Cohort-based Clinical Trial Retrieval

Bevan Koopman
bevan.koopman@csiro.au
CSIRO
Brisbane, Queensland, Australia

Guido Zuccon
g.zuccon@uq.edu.au
University of Queensland
Brisbane, Queensland, Australia

ABSTRACT

Clinical Trials are a critical step for medical advancement; key to success is recruiting eligible patients to a trial. Retrieval methods are used to identify relevant trials given a *single* patient/query. After careful consideration of the clinical setting, this paper takes a different approach: *cohort-based trial retrieval*. We consider ranking trials that maximise recruitment opportunities across the whole patient cohort, instead of a single patient. This resolves into optimising a ranking for the whole query set formed by the patient cohort, rather than treating each query independently — and thus considering an evaluation measure based on cohort coverage. We study the adaptation of rank fusion methods and diversity reranking to this problem. Empirically, we show the surprising impact of initial ranking effectiveness (underlying initial retrieval) on cohort coverage when adapting rank fusion methods. We further highlight that devising cohort-aware methods would have a far greater impact on patient recruitment.

KEYWORDS

clinical trials, clinical information retrieval, query cohorts

ACM Reference Format:
Bevan Koopman and Guido Zuccon. 2021. Cohort-based Clinical Trial Retrieval. In *Australasian Document Computing Symposium (ADCS '21), December 9, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503516.3503529

1 INTRODUCTION

Clinical trials are experiments done in the development of new treatments, drugs or devices. They are a critical step for medical advancement and are essential before new advances can be used in practice. However, recruiting sufficient eligible patients to a trial can be a major obstacle [15]. It can lead to trials being delayed or even cancelled. Even if successful, recruitment is costly and time consuming.

Large collections of clinical trials are published online (ClinicalTrials.gov contained 330,113 trials in 2020). Treated as a document collection, these can be searched using a description of a patient (for example, a patient's electronic patient records). In traditional clinical trial matching, there is a cohort (i.e., query set) of patients; each query representing the patient is issued to the IR system and a ranking of trials is retrieved. There are explicit test collections for this task [12], including the TREC Precision Medicine Track [16, 17]. Evaluation is done in the traditional IR manner: evaluate each *single* query independently according to some evaluation measure; then average across all the queries to get an overall measure of effectiveness. While the workflow is valid it does not capture the actual clinical setting.

In the clinical setting, the patients are not independent and hence queries should also not be treated so. This is because many patients are likely to show similar clinical profiles and diagnosis, even more so when confined to a specialist practice. Across the rankings of clinical trials for the whole patient cohort (i.e., query set), there may be common trials, retrieved for a number of queries and thus relevant to a number of patients. As we will show, there is a real advantage in retrieving trials for which many patients (i.e., queries) are relevant. Thus this paper investigates **cohort-based clinical trial retrieval**: devising a ranking of trials that covers as much of the patient cohort as possible.

We formalise this problem and identify suitable evaluation measures and settings which model the real-world clinical situation of matching patients to eligible trials. We adapt a number of rank fusion techniques to combine rankings from single-queries into a final cohort-based ranking. Empirical evaluation shows that high quality single-query rankings do not necessarily correlate with the final cohort-based ranking. We show the limitations of optimising retrieval for a single-query. Based on this we utilise a diversity-based reranking method that accounts for patient coverage; this does improve cohort-based ranking. Establishing a strong baseline for a greedy selection of final cohort ranking, we show the potential for new models that aim to optimise the final cohort ranking.

2 THE USE CASE FOR COHORT-BASED TRIAL RETRIEVAL

Matching patients to clinical trials happens in the four different settings outlined in Table 1. Automated methods for both pT and tP have been considered; in particular, by researchers in the Text Retrieval Conferences (TREC). However, this was always done for either a single patient (pT) or single trials (tP). Matching trials to cohorts of patients has not been tackled to our knowledge.

In a real-world setting, given the above situations, people are limited in their capacity to review trials or patients. Note that even with a highly effective information retrieval system, there is always a manual review step to determine eligibility. (This could involve, for example, further medical tests for a particular patient so beyond the scope of an automated matching system.) The limited capacity of manual reviews imposes two types of constraint:

Fixed Number of Patients: First, where someone is responsible for a cohort P patients (e.g., a hospital or individual clinician) and

ADCS '21, December 9, 2021, Virtual Event, Australia

ACM ISBN 978-1-4503-9599-1/21/12…$15.00
https://doi.org/10.1145/3503516.3503529

Id	Input	Output	User/Example scenario
pT	Single patient, p	Eligible trials, T.	Individual clinician trying to find a trial for their patient.
tP	Single trial, t	Eligible patients, P.	Researchers or pharmaceutical companies conducting a particular clinical trial for which they want to recruit patients.
PT	Cohort of patients, P	Eligible trials, T.	Clinician or organisation (e.g., hospital) responsible for a number of patients they want to enrol in to clinical trials.
TP	Set of trials, T	Cohort of eligible patients, P.	Either similar to tP where researchers or pharmaceutical company are trying to recruit to multiple trials; or where a health provider (clinician or hospital) is supporting multiple trials.

Table 1: Different settings of the patient-trial matching problem. $p \in P$ is a patient in a set of patients; $t \in T$ is a trial in a set of trials.

wants to enrol them into trials. They would like to enrol as many of the P patients as possible. It is particularly beneficial to enrol multiple patients to a trial, thus reducing the overall number of trials one has to deal with. This is because each trial incurs a certain fixed cost, both from an administrative and cognitive overhead perspective.

Fixed Number of Trials: The second situation is for people strictly limited on the number of trials they can review; for example, a busy clinician. Here they are managing a fixed set of T trials — again resource limitations dictate the amount of trials (size of T).

This paper considers the case of matching a cohort of patients to a set of trials, rather than individual patients. There are a number of advantages of a cohort-based approach:

(1) Less trials overall are needed to cover the whole patient cohort.
(2) From a practical perspective, this means a clinician (e.g., doctor) has to review less trials. In many cases, a busy clinician will have a strict limit on the number of trials they will review, thus retrieving trials that cover more patients means more of the overall patient cohort are covered, which may result in better health outcomes for more patients.
(3) Recruitment effort reduces with less trials — the overhead of recruiting 10 patients to 1 trial is far less than recruiting 10 patients to 10 different trials.
(4) By retrieving trials that increase coverage of the patient cohort, more patients have access to new and emergency treatments — in the cancer space, for example, this can be life saving.

As practical illustration of the benefit, consider the small sample case in Figure 1. The left ranking is not optimised for patient coverage: if someone was reviewing trials in order, top to bottom, they would need to review all five trials to cover each of the five patients. In contrast, the right ranking is optimised for patient coverage: someone reviewing trials would only need consider the top two trials to ensure complete coverage.

3 RELATED WORK

There have been a number of initiatives to foster research in matching patients to trials. While research has been hampered by a lack of publicly available patient records (understandable given the privacy constraints of releasing such data), there have been four key test collections available for empirical evaluation: TREC MedTrack [20, 21], TREC Precision Medicine (PM) Tracks, MIMIC III [8] and a custom collection we released [12].

TREC MedTrack ran in 2011 and 2012 and represented the single trial, eligible patients setting (**tP** of Table 1). The document collection was a set of de-identified patient records, including discharge summaries, surgery notes and laboratory reports. The queries, while not explicitly a clinical trial, were a general description of the eligibility criteria; for example, "Adult patients who are admitted with an asthma exacerbation" (TREC 2011 topic# 15).

Effective methods in TREC MedTrack were mostly well known statistical, IR approaches from the time with a number of domain-specific enhancements added [3]. These included:

- Normalising vocabulary to the particular clinical domain.
- Query expansion; either via controlled medical vocabulary such as Unified Medical Language System Metathesaurus, or via external corpora [24].
- Recognition and handling of negation in text [11] (e.g., "patient had no fever"), which is particularly prevalent in clinical language.
- More recently, learning-to-rank methods have showed to be very effective on this task [7].

Post TREC MedTrack error analysis [5] revealed that vocabulary mismatch was not the main cause of errors. Instead, the main challenge was effectively ranking relevant documents above non-relevant documents, where both contained the query terms.

The single patient, eligible trials setting (**pT** of Table 1), was the focus of the TREC Precision Medicine (PM) Tracks [16–18]. The document collection was a snapshot of trials from ClinicalTrials.gov. The queries were a description of a particular patient, including symptoms, diagnoses, demographics and genetics. These were terse and thus not directly equivalent to the type of notes found in an Electronic Health Record system, but at least provide a surrogate patient description that can be used for experimentation with different retrieval systems.

The most effective methods in TREC PM were learning-to-rank based. These had specific features for genes and diseases and often did significant query and document preprocessing to extract this information [6]. Effective query handling (e.g., preprocessing or handling) proved important [4, 22].

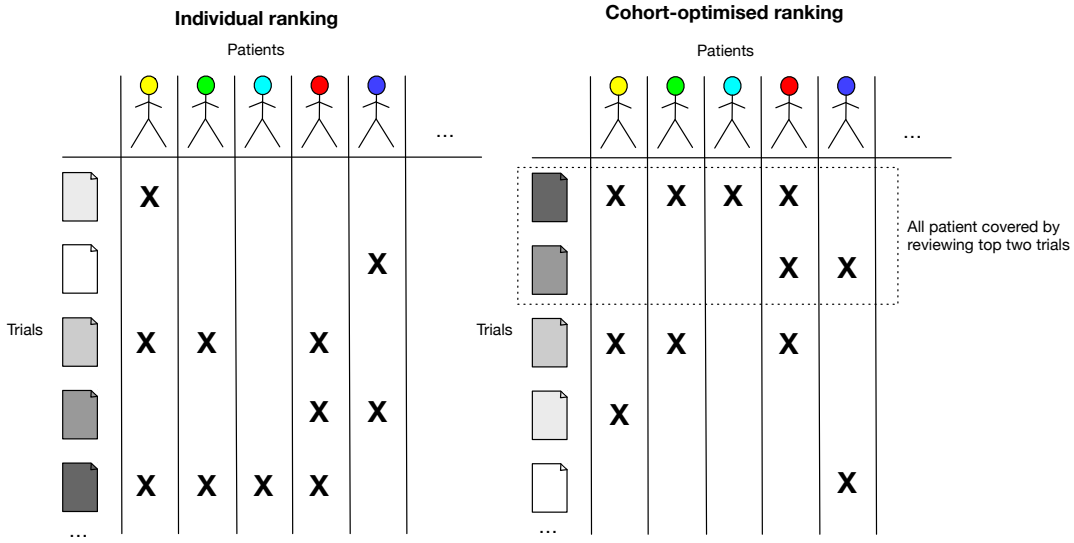

Figure 1: Sample rankings for two different approaches: cohort-aware (right) and individual patient ranking (left). The X indicates that the patient is relevant to the corresponding trial. For cohort-optimised ranking, full coverage of the five patients is achieved by reviewing only the top two trials. In contrast, for individual ranking, full coverage is only achieved by having to review five trials.

A key observation with all the above methods is that they focus on matching a single patient to trials (e.g., TREC PM); or matching a single trial to patients (TREC MedTrack); they do not consider the case of cohort retrieval — matching a cohort of patients to a set of trials. In the previous section, we have outlined the case for cohort retrieval. To our knowledge, there has been no investigation into how cohort retrieval might be applied to or impact search for clinical trials — thus cohort retrieval is the focus of this study.

4 FORMALISING THE PATIENT COHORT COVERAGE PROBLEM

Figure 2 presents graphically how the proposed cohort-based retrieval problem compares with the traditional, single-query approach. We formally describe the various aspects below.

Let Q represent the set of n patients/queries $q \in Q$ for which we would like to find eligible trials. If each query is answered independently, as in traditional IR evaluation settings, then a result set R_Q is formed by considering the individual ranked lists $r_1, \ldots, r_n$ for each of the single queries q (thus $R_Q = \{r_1, \ldots, r_n\}$). Each result list would contain clinical trials $t \in T$ for a large collection of clinical trials (e.g., from ClinicalTrials.gov).

Now let T_Q instead represent the final cohort-based ranking of trials $t \in T$ for the set of patients Q. Unlike R_Q, that contains a ranked list for each query (and thus n rankings in total), T_Q is a unique ranking that is meant to consider all patients (queries) in the cohort. Ideally, T_Q should provide a ranking of relevant trials that *maximises* the number of patients that can be recruited to the trials.

In the clinical setting, a clinician needs to manually review the retrieved trials. Busy clinicians will only have the capacity to process a limited numbers of trials and they may thus impose the constraint of viewing no more than r trials such that $|T_Q| < r$.

Note that it is important that a patient be enrolled in at least one trial, and there is no penalty nor gain in suggesting multiple relevant trials for a single patient. Formally, this problem is akin to that of *maximum coverage problem* in computer science [9]. Refinements to the cohort coverage problem we have outline would be to account for the non-uniform cost of assessing trials (some trials may require more effort to assess), and non-uniform gains (a trial for life-threatening conditions or a rare diseases); however, we leave them to future work.

In cohort-based retrieval, the effectiveness of a system is then a measure of what portion of the patient cohort (Q) is covered by the ranking of trials (T_Q). We call this measure *recruitment coverage* and define it for the whole patient cohort as:

$$\texttt{rec\_cov}(T_Q) = \frac{\sum_{q \in Q} \text{rel}(q, T_Q)}{|Q|}, \tag{1}$$

$$\text{rel}(q, T_Q) = \begin{cases} 1, & \text{if } \sum_{t \in T_Q} \text{rel}(q, t) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

and $\text{rel}(q, t)$ is 1 if trial t is relevant to patient q, and 0 otherwise. Thus rec_cov is 1.0 when T_Q contains at least one relevant trial for every patient in Q.

5 METHODOLOGY

In this paper, we aim to understand the applicability of traditional, single-query retrieval methods to the problem of cohort-based clinical trial retrieval, and understand what advantages new cohort-aware methods may provide.

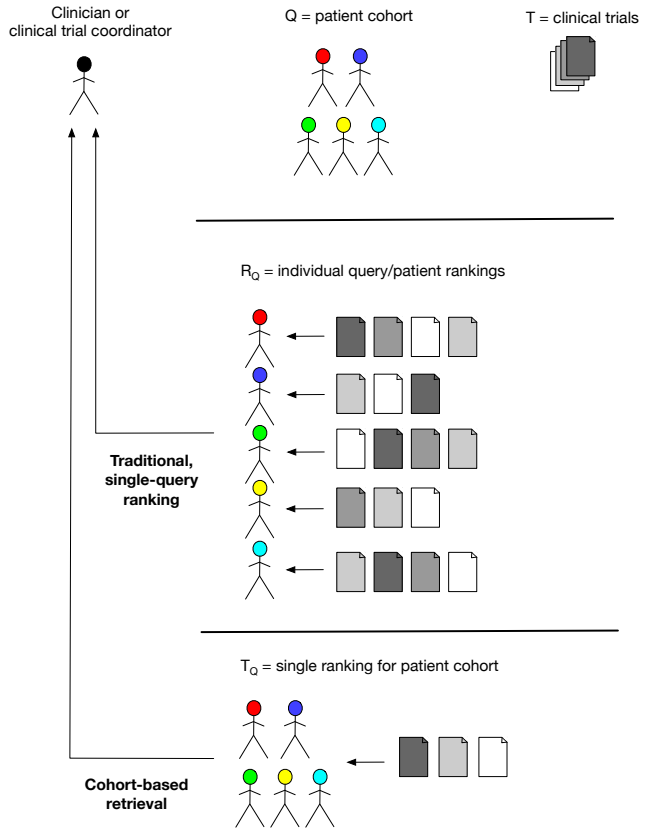

Figure 2: Comparing the common approach of clinical trial retrieval, which is traditional, single-query ranking with the alternative of cohort-based retrieval.

This study uses data from the TREC Precision Medicine Track (TREC PM) 2017 [17] and 2018 [16]. For both years, participants were provided synthetic patient records as queries and tasked with retrieving clinical trials crawled from ClinicalTrials.gov. Relevance assessment was done by clinicians and the focus was on cancer patients and trials. The TREC PM test collection was not devised with cohort-based retrieval in mind and the patients were specific to the cancer space. This, though, represents a realistic scenario for cohort-based retrieval where there would, in fact, be a single, similar source of patients — a specific cancer treatment clinic, for example. Furthermore, the specific domain of cancer trials is large enough and pressing enough in terms of critical access to new trials, to warrant it as a focus in its own right.

In the next sub-sections, we describe a number of different methods to investigate cohort-based retrieval. Figure 3 is used as a visual aid and we refer to different parts of the figure in subsequent sections.

5.1 Single-Query Ranking Systems

Given the TREC PM task, we established two single-query retrieval systems: baseline and state-of-the-art. The baseline system used BM25F as clinical trials are provided as XML files with separate

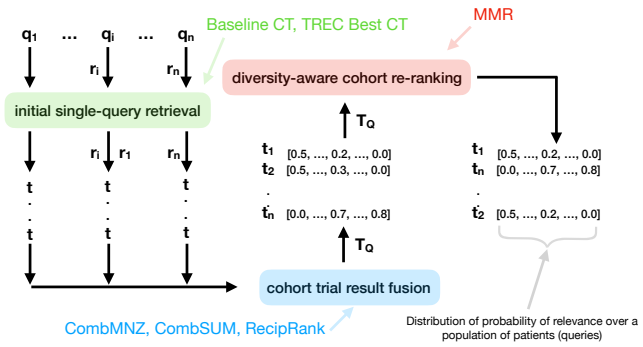

Figure 3: Various methods and experiments used in investigating cohort-based retrieval.

fields.[1] As patient record queries are often verbose we adopt the query reduction technique of [10], shown to be effective in this task. In addition, filtering on demographics (age, gender, pregnancy) and trials status (recruiting status, location) was implemented via boolean field restrictions. This system, denoted 'Baseline CT', represents a baseline IR system specific for clinical trials search.

To form a state-of-the-art system we used the best performing runs from TREC PM, combing the best run from 2017 [14] with the best run from 2018 [23]. This represents the most effective ranking of trials for single queries; we denote it 'TREC Best CT'. Figure 3 shows the single-query ranking systems in green.

5.2 Cohort Trial Ranking via Results Fusion

Given the single-query rankings provided by the systems outlined above, we now consider cohort-based retrieval: providing a single ranking T_Q for the entire query set Q. We adapt result fusion techniques for this purpose [13]. In result fusion, typically rankings from different *systems* — but for the same query — are fused to provide a single ranking for that query. Instead, we fuse the rankings from different *queries* into a single ranking for the query set.

Assuming we have a set of ranked lists R_Q, with r_q representing the ranked list for a query $q \in Q$, fusion methods would compute a cohort retrieval score (CRS) for a trial t by considering the scores t obtained for each query (patient) q. The three fusion methods used were:

CombSUM: Sums the retrieval scores of documents contained in more than one rank list and rearranges the order:

$$CRS_{\text{CombSUM}}(t) = \sum_{t \in R_q} score_{r_q}(t). \quad (3)$$

CombMNZ: Sums the retrieval scores of documents contained in more than one list, and multiplies their sum by the number of lists where the document occurs:

$$CRS_{\text{CombMNZ}}(t) = |\{r_q : t \in r_q\}| \cdot \sum_{t \in R_q} score_{r_q}(t). \quad (4)$$

Recip. Rank: Sums the reciprocal rank of documents contained in each rank list:

$$CRS_{\text{RR}}(t) = \sum_{t \in R_q} \frac{1}{rank_{r_q}(t)}. \quad (5)$$

[1] Implemented in Elasticsearch v5.3; no stopping or stemming.

Both CombSUM and CombMNZ involve combing scores from different distributions. We normalise scores for each query using minmax score normalisation:

$$\text{norm\_score(t)} = \frac{\text{score(t)} - \min(r_q)}{\max(r_q) - \min(r_q)}. \quad (6)$$

Figure 3 shows results fusion in blue.

5.3 Cohort-aware Greedy Ranking

The fusion techniques outlined so far do not take into account the coverage and, therefore, do not optimise the final ranking for `rec_cov`. To demonstrate whether this is a problem, next we utilise the qrels to form a strong cohort-aware benchmark that ranks trials according to `rec_cov`. Note that producing a true optimal ranking of trials for a query set is an NP-hard problem, as the task reduces to the maximum coverage problem [9]. We thus consider heuristics for this problem. This involves first selecting the trial that is relevant to (i.e., best covers) the most patients/queries, removing these queries/patients from the list of candidates, then selecting the next trial that best covers the remaining queries. The process continues iteratively until the max number of trials r is reached (or until no more queries can be covered). Pseudo code for this process is provided in Algorithm 1. This greedy-based algorithm, which uses qrels, represents a sub-optimal, but computable, upper bound effectiveness for cohort-based clinical trial retrieval; we denote it 'GreedySetCover'.

A more naive cohort-based retrieval method is to simply rank trials according to how many queries they are relevant to (again using qrels). This differs from GreedySetCover in that it does not consider which queries have been covered already by previous trials. Thus two trials may be ranked highly because they cover a large number of queries, but they happen to have high overlap — they cover the same set of queries — so do not contribute to a higher `rec_cov`. This naive greedy approach is denoted 'GreedyNaive'.

Both these greedy approaches used the qrels to know which trials were relevant to a patient; thus they were an oracle approach rather than a real cohort-based retrieval method. Their purpose was to study and understand how single-query retrieval compared against cohort-aware approaches and understand the potential of new cohort-aware methods.

Algorithm 1 Pseudo code for set cover algorithm to produce cohort-aware final ranking of trials.

Require: Q, T, r ▷ Set of queries/patients, Trials, Max trials
Ensure: T_Q, Q' ▷ Final ranking of trials, Patient cohort covered

```
T_Q, Q' = {}
while |T_Q| < r do
    select t_i ∈ T that covers the most queries in Q
    T_Q = T_Q ∪ {t_i}
    T = T − {t_i}
    Q' = Q' ∪ {queries covered by t_i}
return rec_cov = |Q'| / |Q|
```

5.4 Cohort-aware Diversity Reranking

The cohort-aware greedy ranking methods above explicitly model coverage using the qrels; thus they cannot be used in practice. Next we present a possible automatic cohort-aware method that accounts for coverage without using qrels. We adapt the Maximal Marginal Relevance (MMR) method [1], commonly used to diversify search results [19], to the problem of cohort-aware ranking. MMR computes the final score of a candidate document to be ranked by interpolating the standard relevance score with a diversity score computed with respect to the previous documents ranked so far.

In our adaption of MMR, we apply diversification after single-query ranking, namely we first use fusion to produce a single list of trials for the patient cohort, and then rerank this list via MMR. We compute the diversity score by comparing a candidate trial with a condensed representation of all the trials already ranked thus far. Figure 3 shows diversity reranking in red. Specifically, for the purpose of computing diversity scores, we represent a trial t_i with a vector v_i containing a likelihood distribution defined over the patient cohort. Each element j of the distribution refers to the likelihood that trial t_i is relevant to patient q_j. (This is taken from whether the trial was retrieved for that query/patient in the single-ranking system.) When iteratively building a ranking, we construct a cumulative vector v_p that represents the k trials $t_1, \ldots, t_k$ ranked so far, by summing the individual trial vectors $v_1, \ldots, v_k$. We then compare, using Jensen–Shannon divergence (JSD), the cumulative trial vector v_p with each of the candidate trial vectors v_c for the trials yet to be ranked. The candidate trial to be ranked at rank position $k+1$ is then selected according to:

$$\arg \max_{t_c \in T \setminus t_1,\ldots,t_k} [\alpha \cdot CRS(t_c) + (1-\alpha) \cdot \text{JSD}(v_p, v_c)] \quad (7)$$

where $CRS(t_c)$ is the cohort relevance score of a trial (according to one of the fusion methods outline in Section 5.2) and α is a hyper-parameter that controls the mix of relevance and diversity.

Plainly put, we rank the trial by combining its relevance score (*CRS*) and how diverse it is with a cumulative representation of the previous trials ranked so far, favouring trials that relate to patients less covered in previous trials.

Note that diversity reranking via MMR is done as a means to understand the influence of diversity rather than as a full-fledged method we propose to solve the cohort-ranking problem. For the latter, we would need sufficient training data in the form of multiple patient cohorts in order to both properly set α and to conduct statistical tests of effectiveness. Since we do not have this data, we instead explore how α (i.e., diversity) influences recruitment coverage and leave proper α estimation to future work.

5.5 Evaluation Settings

Empirical evaluation was done using TREC PM 2017 and 2018 (30 + 50 topics), where documents were clinical trials from an April 2017 snapshot of ClinicalTrials.gov [16, 17]. For each topic, a maximum of 1,000 clinical trials were retrieved. The BM25F-based single-query ranking system had two free parameters: $b = 0.75$, $k_1 = 1.2$ (we set fields to have equal importance). The cohort trial ranking methods based on result fusion were parameter-free. The cohort-aware diversity method had a parameter α which was tuned on the

System	Recip. Rank	Prec@10	Prec@100	NDCG
Baseline CT	0.2944	0.2354	0.1099	0.2898
TREC Best CT	0.7745	0.5620	0.1914	0.5536

Table 2: Retrieval effectiveness for single-query ranking systems. Results between systems statistically significant (paired two-tails t-test, $p < 0.001$) on all four measures. NDCG cut 1000.

evaluation corpus by sweeping $\alpha = 0.0$ (full relevance) to 1.0 (full diversity), with step 0.1.

The two single-query ranking systems, Baseline CT and TREC Best CT, were evaluated by averaging their effectiveness across the query set. The evaluation measures were precision at 10, precision at 100 and NDCG (cut 1000). P@10 models the use case of a clinician accessing a patient's record as part of a consultation. An IR system can automatically initiate a search to find relevant clinical trials. The clinician is time-pressured and would likely only review a small number of trials. P@100 models the use case where the clinician is specifically searching for clinical trials, may dedicate more time and be willing to evaluate in the order of 100 trials (hence P@100). NDCG (cut 1000) accounts for the rank position of relevant trials and was an official measure at TREC PM. These measures show the effectiveness of the single-query rankings but not the final, cohort-based fused ranking.

The ultimate evaluation is what portion of the patient cohort is covered by a final ranking of trials. The effectiveness of the cohort-based retrieval is done according to the `rec_cov` measure. Evaluation is done for the different fusion methods (CombSUM, CombMNZ and RecipRank) for the Baseline CT and TREC Best CT systems, as well as for the two cohort-aware rankings of Greedy-Naive and GreedySetCover. Remember we stated that clinicians would only consider trials up to a set depth r; thus we experiment with a depth cutoff for `rec_cov` of r = [1, . . . , 10, 15, 20, 25, 30, 40, 50, 70, 90, 100, 150, 200].

6 RESULTS & ANALYSIS

6.1 How effective was single-query ranking?

First, we consider the single-query ranking system effectiveness, reported in Table 2. The TREC Best runs were significantly better than Baseline CT on all four measures. TREC Best CT returned many more relevant trials within higher ranked positions than Baseline CT. Using fusion, it would then be more likely that these relevant documents would be included in the final cohort-based ranking. More relevant documents in the cohort-based ranking would likely have led to a higher `rec_cov`. This was the intuition at least.

6.2 How effective was cohort-based retrieval via single-query ranking fusion?

Figure 4 shows an aggregation of results for a number of facets of cohort-based ranking. In the next sections, we consider different aspects of this figure.

First, we consider how the fusion of the single-query rankings impacted the ultimate recruitment coverage (`rec_cov`). This is shown as black lines for each fusion method in the first three plots (CombMNZ, CombSUM and RecipRank). (We leave the discussion on diversity reranking, shown in red, for later in Section 6.3.) For score-based fusion of CombMNZ and CombSUM, there was no large difference in `rec_cov` for the two single-query rankings — TREC Best CT was not much better then Baseline CT. This was surprising given TREC Best CT was considerably better than Baseline CT for the single-query setting already outlined in the previous section. In fact, when the number of trials was less than 10, Baseline CT was actually superior: e.g., for $r = 5$, `rec_cov`(Baseline CT) = 0.2405 vs `rec_cov`(TREC Best CT) = 0.1899.

While different single-query rankings did not impact score-based fusion, they did, instead impact rank-based fusion (as shown in the third, RecipRank plot). Here, `rec_cov` was much greater when fusion was applied to TREC Best CT. Two factors were at play for RecipRank: First, the ranking scores were ignored. Scores may not have been a good measure of the likelihood of relevance and thus ignoring them in RecipRank may have been beneficial. Further, even with minmax normalisation, scores may not have translated well when fusing rankings from different queries, e.g., because of different score distributions. Second, RecipRank fusion applied an exponential decay in the weighting as it moved down the ranking. The fact that a document was in a top-rank position had far more influence on the final cohort-ranking than if the document appeared in many single-query rankings. Said another way, trials that much more closely matched a single patient were better than trials that weakly matched multiple patients. This characteristic of RecipRank may have also helped to ensure that the documents that did end up in the final cohort-ranking were, in fact, relevant.

6.3 How effective was cohort-based diversity reranking?

Now we consider the effect of diversity reranking, comparing each red curve with its corresponding black equivalent. Diversity reranking was beneficial but in different ways. Diversity strongly improved trials fused using TREC Best CT while having a much smaller improvement on Baseline CT. Previously, we showed TREC Best CT retrieved many more relevant trials than Baseline CT (Table 2). (While these trials were relevant to one patient, they were not diverse; thus they did not translate to high `rec_cov`.) Now consider what happens when diversity is applied to TREC Best CT and to Baseline CT. Using TREC Best CT you have a large pool of relevant documents to draw on to rerank via diversity; using Baseline CT you risk bringing in non-relevant documents as you rerank via diversity. Thus applying some cohort-based diversity to the good quality initial retrieval of TREC Best CT yields much better recruitment coverage.

Diversity reranking was done by interpolating the relevance and diversity scores using the mixing parameter α. Figure 5 shows how applying diversity (α) to trial ranking impacted recruitment coverage ($\alpha = 0.0$ only relevance, $\alpha = 1.0$ only diversity). We can see that diversity reranking was beneficial but how so depended on the fusion method: score-based fusion methods CombMNZ and

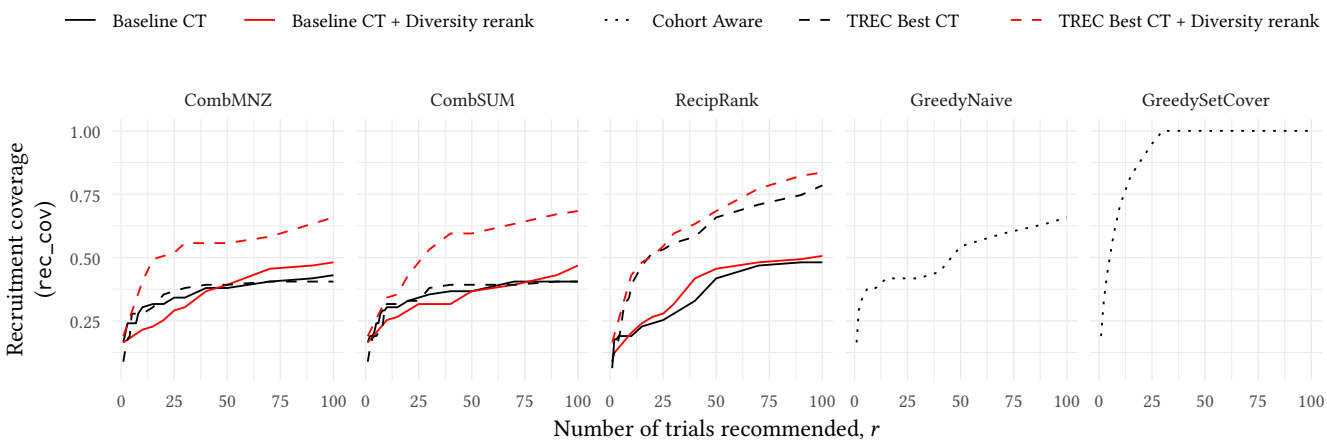

Figure 4: The effectiveness of cohort-based retrieval as measured by `rec_cov` (y-axis) for different fusion methods (CombMNZ, CombSUM, RecipRank) applied to different single-query rankings (Baseline CT and TREC Best CT). Cohort-aware rankings used qrels to provide comparison: GreedyNaive ignoring query overlap and GreedySetCover showing benefits of complete knowledge of overlap. The x-axis shows the number of trials a clinician would review.

CombSUM benefitting from more diversity reranking than the most effective fusion method, RecipRank.

6.4 What is the relationship between single-query effectiveness and cohort recruitment coverage?

Previously we observed that single-query ranking effectiveness does not always translate directly to recruitment coverage effectiveness. To better understand the effect of single-query ranking effectiveness we manually produced single-query rankings that had set precision levels of precision at 10 of [0.1, .., 1.0]. (This was done by producing rankings of 10 documents containing 1..10 relevant documents from the qrels.) The rankings were then fused and the corresponding `rec_cov` score calculated. The results of this experiment is shown in Figure 6.

Without diversity reranking, as we manually increase precision at 10 from 0.0, `rec_cov` increases. However, beyond 0.4 in precision (i.e., 4 relevant documents in top 10) `rec_cov` does not increase. This shows that even though more relevant clinical trials are being returned by the single-query ranking systems, these trials do not help to increase coverage. For example, they may just be more relevant trials for a single patient rather than matching multiple patients. Instead, when diversity with respect to the patient cohort is injected via diversity reranking, `rec_cov` continues to increases because the growing pool of relevant trials is being reranked to favour those that match multiple patients/queries.

6.5 How effective was cohort-aware greedy ranking?

We now consider the two cohort-aware greedy ranking methods, GreedyNaive and GreedySetCover, in Figure 4. Recall that GreedyNaive simply ranked trials by the number of queries that it matched and did not take into account any overlap of these queries. First, we obverse that GreedyNaive was actually not as effective as RecipRank. GreedyNaive's lack of effectiveness tells us that the trials preferred by GreedyNaive were those that matched a common set of overlapping queries/patients. These may have been very general trials, for example, for which many patients could have been eligible. However, they still did not cover that much of the overall patient cohort, hence the poor `rec_cov` for GreedyNaive. These results tell us that if we were to build a specialised cohort-based ranking model, not accounting for overlap would severely hamper effectiveness.

The GreedySetCover results tell us what was possible when we had complete knowledge of overlap. The marked difference in effectiveness between GreedySetCover and GreedyNaive shows the importance of considering overlap.

From the figure, we see it was possible to obtain complete coverage (i.e., `rec_cov` = 1.0), across the 80 patients, with just 30 trials. The fact that only 30 trials were needed highlights that it was the case that some trials match multiple patients. In turn, this further motivates taking a cohort-based retrieval approach, with all its benefits: reduced recruitment effort, less clinician screening time and greater access for patients to new treatments.

How well can we do with automated cohort-based ranking? GreedySetCover sets the local optimal[2] for `rec_cov` when ranking iteratively (greedy) with respect to cohort coverage. We compare this to the best fusion-based method of RecipRank (Figure 4). RecipRank actually performs close to the optimal GreedySetCover for early rank positions (i.e., low value of r). If we return to one of our original use cases — a busy doctor able to review only a small

[2] A global optimal is only attainable by solving the NP-hard maximum coverage problem [2, 9].

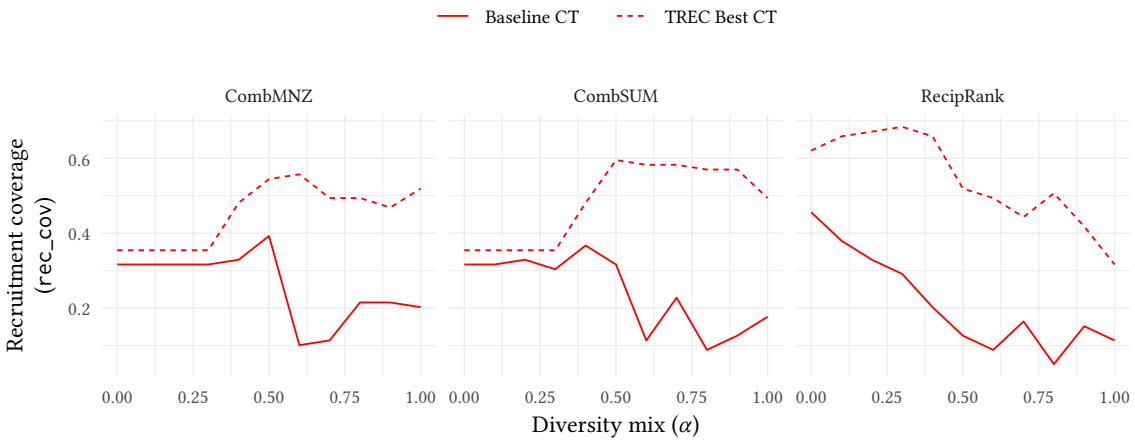

Figure 5: The impact of diversity scoring on cohort-based retrieval. Fusion is done on the two single-ranking system (Baseline CT & TREC Best CT) using the three fusion methods. The diversity mix (α) is varied from 0.0 (only relevance, no diversity) to 1.0 (no relevance vs. only diversity). Results show diversity reranking improved `rec_cov`.

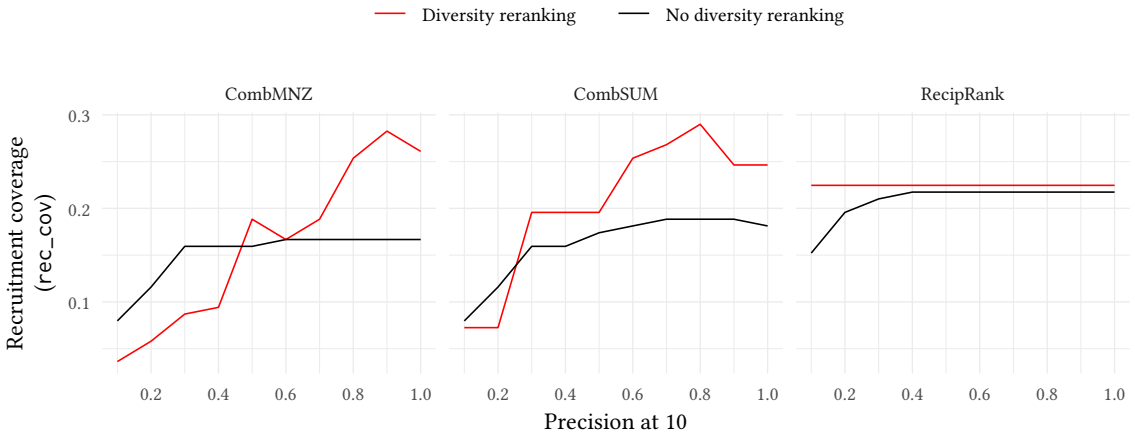

Figure 6: The impact of single-rankings of set effectiveness on `rec_cov`. Manually produced rankings for set precision at 10 levels (x-axis) and the impact of these on `rec_cov` (y-axis).

number of trials — we can see that RecipRank would be highly effective. However, RecipRank begins to diverge from the optimal as $r > 10$. GreedySetCover achieves complete coverage (`rec_cov`=1.0) when $r \approx 25$. While RecipRank continues to improve past this point, it does not achieve complete coverage within the first 100 trials. Thus for the use case of really wanting to maximising full recruitment coverage, RecipRank would not be sufficient and more work is required.

7 CONCLUSION & FUTURE WORK

This study provides an initial foray into the problem of cohort-based clinical trials retrieval; that is, devising a ranking of clinical trials that covers as much of the patient cohort of queries as possible. First, we formalised this problem, showing how cohort-based ranking contrasted with single query ranking. Furthermore, we highlighted how the objective of cohort-based is that of coverage across the query set — we introduced `rec_cov` as a means to measure coverage.

Cohort-based retrieval can be achieved through result fusion. We applied fusion techniques to fuse the rankings from different *queries* (as opposed to different systems), thus producing a cohort-based ranking for the patient cohort. We implemented three result fusion techniques, applied to two single-query rankings: a Baseline clinical trials system and the TREC Best clinical trial systems. Empirical evaluation on the TREC Precision Medicine track showed that the TREC Best system was significantly better in terms of single-query evaluation. However, and perhaps surprisingly, this did not translate to so large a corresponding improvement in cohort-based

retrieval as measured by rec_cov. The differences in effectiveness were also highly dependent on how many trials were considered in the cohort-based ranking, equating to how many trials a clinician may manually review as part of a patient recruitment exercise.

To provide further insights into cohort-based retrieval, we implemented two oracle, greedy cohort-aware methods that used qrels to rank trials. These highlighted that trials that match many patients often overlap in the patients they match, suggesting that it is necessary to explicitly model the coverage and overlap when producing an effective cohort-based ranking. Also, through the exploration of greedy heuristics, we were able to show that the entire 80 patients cohort could be fully covered by retrieving a small set of only 30 trials. We tied this back to the benefits of cohort-based retrieval in terms of reducing the effort in clinical trials recruitment and providing more patients with access to new and emerging treatments.

With the greedy methods providing evidence for the benefit of cohort-aware methods, we adapted existing diversity ranking [19] methods to rank trials that increase patient coverage. Diversity reranking did show promise, while multiple cohort datasets were needed to properly estimate the mix of diversity vs. relevance (α).

While TREC Precision Medicine provided the right data and use case for cohort-based retrieval, it had some limitations. TREC PM mainly focused on matching cancer patients to oncology trials, where there is additional genetic information about the patient. None of the methods, experiments or analysis of this paper was specific to this cancer related setting; however, the results may differ outside of the cancer space. That said, matching patients to trials in the cancer space is critical in its own right and even isolating cohort-based retrieval to just this area is still justification enough for such work. An additional limitation of our empirical evaluation is that only one cohort of patients was used. Ideally, we would have multiple cohorts of different patient populations. This would allow statistical comparisons between these to determine if findings generalise. For example, to determine if the single-query rankings (Baseline CT vs. TREC Best CT) led to statistically different rec_cov effectiveness, we would need a set of patient cohorts to evaluate on.

Having provided the motivation and evidence for cohort-based retrieval, we now intend to investigate specific retrieval methods that rank trials directly from a patient cohort rather than fusing single-query rankings. Beyond this, a learning-to-rank approach, that takes as input a cohort of patients, and is able to extract multiple features from trials, would be our goal. As part of this, it would be necessary to encode features related to overlap and coverage. How to do this generically for any cohort of patients is an open line of enquiry.

REFERENCES

[1] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.

[2] Ben Carterette. 2011. An analysis of NP-completeness in novelty and diversity ranking. *Information Retrieval* 14, 1 (2011), 89–106.

[3] Steven R Chamberlin, Steven D Bedrick, Aaron M Cohen, Yanshan Wang, Andrew Wen, Sijia Liu, Hongfang Liu, and William R Hersh. 2020. Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. *JAMIA open* 3, 3 (2020), 395–404.

[4] Giorgio Maria Di Nunzio, Stefano Marchesin, and Maristella Agosti. 2019. Exploring how to Combine Query Reformulations for Precision Medicine.. In *TREC*.

[5] Tracy Edinger, Aaron M Cohen, Steven Bedrick, Kyle Ambert, and William Hersh. 2012. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC medical records track. In *AMIA annual symposium proceedings*, Vol. 2012. American Medical Informatics Association, 180.

[6] Erik Faessler, Udo Hahn, and Michel Oleynik. 2019. JULIE Lab & Med Uni Graz@ TREC 2019 Precision Medicine Track.. In *TREC*.

[7] Travis R Goodwin and Sanda M Harabagiu. 2018. Learning relevance models for patient cohort retrieval. *JAMIA open* 1, 2 (2018), 265–275.

[8] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[9] Samir Khuller, Anna Moss, and Joseph (Seffi) Naor. 1999. The budgeted maximum coverage problem. *Inform. Process. Lett.* 70, 1 (1999), 39 – 45.

[10] Bevan Koopman, Liam Cripwell, and Guido Zuccon. 2017. Generating Clinical Queries from Patient Narratives: A Comparison between Machines and Humans. In *SIGIR*. Tokyo, Japan.

[11] Bevan Koopman and Guido Zuccon. 2014. Understanding Negation and Family History to Improve Clinical Information Retrieval. In *Proceedings of the 37th annual international ACM SIGIR conference on research and development in information retrieval*. ACM.

[12] Bevan Koopman and Guido Zuccon. 2016. A test collection for matching patients to clinical trials. In *SIGIR*. 669–672.

[13] Oren Kurland and J. Shane Culpepper. 2018. Fusion in Information Retrieval: SIGIR 2018 Half-Day Tutorial. In *SIGIR*. 1383–1386.

[14] ASM Ashique Mahmood, Gang Li, Shruti Rao, Peter B McGarvey, Cathy H Wu, Subha Madhavan, and K Vijay-Shanker. 2017. UD_GU_BioTM at TREC 2017: Precision Medicine Track. In *TREC*.

[15] Lynne T Penberthy, Bassam A Dahman, Valentina I Petkov, and Jonathan P DeShazo. 2012. Effort required in eligibility screening for clinical trials. *Journal of Oncology Practice* 8, 6 (2012), 365–370.

[16] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, and Alexander J. Lazar. 2018. Overview of the TREC 2018 Precision Medicine Track. In *TREC*.

[17] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2017. Overview of the TREC 2017 Precision Medicine Track. In *TREC*.

[18] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, Shubham Pant, and Funda Meric-Bernstam. 2019. Overview of the TREC 2019 Precision Medicine Track. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019 (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. 1250. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.PM.pdf

[19] Rodrygo LT Santos, Craig Macdonald, Iadh Ounis, et al. 2015. Search result diversification. *FnTIR* 9, 1 (2015), 1–90.

[20] Ellen M Voorhees and William R Hersh. 2012. Overview of the TREC 2012 Medical Records Track. In *TREC*.

[21] Ellen M Voorhees and Richard M Tong. 2011. Overview of the TREC 2011 Medical Records Track. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*. Gaithersburg, Maryland, USA.

[22] Qi Zheng, Yong Li, Jiaying Hu, Yan Yang, Liang He, and Yi Xue. 2019. ECNU-ICA team at TREC 2019 Precision Medicine Track.. In *TREC*.

[23] Xuesi Zhou, Xin Chen, Jian Song, Gang Zhao, and Ji Wu. 2018. Team Cat-Garfield at TREC 2018 Precision Medicine Track. In *TREC*.

[24] Dongqing Zhu, Stephen Wu, Ben Carterette, and Hongfang Liu. 2014. Using large clinical corpora for query expansion in text-based cohort identification. *Journal of biomedical informatics* 49 (2014), 275–281.