CLEF 2019 technology assisted reviews in empirical medicine overview

Kanoulas, E.; Li, D.; Azzopardi, L.; Spijker, R.

Publication date
2019

Document Version
Final published version

Published in
Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum

License
CC BY

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview

Evangelos Kanoulas¹, Dan Li¹, Leif Azzopardi², and Rene Spijker³

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
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 prospec-tively 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

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:

<table>
<thead>
<tr>
<th>Topic: CD008122</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title: Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries</td>
</tr>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

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. René 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/ovidseptools/launcher.htm](http://demo.ovid.com/demo/ovidseptools/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 PMID’s 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.

<table>
<thead>
<tr>
<th>Topic: CD008122</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title: Rapid diagnostic tests for diagnosing uncomplicated P. falciparum malaria in endemic countries</td>
</tr>
<tr>
<td>Query:</td>
</tr>
<tr>
<td>1. Exp Malaria/</td>
</tr>
<tr>
<td>2. Exp Plasmodium/</td>
</tr>
<tr>
<td>3. Malaria.ti,ab</td>
</tr>
<tr>
<td>4. 1 or 2 or 3</td>
</tr>
<tr>
<td>5. Exp Reagent kits, diagnostic/</td>
</tr>
<tr>
<td>6. rapid diagnos* test*.ti,ab</td>
</tr>
<tr>
<td>7. RDT.ti,ab</td>
</tr>
<tr>
<td>8. Dipstick*.ti,ab</td>
</tr>
<tr>
<td>9. Rapid diagnos* device*.ti,ab</td>
</tr>
<tr>
<td>10. MRDD.ti,ab</td>
</tr>
<tr>
<td>11. OptiMal.ti,ab</td>
</tr>
<tr>
<td>12. Binax NOW.ti,ab</td>
</tr>
<tr>
<td>13. ParaSight.ti,ab</td>
</tr>
<tr>
<td>14. Immunochromatograph*.ti,ab</td>
</tr>
<tr>
<td>15. Antigen detection method*.ti,ab</td>
</tr>
<tr>
<td>16. Rapid malaria antigen test*.ti,ab</td>
</tr>
</tbody>
</table>

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/](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/](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:

\[
\begin{array}{cccc}
\text{Topic} & \text{Iteration} & \text{Document} & \text{Relevance} \\
\end{array}
\]

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 unlimited 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 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 unlimited 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 [7,8]. 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, \( WSS_{sr} = (TN + FN)/N(1 - r) \)
  4. Reliability = \( loss_r + loss_e \) [7], 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](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 present 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 databases, 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.
## Appendix: Tables and Figures

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

*Table 1. The provided to participants set of testing DTA topics.*
<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Topic Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD010239</td>
<td>Lower versus higher oxygen concentrations titrated to target oxygen saturations during resuscitation of preterm infants at birth</td>
</tr>
<tr>
<td>CD012551</td>
<td>Non-pharmacological interventions for treating chronic prostatitis/chronic pelvic pain syndrome</td>
</tr>
<tr>
<td>CD011571</td>
<td>Antistreptococcal interventions for guttate and chronic plaque psoriasis</td>
</tr>
<tr>
<td>CD011140</td>
<td>Implantable miniature telescope (IMT) for vision loss due to end-stage age-related macular degeneration</td>
</tr>
<tr>
<td>CD012455</td>
<td>Melatonin for the promotion of sleep in adults in the intensive care unit</td>
</tr>
<tr>
<td>CD009642</td>
<td>Continuous intravenous perioperative lidocaine infusion for postoperative pain and recovery in adults</td>
</tr>
<tr>
<td>CD007867</td>
<td>Prescribed hypocaloric nutrition support for critically-ill adults</td>
</tr>
<tr>
<td>CD011768</td>
<td>Educational interventions for improving primary caregiver complementary feeding practices for children aged 24 months and under</td>
</tr>
<tr>
<td>CD011977</td>
<td>Blue-light filtering intraocular lenses (IOLs) for protecting macular health</td>
</tr>
<tr>
<td>CD012164</td>
<td>Subfascial endoscopic perforator surgery (SEPS) for treating venous leg ulcers</td>
</tr>
<tr>
<td>CD010038</td>
<td>Face-to-face interventions for informing or educating parents about early childhood vaccination</td>
</tr>
<tr>
<td>CD009069</td>
<td>Prophylactic vaccination against human papillomaviruses to prevent cervical cancer and its precursors</td>
</tr>
<tr>
<td>CD001261</td>
<td>Vaccines for preventing typhoid fever</td>
</tr>
<tr>
<td>CD010753</td>
<td>Antidepressants for insomnia in adults</td>
</tr>
<tr>
<td>CD006468</td>
<td>Anticoagulation for people with cancer and central venous catheters</td>
</tr>
<tr>
<td>CD010558</td>
<td>Psychological therapies for treatment-resistant depression in adults</td>
</tr>
<tr>
<td>CD000996</td>
<td>Inhaled corticosteroids for bronchiectasis</td>
</tr>
<tr>
<td>CD012069</td>
<td>Methylphenidate for attention deficit hyperactivity disorder (ADHD) in children and adolescents $\bar{A}$ assessment of adverse events in non-randomised studies</td>
</tr>
<tr>
<td>CD004414</td>
<td>Interventions for preventing occupational irritant hand dermatitis</td>
</tr>
<tr>
<td>CD012342</td>
<td>Comparison of a therapeutic-only versus prophylactic platelet transfusion policy for people with congenital or acquired bone marrow failure disorders</td>
</tr>
</tbody>
</table>

**Table 2.** The provided to participants set of testing Intervention topics.
### Table 3. The provided to participants set of testing Prognosis topics.

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Topic Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD012661</td>
<td>Development of type 2 diabetes mellitus in people with intermediate hyperglycaemia</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Topic Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD011787</td>
<td>Parents’ and informal caregivers’ views and experiences of communication about routine childhood vaccination: a synthesis of qualitative evidence</td>
</tr>
<tr>
<td>CD011558</td>
<td>Factors that influence the provision of intrapartum and postnatal care by skilled birth attendants in low- and middle-income countries: a qualitative evidence synthesis</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Topic</th>
<th># total PMIDs</th>
<th># abs rel</th>
<th># doc rel</th>
<th>% abs rel</th>
<th>% doc rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD008874</td>
<td>2382</td>
<td>130</td>
<td>121</td>
<td>0.055</td>
<td>0.051</td>
</tr>
<tr>
<td>CD009044</td>
<td>3169</td>
<td>47</td>
<td>8</td>
<td>0.015</td>
<td>0.003</td>
</tr>
<tr>
<td>CD011686</td>
<td>9729</td>
<td>74</td>
<td>3</td>
<td>0.008</td>
<td>0.000</td>
</tr>
<tr>
<td>CD012080</td>
<td>6643</td>
<td>85</td>
<td>85</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>CD012233</td>
<td>472</td>
<td>54</td>
<td>10</td>
<td>0.114</td>
<td>0.021</td>
</tr>
<tr>
<td>CD012567</td>
<td>6735</td>
<td>12</td>
<td>5</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>CD012669</td>
<td>1260</td>
<td>82</td>
<td>31</td>
<td>0.065</td>
<td>0.025</td>
</tr>
<tr>
<td>CD012768</td>
<td>131</td>
<td>100</td>
<td>41</td>
<td>0.763</td>
<td>0.313</td>
</tr>
<tr>
<td>Topic</td>
<td># total PMIDs</td>
<td># abs rel</td>
<td># doc rel</td>
<td>% abs rel</td>
<td>% doc rel</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>------------</td>
<td>-----------</td>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td>CD000996</td>
<td>281</td>
<td>10</td>
<td>6</td>
<td>0.036</td>
<td>0.021</td>
</tr>
<tr>
<td>CD001261</td>
<td>571</td>
<td>85</td>
<td>26</td>
<td>0.149</td>
<td>0.046</td>
</tr>
<tr>
<td>CD004414</td>
<td>336</td>
<td>32</td>
<td>13</td>
<td>0.095</td>
<td>0.039</td>
</tr>
<tr>
<td>CD006468</td>
<td>3874</td>
<td>91</td>
<td>15</td>
<td>0.023</td>
<td>0.004</td>
</tr>
<tr>
<td>CD007867</td>
<td>943</td>
<td>31</td>
<td>15</td>
<td>0.033</td>
<td>0.016</td>
</tr>
<tr>
<td>CD009069</td>
<td>1757</td>
<td>94</td>
<td>6</td>
<td>0.054</td>
<td>0.003</td>
</tr>
<tr>
<td>CD009642</td>
<td>1922</td>
<td>90</td>
<td>72</td>
<td>0.047</td>
<td>0.037</td>
</tr>
<tr>
<td>CD010038</td>
<td>8867</td>
<td>36</td>
<td>12</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>CD010239</td>
<td>224</td>
<td>23</td>
<td>12</td>
<td>0.103</td>
<td>0.054</td>
</tr>
<tr>
<td>CD010558</td>
<td>2815</td>
<td>75</td>
<td>16</td>
<td>0.027</td>
<td>0.006</td>
</tr>
<tr>
<td>CD010753</td>
<td>2539</td>
<td>35</td>
<td>21</td>
<td>0.014</td>
<td>0.008</td>
</tr>
<tr>
<td>CD011140</td>
<td>289</td>
<td>4</td>
<td>0</td>
<td>0.014</td>
<td>0.000</td>
</tr>
<tr>
<td>CD011571</td>
<td>146</td>
<td>21</td>
<td>6</td>
<td>0.144</td>
<td>0.041</td>
</tr>
<tr>
<td>CD011768</td>
<td>9160</td>
<td>81</td>
<td>31</td>
<td>0.009</td>
<td>0.003</td>
</tr>
<tr>
<td>CD011977</td>
<td>195</td>
<td>65</td>
<td>38</td>
<td>0.333</td>
<td>0.195</td>
</tr>
<tr>
<td>CD012069</td>
<td>3479</td>
<td>425</td>
<td>327</td>
<td>0.122</td>
<td>0.094</td>
</tr>
<tr>
<td>CD012164</td>
<td>61</td>
<td>10</td>
<td>3</td>
<td>0.164</td>
<td>0.049</td>
</tr>
<tr>
<td>CD012342</td>
<td>2353</td>
<td>9</td>
<td>0</td>
<td>0.004</td>
<td>0.000</td>
</tr>
<tr>
<td>CD012455</td>
<td>1593</td>
<td>12</td>
<td>5</td>
<td>0.008</td>
<td>0.003</td>
</tr>
<tr>
<td>CD012551</td>
<td>591</td>
<td>86</td>
<td>34</td>
<td>0.146</td>
<td>0.058</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th># total PMIDs</th>
<th># abs rel</th>
<th># doc rel</th>
<th>% abs rel</th>
<th>% doc rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prognosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th># total PMIDs</th>
<th># abs rel</th>
<th># doc rel</th>
<th>% abs rel</th>
<th>% doc rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD012661</td>
<td>3367</td>
<td>527</td>
<td>317</td>
<td>0.157</td>
<td>0.094</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th># total PMIDs</th>
<th># abs rel</th>
<th># doc rel</th>
<th>% abs rel</th>
<th>% doc rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th># total PMIDs</th>
<th># abs rel</th>
<th># doc rel</th>
<th>% abs rel</th>
<th>% doc rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD011558</td>
<td>2168</td>
<td>51</td>
<td>27</td>
<td>0.024</td>
<td>0.012</td>
</tr>
<tr>
<td>CD011787</td>
<td>4369</td>
<td>125</td>
<td>34</td>
<td>0.029</td>
<td>0.008</td>
</tr>
<tr>
<td>Run</td>
<td>L_Rel</td>
<td>MAP</td>
<td>R@5%</td>
<td>R@10%</td>
<td>R@20%</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>ILPS/DTA/abs-hh-ratio-ilps@uva.out</td>
<td>2420</td>
<td>0.493</td>
<td>0.682</td>
<td>0.789</td>
<td>0.834</td>
</tr>
<tr>
<td>ILPS/DTA/abs-th-ratio-ilps@uva.out</td>
<td>2676</td>
<td>0.399</td>
<td>0.536</td>
<td>0.661</td>
<td>0.734</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p10_t400.out</td>
<td>1190</td>
<td>0.229</td>
<td>0.634</td>
<td>0.818</td>
<td>0.895</td>
</tr>
<tr>
<td>Padua/DTA/distributed_effort_p10_t1500.out</td>
<td>1111</td>
<td>0.229</td>
<td>0.445</td>
<td>0.814</td>
<td>0.895</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p10_t1000.out</td>
<td>1141</td>
<td>0.229</td>
<td>0.445</td>
<td>0.814</td>
<td>0.893</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p10_t200.out</td>
<td>1282</td>
<td>0.229</td>
<td>0.445</td>
<td>0.823</td>
<td>0.891</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p10_t500.out</td>
<td>1200</td>
<td>0.229</td>
<td>0.445</td>
<td>0.818</td>
<td>0.893</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p10_t300.out</td>
<td>1280</td>
<td>0.229</td>
<td>0.452</td>
<td>0.627</td>
<td>0.816</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p10_t100.out</td>
<td>1190</td>
<td>0.229</td>
<td>0.445</td>
<td>0.634</td>
<td>0.818</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p10_t1500.out</td>
<td>1141</td>
<td>0.229</td>
<td>0.445</td>
<td>0.814</td>
<td>0.895</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p10_t1000.out</td>
<td>2104</td>
<td>0.229</td>
<td>0.445</td>
<td>0.634</td>
<td>0.818</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p50_t1000.out</td>
<td>1127</td>
<td>0.229</td>
<td>0.445</td>
<td>0.811</td>
<td>0.893</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p50_t100.out</td>
<td>1271</td>
<td>0.204</td>
<td>0.439</td>
<td>0.614</td>
<td>0.77</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p50_t200.out</td>
<td>1291</td>
<td>0.229</td>
<td>0.445</td>
<td>0.634</td>
<td>0.82</td>
</tr>
<tr>
<td>Padua/DTA/baseline_bm25_t500.out</td>
<td>2470</td>
<td>0.119</td>
<td>0.402</td>
<td>0.548</td>
<td>0.65</td>
</tr>
<tr>
<td>Padua/DTA/distributed_effort_p10_t300.out</td>
<td>1111</td>
<td>0.232</td>
<td>0.445</td>
<td>0.634</td>
<td>0.814</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p50_t1000.out</td>
<td>1127</td>
<td>0.229</td>
<td>0.445</td>
<td>0.634</td>
<td>0.811</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p50_t100.out</td>
<td>1271</td>
<td>0.204</td>
<td>0.439</td>
<td>0.614</td>
<td>0.77</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p50_t200.out</td>
<td>1291</td>
<td>0.229</td>
<td>0.445</td>
<td>0.634</td>
<td>0.82</td>
</tr>
<tr>
<td>Padua/DTA/baseline_bm25_t500.out</td>
<td>2470</td>
<td>0.119</td>
<td>0.402</td>
<td>0.548</td>
<td>0.65</td>
</tr>
<tr>
<td>Padua/DTA/distributed_effort_p10_t300.out</td>
<td>1111</td>
<td>0.232</td>
<td>0.445</td>
<td>0.634</td>
<td>0.814</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p50_t1000.out</td>
<td>1127</td>
<td>0.229</td>
<td>0.445</td>
<td>0.634</td>
<td>0.811</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p50_t100.out</td>
<td>1271</td>
<td>0.204</td>
<td>0.439</td>
<td>0.614</td>
<td>0.77</td>
</tr>
<tr>
<td>Padua/DTA/2018_stem_original_p50_t200.out</td>
<td>1291</td>
<td>0.229</td>
<td>0.445</td>
<td>0.634</td>
<td>0.82</td>
</tr>
<tr>
<td>Padua/DTA/baseline_bm25_t500.out</td>
<td>2470</td>
<td>0.119</td>
<td>0.402</td>
<td>0.548</td>
<td>0.65</td>
</tr>
<tr>
<td>Padua/DTA/distributed_effort_p10_t300.out</td>
<td>1111</td>
<td>0.232</td>
<td>0.445</td>
<td>0.634</td>
<td>0.814</td>
</tr>
<tr>
<td>Sheffield/DTA/DTA_sheffield-Chi-Squared.out</td>
<td>2133</td>
<td>0.222</td>
<td>0.305</td>
<td>0.45</td>
<td>0.641</td>
</tr>
<tr>
<td>Sheffield/DTA/DTA_sheffield-baseline.out</td>
<td>2250</td>
<td>0.175</td>
<td>0.22</td>
<td>0.336</td>
<td>0.525</td>
</tr>
<tr>
<td>Sheffield/DTA/DTA_sheffield-Odds_Ratio.out</td>
<td>2184</td>
<td>0.248</td>
<td>0.382</td>
<td>0.561</td>
<td>0.707</td>
</tr>
<tr>
<td>Sheffield/DTA/DTA_sheffield-Log_Likelihood.out</td>
<td>1972</td>
<td>0.234</td>
<td>0.35</td>
<td>0.527</td>
<td>0.668</td>
</tr>
</tbody>
</table>
### Table 10. Intervention studies with abstract-level QRELs

<table>
<thead>
<tr>
<th>Run</th>
<th>L_Rel</th>
<th>MAP</th>
<th>R@5%</th>
<th>R@10%</th>
<th>R@20%</th>
<th>R@30%</th>
<th>WSS95</th>
<th>WSS100</th>
<th>Rely</th>
<th>R@k</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILPS/Int/abs-hh-ratio-ilps@uva.out</td>
<td>958 0.567 0.518 0.628</td>
<td>0.736 0.813 0.526</td>
<td>0.48 0.213</td>
<td>0.915 773</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ILPS/Int/abs-th-ratio-ilps@uva.out</td>
<td>986 0.556 0.478 0.576</td>
<td>0.692 0.774 0.535</td>
<td>0.45 0.197</td>
<td>0.868 555</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p10_t400.out</td>
<td>985 0.28 0.307 0.502</td>
<td>0.663 0.744 0.632</td>
<td>0.511 0.334</td>
<td>0.941 487</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/distributed_effort_p10_t1500.out</td>
<td>981 0.28 0.306 0.499</td>
<td>0.664 0.745 0.633</td>
<td>0.517 0.247</td>
<td>0.968 1349</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p10_t1000.out</td>
<td>977 0.28 0.306 0.499</td>
<td>0.664 0.745 0.63</td>
<td>0.51 0.245</td>
<td>0.973 870</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p10_t200.out</td>
<td>1180 0.28 0.312 0.501</td>
<td>0.671 0.775 0.617</td>
<td>0.488 0.267</td>
<td>0.901 301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p10_t500.out</td>
<td>975 0.28 0.306 0.502</td>
<td>0.662 0.742 0.63</td>
<td>0.514 0.353</td>
<td>0.946 560</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p10_t1500.out</td>
<td>952 0.28 0.306 0.499</td>
<td>0.664 0.745 0.63</td>
<td>0.522 0.474</td>
<td>0.984 1117</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/distributed_effort_p10_t1000.out</td>
<td>991 0.279 0.306 0.499</td>
<td>0.664 0.745 0.62</td>
<td>0.492 0.157</td>
<td>0.921 975</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p10_t100.out</td>
<td>1153 0.274 0.306 0.483</td>
<td>0.639 0.737 0.54</td>
<td>0.474 0.292</td>
<td>0.711 402</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/baseline_bm25_t500.out</td>
<td>1233 0.222 0.191 0.282</td>
<td>0.41 0.515 0.435</td>
<td>0.394 0.481</td>
<td>0.794 441</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/distributed_effort_p10_t300.out</td>
<td>974 0.276 0.306 0.499</td>
<td>0.664 0.733 0.592</td>
<td>0.481 0.122</td>
<td>0.794 441</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p50_t1000.out</td>
<td>836 0.29 0.306 0.498</td>
<td>0.688 0.795 0.643</td>
<td>0.542 0.493</td>
<td>0.988 1139</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/distributed_effort_p10_t100.out</td>
<td>1114 0.248 0.315 0.444</td>
<td>0.604 0.704 0.458</td>
<td>0.372 0.402</td>
<td>0.45 156</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p50_t200.out</td>
<td>1185 0.29 0.312 0.499</td>
<td>0.693 0.792 0.63</td>
<td>0.481 0.331</td>
<td>0.911 334</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/baseline_bm25_t1000.out</td>
<td>1241 0.222 0.191 0.282</td>
<td>0.408 0.524 0.446</td>
<td>0.392 0.471</td>
<td>0.827 682</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/distributed_effort_p10_t500.out</td>
<td>991 0.278 0.306 0.499</td>
<td>0.664 0.743 0.606</td>
<td>0.483 0.115</td>
<td>0.842 594</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/baseline_bm25_t300.out</td>
<td>1262 0.222 0.187 0.286</td>
<td>0.41 0.523 0.44</td>
<td>0.398 0.506</td>
<td>0.664 270</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/baseline_bm25_t100.out</td>
<td>1397 0.223 0.186 0.291</td>
<td>0.429 0.557 0.414</td>
<td>0.368 0.485</td>
<td>0.507 99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p50_t400.out</td>
<td>985 0.29 0.307 0.501</td>
<td>0.685 0.767 0.646</td>
<td>0.514 0.374</td>
<td>0.949 572</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p50_t300.out</td>
<td>1144 0.29 0.313 0.495</td>
<td>0.682 0.788 0.569</td>
<td>0.497 0.355</td>
<td>0.933 462</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p50_t1000.out</td>
<td>1150 0.284 0.306 0.483</td>
<td>0.653 0.752 0.556</td>
<td>0.481 0.362</td>
<td>0.728 188</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/distributed_effort_p10_t200.out</td>
<td>965 0.271 0.306 0.482</td>
<td>0.651 0.752 0.56</td>
<td>0.445 0.165</td>
<td>0.714 312</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/baseline_bm25_t400.out</td>
<td>1242 0.222 0.191 0.286</td>
<td>0.412 0.523 0.434</td>
<td>0.393 0.485</td>
<td>0.713 337</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p50_t1500.out</td>
<td>796 0.29 0.306 0.498</td>
<td>0.688 0.785 0.642</td>
<td>0.553 0.541</td>
<td>0.999 1425</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/2018_stem_original_p50_t500.out</td>
<td>1001 0.29 0.306 0.501</td>
<td>0.691 0.779 0.65</td>
<td>0.505 0.395</td>
<td>0.961 677</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/baseline_bm25_t1500.out</td>
<td>1203 0.222 0.191 0.282</td>
<td>0.411 0.533 0.453</td>
<td>0.399 0.461</td>
<td>0.933 932</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/baseline_bm25_t200.out</td>
<td>1263 0.222 0.189 0.284</td>
<td>0.417 0.535 0.438</td>
<td>0.396 0.466</td>
<td>0.624 191</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Padua/Int/distributed_effort_p10_t400.out</td>
<td>981 0.277 0.306 0.499</td>
<td>0.663 0.734 0.595</td>
<td>0.483 0.116</td>
<td>0.822 518</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheffield/Int/Int_sheffield-Log_likelihood.out</td>
<td>1132 0.293 0.258 0.378</td>
<td>0.583 0.695 0.458</td>
<td>0.381 0.599</td>
<td>1 2100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheffield/Int/Int_sheffield-Odds_Ratio.out</td>
<td>1070 0.261 0.267 0.404</td>
<td>0.569 0.462 0.384</td>
<td>0.599</td>
<td>1 2100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheffield/Int/Int_sheffield-baseline.out</td>
<td>1276 0.245 0.22 0.334</td>
<td>0.507 0.653 0.47</td>
<td>0.386 0.599</td>
<td>1 2100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheffield/Int/Int_sheffield-Chi_Squared.out</td>
<td>1149 0.262 0.238 0.36</td>
<td>0.537 0.687 0.469</td>
<td>0.415 0.599</td>
<td>1 2100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run</td>
<td>L_Rel</td>
<td>MAP</td>
<td>R@5%</td>
<td>R@10%</td>
<td>R@20%</td>
<td>R@30%</td>
<td>WSS95</td>
<td>WSS100</td>
<td>Rely</td>
<td>R@k</td>
<td>k</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>-------</td>
<td>-----</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>ILPS/Pro/abs/abs-hh-ratio-ilps@uva</td>
<td>2885</td>
<td>0.673</td>
<td>0.562</td>
<td>0.714</td>
<td>0.875</td>
<td>0.911</td>
<td>0.591</td>
<td>0.143</td>
<td>0.018</td>
<td>0.948</td>
<td>1221</td>
</tr>
<tr>
<td>ILPS/Pro/abs/abs-th-ratio-ilps@uva</td>
<td>2537</td>
<td>0.628</td>
<td>0.521</td>
<td>0.682</td>
<td>0.818</td>
<td>0.927</td>
<td>0.566</td>
<td>0.247</td>
<td>0.014</td>
<td>0.922</td>
<td>867</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t400</td>
<td>2967</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.901</td>
<td>0.567</td>
<td>0.119</td>
<td>0.035</td>
<td>0.828</td>
<td>735</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t1500</td>
<td>2594</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.896</td>
<td>0.554</td>
<td>0.23</td>
<td>0.049</td>
<td>0.99</td>
<td>2165</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t11000</td>
<td>2644</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.896</td>
<td>0.554</td>
<td>0.215</td>
<td>0.022</td>
<td>0.943</td>
<td>1332</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t1200</td>
<td>2911</td>
<td>0.242</td>
<td>0.214</td>
<td>0.536</td>
<td>0.812</td>
<td>0.901</td>
<td>0.53</td>
<td>0.135</td>
<td>0.162</td>
<td>0.599</td>
<td>398</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t1500</td>
<td>2920</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.891</td>
<td>0.56</td>
<td>0.133</td>
<td>0.027</td>
<td>0.859</td>
<td>832</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t300</td>
<td>2955</td>
<td>0.239</td>
<td>0.214</td>
<td>0.547</td>
<td>0.818</td>
<td>0.891</td>
<td>0.556</td>
<td>0.122</td>
<td>0.054</td>
<td>0.776</td>
<td>597</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t11000</td>
<td>2578</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.896</td>
<td>0.554</td>
<td>0.234</td>
<td>0.035</td>
<td>0.984</td>
<td>1831</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t1000</td>
<td>2563</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.896</td>
<td>0.554</td>
<td>0.239</td>
<td>0.026</td>
<td>0.974</td>
<td>1566</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t200</td>
<td>2911</td>
<td>0.242</td>
<td>0.214</td>
<td>0.536</td>
<td>0.812</td>
<td>0.901</td>
<td>0.53</td>
<td>0.135</td>
<td>0.162</td>
<td>0.599</td>
<td>398</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t500</td>
<td>2920</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.896</td>
<td>0.554</td>
<td>0.239</td>
<td>0.026</td>
<td>0.974</td>
<td>1831</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t1000</td>
<td>2563</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.896</td>
<td>0.554</td>
<td>0.239</td>
<td>0.026</td>
<td>0.974</td>
<td>1566</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t500</td>
<td>2967</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.896</td>
<td>0.554</td>
<td>0.239</td>
<td>0.026</td>
<td>0.974</td>
<td>1566</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p10_t1000</td>
<td>2556</td>
<td>0.221</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.896</td>
<td>0.554</td>
<td>0.239</td>
<td>0.026</td>
<td>0.974</td>
<td>1566</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p50_t1000</td>
<td>2556</td>
<td>0.221</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.896</td>
<td>0.554</td>
<td>0.239</td>
<td>0.026</td>
<td>0.974</td>
<td>1566</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p50_t11000</td>
<td>2789</td>
<td>0.252</td>
<td>0.25</td>
<td>0.568</td>
<td>0.786</td>
<td>0.875</td>
<td>0.594</td>
<td>0.172</td>
<td>0.288</td>
<td>0.464</td>
<td>248</td>
</tr>
<tr>
<td>Padua/Pro/abs/2018_stem_original_p50_t1200</td>
<td>2911</td>
<td>0.242</td>
<td>0.214</td>
<td>0.536</td>
<td>0.812</td>
<td>0.901</td>
<td>0.53</td>
<td>0.135</td>
<td>0.162</td>
<td>0.599</td>
<td>398</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t1000</td>
<td>3346</td>
<td>0.071</td>
<td>0.057</td>
<td>0.13</td>
<td>0.281</td>
<td>0.422</td>
<td>0.084</td>
<td>0.007</td>
<td>0.621</td>
<td>0.214</td>
<td>501</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t300</td>
<td>2964</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.906</td>
<td>0.567</td>
<td>0.12</td>
<td>0.038</td>
<td>0.818</td>
<td>709</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t500</td>
<td>2556</td>
<td>0.221</td>
<td>0.214</td>
<td>0.484</td>
<td>0.74</td>
<td>0.87</td>
<td>0.571</td>
<td>0.241</td>
<td>0.041</td>
<td>0.995</td>
<td>1981</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t1000</td>
<td>2789</td>
<td>0.252</td>
<td>0.25</td>
<td>0.568</td>
<td>0.786</td>
<td>0.875</td>
<td>0.594</td>
<td>0.172</td>
<td>0.288</td>
<td>0.464</td>
<td>248</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t100</td>
<td>2911</td>
<td>0.242</td>
<td>0.214</td>
<td>0.536</td>
<td>0.812</td>
<td>0.901</td>
<td>0.53</td>
<td>0.135</td>
<td>0.162</td>
<td>0.599</td>
<td>398</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t500</td>
<td>2708</td>
<td>0.235</td>
<td>0.214</td>
<td>0.484</td>
<td>0.812</td>
<td>0.891</td>
<td>0.566</td>
<td>0.196</td>
<td>0.026</td>
<td>0.87</td>
<td>955</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t1000</td>
<td>3350</td>
<td>0.071</td>
<td>0.057</td>
<td>0.13</td>
<td>0.276</td>
<td>0.396</td>
<td>0.057</td>
<td>0.006</td>
<td>0.382</td>
<td>0.391</td>
<td>1001</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t500</td>
<td>3350</td>
<td>0.066</td>
<td>0.047</td>
<td>0.13</td>
<td>0.276</td>
<td>0.385</td>
<td>0.104</td>
<td>0.005</td>
<td>0.794</td>
<td>0.109</td>
<td>301</td>
</tr>
<tr>
<td>Padua/Pro/abs/distributed_effort_p10_t1000</td>
<td>2955</td>
<td>0.231</td>
<td>0.214</td>
<td>0.484</td>
<td>0.807</td>
<td>0.896</td>
<td>0.556</td>
<td>0.122</td>
<td>0.033</td>
<td>0.839</td>
<td>798</td>
</tr>
<tr>
<td>Sheffield/Pro/abs/Pro_sheffield-baseline</td>
<td>2990</td>
<td>0.126</td>
<td>0.146</td>
<td>0.255</td>
<td>0.448</td>
<td>0.594</td>
<td>0.247</td>
<td>0.112</td>
<td>0.018</td>
<td>0.948</td>
<td>1221</td>
</tr>
<tr>
<td>Sheffield/Pro/abs/Pro_sheffield-relevance_feedback</td>
<td>2775</td>
<td>0.141</td>
<td>0.151</td>
<td>0.307</td>
<td>0.484</td>
<td>0.646</td>
<td>0.305</td>
<td>0.176</td>
<td>0.117</td>
<td>1</td>
<td>3367</td>
</tr>
</tbody>
</table>
Table 12. Qualitative studies with abstract-level QRELs

<table>
<thead>
<tr>
<th>Run</th>
<th>L_Rel</th>
<th>MAP</th>
<th>R@5%</th>
<th>R@10%</th>
<th>R@20%</th>
<th>R@30%</th>
<th>WSS95</th>
<th>WSS100</th>
<th>Rely</th>
<th>R@k</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILPS/Qual/abs/abs-hh-ratio-ilps@uva.out</td>
<td>1796.0</td>
<td>0.204</td>
<td>0.478</td>
<td>0.655</td>
<td>0.876</td>
<td>0.929</td>
<td>0.417</td>
<td>0.397</td>
<td>0.326</td>
<td>0.919</td>
<td>1247</td>
</tr>
<tr>
<td>ILPS/Qual/abs/abs-th-ratio-ilps@uva.out</td>
<td>2564.0</td>
<td>0.187</td>
<td>0.487</td>
<td>0.628</td>
<td>0.805</td>
<td>0.92</td>
<td>0.398</td>
<td>0.215</td>
<td>0.341</td>
<td>0.878</td>
<td>1158</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p10_t400.out</td>
<td>2547.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.717</td>
<td>0.779</td>
<td>0.914</td>
<td>0.302</td>
<td>0.295</td>
<td>0.199</td>
<td>0.638</td>
<td>1515</td>
</tr>
<tr>
<td>Padua/Qual/abs/distributed_effort_p10_t1500.out</td>
<td>2544.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.885</td>
<td>0.268</td>
<td>0.168</td>
<td>0.37</td>
<td>0.745</td>
<td>2098</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p10_t1000.out</td>
<td>2662.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.885</td>
<td>0.273</td>
<td>0.141</td>
<td>0.29</td>
<td>0.714</td>
<td>1320</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p10_t200.out</td>
<td>2934.0</td>
<td>0.089</td>
<td>0.478</td>
<td>0.522</td>
<td>0.699</td>
<td>0.805</td>
<td>0.216</td>
<td>0.215</td>
<td>0.341</td>
<td>0.878</td>
<td>1247</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p10_t500.out</td>
<td>2535.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.894</td>
<td>0.301</td>
<td>0.185</td>
<td>0.578</td>
<td>0.396</td>
<td>820</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p10_t500.out</td>
<td>2660.0</td>
<td>0.103</td>
<td>0.496</td>
<td>0.655</td>
<td>0.752</td>
<td>0.858</td>
<td>0.29</td>
<td>0.159</td>
<td>0.582</td>
<td>0.338</td>
<td>554</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p10_t1000.out</td>
<td>2534.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.885</td>
<td>0.268</td>
<td>0.17</td>
<td>0.447</td>
<td>0.732</td>
<td>1819</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p50_t1000.out</td>
<td>2660.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.894</td>
<td>0.301</td>
<td>0.185</td>
<td>0.578</td>
<td>0.396</td>
<td>820</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p50_t300.out</td>
<td>2658.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.894</td>
<td>0.301</td>
<td>0.185</td>
<td>0.578</td>
<td>0.396</td>
<td>820</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p50_t1500.out</td>
<td>2641.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.894</td>
<td>0.301</td>
<td>0.185</td>
<td>0.578</td>
<td>0.396</td>
<td>820</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p50_t1000.out</td>
<td>2543.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.894</td>
<td>0.301</td>
<td>0.185</td>
<td>0.578</td>
<td>0.396</td>
<td>820</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p50_t200.out</td>
<td>2934.0</td>
<td>0.089</td>
<td>0.478</td>
<td>0.522</td>
<td>0.699</td>
<td>0.805</td>
<td>0.216</td>
<td>0.215</td>
<td>0.341</td>
<td>0.878</td>
<td>1247</td>
</tr>
<tr>
<td>Padua/Qual/abs/distributed_effort_p10_t1000.out</td>
<td>2534.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.885</td>
<td>0.268</td>
<td>0.17</td>
<td>0.447</td>
<td>0.732</td>
<td>1819</td>
</tr>
<tr>
<td>Padua/Qual/abs/2018_stem_original_p50_t200.out</td>
<td>2687.0</td>
<td>0.103</td>
<td>0.496</td>
<td>0.655</td>
<td>0.752</td>
<td>0.858</td>
<td>0.29</td>
<td>0.159</td>
<td>0.582</td>
<td>0.338</td>
<td>554</td>
</tr>
<tr>
<td>Padua/Qual/abs/distributed_effort_p10_t1000.out</td>
<td>2534.0</td>
<td>0.109</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.885</td>
<td>0.268</td>
<td>0.17</td>
<td>0.447</td>
<td>0.732</td>
<td>1819</td>
</tr>
<tr>
<td>Padua/Qual/abs/baseline_bm25_t500.out</td>
<td>2700.0</td>
<td>0.051</td>
<td>0.274</td>
<td>0.425</td>
<td>0.469</td>
<td>0.611</td>
<td>0.412</td>
<td>0.256</td>
<td>0.683</td>
<td>0.221</td>
<td>501</td>
</tr>
<tr>
<td>Padua/Qual/abs/baseline_bm25_t1000.out</td>
<td>2697.0</td>
<td>0.049</td>
<td>0.274</td>
<td>0.372</td>
<td>0.451</td>
<td>0.628</td>
<td>0.294</td>
<td>0.257</td>
<td>0.626</td>
<td>0.171</td>
<td>301</td>
</tr>
<tr>
<td>Padua/Qual/abs/baseline_bm25_t300.out</td>
<td>2762.0</td>
<td>0.104</td>
<td>0.496</td>
<td>0.743</td>
<td>0.77</td>
<td>0.894</td>
<td>0.301</td>
<td>0.185</td>
<td>0.578</td>
<td>0.396</td>
<td>820</td>
</tr>
<tr>
<td>Sheffield/Qual/abs/Qual_sheffield-relevance_feedback.out</td>
<td>2940.0</td>
<td>0.06</td>
<td>0.274</td>
<td>0.549</td>
<td>0.717</td>
<td>0.832</td>
<td>0.185</td>
<td>0.103</td>
<td>0.593</td>
<td>1</td>
<td>3268</td>
</tr>
<tr>
<td>Sheffield/Qual/abs/Qual_sheffield-baseline</td>
<td>3031.0</td>
<td>0.051</td>
<td>0.274</td>
<td>0.549</td>
<td>0.717</td>
<td>0.832</td>
<td>0.185</td>
<td>0.103</td>
<td>0.593</td>
<td>1</td>
<td>3268</td>
</tr>
</tbody>
</table>
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.