Modelling Relevance towards Multiple Inclusion Criteria when Ranking Patients

Nut Limsopatham, Craig Macdonald, and Iadh Ounis
firstname.lastname@glasgow.ac.uk

School of Computing Science
University of Glasgow, Glasgow, UK

ABSTRACT

In the medical domain, information retrieval systems can be used for identifying cohorts (i.e. patients) required for clinical studies. However, a challenge faced by such search systems is to retrieve the cohorts whose medical histories cover the inclusion criteria specified in a query, which are often complex and include multiple medical conditions. For example, a query may aim to find patients with both 'lupus nephritis' and 'thrombotic thrombocytopenic purpura'. In a typical best-match retrieval setting, any patient exhibiting all of the inclusion criteria should naturally be ranked higher than a patient that only exhibits a subset, or none, of the criteria. In this work, we extend the two main existing models for ranking patients to take into account the coverage of the inclusion criteria by adapting techniques from recent research into coverage-based diversification. We propose a novel approach for modelling the coverage of the query inclusion criteria within the records of a particular patient, and thereby rank highly those patients whose medical records are likely to cover all of the specified criteria. In particular, our proposed approach estimates the relevance of a patient, based on the mixture of the probability that the patient is retrieved by a patient ranking model for a given query, and the likelihood that the patient's records cover the query criteria. The latter is measured using the relevance towards each of the criteria stated in the query, represented in the form of sub-queries. We thoroughly evaluate our proposed approach using the test collection provided by the TREC 2011 and 2012 Medical Records track. Our results show significant improvements over existing strong baselines.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

1. INTRODUCTION

Electronic medical records (EMRs) have recently been deployed to improve healthcare services [12]. One of the important applications is to leverage the EMRs within a search system to identify cohorts for clinical trials. An effective patient ranking system plays a crucial role in identifying such relevant cohorts. Specifically, when conducting comparative effectiveness studies for a particular healthcare procedure (e.g. a diagnostic test), healthcare practitioners can use the system to search for cohorts (i.e. patients) having a medical history (in the form of medical records) relevant to a set of inclusion criteria [8, 30, 31], which are used as a query. It is essential to identify the patients whose medical histories are relevant to (i.e. cover) all of the inclusion criteria [8, 12]. Edinger et al. [8] showed that existing patient ranking systems (e.g. [7, 14, 19]) fail in retrieving the patients whose medical records cover all of the criteria specified in a given query. In this paper, we propose to rank the patients whose medical histories are likely to cover all or most of the inclusion criteria higher. For example, for a query identifying patients with 'heart disease', 'diabetes' and 'alzheimer's', an effective patient ranking system should rank the patients who are diagnosed with all of the three diseases higher than those who suffer from only two or one of the conditions.

Existing works on patient ranking mostly focus on estimating the relevance of the patients without considering how many of the inclusion criteria the retrieved patients cover. There are two main groups of patient ranking models [18]. First, the so-called *patient model* ranking approaches (e.g. [7, 14]) represent a patient by combining the associated medical records into the form of a single *patient document*, and use the latter as a unit of retrieval. For example, Demner-Fushman et al. [7] effectively deployed the patient model by using a term weighting model (e.g. BM25 [25] or a language model [13]) to rank the patient documents. On the other hand, the so-called *two-stage model* ranking approaches (e.g. [19, 34]) initially rank the medical records based on their relevance towards the query, and then calculate the relevance of patients by aggregating the relevance scores of their associated medical records that have been retrieved for the query. For example, Limsopatham et al. [19] used the expCombSUM voting technique from the Voting Model [20] to effectively aggregate the relevance scores for a patient. Later, Limsopatham et al. [18] showed that while in most cases the two-stage model approach based on the expCombSUM voting technique outperformed the patient model, the differences in the retrieval performances were not significant.

However, none of the aforementioned approaches explicitly aims to rank patients based on the probability that they are relevant to all or most of the query criteria. To deal with such a challenge, in this work, we propose to rank patients based on the relevance of their medical history towards a query, while also maximising the relevance towards a greater number of the specified inclusion criteria, within both the patient and the two-stage models. Inspired by recent work on coverage-based search result diversification [1, 26], we estimate the relevance of a patient towards the inclusion

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright 2014 ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661958.

criteria based on the likelihood that a patient's medical history covers the set of inclusion criteria, measured using the notion of sub-queries [26]. To do so, we make use of the probability that a patient's medical history is relevant to (i.e. covers) a sub-query representing a criterion extracted from the original query. While existing search result diversification techniques aim to generate a ranking that covers the possible interpretations of information needs, the goal of our proposed approach is to retrieve patients who are highly relevant to multiple inclusion criteria. In particular, we propose to extend both the patient and the two-stage models to measure the relevance towards multiple inclusion criteria, while ranking patients. We demonstrate the effectiveness of our proposed approach in the context of the TREC 2011 & 2012 Medical Records track. Our results show that our proposed approach significantly outperforms existing effective patient ranking baselines. The main contributions of this paper are four-fold:

1. We propose a novel extension for existing patient ranking models to consider the relevance towards multiple inclusion criteria in the query.
2. We extract the inclusion criteria from a query by using a domain-specific resource and represent each obtained criterion as a sub-query.
3. We describe different techniques to estimate a parameter required in our proposed approach to trade-off between the relevance towards the query itself and the relevance towards multiple inclusion criteria.
4. We thoroughly evaluate our proposed approach using the standard experimental setup provided by the TREC 2011 & 2012 Medical Records track.

The remainder of the paper is organised as follows. Section 2 further discusses related work. Section 3 illustrates the problem that existing approaches could not effectively rank higher those patients who are relevant to more inclusion criteria stated in the query. Sections 4 and 5 introduce our approach to model relevance towards multiple inclusion criteria for a particular patient by measuring the relevance towards each of the inclusion criteria using sub-queries, and our approach for extracting the inclusion criteria from a given query using a well-established domain-specific resource, respectively. Sections 6 and 7 discuss our experimental setup and the obtained results when the trade-off parameter is uniformly set. Section 8 discusses and evaluates our proposed approach when using a regression technique to automatically set the trade-off between the relevance probability towards a query and the likelihood of covering the multiple inclusion criteria extracted from the query using training data. Section 9 analyses the retrieval performance of our approach. Finally, we provide concluding remarks in Section 10.

2. RELATED WORK

Existing patient ranking approaches do not explicitly model the relevance towards the multiple inclusion criteria that are stated in the query. For example, Demner-Fushman et al. [7] and King et al. [14] whose systems achieved the best retrieval performances at the TREC 2011 Medical Records track used the patient model to retrieve patients after enriching the medical records and/or queries using medical resources (e.g. UMLS Metathesaurus). Meanwhile, Limsopatham et al. [19] used the expCombSUM voting technique [20] to effectively retrieve patients based on the relevance of their medical records. The expCombSUM voting technique ranks the patients who have a few medical records that are highly relevant to the query higher than those who have many partially relevant medical records. However, Edinger et al. [8] showed that these approaches could not effectively retrieve patients whose medical records were relevant to the multiple inclusion criteria stated in the queries. Later, Zhu and Carterette [34], whose system achieved the best performance at TREC 2012, showed that combining the relevance scores computed using both the patient model and the two-stage model further improved retrieval performance. Still, their approach did not take into account the relevance towards multiple inclusion criteria. A conceptual representation approach (e.g. [16, 17, 23]) that represents documents and queries using medical concepts may be considered as implicitly ranking patients based on the relevance towards the inclusion criteria, assuming that the medical concepts in the query are the inclusion criteria. However, the conceptual representation approach does not explicitly model the probability that the set of the inclusion criteria stated in the query are covered by the medical records of a patient. In contrast, in this work, we introduce a novel approach to explicitly model the relevance towards the multiple inclusion criteria, which is inspired by existing works in the area of search results diversification. To the best of our knowledge, this is the first study on explicitly modelling the relevance towards several inclusion criteria when ranking patients.

Coverage-based search result diversification approaches aim to maximise the coverage of the possible interpretations of information needs within a set of retrieved documents. For example, Agrawal et al. [1] and Santos et al. [26] reranked documents to promote the maximum coverage of the predefined interpretations of the query, while minimising the redundancy of the documents. Unlike these approaches, we model the coverage of the multiple inclusion criteria within the subset of the retrieved medical records that are also associated to a particular patient, in order to promote the patients whose medical records are likely to cover a higher number of the inclusion criteria.

Another research area related to this work is the term weighting regularisation approach to promote the coverage of query aspects within a given retrieved document [33]. Indeed, the approach to regularise term weighting based on the semantic relationship among the query terms increases the weight of the query terms that are not associated to the other query terms, while decreasing the weight of the terms that highly relate to the other terms. Different from this approach, we highly rank patients whose retrieved medical records cover the multiple inclusion criteria stated in a query, which are measured based on the relevance towards each inclusion criterion extracted from the query.

3. MOTIVATION & PROBLEM DEFINITION

The aforementioned existing approaches rank patients based on their relevance to the query; however, they do not explicitly promote the patients whose medical records are relevant to multiple query criteria. Indeed, the *patient model* (e.g. [7, 14]) estimates the relevance of a patient p for a query q, $P(p|q)$, as follows:

$$P(p|q) \propto P(D_p|q) \tag{1}$$

where the patient document D_p is created by concatenating the medical records associated to the patient p. $P(D_p|q)$, which is the probability that D_p is relevant to the query q, can be estimated using any probabilistic retrieval model, such as a language model [13].

Alternatively, the *two-stage model* (e.g. [19, 34]) estimates the relevance of a patient p, by suitably aggregating the relevance probabilities of the medical records associated to the patient p, as follows:

$$P(p|q) \propto aggregate_{d_i \in R_p}\left[P(d_i|q)\right] \tag{2}$$

where d_i is a medical record in R_p, which is the set of retrieved medical records that are also associated to the patient p. $P(d_i|q)$ is the probability that the medical record d_i is relevant to the query q (e.g. estimated using a language model [13]), while $aggregate_{d_i \in R_p}[\cdot]$ can be calculated using any aggregate function, such as a voting technique [20].

Nevertheless, both the patient and the two-stage models may fail in ranking the patients for a query searching for patients with multiple health conditions, as discussed in the previous sections. We use Figure 1 to illustrate this problem. Consider that a query q is to find patients with 'heart disease' (i.e. criterion q_1), 'diabetes' (i.e. criterion q_2) and 'alzheimer's' (i.e. criterion q_3), and that medical records d_1 and d_2 are associated with the patient p_1, while the medical records d_3 and d_4 are related to the patient p_2. In Figure 1(a), the patient model (as in Equation (1)), which estimates the relevance of each patient using the concatenation of the medical records of that patient (e.g. D_{p1} and D_{p2}), ranks the patient p_1 higher than the patient p_2, according to their relevance probabilities, towards the query q (0.9 vs. 0.8). Meanwhile, in Figure 1(b), the two-stage model, which estimates the relevance of patients by suitably aggregating the relevance of their associated medical records (as in Equation (2)), also ranks the patient p_1 higher than the patient p_2. For instance, CombSUM estimates the relevance probability of a patient by summing up and normalising[1] the relevance probabilities of their associated medical records (e.g. after normalising, the relevance probabilities of patient p_1 and p_2 are 0.57 and 0.43, respectively). However, as previously discussed, an effective patient ranking approach should rank the patient p_2 higher than the patient p_1, as the patient p_2 is relevant to more criteria in the query q than the patient p_1 (q_1, q_2, q_3 vs. q_1, q_2).

Denoting the probability that a patient p is relevant to a query q as $P(p|q)$, and the likelihood that p is relevant to the multiple inclusion criteria stated in the query q as $P_c(p|q)$, an effective patient ranking model denoted by $F(p|q)$ must have the following two properties, which promote patients who are likely to be relevant to multiple query criteria:

Property 1: If $P(p_1|q) = P(p_2|q)$, then a patient p_1 should be ranked higher than a patient p_2, $F(p_1|q) > F(p_2|q)$, when $P_c(p_1|q) > P_c(p_2|q)$.

Property 2: If $P(p_1|q) \neq P(p_2|q)$, then $F(p_1|q) > F(p_2|q)$ when $P(p_1|q) \bigoplus P_c(p_1|q) > P(p_2|q) \bigoplus P_c(p_2|q)$ where $\bigoplus$ is an appropriate mixture of the two probabilities.

4. MODELLING RELEVANCE TOWARDS MULTIPLE INCLUSION CRITERIA

In this section, we propose to build a probabilistic model that satisfies the two properties previously discussed in Section 3. Specifically, our proposed approach models the mixture of the relevance towards the query and the likelihood of covering the inclusion criteria extracted from the query, to promote the patients whose medical records are relevant to a higher number of the inclusion criteria. Our proposed approach can be calculated as follows:

$$F(p|q) \propto (1-\lambda)\cdot P(p|q) + \lambda P_c(p|q) \tag{3}$$

where $P(p|q)$ is the probability that the patient p is relevant to the query q (i.e. *the relevance probability*), which can be measured using any existing patient ranking approach (e.g. Equations (1) or (2)); $P_c(p|q)$ is the likelihood that the medical records of the patient p cover the multiple inclusion criteria stated in the query q (we refer to this as the *coverage likelihood*), and λ is a mixture parameter to weight the importance of $P(p|q)$ and $P_c(p|q)$. $P_c(p|q)$ enables our approach to promote the patients whose medical records cover several query criteria, which may not be measured when using existing patient ranking approaches.

Given a set $Q = \{q_1, q_2, ..., q_n\}$ containing the inclusion criteria stated in the query q, we propose to estimate the *coverage likelihood* $P_c(p|q)$ as the combination of the beliefs that each criterion (i.e. akin to a sub-query) q_i in Q is covered by the medical records of the patient p, as follows:

$$P_c(p|q) = bel_{q_i \in Q}\left(P(p|q_i)\right) \tag{4}$$

where $P(p|q_i)$ is the probability that the patient p is relevant to the criterion q_i in Q. bel is a belief combination function, such as AND and OR, to combine the probabilities that the medical records of the patient p cover each inclusion criterion. Indeed, the belief combination functions have been extensively deployed within search approaches (e.g. [21, 24, 29]) to combine the probabilities that a particular document is relevant to each query term. For instance, bel^{AND} can be calculated as [21]:

$$bel^{AND}_{q_i \in Q}\left(P(p|q_i)\right) = \prod_{q_i \in Q} P(p|q_i) \tag{5}$$

In the remainder of this section, we discuss how our proposed approach can be applied within the existing patient and two-stage ranking models, respectively. Then, in Section 5, we describe our technique to extract the inclusion criteria from a query.

4.1 The Extended Patient Model

Within the patient model, our proposed approach can be adapted by inserting Equations (1) and (4) into Equation (3), as follows:

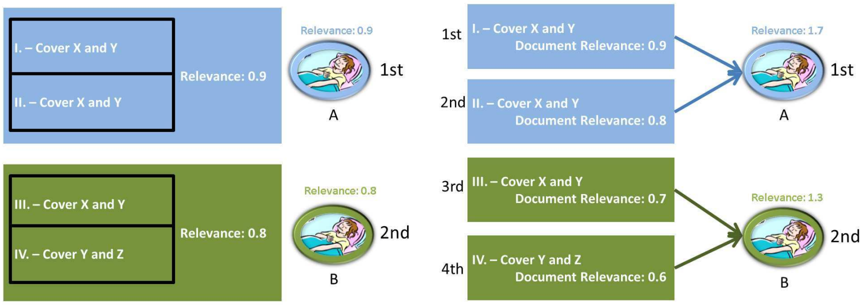
$$F(p|q) \propto (1-\lambda)\cdot P(D_p|q) + \lambda \cdot bel_{q_i \in Q}\left(P(D_p|q_i)\right) \tag{6}$$

where λ is a mixture parameter that weights the importance of the relevance probability and the coverage likelihood. Specifically, the first part of the equation, $(1-\lambda)\cdot P(D_p|q)$, focuses on the relevance probability of the patient document D_p towards the query q. The second part of the equation calculates the coverage likelihood of D_p. We use $P(D_p|q_i)$ to measure the likelihood that D_p covers a particular inclusion criterion q_i.

4.2 The Extended Two-Stage Model

As shown in Figure 1(b), the two-stage model firstly ranks the medical records, and then suitably aggregates their relevance probabilities to rank the associated patients. Hence, we can model the mixture of the relevance probability and the coverage likelihood either at the stage of ranking patients (Section 4.2.1), or at the stage of ranking medical records (Section 4.2.2).

[1] Note that there is a need to normalise the aggregated relevance probabilities to maintain probability estimates.



(a) A patient model (b) A two-stage model (CombSUM)

Figure 1: Illustrative examples of ranking patients using the two main existing approaches. Note that the relevance scores computed using CombSUM are normalised to maintain probability estimates.

4.2.1 *Ranking Patients*

First, we can model the mixture of the relevance probability and coverage likelihood at the patient ranking stage of a two-stage model by inserting Equations (2) and (4) into Equation (3), as follows:

$$F(p|q) \propto (1-\lambda)\cdot aggregate_{d_i \in R_p}\Big[P(d_i|q)\Big] + \lambda \cdot bel_{q_i \in Q}\left(aggregate_{d_i \in R_p}\Big[P(d_i|q_i)\Big]\right) \qquad (7)$$

4.2.2 *Ranking Medical Records*

In contrast, at the medical record ranking stage, the two-stage model considers each medical record individually, before suitably aggregating the relevance probabilities of the medical records to estimate the relevance of their associated patients. The existing aggregation techniques, such as the voting techniques, cannot take into account the coverage of the multiple inclusion criteria among the medical records of a particular patient. Indeed, without alteration, the medical record ranking stage of the two-stage model cannot examine the fact that a particular medical record may cover an inclusion criterion that the other medical records associated to the same patient do not cover. Thus, to highly rank the patients whose medical records cover the multiple inclusion criteria of the query, we need a mechanism to measure how well each of the inclusion criteria stated in the query is covered by different medical records of a particular patient. To achieve this, we introduce the notion of *criterion novelty*, which is the probability that a criterion is not well covered by the medical records that are also associated to the same patient. For instance, in the example of Figure 1(b), after considering the criterion novelty, the coverage likelihoods of the medical records d_3 and d_4 are boosted, since both of them cover at least one new criterion that is not covered by the other medical records associated to the same patient. Consequently, the patient p_2, which is associated to the medical records d_3 and d_4, is likely to be ranked higher than the patient p_1. To integrate the criterion novelty at the medical record ranking stage of the two-stage model, we adapt Equations (3) and (4) to estimate the relevance probability and the coverage likelihood of the medical record d_i, before using Equation (2), as follows:

$$F(p|q) \propto aggregate_{d_i \in R_p}\Bigg[(1-\lambda)\cdot P(d_i|q) + \lambda \cdot bel_{q_i \in Q}\left(P(d_i|q_i)\cdot \overline{P(R_p \setminus d_i|q_i)}\right)\Bigg] \qquad (8)$$

where $\overline{P(R_p \setminus d_i|q_i)}$ is the criterion novelty of q_i. To estimate the criterion novelty, we resort to techniques (e.g. [1, 4]) from web search result diversification that measure the novelty of a document within a set of web search results, based on the probability that the document covers an interpretation of information need that is not well covered by the other documents in the result set. In this work, we adapt an existing state-of-the-art technique for search result diversification (namely, xQuAD [26]) to estimate the criterion novelty within our approach, since it has been shown to be effective for diversifying web search results over several successive TREC tracks (e.g. [26, 27]). However, other techniques that explicitly model the novelty of a document within a set of search results (e.g. IA-Select [1]) can also be adapted.

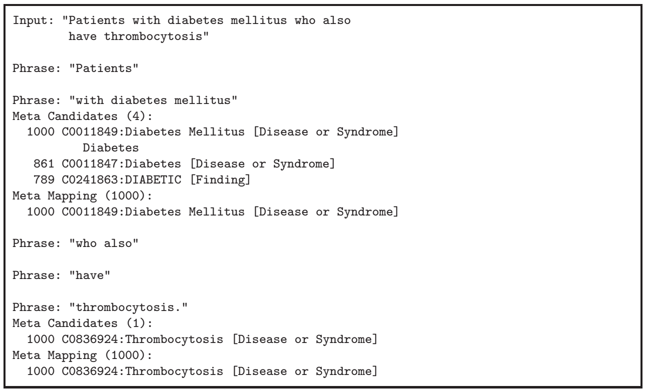
Specifically, for a given query q and a patient p, we adapt xQuAD [26] to iteratively rerank the medical records in R_p by maximising the mixture of the relevance probability and the coverage likelihood within Equation (8). By assuming the independence of the inclusion criteria within the query, $\overline{P(R_p \setminus d_i|q_i)}$ can be estimated as $\prod_{d_j \in {R_p}^*}(1 - P(d_j|q_i))$ when calculating the mixture probability:

$$F(p|q) \propto aggregate_{d_i \in R_p}\Bigg[(1-\lambda)\cdot P(d_i|q) + \lambda \cdot bel_{q_i \in Q}\left(P(d_i|q_i)\cdot \prod_{d_j \in {R_p}^*}(1 - P(d_j|q_i))\right)\Bigg] \qquad (9)$$

where $\prod_{d_j \in {R_p}^*}(1 - P(d_j|q_i))$ estimates the probability that the criterion q_i is not well covered by any medical records in ${R_p}^*$, the set of medical records that are ranked higher than d_j and that are also associated to the patient p.

5. INCLUSION CRITERIA EXTRACTION

As discussed in Section 4, to measure the coverage likelihood, we need to extract the set of the inclusion criteria Q (as used in Equations (6), (7), and (9)) from a query q and use the extracted inclusion criteria as sub-queries. Importantly, the quality of the extracted sub-queries can affect the effectiveness of our proposed approach. For example, if the sub-queries are not a good representative of all of the inclusion criteria stated in the query, our approach may not be able to highly rank the patients whose medical records cover all or at least most of the inclusion criteria expected by the searchers. When diversifying web search results, Santos et al. [26] use as sub-queries the recommended queries suggested by a commercial web search engine for the original query. It has been shown in the literature that the inclusion criteria that healthcare practitioners focus on when



```
Input: "Patients with diabetes mellitus who also
        have thrombocytosis"

Phrase: "Patients"

Phrase: "with diabetes mellitus"
Meta Candidates (4):
  1000 C0011849:Diabetes Mellitus [Disease or Syndrome]
          Diabetes
   861 C0011847:Diabetes [Disease or Syndrome]
   789 C0241863:DIABETIC [Finding]
Meta Mapping (1000):
  1000 C0011849:Diabetes Mellitus [Disease or Syndrome]

Phrase: "who also"

Phrase: "have"

Phrase: "thrombocytosis."
Meta Candidates (1):
  1000 C0836924:Thrombocytosis [Disease or Syndrome]
Meta Mapping (1000):
  1000 C0836924:Thrombocytosis [Disease or Syndrome]
```

Figure 2: Medical concepts extracted by the MetaMap tool from query 102: 'Patients with diabetes mellitus who also have thrombocytosis', as the query inclusion criteria.

Table 1: Statistics of the inclusion criteria extracted from the queries

# of Criteria	TREC 2011	TREC 2012
Average	4.32	3.36
Standard deviation	2.90	3.06
Minimum	1	1
Maximum	13	17

searching the medical records often relate to four types of the medical conditions of patients (namely, symptom, diagnostic test, diagnosis, and treatment) [16, 17]. Consequently, we follow [16] and deploy MetaMap [3] to extract the medical concepts related to these four types of medical conditions from the query. Importantly, we use the textual definitions of the extracted concepts as sub-queries. As MetaMap can generate a number of candidate concepts when mapping a given phrase, we select only those that are defined as 'Meta Mapping', which are the concepts identified with the highest confidence for a particular phrase in order to improve the accuracy and to avoid the redundancy of the extracted inclusion criteria. For example, in Figure 2, for the query "Patients with diabetes mellitus who also have thrombocytosis", we extract two inclusion criteria, 'Diabetes Mellitus' and 'Thrombocytosis', both of which are diagnosis concepts identified using MetaMap.

6. EXPERIMENTAL SETUP

In this section, we discuss our experimental setup to evaluate the effectiveness of our proposed approach for modelling the coverage of the inclusion criteria. In particular, Section 6.1 describes the used test collection, while Section 6.2 discusses the ranking approaches used in our experiment.

6.1 Test Collection

We evaluate our proposed approach using the test collection provided by the TREC 2011 and 2012 Medical Records track [30, 31], which aims to retrieve patient *visits* based on the relevance of their associated medical records towards a query. To avoid privacy issues, a visit, which contains a set of medical records associated to a patient during a visit to the hospital, is used as a representative of a patient [30, 31]. The test collection consists of 101,710 medical records, which are associated to 17,265 patient visits, and includes 34 and 47 queries from TREC 2011 and 2012, respectively. A query describes the compulsory medical conditions of the targeted patients. For example, query 149 is "Patients with delirium, hypertension, and tachycardia".

Table 1 shows the statistics of the inclusion criteria extracted from the queries, using our approach discussed in Section 5. For example, we find that on average we extract more inclusion criteria from the queries from TREC 2011 than those from TREC 2012 (i.e. 4.32 vs. 3.36 inclusion criteria per a query). The highest numbers of inclusion criteria extracted from a query are 13 (2 queries) and 17 (1 query), for TREC 2011 and 2012, respectively.

We evaluate the retrieval effectiveness using the track's official measures, namely bpref for TREC 2011 and infAP & infNDCG for TREC 2012 [30, 31]. In addition, we measure significant differences between the retrieval performance achieved by our approach and the existing patient ranking baselines using the paired t-test ($p < 0.05$).

6.2 Ranking Approaches

We conduct experiments using the Terrier retrieval platform[2] [22], applying Porter's English stemmer and removing stopwords. In addition, as handling negated language is a common, effective practice for this patient ranking task [7, 14, 15], we follow Limsopatham et al. [15] and tokenise term occurrences with a positive (e.g. patient has nausea) or negative context (e.g. patient has no nausea) differently, so that our search system can distinguish between terms with positive and negative contexts in both the medical records and the queries. For example, a term 'nausea' is tokenised as 'nausea' or 'n$nausea', if it occurs in a positive or negative context, respectively.

6.2.1 Patient Model Baselines

As representatives of the patient model, we follow King et al. [14] and concatenate the medical records associated to the same patient to create a patient document (discussed in Section 3). We estimate the relevance towards a given query of the patient documents using DPH from the Divergence from Randomness (DFR) framework [2], and BM25.

6.2.2 Two-Stage Model Baselines

Since the voting techniques have been shown to be effective for this TREC patient ranking task [18, 19], we use the two-stage model approaches based on the CombSUM, expCombMNZ, and expCombSUM voting techniques [20], as alternative baselines. Table 2 describes how each of the three used voting techniques estimates the relevance of a given patient. We follow Limsopatham et al. [19] and use DFR DPH to rank medical records at the first stage of the model, and limit the number of voting medical records to 5,000.

Another possible baseline is to use a boolean model to retrieve patients whose medical records contain all of the extracted inclusion criteria (i.e. the textual definitions of the medical concepts extracted from the query, as discussed in Section 5); however, we find that the boolean model is not effective, as it retrieves patients for only 6 out of the 34 queries of TREC 2011 and 16 out of the 47 queries of TREC 2012 (i.e. for some queries, the medical records of relevant patients may not contain all of the textual definitions of the extracted query criteria). Hence, it is excluded from the paper.

6.2.3 The Setup of Our Approach

We evaluate our proposed Inclusion Criteria Coverage (IC-Cover) approach within both the patient and the two-stage models by deploying the same ranking techniques as those of the corresponding baselines. Indeed, we apply our proposed approach within the patient model (denoted *IC-Cover-P*)

[2] `http://terrier.org`

Table 2: Voting techniques used in our experiments.

Voting technique	Description
CombSUM	sum of the relevance probabilities of retrieved records associated to a given patient.
expCombSUM	sum of the exponential of the relevance probabilities of the retrieved records associated to a given patient.
expCombMNZ	the product of expCombSUM and the number of retrieved records associated to a given patient.

as in Equation (6) and use either BM25 or DFR DPH to estimate the relevance of a patient. Next, for the two-stage model, we apply our proposed approach either within the patient ranking stage (denoted *IC-Cover-2P*), as in Equation (7), or within the medical record ranking stage (denoted *IC-Cover-2R*), as in Equation (9). Specifically, we deploy DFR DPH to estimate the relevance of the medical records, while using the CombSUM, expCombSUM, or expCombMNZ voting technique to estimate the relevance of a particular patient. In addition, similar to the baselines, we also limit the number of voting medical records to 5,000.

Mixture Parameter Setting: Initially, we uniformly set the mixture parameter λ in Equations (6), (7) and (9) to 0.5 for every query. This gives an equal emphasis on both the relevance probability and the coverage likelihood, of a particular patient. Later, in Sections 7.2 and 8, we analyse the effect of λ and discuss how to automatically learn a suitable λ value for a given query from training data, respectively.

Belief Combination Function: We evaluate our proposed approach using three different belief combination functions (namely, *AND*, *OR*, and *SUM*), which have been shown to be effective for different search tasks [21, 24, 29]. Specifically, the belief combination functions AND, OR, and SUM combine the coverage likelihood as follows [21, 29]:

$$bel^{AND}_{q_i \in Q}\left(P(p|q_i)\right) = \prod_{q_i \in Q} P(p|q_i) \tag{10}$$

$$bel^{OR}_{q_i \in Q}\left(P(p|q_i)\right) = 1 - \prod_{q_i \in Q} (1 - P(p|q_i)) \tag{11}$$

$$bel^{SUM}_{q_i \in Q}\left(P(p|q_i)\right) = \sum_{q_i \in Q} \frac{P(p|q_i)}{|Q|} \tag{12}$$

where $|Q|$ is the number of inclusion criteria in the set $Q = \{q_1, q_2, ..., q_n\}$.

7. EXPERIMENTAL RESULTS

This section presents the experimental results obtained using our proposed approach. Specifically, Section 7.1 examines the effectiveness of our proposed approach compared to the baselines, when our mixture parameter λ is set to equally weight the relevance probability and the coverage likelihood. Section 7.2 discusses the robustness of our proposed approach, as we vary the parameter λ. Then, in Section 8, we introduce and evaluate a technique to automatically set λ using training data. We further discuss and analyse the retrieval performance of our approach in Section 9.

7.1 Uniformly Setting the Parameter λ

We first compare the retrieval performance of our proposed approach, when the mixture parameter λ is uniformly set to 0.5 to balance the importance of the relevance probability and the coverage likelihood, with existing effective baselines, including the patient and the two-stage models, discussed in Section 6.2. As we will show later in Sections 7.2 and 8.3, this will prove to be a very effective parameter value for λ. Table 3 compares the retrieval effectiveness of our approach with the baselines, in terms of bpref for TREC 2011 and infNDCG & infAP for TREC 2012. Moreover, the number of queries that our proposed approach improves or harms compared to the corresponding baselines is also reported. The remainder of the queries are unaffected. For ease of notation, in Table 3, the used belief combination function along our proposed approach is indicated between parentheses. For instance, for IC-Cover-P(AND), we apply our approach within a patient model and use the belief combination function AND to combine the probabilities that the medical records of a patient cover the multiple inclusion criteria extracted from the query.

From Table 3, we observe that applying our proposed approach within the medical record ranking stage of the two-stage model (i.e. IC-Cover-2R) is effective. For example, when using the CombSUM voting technique, IC-Cover-2R(OR) and IC-Cover-2R(SUM), which use the belief combination functions OR and SUM, respectively, significantly (paired t-test, $p < 0.05$) outperform the CombSUM-based baseline, for every reported measure across both TREC 2011 and 2012. In addition, when using the expCombSUM voting technique, IC-Cover-2R(OR) and IC-Cover-2R(SUM) improve the retrieval performances over the expCombSUM-based baseline by up to 4.77%. The performance improvement is statistically significant (paired t-test, $p < 0.05$) for TREC 2011. On the other hand, we find that the belief combination function AND is not effective for our proposed approach. We also observe that the expCombMNZ voting technique is not effective when used in conjunction with our approach. This might be due to the fact that the expCombMNZ voting technique also considers the number of retrieved medical records associated to a particular patient when estimating the relevance of that patient, which could outweigh the coverage likelihood within our approach.

Next, when applied within the patient model (i.e. IC-Cover-P), our proposed approach markedly improves the retrieval performance by up to 9.13% compared to the corresponding baselines. For example, when using BM25, the retrieval performances of IC-Cover-P(SUM) are bpref 0.5315, infNDCG 0.4286 and infAP 0.1959, while the retrieval performances of the BM25 baseline are bpref 0.4870, infNDCG 0.4080, and infAP 0.1922. Even though the performance improvements are not statistically significant, our approach improves the retrieval performances for about half of the queries. For example, when using BM25, IC-Cover-P(SUM) improves the retrieval performance over the BM25 baseline, in terms of bpref, for 19 out of 34 queries from TREC 2011, while for TREC 2012 it benefits 24 and 23 out of 47 queries, in terms of infNDCG and infAP, respectively.

When applied within the patient ranking stage of the two-stage model, we observe that our approach (i.e. IC-Cover-2P) is not effective. For example, when using the CombSUM voting technique, our IC-Cover-2P(OR) performs only comparably to the CombSUM-based baseline (e.g. infNDCG 0.3361 vs. 0.3304). This is likely because when applied within the patient ranking stage of the two-stage model, the used voting techniques tend to give very high relevance probabilities to the patients who are relevant to a single particular criterion. Hence, when using a belief function (i.e. AND, OR, or SUM) to combine the likelihoods that the medical records of the patients cover the multiple inclusion criteria, the highly ranked patients could be relevant to only one or few criteria.

Table 3: Comparison of the retrieval performances of our proposed approach (λ=0.5) against the baselines on the TREC Medical Records Track 2011 and 2012. Statistical significance (paired t-test, $p < 0.05$) over the corresponding baseline is denoted $^\bullet$. The column denoted △ (resp. ▽) shows the number of queries improved (resp. harmed) in relation to the corresponding baseline.

Approach	TREC 2011 (34 queries)			TREC 2012 (47 queries)					
	bpref	△	▽	infNDCG	△	▽	infAP	△	▽
Patient Models									
DPH	0.4968			0.4392			0.1845		
+IC-Cover-P(AND)	0.4985	1	1	0.4381	4	5	0.1831	4	5
+IC-Cover-P(OR)	0.5165	19	13	0.4507	23	18	0.1909	22	19
+IC-Cover-P(SUM)	0.5165	19	13	0.4507	23	18	0.1909	22	19
BM25	0.4870			0.4080			0.1922		
+IC-Cover-P(AND)	0.4879	2	1	0.4092	5	4	0.1916	5	3
+IC-Cover-P(OR)	0.5227	19	13	0.4269	24	13	0.1958	23	14
+IC-Cover-P(SUM)	**0.5315**	19	13	0.4286	24	13	**0.1959**	23	14
Two-Stage Models									
CombSUM	0.3771			0.3304			0.0969		
+IC-Cover-2P(AND)	0.3769	2	1	0.3389	8	1	0.1008	8	1
+IC-Cover-2P(OR)	0.3734	13	17	0.3361	18	22	0.0983	16	23
+IC-Cover-2P(SUM)	0.3735	13	17	0.3361	18	22	0.0983	16	23
+IC-Cover-2R(AND)	0.3731	5	21	0.3342	16	28	0.1005	13	28
+IC-Cover-2R(OR)	0.3859$^\bullet$	20	11	0.3496$^\bullet$	28	16	0.1098$^\bullet$	30	13
+IC-Cover-2R(SUM)	0.3859$^\bullet$	20	11	0.3496$^\bullet$	28	16	0.1098$^\bullet$	30	13
expCombMNZ	0.5007			0.4506			0.1822		
+IC-Cover-2P(AND)	0.4989	3	2	0.4484	4	6	0.1775	5	5
+IC-Cover-2P(OR)	0.3582	9	23	0.3445	6	36	0.1218	10	32
+IC-Cover-2P(SUM)	0.3582	9	23	0.3445	6	36	0.1218	10	32
+IC-Cover-2R(AND)	0.4602	13	21	0.4264	17	28	0.1620	17	28
+IC-Cover-2R(OR)	0.5015	17	12	0.4469	21	23	0.1819	20	24
+IC-Cover-2R(SUM)	0.5015	17	12	0.4469	21	23	0.1819	20	24
expCombSUM	0.5055			0.4355			0.1833		
+IC-Cover-2P(AND)	0.5100	3	4	0.4382	12	6	0.1738	10	6
+IC-Cover-2P(OR)	0.3738	12	21	0.3215	11	33	0.1129	11	32
+IC-Cover-2P(SUM)	0.3738	12	21	0.3215	11	33	0.1129	11	33
+IC-Cover-2R(AND)	0.5080	18	13	0.4453	23	21	0.1785	22	22
+IC-Cover-2R(OR)	0.5296$^\bullet$	21	9	**0.4515**	22	22	0.1913	23	21
+IC-Cover-2R(SUM)	0.5296$^\bullet$	21	9	**0.4515**	22	22	0.1913	23	21

Overall, we conclude that our approach to model the relevance towards multiple inclusion criteria applied either within the patient model or within the medical record ranking stage of the two-stage model is effective for the patient ranking task. In addition, we find that the belief combination functions SUM and OR are effective for combining the likelihoods that the medical records of a patient cover each of the inclusion criteria stated in the query.

7.2 Model Robustness

This section investigates the robustness of our proposed approach by varying the mixture parameter λ that weights the importance of the relevance probability and the coverage likelihood. To analyse the impact of the parameter λ within our approach, we experiment setting λ within a range of values between 0 and 1, with an interval of 0.1. When $\lambda = 0$, our proposed approach considers only the relevance probability towards the query, while when $\lambda = 1$ our approach takes into account only the coverage likelihood. Due to space limitation, in Figure 3, we report only the experiment on TREC 2011. We found that the experimental results on TREC 2012 follow the same pattern. In addition, for readability purposes, while we only show the retrieval performances of our approach using the most effective belief combination function (namely SUM as shown in Table 3), the results with the belief combination function OR are consistently similar, although slightly less effective in magnitude.

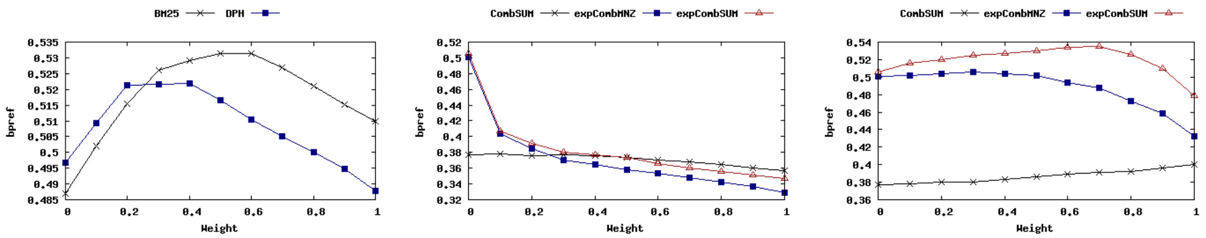
From Figure 3(a), we observe that when applied within the patient model (using either BM25 or DPH), our approach (i.e. IC-Cover-P) performs better than the baseline (i.e. $\lambda = 0$), when $0.1 \leq \lambda \leq 0.8$. Hence, our approach is robust when applied within the patient model, as it is more effective than the baseline for a very wide range of λ values (in fact for all λ values when using BM25). In addition, the most effective performance is achieved when λ is set to 0.2 and 0.6 when using BM25 and DPH, respectively. Next, Figure 3(b) shows the retrieval performances of our approach when applied within the patient ranking stage of the two-stage model (i.e. IC-Cover-2P). We observe that our approach is not effective, which is in line with the observation in Section 7.1. Meanwhile, in Figure 3(c), when applied within the medical record ranking stage (i.e. IC-Cover-2R), our proposed approach markedly outperforms the baseline ($\lambda = 0$) for a wide range of λ values. For CombSUM, when $0 < \lambda \leq 1$, our approach outperforms the baseline, especially for λ values closer to 1. For expCombSUM, our approach performs better than the baseline when $0.1 \leq \lambda \leq 0.9$, while the most effective performance is obtained when $\lambda = 0.7$. On the other hand, Figure 3(c) shows that applying expCombMNZ in conjunction with our approach is not effective, which supports the finding discussed in Section 7.1.

To summarise, we find that our proposed approach is effective and robust when applied within the patient model or when applied within the medical record ranking stage of the two-stage model. This shows the importance of promoting patients relevant to multiple query criteria for the patient ranking task. Specifically, as shown in Figure 3, our approach is effective when setting λ uniformly for a wide range of values (particularly $0.2 \leq \lambda \leq 0.7$). In the next section, we show how to automatically set the parameter λ for each query, using training data.

8. MIXTURE PARAMETER ESTIMATION

This section discusses an automatic technique to set the mixture parameter λ. We hypothesise that queries benefit from different levels of emphasis on the relevance probability and the coverage likelihood. For example, a query that contains several inclusion criteria may benefit from more emphasis on the coverage likelihood. Within the xQuAD framework, Santos et al. [27] also suggested to selectively set the level of diversification based on the ambiguity of the query.

In this work, we deploy Gradient Boosted Regression Trees (GBRT) [28] to learn the parameter λ from a set of training queries, since GBRT has been shown to be effective for several regression tasks (e.g. [9, 18, 28]). We use the jforest



(a) IC-Cover-P (b) IC-Cover-2P (c) IC-Cover-2R

Figure 3: The retrieval performances of our proposed approach when applied within the patient model (IC-Cover-P), within the patient ranking stage (IC-Cover-2P) and within the medical record ranking stage (IC-Cover-2R) of the two-stage model, in terms of bpref for TREC 2011, as we vary the mixture parameter λ between 0 and 1.

Table 4: List of the used features.

Query Performance Predictors		Inclusion Criteria Similarity	
Clarity Score [6]	AvIDF [5]	WUPALMER [10]	PATH [10]
Query Scope [11]	EnIDF [5]	INTRINSIC_PATH [10]	RADA [10]
AvICTF [5]	γ_1 [11]	INTRINSIC_LIN [10]	LCH [10]
AvPMI [5]	γ_2 [11]	INTRINSIC_LCH [10]	SOKAL [10]
MAXCQ [32]	SCQ [32]	INTRINSIC_RADA [10]	LIN [10]
# of medical concepts [18]	NSCQ [32]	JACCARD [10]	

package implementation [9][3] with the default settings. However, any regression technique can be deployed. To train a regression model, the root-mean-square error (RMSE) is used as a loss function when learning λ.

8.1 Estimating an Effective Mixture Parameter

To estimate an effective mixture parameter λ, we identify the λ that attains the best retrieval performance in terms of a particular retrieval measure (e.g. bpref or infNDCG) for each training query. In particular, for a given query, we sweep the λ between 0 and 1 (with an interval of 0.1) to find the best setting of λ. Then, the set of the identified λ from the training queries are used as the labelled data to train the regression model for choosing λ for an unseen query.

8.2 Learning Features

Next, we define the features used for choosing the effective parameter λ for an unseen query. An effective feature should indicate the level of emphasis on the relevance probability and the coverage likelihood for each query. In this work, we use 23 features, which measure the predicted difficulty of the query. A query with multiple inclusion criteria tends to be complex and long; hence, it can be assumed to be difficult. If the query is difficult, then it might be beneficial to focus on the coverage likelihood. Table 4 lists the two groups of features used in this experiments. The first group of features are the 12 query performance predictors [5, 6, 32] computed on the original query, which are well-known for measuring the difficulty of a query. The second group of features are the 11 semantic similarities [10], which can estimate the similarity between the inclusion criteria extracted from the original query. The more dissimilar the inclusion criteria in the query, the more difficult the query is likely to be, since it may be difficult to find a patient with such unrelated conditions specified by the inclusion criteria. We use YTEX[4] to calculate 11 recent semantic similarity measures [10] and average the similarity scores among every pair of the medical concepts extracted as inclusion criteria by the approach described in Section 5.

8.3 Experiment with the Learned λ

Due to the difference between the used methods for the relevance assessment in TREC 2011 and 2012 [30, 31], we deploy a 5-fold cross-validation on the 34 and 47 queries of TREC 2011 and 2012, respectively, where each fold has completely separated training and test query sets.

We compare the retrieval performance of our approach when using the cross-validation setting to learn the mixture parameter λ (i.e. 5-fold) with uniform settings of λ, namely $\lambda = 0$ (i.e. the focus is only on the relevance probability towards the query), $\lambda = 1$ (i.e. the focus is only on the coverage likelihood), and $\lambda = 0.5$ (i.e. equally weight the importance of both the relevance probability and the coverage likelihood). In addition, the best possible retrieval performance, when the mixture parameter λ is optimally set for every query (i.e. an oracle), is also reported. For space reasons, we show only the experiments with our proposed approach when applied within the patient model (i.e. BM25) and using the belief combination function SUM, since it is the most effective approach as shown in Figure 3(a); however, we were also able to find effective λ values when using the learned technique on the other variants of our proposed approach.

Table 5 reports the retrieval performance of our proposed approach when the mixture parameter is set in various manners. From Table 5, we observe that our approach with the learned λ (5-fold) is effective, as it outperforms all of the uniform setting baselines (i.e. when λ is set to 0, 1, or 0.5). In particular, for TREC 2011, our cross-validation setting improves the retrieval performance over the baseline where $\lambda = 0$ by up to 9.8% (bpref 0.5346 vs. 0.4870). Indeed, it improves the retrieval performance for the majority of the queries (i.e. 18 of 34 queries). For TREC 2012, in terms of infNDCG, our cross-validation setting significantly outperforms the settings where $\lambda = 0$ and $\lambda = 1$ (paired t-test, $p < 0.05$). The performance improvements are up to 7.5% and 16.5%, respectively. Specifically, our cross-validation setting outperforms the baseline that does not take into account the coverage likelihood (i.e. $\lambda = 0$) for 24 out of 47 queries. Meanwhile, in terms of infAP, our cross-validation setting (infAP 0.2066) significantly outperforms the baseline where $\lambda = 1$ (infAP 0.1585). However, even though the cross-validation setting outperforms the setting where $\lambda = 0.5$ for all of the reported measures, the improvements are not significant. We find that setting $\lambda = 0.5$ is an effective baseline, since it can improve the retrieval performance for most of the queries improved by the oracle. For instance, for TREC 2012, the setting where $\lambda = 0.5$ improves the retrieval performance, in terms of infNDCG, over the baseline where $\lambda = 0$ for 24 out of 47 queries, while the oracle setting improves the retrieval performance for 30 out of 47 queries.

In addition, our approach, when λ is set either to 0.5 (i.e. $\lambda = 0.5$) or to a learned value (i.e. 5-fold), markedly outperforms the median of 127 and 88 participating systems at TREC 2011 and 2012, respectively. In particular, our approach performs comparably to the best system of

[3] http://code.google.com/p/jforests/

[4] http://code.google.com/p/ytex/wiki/SemanticSim_V06

Table 5: Comparison of the retrieval performances using various λ values within our proposed approach. Statistical significances (paired t-test, $p < 0.05$) over the settings when $\lambda = 0$, $\lambda = 1$, $\lambda = 0.5$, and when using a learned technique (5-fold) are denoted $^{\oplus}$, $^{\ominus}$, $^{\odot}$, and $^{\otimes}$, respectively. The column denoted $\triangle$ (resp. $\triangledown$) shows the number of queries improved (resp. harmed) in relation to the baseline where $\lambda = 0$.

Approach	TREC 2011 (34 queries)			TREC 2012 (47 queries)					
	bpref	$\triangle$	$\triangledown$	infNDCG	$\triangle$	$\triangledown$	infAP	$\triangle$	$\triangledown$
IC-Cover-P(SUM)									
$+\lambda = 0$	0.4870			0.4080			0.1922		
$+\lambda = 1$	0.5098	17	15	0.3764	20	21	0.1585	15	24
$+\lambda = 0.5$	0.5315	19	13	$0.4286^{\ominus}$	24	13	$0.1959^{\ominus}$	23	14
+5-fold	0.5346	18	11	$0.4384^{\oplus,\ominus}$	24	11	$0.2066^{\ominus}$	23	14
+oracle	$0.5872^{\oplus,\ominus,\odot,\otimes}$	24	0	$0.4637^{\oplus,\ominus,\odot,\otimes}$	30	0	$0.2208^{\oplus,\ominus,\odot,\otimes}$	29	0
TREC Median	0.4219	NA	NA	0.4243	NA	NA	0.1695	NA	NA
TREC Best	0.5520	NA	NA	0.5780	NA	NA	0.2860	NA	NA

TREC 2011 (bpref 0.5346 vs 0.5520). This is despite the fact that we only focus on investigating an effective modelling of the relevance towards inclusion criteria, while most of the TREC participating systems deployed several sophisticated techniques (e.g. information extraction, query expansion) to deal with other unique characteristics (e.g. the use of acronyms) of the medical records in the patient ranking task. For TREC 2012, we observe that the best TREC system performs better than our approach. This system applied a query expansion technique, which is typically used by effective systems at TREC. However, we do not use query expansion in our current approach. We leave for future work the extension of our approach to use query expansion to improve the representation of each of the query inclusion criteria.

Next, we discuss the retrieval performances assuming we can effectively set the λ for all of the queries (i.e. oracle). We observe that with the oracle setting, our approach further improves the retrieval performances markedly. It significantly (paired t-test, $p < 0.05$) outperforms all other settings in Table 5. In addition, we find that it improves the retrieval performances over the corresponding baselines where the coverage likelihood is not taken into account (i.e. $\lambda = 0$) for the majority of the queries (i.e. 24 out of 34 queries for TREC 2011, 30 out of 47 queries for infNDCG TREC 2012, and 29 out of 47 queries for infAP TREC 2012). These results show the potential of our proposed approach, and suggest that there is room for further improvements.

9. ANALYSIS AND DISCUSSION

In this section, we analyse the retrieval performance of our approach for each query, when using the cross-validation setting discussed in Section 8.3. Table 6 shows, for various numbers of extracted inclusion criteria, the numbers and percentages of the queries impacted (either positively or negatively) by our proposed approach. The impacted performances are measured based on bpref and infNDCG because they are the primary measures for TREC 2011 and TREC 2012, respectively [31]. We divide the 81 queries from TREC 2011 (34 queries) and 2012 (47 queries) into 4 groups, according to the number of the extracted inclusion criteria in the queries, and report the percentages of queries for each group. The sizes of the groups vary from 14 to 30 (average of 20.25) queries. From the cross-validation setting, we observe that our approach is most likely to benefit (63%) the queries with 3 inclusion criteria, followed by the queries having a number of inclusion criteria between 4 and 17 (57%). This is in line with the oracle setting where the queries for which the number of extracted criteria is at least 3 are most likely (more than 70%) to benefit. On the other hand, for the queries with 1 or 2 inclusion criteria, our approach is less likely to be beneficial (e.g. 43% and 39% for the queries with 1 and 2 inclusion criteria, respectively), which is intuitive. Indeed, as our approach aims to promote the relevance towards multiple inclusion criteria, queries that contain only very few criteria may not benefit much from our approach (e.g. queries 109, 143, 147 and 154). However, we also observe improvements for queries with one inclusion criterion, since our approach enables the retrieval system to focus on the inclusion criterion instead of the non-important terms in the queries. In contrast, our proposed approach tends to be effective for long and complex queries (i.e. the queries with at least 3 inclusion criteria), such as queries 111, 113, 121 and 176. For example, for query 121:'patients with CAD who presented to the Emergency Department with Acute Coronary Syndrome and were given Plavix', our approach can improve the retrieval performance to bpref 0.4088, while the performance of the patient model baseline is bpref 0.1869. This is because the patient model, which uses a patient ranking model to estimate the relevance probability of the patients, tends to focus on the occurrences of informative terms (e.g. Plavix) within the medical records. However, the relevant patients are also required to be relevant to the other conditions, indicated by occurrences of the other query terms, such as acute, coronary, and syndrome. Meanwhile, our proposed approach is effective since it promotes the patients who are relevant to multiple inclusion criteria (e.g. 'CAD', 'Acute Coronary Syndrome', and 'Plavix').

Table 6: Analysis of our approach w.r.t. the number of inclusion criteria extracted from the queries. The numbers between the parentheses indicate the percentage compared to the total number of queries.

# of Criteria	Cross-validation			Oracle[5]	
	#Benefiting	#Harmed	#Stable	#Benefiting	#Stable
1 (14 queries)	6 (43%)	3 (21%)	5 (36%)	8 (57%)	6 (43%)
2 (18 queries)	7 (39%)	5 (28%)	6 (33%)	10 (56%)	8 (44%)
3 (19 queries)	12 (63%)	6 (32%)	1 (5%)	15 (79%)	4 (21%)
4-17 (30 queries)	17 (57%)	12 (40%)	1 (3%)	21 (70%)	9 (30%)

On the other hand, when looking at the harmed queries, we observe that the queries with 3 and 4-17 inclusion criteria are also more likely to be harmed by our proposed approach than the queries with 1 or 2 inclusion criteria. Because of the limited number of queries (i.e. we use a 5-fold cross-validation on two separate sets of only 34 and 47 queries when learning the regression model), we find that the learned model could not always generalise, as it tends to favour the coverage likelihood for the queries with several inclusion criteria, while focusing on the relevance probability for the queries with a very few inclusion criteria.

Moreover, when investigating the effectiveness of each query, we find that some queries do not benefit from our approach, because the used MetaMap tool could not effectively extract the inclusion criteria (e.g. queries 104, 107, 146 and

[5] Note that there is no harmed query for oracle setting because the λ is set to 0, which is the baseline, if the coverage likelihood could not improve the retrieval performance.

149). For example, instead of only 2, 13 inclusion criteria are incorrectly extracted from the query 104: 'patients diagnosed with localized prostate cancer and treated with robotic surgery'. Consequently, this misleads the estimation of the coverage likelihood. We leave for future work the investigation of a more effective inclusion criteria extraction technique.

10. CONCLUSIONS

We have discussed the importance of ranking highly patients whose medical records cover the multiple inclusion criteria stated in the query when retrieving patients for clinical studies. To achieve this, we have proposed a novel approach for modelling the mixture of the relevance probability towards the query and the likelihood that the medical records of a patient are relevant to the multiple inclusion criteria occurring in the query (i.e. the coverage likelihood). We measured the coverage likelihood using the relevance probability towards each of the inclusion criteria, represented as medical concepts extracted from the query using an existing medical resource. We showed how our approach can be applied within the state-of-the-art patient ranking models (i.e. the patient and the two-stage models). When applied within the patient model or at the medical record ranking stage of the two-stage model, our approach significantly improved the retrieval performance over both of the state-of-the-art patient ranking models. Moreover, we showed that our proposed approach was effective, either by weighting equally the importance of the relevance probability and the coverage likelihood or by deploying a regression technique to learn an effective setting of the mixture parameter.

Through the experiments, we showed that our proposed approach was robust, as it improved the retrieval performances for a wide range of the mixture parameter's values, especially when the parameter is set between 0.2 and 0.7. When analysing the retrieval performances for each query, we found that our approach tended to particularly improve the retrieval performances for queries from which at least 3 inclusion criteria could be extracted.

11. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying Search Results. In *Proc. of WSDM*, 2009.

[2] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track. In *Proc. of TREC*, 2007.

[3] A. R. Aronson and F. Lang. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*, 17(3), 2010.

[4] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of SIGIR*, 1998.

[5] D. Carmel and E. Yom-Tov. Estimating the Query Difficulty for Information Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1), 2010.

[6] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting Query Performance. In *Proc. of SIGIR*, 2002.

[7] D. Demner-Fushman, S. Abhyankar, A. Jimeno-Yepes, R. Loane, B. Rance, F. Lang, N. Ide, E. Apostolova, and A. R. Aronson. A Knowledge-Based Approach to Medical Records Retrieval. In *Proc. of TREC*, 2011.

[8] T. Edinger, A. M. Cohen, S. Bedrick, K. Ambert, and W. Hersh. Barriers to Retrieving Patient Information from Electronic Health Record Data: Failure Analysis from the TREC Medical Records Track. In *Proc. of AMIA*, 2012.

[9] Y. Ganjisaffar, R. Caruana, and C. V. Lopes. Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proc. of SIGIR*, 2011.

[10] V. N. Garla and C. Brandt. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinformatics*, 13, 2012.

[11] B. He and I. Ounis. Query performance prediction. *Inf. Syst.*, 31(7), 2006.

[12] W. Hersh, M. Weiner, P. J. Embi, J. R. Logan, P. R. O. Payne, E. V. Bernstam, H. P. Lehmann, G. Hripcsak, T. H. Hartzog, J. Cimino, and J. H. Saltz. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care.* 51, 2013.

[13] D. Hiemstra. Using Language Models for Information Retrieval. *PhD thesis.* University of Twente, 2001.

[14] B. King, L. Wang, I. Provalov, and J. Zhou. Cengage Learning at TREC 2011 Medical Track. In *Proc. of TREC*, 2011.

[15] N. Limsopatham, C. Macdonald, R. McCreadie, and I. Ounis. Exploiting Term Dependence while Handling Negation in Medical Search. In *Proc. of SIGIR*, 2012.

[16] N. Limsopatham, C. Macdonald, and I. Ounis. A Task-Specific Query and Document Representation for Medical Records Search. In *Proc. of ECIR*, 2013.

[17] N. Limsopatham, C. Macdonald, and I. Ounis. Inferring Conceptual Relationships to Improve Medical Records Search. In *Proc. of OAIR* 2013.

[18] N. Limsopatham, C. Macdonald, and I. Ounis. Learning to Selectively Rank Patients' Medical History. In *Proc. of CIKM*, 2013.

[19] N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, and M. Bouamrane. University of Glasgow at Medical Records track 2011: Experiments with Terrier. In *Proc. of TREC*, 2011.

[20] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *Proc. of CIKM*, 2006.

[21] D. Metzler and W. B. Croft. Combing the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40, 2004.

[22] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proc. of OSIR at SIGIR*, 2006.

[23] Y. Qi and P. F. Laquerre. Retrieving Medical Records with "sennamed": NEC Labs America at TREC 2012 Medical Record Track. In *Proc. of TREC*, 2012.

[24] B. A. N. Ribeiro and R. Muntz. A Belief Network Model for IR. In *Proc. of SIGIR*, 1996.

[25] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. of TREC*, 1994.

[26] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proc. of WWW*, 2010.

[27] R. L. T. Santos, C. Macdonald, and I. Ounis. Selectively diversifying Web search results. In *Proc. of CIKM*, 2010.

[28] S. Tyree, K. Q. Weinberger, K. Agrawal, and J. Paykin. Parallel boosted regression trees for web search ranking. In *Proc. of WWW*, 2011.

[29] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9(3), 1991.

[30] E. M. Voorhees and W. Hersh. Overview of the TREC 2012 Medical Records Track. In *Proc. of TREC*, 2012.

[31] E. M. Voorhees and R. Tong. Overview of the TREC 2011 Medical Records Track. In *Proc. of TREC*, 2011.

[32] Y. Zhao, F. Scholer, and Y. Tsegay. Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In *Proc. of ECIR*, 2008.

[33] W. Zheng and H. Fang. Query Aspect Based Term Weighting Regularization in Information Retrieval. In *Proc. of ECIR*, 2010.

[34] D. Zhu and B. Carterette. Exploring Evidence Aggregation Methods and External Expansion Sources for Medical Record Search. In *Proc. of TREC*, 2012.