.536	0.812	0.901	0.53	0.135	0.162	0.599	398
Padua/Pro/abs/2018_stem_original_p10_t500	2920	0.235	0.214	0.484	0.812	0.891	0.56	0.133	0.027	0.859	832
Padua/Pro/abs/2018_stem_original_p10_t300	2955	0.239	0.214	0.547	0.818	0.891	0.556	0.122	0.054	0.776	597
Padua/Pro/abs/2018_stem_original_p10_t1500	2578	0.235	0.214	0.484	0.812	0.896	0.554	0.234	0.035	0.984	1831
Padua/Pro/abs/distributed_effort_p10_t1000	2563	0.235	0.214	0.484	0.812	0.896	0.554	0.239	0.026	0.974	1566
Padua/Pro/abs/2018_stem_original_p10_t100	2802	0.259	0.286	0.562	0.797	0.891	0.6	0.168	0.411	0.359	198
Padua/Pro/abs/baseline_bm25_t500	3343	0.071	0.057	0.13	0.281	0.422	0.084	0.007	0.621	0.214	501
Padua/Pro/abs/distributed_effort_p10_t300	2964	0.235	0.214	0.484	0.812	0.906	0.567	0.12	0.038	0.818	709
Padua/Pro/abs/2018_stem_original_p50_t1000	2556	0.221	0.214	0.484	0.74	0.87	0.571	0.241	0.041	0.995	1981
Padua/Pro/abs/distributed_effort_p10_t100	2789	0.252	0.25	0.568	0.786	0.875	0.594	0.172	0.288	0.464	248
Padua/Pro/abs/2018_stem_original_p50_t200	2911	0.242	0.214	0.536	0.812	0.901	0.53	0.135	0.162	0.599	398
Padua/Pro/abs/baseline_bm25_t1000	3346	0.07	0.057	0.13	0.276	0.396	0.057	0.006	0.382	0.391	1001
Padua/Pro/abs/distributed_effort_p10_t500	2708	0.235	0.214	0.484	0.812	0.891	0.566	0.196	0.026	0.87	955
Padua/Pro/abs/baseline_bm25_t300	3350	0.071	0.057	0.135	0.276	0.385	0.104	0.005	0.794	0.109	301
Padua/Pro/abs/baseline_bm25_t100	3350	0.066	0.047	0.13	0.255	0.365	0.059	0.005	0.939	0.031	101
Padua/Pro/abs/2018_stem_original_p50_t400	2955	0.231	0.214	0.484	0.807	0.896	0.556	0.122	0.033	0.839	798
Padua/Pro/abs/2018_stem_original_p50_t300	2955	0.239	0.214	0.547	0.818	0.891	0.556	0.122	0.054	0.776	597
Padua/Pro/abs/2018_stem_original_p50_t100	2802	0.259	0.286	0.562	0.797	0.891	0.6	0.168	0.411	0.359	198
Padua/Pro/abs/distributed_effort_p10_t200	2968	0.24	0.214	0.542	0.807	0.906	0.548	0.119	0.079	0.724	501
Padua/Pro/abs/baseline_bm25_t400	3347	0.071	0.057	0.13	0.281	0.417	0.109	0.006	0.696	0.167	401
Padua/Pro/abs/2018_stem_original_p50_t1500	1975	0.219	0.214	0.484	0.74	0.828	0.5	0.413	0.091	1	2966
Padua/Pro/abs/2018_stem_original_p50_t500	2660	0.228	0.214	0.484	0.807	0.891	0.576	0.21	0.022	0.891	993
Padua/Pro/abs/baseline_bm25_t1500	3346	0.07	0.057	0.13	0.276	0.396	0.05	0.006	0.258	0.516	1501
Padua/Pro/abs/baseline_bm25_t200	3350	0.069	0.057	0.125	0.266	0.385	0.111	0.005	0.86	0.073	201
Padua/Pro/abs/distributed_effort_p10_t400	2920	0.235	0.214	0.484	0.812	0.891	0.56	0.133	0.028	0.854	830
Sheffield/Pro/abs/Pro_sheffield-baseline	2990	0.126	0.146	0.255	0.448	0.594	0.247	0.112	0.117	1	3367
Sheffield/Pro/abs/Pro_sheffield-relevence_feedback	2775	0.141	0.151	0.307	0.484	0.646	0.305	0.176	0.117	1	3367

Table 12. Qualitative studies with abstract-level QRELs

Run	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Rely	R@k	k
ILPS/Qual/abs/abs-hh-ratio-ilps@uva.out	1796	0.204	0.478	0.655	0.876	0.929	0.417	0.397	0.326	0.919	1247
ILPS/Qual/abs/abs-th-ratio-ilps@uva.out	2564	0.187	0.487	0.628	0.805	0.92	0.398	0.215	0.341	0.878	1158
Padua/Qual/abs/2018_stem_original_p10_t400.out	2547	0.109	0.496	0.717	0.779	0.894	0.302	0.183	0.568	0.387	704
Padua/Qual/abs/distributed_effort_p10_t1500.out	2544	0.109	0.496	0.743	0.77	0.885	0.268	0.168	0.37	0.745	2098
Padua/Qual/abs/2018_stem_original_p10_t1000.out	2662	0.109	0.496	0.743	0.77	0.885	0.273	0.141	0.29	0.714	1320
Padua/Qual/abs/2018_stem_original_p10_t200.out	2934	0.089	0.478	0.522	0.699	0.805	0.216	0.101	0.627	0.266	397
Padua/Qual/abs/2018_stem_original_p10_t500.out	2535	0.109	0.496	0.743	0.77	0.894	0.301	0.185	0.578	0.396	820
Padua/Qual/abs/2018_stem_original_p10_t300.out	2660	0.103	0.496	0.655	0.752	0.858	0.303	0.159	0.582	0.338	554
Padua/Qual/abs/2018_stem_original_p10_t1500.out	2534	0.109	0.496	0.743	0.77	0.885	0.268	0.17	0.447	0.732	1819
Padua/Qual/abs/distributed_effort_p10_t1000.out	2469	0.109	0.496	0.743	0.77	0.885	0.295	0.199	0.628	0.491	1515
Padua/Qual/abs/2018_stem_original_p10_t100.out	2996	0.071	0.327	0.416	0.637	0.796	0.186	0.09	0.726	0.167	198
Padua/Qual/abs/baseline_bm25_t500.out	2700	0.051	0.274	0.425	0.469	0.611	0.412	0.256	0.683	0.221	501
Padua/Qual/abs/distributed_effort_p10_t300.out	2518	0.109	0.496	0.743	0.77	0.894	0.309	0.193	0.547	0.396	684
Padua/Qual/abs/2018_stem_original_p50_t1000.out	2438	0.116	0.496	0.743	0.92	0.947	0.357	0.194	0.545	0.745	1977
Padua/Qual/abs/distributed_effort_p10_t100.out	2920	0.083	0.416	0.469	0.681	0.814	0.258	0.106	0.659	0.221	244
Padua/Qual/abs/2018_stem_original_p50_t200.out	2934	0.089	0.478	0.522	0.699	0.805	0.216	0.101	0.627	0.266	397
Padua/Qual/abs/baseline_bm25_t1000.out	3040	0.055	0.274	0.425	0.496	0.788	0.239	0.101	0.278	0.601	1001
Padua/Qual/abs/distributed_effort_p10_t500.out	2641	0.109	0.496	0.743	0.77	0.894	0.295	0.162	0.553	0.446	924
Padua/Qual/abs/baseline_bm25_t300.out	2697	0.049	0.274	0.372	0.451	0.628	0.294	0.257	0.726	0.171	301
Padua/Qual/abs/baseline_bm25_t100.out	2700	0.056	0.301	0.389	0.637	0.743	0.399	0.256	0.845	0.086	101
Padua/Qual/abs/2018_stem_original_p50_t400.out	2566	0.109	0.496	0.717	0.779	0.894	0.293	0.174	0.594	0.387	795
Padua/Qual/abs/2018_stem_original_p50_t300.out	2687	0.103	0.496	0.655	0.752	0.858	0.29	0.147	0.591	0.338	595
Padua/Qual/abs/2018_stem_original_p50_t100.out	2996	0.071	0.327	0.416	0.637	0.796	0.186	0.09	0.726	0.167	198
Padua/Qual/abs/distributed_effort_p10_t200.out	2762	0.104	0.496	0.673	0.761	0.867	0.303	0.135	0.56	0.347	486
Padua/Qual/abs/baseline_bm25_t400.out	2700	0.052	0.274	0.434	0.469	0.619	0.417	0.256	0.694	0.203	401
Padua/Qual/abs/2018_stem_original_p50_t1500.out	1970	0.116	0.496	0.743	0.92	0.965	0.356	0.301	0.532	1	2568
Padua/Qual/abs/2018_stem_original_p50_t500.out	2576	0.11	0.496	0.743	0.788	0.894	0.283	0.168	0.624	0.405	991
Padua/Qual/abs/baseline_bm25_t1500.out	3039	0.055	0.274	0.425	0.496	0.779	0.24	0.101	0.382	0.669	1501
Padua/Qual/abs/baseline_bm25_t200.out	2698	0.053	0.274	0.381	0.619	0.726	0.395	0.256	0.764	0.14	201
Padua/Qual/abs/distributed_effort_p10_t400.out	2636	0.109	0.496	0.743	0.77	0.894	0.301	0.165	0.545	0.432	804
Sheffield/Qual/abs/Qual_sheffield-relevance_feedback.out	2940	0.06	0.274	0.549	0.717	0.832	0.185	0.103	0.593	1	3268
Sheffield/Qual/abs/Qual_sheffield-baseline	3031	0.051	0.265	0.451	0.619	0.743	0.135	0.082	0.593	1	3268

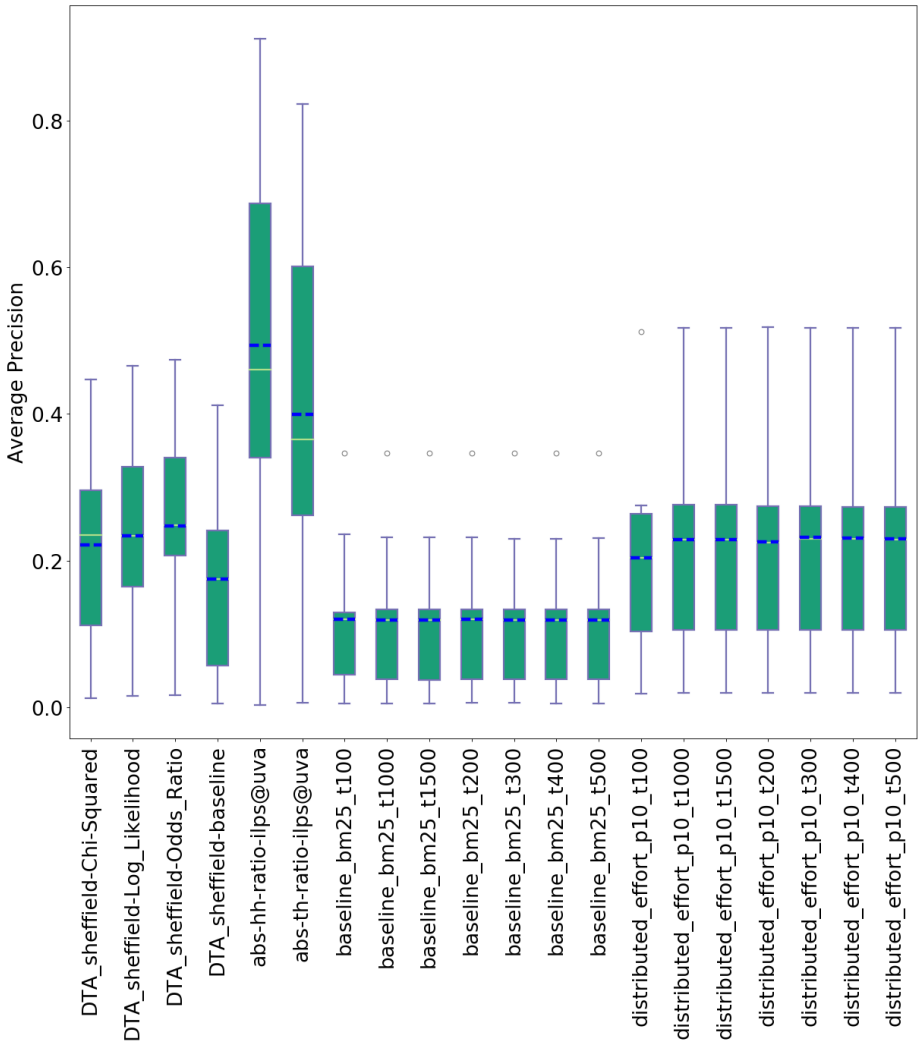

Fig. 1. Average precision using the abstract level relevance judgments for DTA reviews.

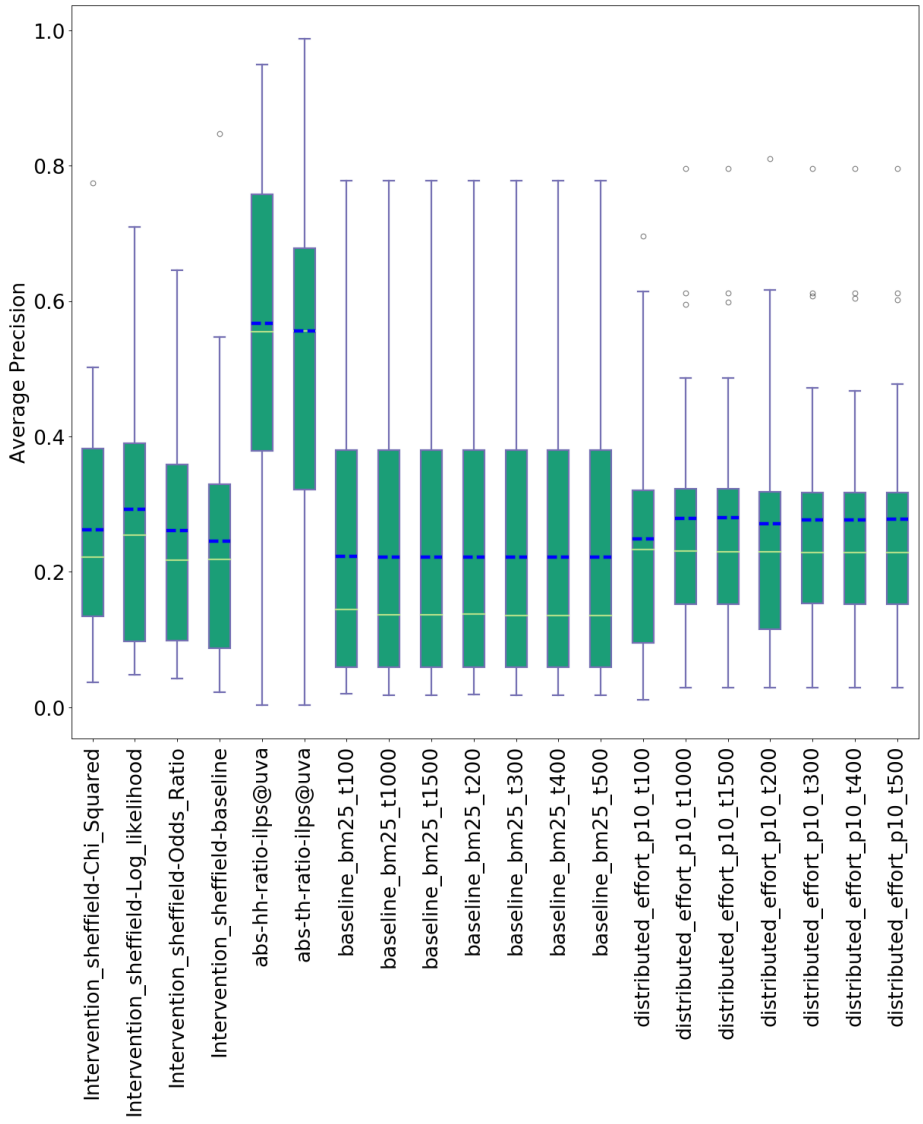

Fig. 2. Average precision using the abstract level relevance judgments for Intervention reviews.

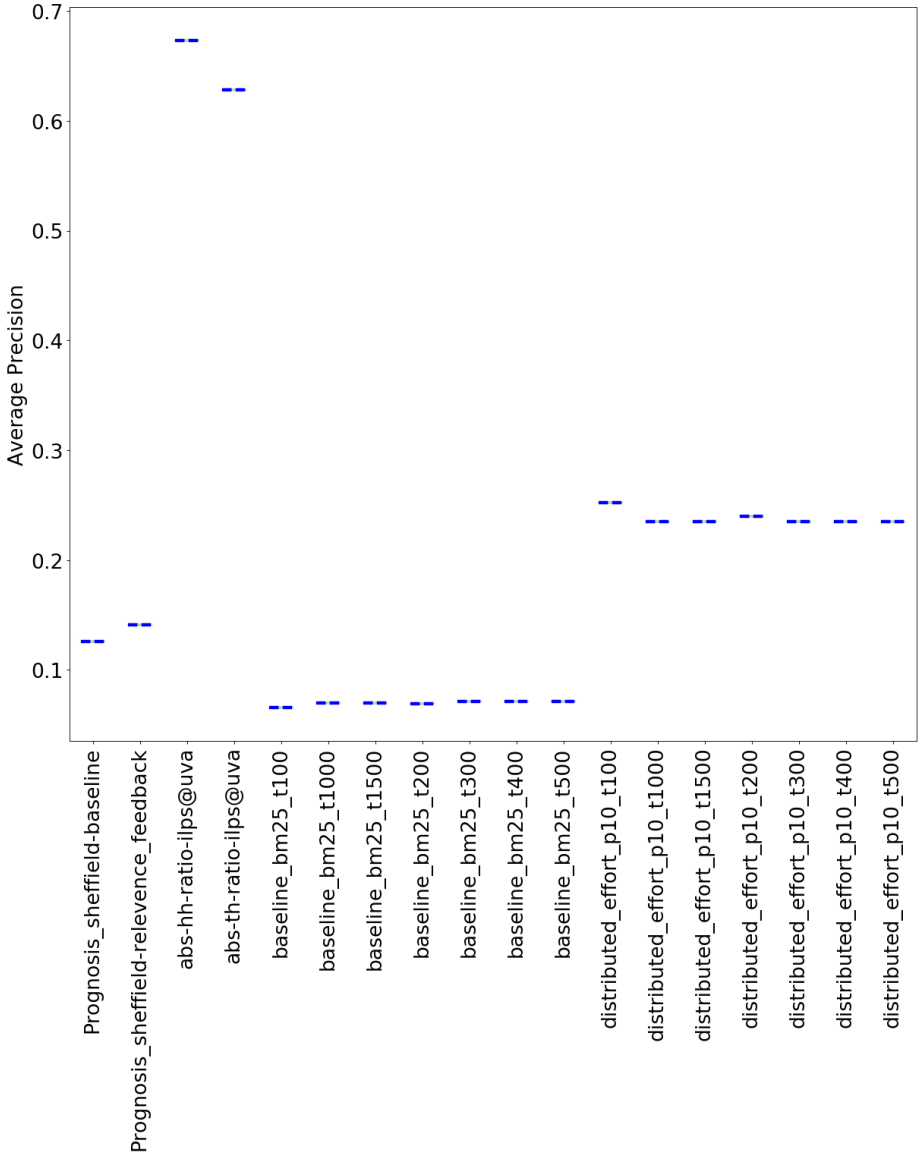

Fig. 3. Average precision using the abstract level relevance judgments for Prognosis reviews.

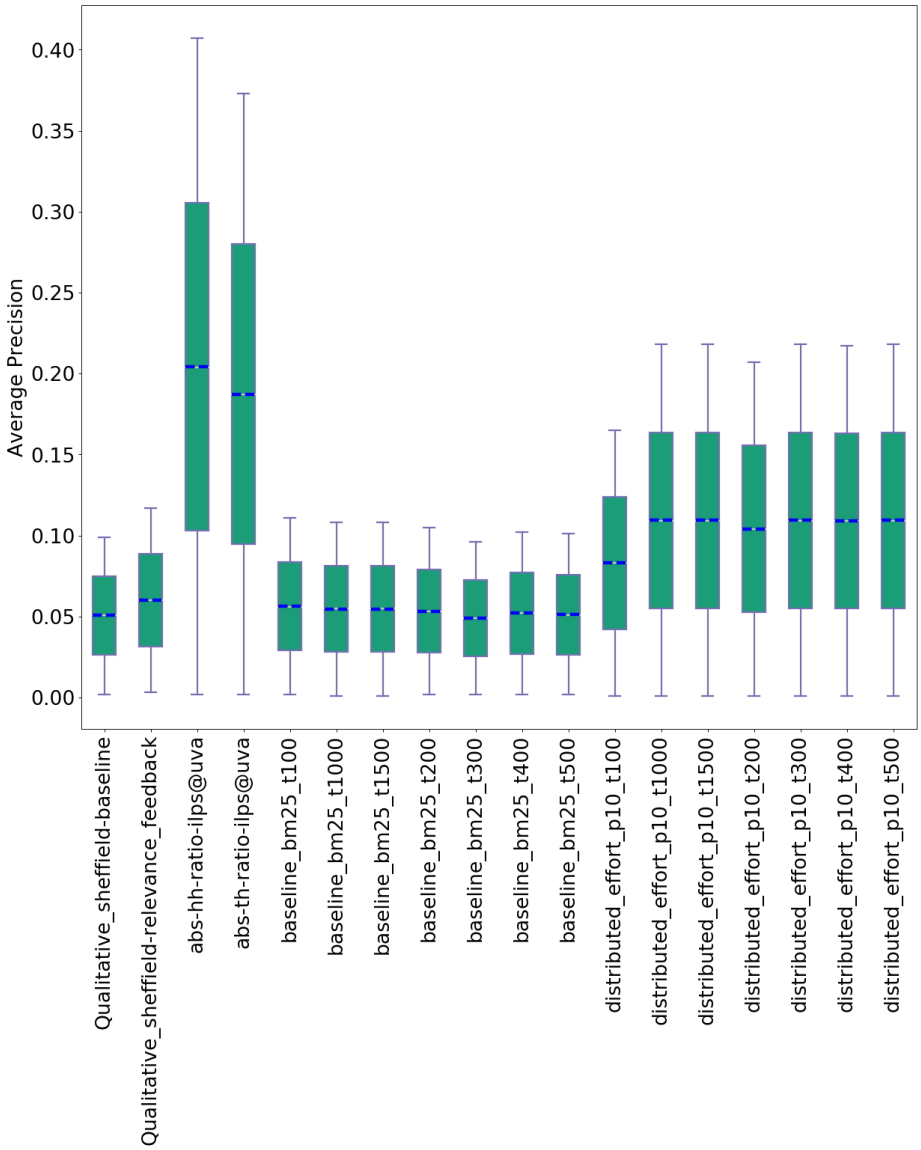

Fig. 4. Average precision using the abstract level relevance judgments for Qualitative reviews.

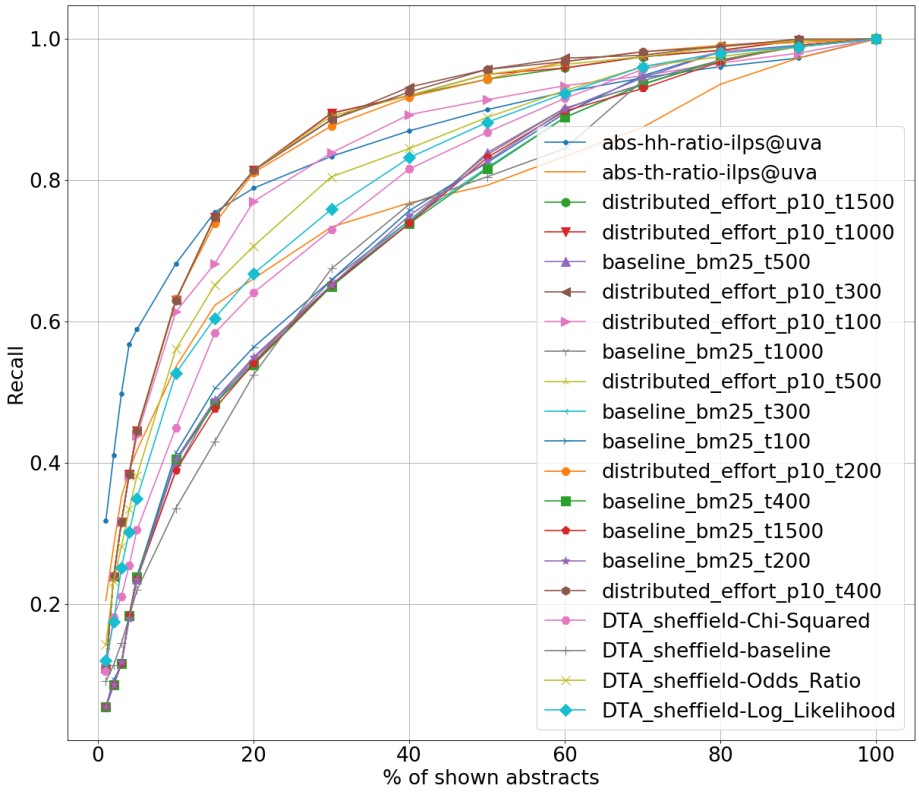

Fig. 5. Recall at different ranks for DTA reviews.

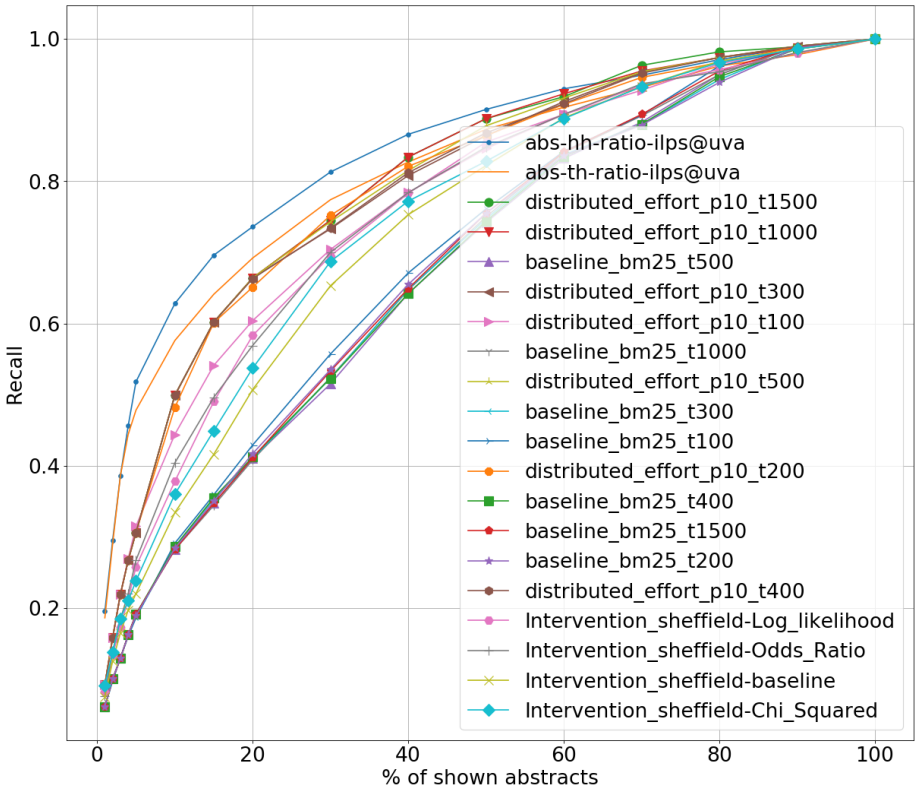

Fig. 6. Recall at different ranks for Intervention reviews.

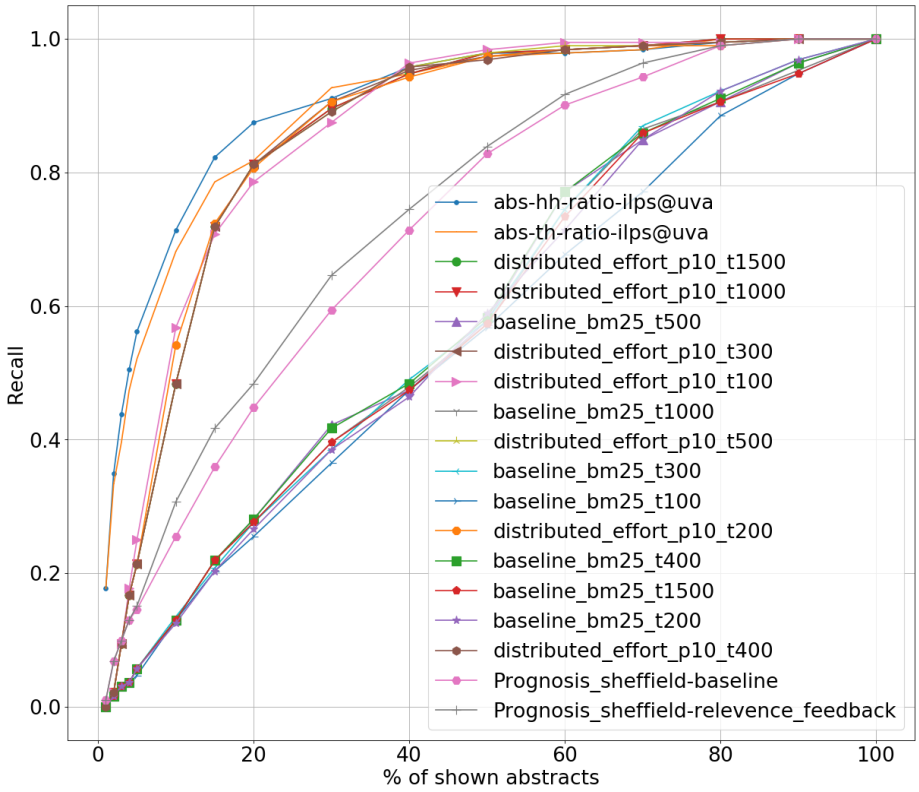

Fig. 7. Recall at different ranks for Prognosis reviews.

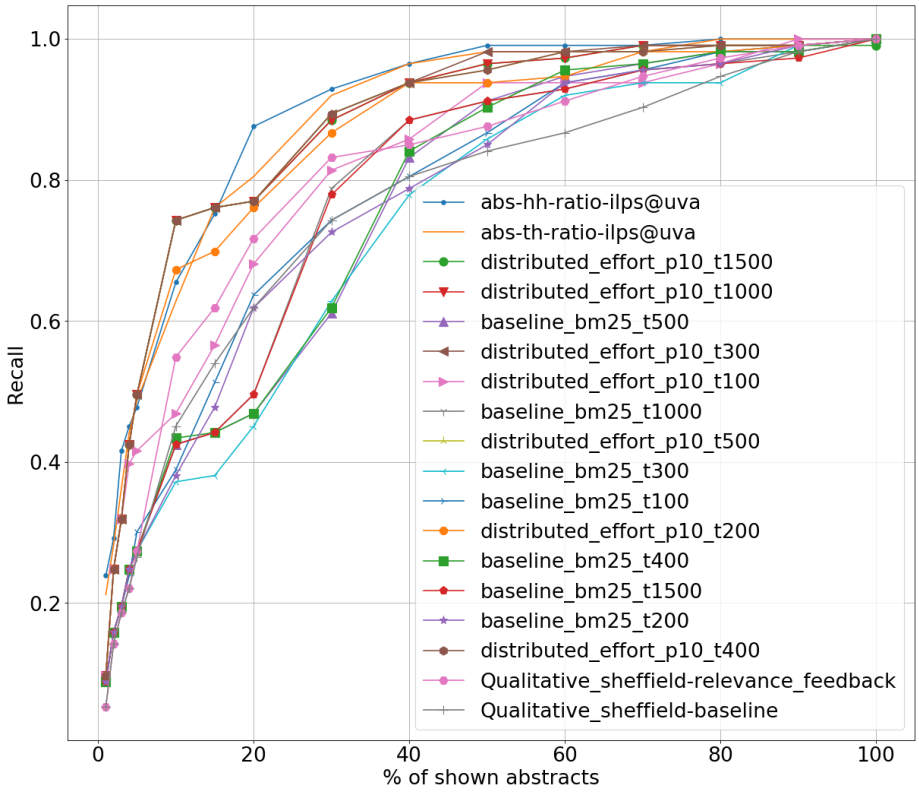

Fig. 8. Recall at different ranks for Qualitative reviews.

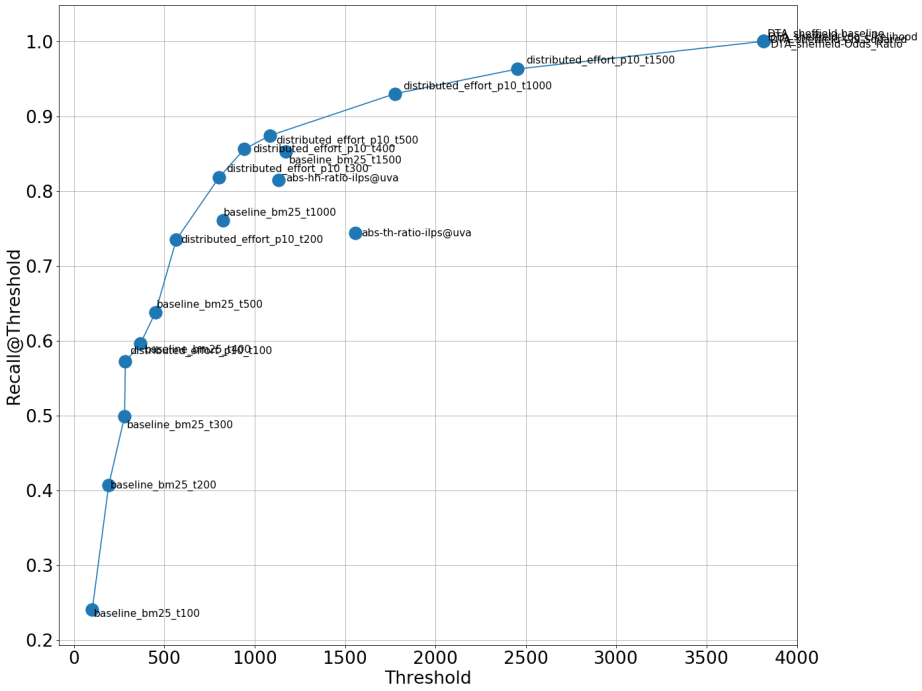

Fig. 9. Recall at the threshold rank as a function of the number of abstracts shown to the user for DTA reviews.

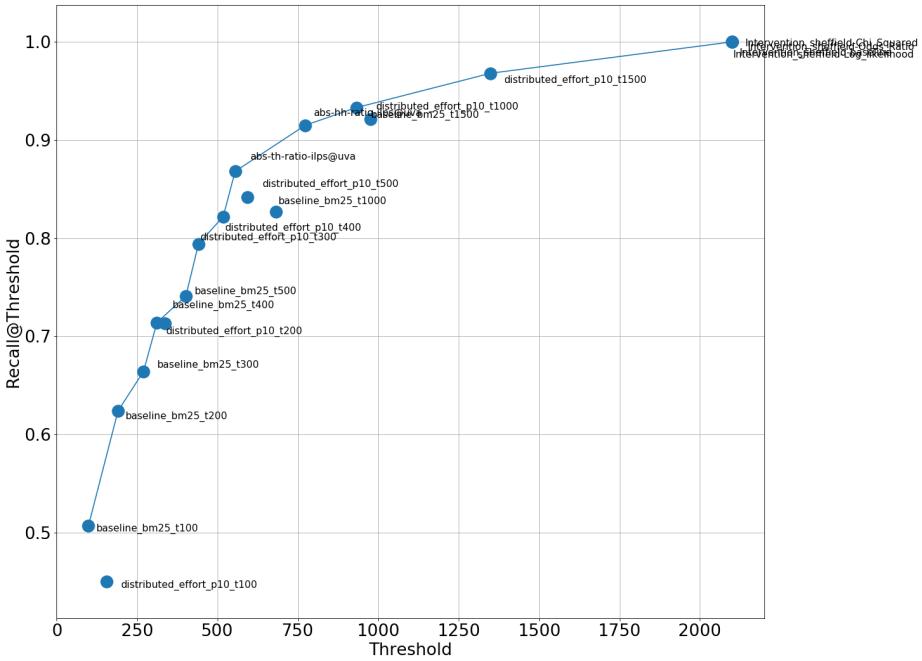

Fig. 10. Recall at the threshold rank as a function of the number of abstracts shown to the user for Intervention reviews.

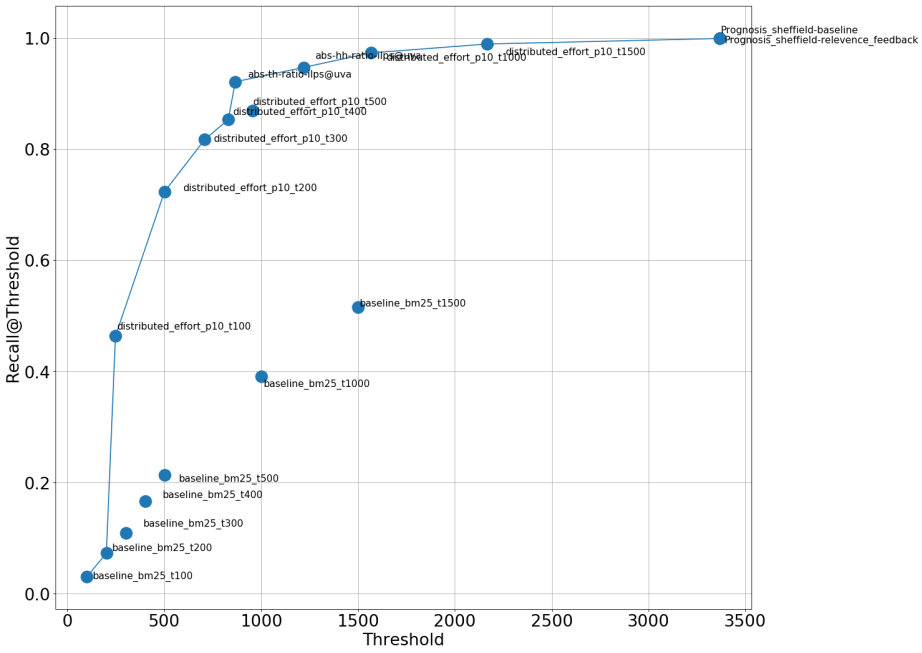

Fig. 11. Recall at the threshold rank as a function of the number of abstracts shown to the user for Prognosis reviews.

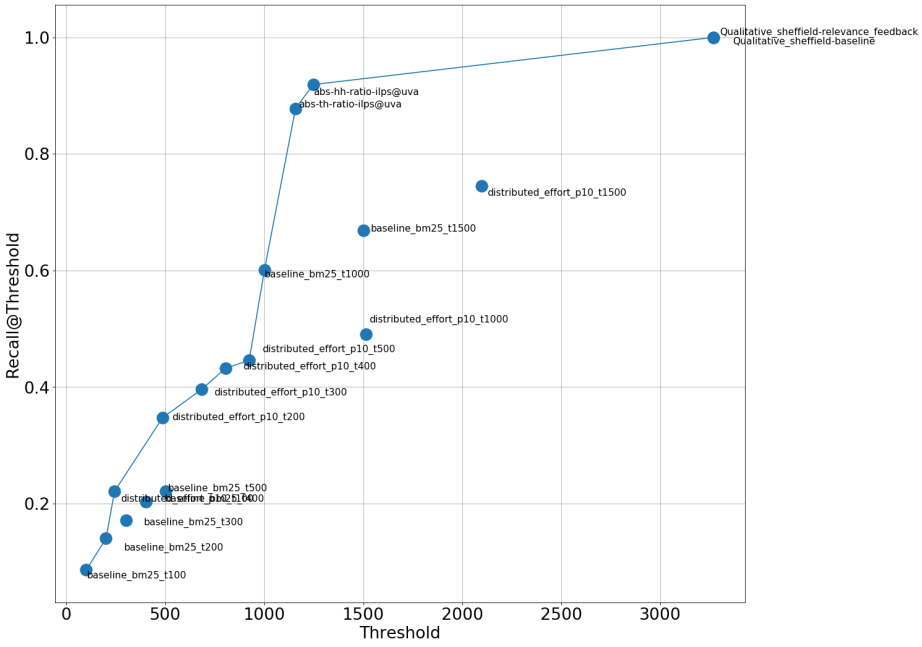

Fig. 12. Recall at the threshold rank as a function of the number of abstracts shown to the user for Qualitative reviews.