A Test Collection for Matching Patients to Clinical Trials

Bevan Koopman[1,2], Guido Zuccon[2]
[1]Australian e-Health Research Centre, CSIRO, Brisbane, Australia
[2]Faculty of Science & Technology, Queensland University of Technology, Brisbane, Australia
bevan.koopman@csiro.au, g.zuccon@qut.edu.au

ABSTRACT

We present a test collection to study the use of search engines for matching eligible patients (the query) to clinical trials (the document). Clinical trials are experiments conducted in the development of new medical treatments, drugs or devices. Recruiting candidates for a trial is often a time-consuming and resource intensive effort, and imposes delays or even the cancellation of trials.

The collection described in this paper provides: i) a large corpus of clinical trials; ii) 60 patient case reports used as topics; iii) multiple query representations for a single topic (long, short and ad-hoc); iv) a user provided estimate of how many trials they expect each patient topic would be eligible for; and v) relevance assessments by medical professionals. The availability of such a collection allows researchers to investigate, among other questions: (1) the effectiveness of retrieval methods for this task, (2) how multiple representations of an information affect retrieval (3) what influences relevance assessments in this context, (4) whether automated matching of patients to trials improves patient recruitment.

The collection is available at http://anonymised.

Categories and Subject Descriptors: H.3 [Information Systems]: Information Storage and Retrieval
General Terms: Evaluation, Experimentation.

1. INTRODUCTION

Clinical trials are experiments done in the development of new treatments, drugs or medical devices. They are a critical step for medical advancement and are a regulatory requirement before new medical advances can be used in practise. However, recruiting a sufficient number of eligible patients to participate in a trial can be a major obstacle [9]. If suitable patients cannot be found then trials may be cancelled or significantly delayed. Even if sufficient patients are found, the recruitment process can be time consuming and resource intensive. Automating and improving this difficult manual process has the potential to improve the running of a clinical trial. In addition, certain patients can benefit from finding and being included into specific trials, e.g., to have access to potentially life-saving treatment options. However, often treating doctors are not aware of trials that may benefit specific patients.

Large collections of clinical trials are published online (e.g., ClinicalTrials.gov contained approx. 200,000 trails in 2015), with details of the inclusion and exclusion criteria of eligible patients. At the same time, a patient's conditions are also documented in electronic form (for example, in electronic patient records). Matching patients to clinical trials is essentially an information retrieval task: the query is the patient details (either in the form of electronic patient records or ad-hoc queries) and the documents are the clinical trials currently recruiting patients.

While research exists on automated matching of patients to trials [10], much of the evaluation is done on small, private datasets and on specific diseases. This paper aims to address this gap by developing a large-scale, heterogeneous test collection of clinical trial documents and associated patient queries. The availability of such a collection allows researchers to investigate: (1) the effectiveness of retrieval methods for this task, (2) how multiple representations of an information affect retrieval (3) what influences relevance assessments in this context, (4) whether automated matching of patients to trials improves patient recruitment.

2. RELATED WORK

Sustained focus on medical information retrieval has led to the development of a number of other relevant test collections. Within TREC, there have been two medical related tracks relevant to this work: the Medical Records Track (MedTrack) and the Clinical Decision Support (CDS Track).

The MedTrack task involved searching a collection of electronic patient records for patients that meet a certain criteria (the query) [12]. One use case was that the query indicated the inclusion criteria for a clinical trial, while the documents were patients to be retrieved that matched that criteria. Thus, Medtrack could be viewed as the opposite, trial-centric (trial is the query and patient is the document) to the patient-centric task considered here. Another important difference that sets this work apart is that real clinical trials were used; instead Medtrack used only ad-hoc queries to describe the patient (e.g., topic# 115 "Adult patients who are admitted with an asthma exacerbation").

In the TREC CDS task the topics were patient case reports and were used to search for medical journal articles

```
A 51-year-old woman is seen in clinic for advice on os-
teoporosis. She has a past medical history of signifi-
cant hypertension and diet-controlled diabetes mellitus.
She currently smokes 1 pack of cigarettes per day. She
was documented by previous LH and FSH levels to be in
menopause within the last year. She is concerned about
breaking her hip as she gets older and is seeking advice
on osteoporosis prevention.
```

Figure 1: Example patient case (topic# 201429).

that would help uncover the diseases, tests and treatments relevant to the patient case [11]. The patient case reports were verbose: on average 78 words per topic (an example report is shown in Figure 1). TREC CDS is intended for searching medical literature for clinical decision support; however, the patient case reports are general descriptions of a patient past and current medical history. The patient case reports can, therefore, also be used to search for clinical trials. For this reason, we use the same patient case reports from TREC CDS as our topics in this test collection to search for clinical trials. This also has the added advantage of being able to link a patient with both clinical trials from this collection and associated medical literature from the TREC CDS collection.

Other medical collections do exist in the TREC Genomics Track and in the CLEF eHealth Lab; however, these are focused on genomic search and consumer health search and, therefore, not detailed here.

3. CREATION OF THE COLLECTION

3.1 Document Collection

A collection of 204,855 publicly available clinical trails was crawled from ClinicalTrials.gov.[1] Trials are made available in a specific XML format, however, large portions (including the inclusion and exclusion criteria) are free-text.[2] These represent the documents to be searched.

3.2 Query Topics

As query topics, we adopt the topics previously used by the TREC CDS [11], comprising 60 patient case reports (30 from 2014 and 30 from 2015). Each topic describes a patient with certain conditions and observations. Each patient case topic has two forms: a description (on average 78 words) and a shorter summary (on average 22 words).

As noted above, the topics were verbose patient case reports. Automatically matching these case reports to clinical trials is the first use case — here the user simply supplies the case report and does not author a query. However, an alternative use case exists as a traditional ad-hoc retrieval scenario where the user authors a short keyword query. To cover this second use case we showed four medical assessors each patient case report and asked them to provide ad-hoc keyword queries that they would issue to a search engine to find clinical trials for the given patient. A total of 489 unique queries were produced, on average 8.2 (sd=3.2) keyword queries per topic. In addition, assessors were asked the following question for each topic: "How many clinical trials do you expect this patient would be eligible for?" The answer to this was recorded and is used in the INST evaluation measure [5] we will detail in Section 3.4.

3.3 Pooling and Judging

A number of baseline retrieval models were run to form the pool. These included: BM25, Language Model (Direchlet and Jelinek-Mercer), Divergence From Randomness (BB2 and DLH) and TF-IDF.[3] While this was only a small number of systems, we note that Moffat et al. found that query variations are as strong as system variations in producing a diverse document pool [6]; thus, we overcame the limit of having a small number of systems by including a large number of query variations. Specifically, each baseline system listed above was run with the following queries for each topic: i) the patient case report description; ii) the patient case report summary; and iii) the ad-hoc keyword queries provided by our medical assessors (8.2 queries on average). This equates to an average of 61 runs per topic (10.2 queries per topic * 6 baseline methods). This provided a diverse set of retrieved documents to form the pool.

To maximise the time and minimise costs associated with employing medical assessors it was important to maximise the chance of sampling important documents for assessment. A standard approach to form the pool is to include all documents that are highly ranked by participating systems. However, Moffat et al. [7] noted that not all documents provide the same benefit and instead propose an alternative method based on the Ranked Biased Precision (RBP) evaluation measure. Documents were ranked according to RBP across all queries; documents that were retrieved by multiple, different systems in top-ranked positions would appear higher in the RBP ranking. The pool was then formed based on the available assessment budget by setting a cut-off point of 4,000 documents in the RBP ranking — documents above the cut-off were included in the pool.

The documents and queries were uploaded to the Relevation! relevance assessment system [2] and four medical assessors were engaged to conduct the relevance assessment. Queries were divided amongst the four assessors; a control query (topic #20158) was used to familiarise assessors with the task and to record inter-coder reliability (agreement found to be 65%). This highlights the difficulty intrinsic in judging relevance in the medical domain, as identified by other studies [3]. Reasons for assessors disagreement will be investigated in future work.

3.4 The Task and Evaluation Measures

The task of matching patients to clinical trials has three specific use cases; we use these to set the evaluation measures for the task.

The first use case is in a General Practitioner (GP) setting where the GP opens a patient's record as part of a normal consultation and a search is automatically initiated to find relevant clinical trials that the GP may refer the patient to. In this scenario the GP is time-pressured and would likely only review a small number of results; they would likely stop when a single relevant trial is found. Thus for this scenario we adopted Mean Reciprocal Rank (MRR) as the evaluation measure.

The second use case is also set within a general medical professional (GP or other) but where the user is specifically

[1]This represents all the trials available on 16th Dec., 2015.
[2]More details on the format and download options can be found at:
https://clinicaltrials.gov/ct2/resources/download.

[3]The Terrier IR system was used for all models and parameters left to Terrier defaults [4].

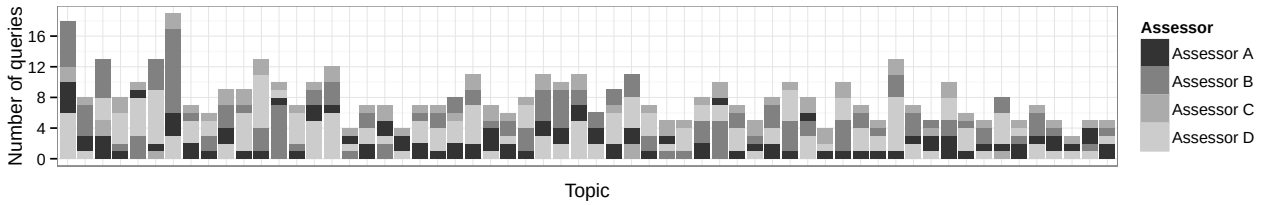
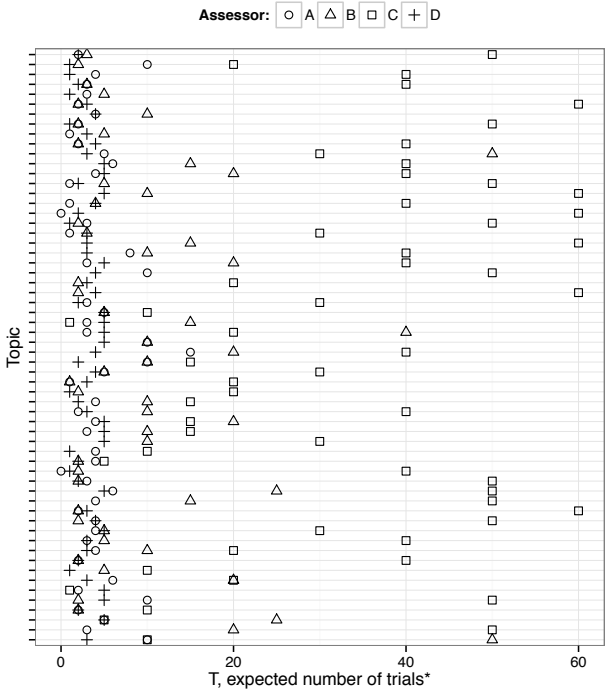


Figure 2: Number of queries supplied by each assessors for each topic.

searching for clinical trials and may dedicate more time and effort to the task. In this case they may issue an ad-hoc query themselves and be willing to evaluate a few more results. For this scenario we adopted Precision at 5 (P@5) as the evaluation measure.

The final use case is for medical specialists or patients themselves searching for trials. Here both types of users may conduct longer search sessions and review far more results. They may use both short ad-hoc queries and more verbose patient case reports. In addition, both users would have an expectation about how many clinical trials they would be eligible for. This would influence their search behaviour: for rare diseases, they may expect to find a very small number of trials and would therefore not persist in examining results at greater rank depths. In contrast, for common diseases, they would expect to find many relevant trials and would therefore persist to greater rank depths. This notion of expected number of (relevant) results is directly modelled by T in the INST evaluation measure [5]; thus we adopted INST for this scenario. INST is a weighted precision metric where the likelihood of the user assessing a document at a specific rank depends on the rank position, the expected number of relevant documents, and the actual number of relevant documents encountered up to that rank. According to INST, the expected depth at which the user would stop viewing documents falls between approximately $T+0.25$ (all encountered documents are relevant) and $2T + 0.5$ (no encountered documents are relevant) [5].



*x-axis truncated at $T = 60$ excluding outliners $T = 80, 80, 100$.

Figure 3: T, the users' expected number of clinical trials for a patient topic.

4. ANALYSIS OF THE COLLECTION

4.1 Test Collection Statistics

The collection contains 204,855 clinical trial documents. There are 60 topics made up of three types: patient case descriptions, patient case summaries and assessor provided ad-hoc queries, totalling an average of 10.2 queries per topic. A total of 4,000 documents were judged (67 per topic, sd=27).

The number of ad-hoc queries provided by the assessors differed per topic and per assessor, as shown in Figure 2. Some assessors entered multiple short queries, while others preferred single longer queries. The average query length was 4.5 words, sd=2.5 words.

Assessors were also asked how many clinical trials they expected a patient would be eligible for. This was represented as T in the INST evaluation measure. The values of T for each topic, across the four assessors, is shown in Figure 3. Values of T varied across topics, thus indicating the different information needs assessors derived from different patients. Although T varied across topics, individual assessors displayed similar trends across topics; e.g., assessor D typically chose lower values for T and assessor C displayed higher values of T. This resulted in values of T that varied across assessors for a single topic. Qualitative feedback from assessors indicated estimating T was challenging and subjective. The assessors were asked about their rationale for determining values of T. We found that regardless of the value of T, assessors indicated that the main rationale was how rare or common the patient's medical condition was; secondary to that was the likelihood that clinical trails were currently being conducted on the patient's condition.

4.2 Retrieval Results Analysis

The relevance assessments we collected were used to evaluate six standard baselines. The purpose of the evaluation was twofold. On one hand, the retrieval systems were used to form the pool for assessments, thus the evaluation reports how effective the systems that contributed to the pool were. On the other hand, this evaluation serves to demonstrate the type of research questions this collection can contribute to investigate, e.g., what type of queries (verbose, summaries, ad-hoc) are most effective for searching for clinical trials.

For each system, runs were created using three different topic types: i) verbose patient case report descriptions; ii) shorter patient case report summaries; and iii) short ad-hoc keyword queries. Note that ad-hoc queries generated more than one run per topic per system, i.e., on average each system generated 8.2 runs per topic. We therefore averaged the effectiveness of a system over all ad-hoc queries for a topic. Retrieval results are shown in Figure 4. We firstly observe

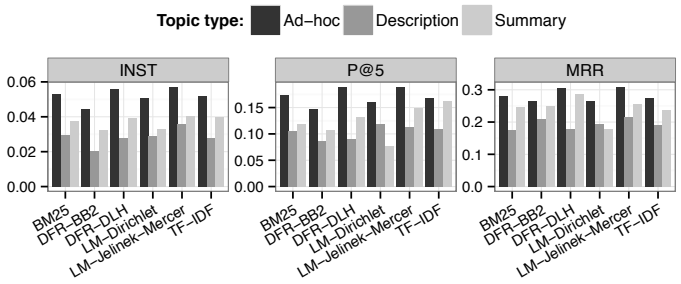


Figure 4: Retrieval results for different baselines and topic representations.

that there was high variability of performance across the topic types. The assessor-provided ad-hoc queries proved most effective overall, followed by summary patient case reports and finally the full description patient case reports were the least effective topic type.

There was also variability across different baseline methods. The best method for ad-hoc queries was clearly the Jelinek-Mercer language model. For the longer summaries and descriptions the best method varied: TF-IDF proving effective for summaries, while no method clearly stood out for descriptions. This observation suggests that different baseline methods are best suited to different use cases.

Overall, we note that there was more variability across topic types than across baseline methods. This is inline with the results of Moffat et al. [6] that found query variability was as significant as system variability.

5. DISCUSSION

The test collection described in this paper is clearly aimed at focusing research on matching eligible patients to clinical trails; however, it also provides the basis for exploring a number of other research aspects:

Query representations and variations. The collection provides multiple representations for a single topics: descriptions, summaries and ad-hoc queries (on average 8.2 per topic). The availability of multiple representations makes it possible to investigate whether specific retrieval methods are more suitable to different representations, e.g., ad-hoc vs verbose. The ad-hoc queries themselves expose different ways of formulating the same information need, each leading to different effectiveness for the same retrieval method. Along with the TREC-8 Query Track [1] and the CLEF 2015 eHealth [8], our collection is a rare example of a test collection with multiple query variations.

Expected number of relevant results (T). Assessors indicated how many trials they expected the patient would have been eligible for. This data can serve the evaluation (e.g., through INST) but also allows exploring the perception about the results assessors expected to obtain. In particular, we observed that T greatly varied across topics and across assessors (this latter result was in contrast to previous studies [6]). Finally, this is the first collection that provides a user's estimate of the number of relevant documents they believe are required for each topic.

What makes a clinical trial relevant. The collection makes available data to understand what characterises relevance when judging the eligibility of a patient to a clinical trial. It also provides evidence to the fact that judging the eligibility of a clinical trial is often challenging when only summary information about patients is available.

6. CONCLUSIONS

This paper presented a test collection aimed at helping the development of systems to automatically match eligible patients to clinical trials. The collection can be used to discriminate between different systems on this task. The limited number of assessments may make the evaluation less reliable for new systems that greatly differ from those used to form the pool. Nevertheless, the value of the collection is the insights it provides into research questions related to, e.g., the effectiveness of different query representations (and variations), how assessors judge relevance of patients to clinical trials, etc. The collection presents several original aspects. This is the first publicly available, large scale collection for matching patients to clinical trials — an important task for medical advancement. The collection is also the first that provides estimates of the number of expected relevant documents for each query topic (T), and is one of the few that provides multiple query representations and variations.

Future work will consider increasing the number of assessed documents, including increasing the number of systems used to form the pool, especially when specialised systems to search clinical trials become available. We also plan to expand the analysis of query variations and the effect query representations have on system effectiveness. Finally, another line of future research will consider the analysis of assessor disagreement (both for relevance and for the value of T) to gain further insights about how users perceive relevance for this task. The collection is available at http://anonymised.

7. REFERENCES

[1] C. Buckley and J. A. Walz. The TREC-8 Query Track. In *TREC*, 1999.

[2] B. Koopman and G. Zuccon. Relevation!: An open source system for information retrieval relevance assessment. In *SIGIR*, Gold Coast, Australia, July 2014.

[3] B. Koopman and G. Zuccon. Why assessing relevance in medical IR is demanding. In *MedIR at SIGIR*, 2014.

[4] C. Macdonald, R. McCreadie, R. L. Santos, and I. Ounis. From puppy to maturity: Experiences in developing terrier. *Proc. of OSIR at SIGIR*, pages 60–63, 2012.

[5] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. INST: An adaptive metric for information retrieval evaluation. In *ADCS*, Sydney, Australia, 2015.

[6] A. Moffat, F. Scholer, P. Thomas, and P. Bailey. Pooled evaluation over query variations: Users are as diverse as systems. In *CIKM*, 2015.

[7] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *SIGIR*, pages 375–382. ACM, 2007.

[8] J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. Jones, M. Lupu, and P. Pecina. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. CLEF, 2015.

[9] L. T. Penberthy, B. A. Dahman, V. I. Petkov, and J. P. DeShazo. Effort required in eligibility screening for clinical trials. *Journal of Oncology Practice*, 8(6):365–370, 2012.

[10] T. R. Pressler, P.-Y. Yen, J. Ding, J. Liu, P. J. Embi, and P. R. O. Payne. Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. *BMC Medical Informatics & Decision Making*, 12:47–47, 2012.

[11] M. S. Simpson, E. M. Voorhees, and W. Hersh. Overview of the TREC clinical decision support track. In *TREC*, 2014.

[12] E. M. Voorhees and W. R. Hersh. Overview of the trec 2012 medical records track. In *TREC*, 2012.