Inf Retrieval J (2018) 21:507–540
https://doi.org/10.1007/s10791-018-9331-4

An analysis of evaluation campaigns in ad-hoc medical information retrieval: CLEF eHealth 2013 and 2014

Lorraine Goeuriot[1] · Gareth J. F. Jones[2] · Liadh Kelly[3] · Johannes Leveling[2] · Mihai Lupu[4] · Joao Palotti[4] · Guido Zuccon[5]

Received: 25 May 2017 / Accepted: 16 April 2018 / Published online: 3 May 2018

Abstract Since its inception in 2013, one of the key contributions of the CLEF eHealth evaluation campaign has been the organization of an ad-hoc information retrieval (IR) benchmarking task. This IR task evaluates systems intended to support laypeople searching for and understanding health information. Each year the task provides registered participants with standard IR test collections consisting of a document collection and topic set. Participants then return retrieval results obtained by their IR systems for each query, which

Johannes Leveling has since moved on to Teckro, Ireland.

Authors listed alphabetically.

✉ Liadh Kelly
liadh.kelly@mu.ie

Lorraine Goeuriot
lorraine.goeuriot@imag.fr

Gareth J. F. Jones
gareth.jones@dcu.ie

Johannes Leveling
johannes.leveling@dcu.ie

Mihai Lupu
lupu@ifs.tuwien.ac.at

Joao Palotti
palotti@ifs.tuwien.ac.at

Guido Zuccon
g.zuccon@qut.edu.au

[1] LIG, Université Grenoble Alpes, Grenoble, France

[2] Dublin City University, Dublin, Ireland

[3] Maynooth University, Maynooth, Ireland

[4] TU Wien, Vienna, Austria

[5] Queensland University of Technology, Brisbane, Australia

are assessed using a pooling procedure. In this article we focus on CLEF eHealth 2013 and 2014s retrieval task, which saw topics created based on patients' information needs associated with their medical discharge summaries. We overview the task and datasets created, and the results obtained by participating teams over these two years. We then provide a detailed comparative analysis of the results, and conduct an evaluation of the datasets in the light of these results. This twofold study of the evaluation campaign teaches us about technical aspects of medical IR, such as the effectiveness of query expansion; the quality and characteristics of CLEF eHealth IR datasets, such as their reliability; and how to run an IR evaluation campaign in the medical domain.

Keywords eHealth · Evaluation · Benchmarking

1 Introduction

The increasing availability of online medical information in recent years has created great interest in the use of these resources to address medical information needs. Online medical information originates from a wide range of sources including professional medical agencies, publishers, informed medical interest groups, commercial organizations, the general public, and less well informed or unreliable sources. Much of this information is freely available on the World Wide Web using general purpose search engines, and is searched for by a wide variety of users ranging from members of the general public with differing levels of knowledge of medical issues to medical professionals such as general practitioners. An important issue when searching these information resources is receiving accurate information relevant to the information need and at a technical level appropriate to the searcher.

The CLEF eHealth benchmark activities,[1] held as part of the Conference and Labs of the Evaluation Forum (CLEF)[2] since 2013, creates annual shared challenges for the evaluation and advancement of medical information extraction, management and retrieval related research. This article analyzes the outcomes of the 2013 and 2014 CLEF eHealth information retrieval (IR) challenges, which provided a platform for the evaluation of search engines to identify items relevant to user information needs as stated in search requests (referred to here as search *topics*). The focus of these tasks was the evaluation of the effectiveness with which search engines could retrieve relevant documents, from an archive collected from the World Wide Web, in response to a set of patient search requests. The tasks provided, to registered task participants, an IR test collection consisting of the document collection harvested from the World Wide Web and the topic set. The registered task participants then returned retrieval results obtained by their IR systems for each query which were then assessed for relevance. Participants detailed descriptions of the IR systems used to create their results in written reports (Working Notes Papers), and then met at the CLEF 2013 and CLEF 2014 conferences to report and discuss their work. While organizers published overview papers in 2013 (Goeuriot et al. 2013a) and 2014 (Goeuriot et al. 2014c), no deeper analysis of these results has so far been reported. This article overviews the creation of these test collections, and summarizes the results obtained by the participants. It then provides, for the first time, a detailed comparative analysis of the results seeking to identify common features of success and failure in the participants' work. Reflections on the general outcomes of the task in terms of experimental design and scientific findings, contributing to improved domain-specific IR benchmark design, are also provided.

[1] http://clef-ehealth.org/.

[2] http://www.clef-initiative.eu/.

The article is organized as follows: we first provide an overview of relevant existing work in the benchmarking of medical IR and analysis of IR benchmark results; we then describe the 2013 and 2014 CLEF eHealth IR challenges; provide a summary of the results obtained by the challenge participants; and detail a comparative analysis of participants' results and the techniques used to produce them. We conclude with the lessons learned from the task results and a summary of the findings.

2 Related work

2.1 Health-related evaluation campaigns

Medical IR evaluation challenges supporting individuals' retrieval needs have historically focused on needs of medical professionals, ignoring the different needs and perspective of laypeople when searching for medical information. Over the last 20 years a large number of evaluation tasks have focused on a wide variety of aspects of the needs of medical professionals and the differing tools needed to support them in their work. OHSUMED, published in 1994, was the first such collection (Hersh et al. 1994), and has subsequently been used in the TREC 2000 Filtering Track and for individual research on health IR (Claveau 2012; Koopman et al. 2012). The TREC Genomics Track (2003–2007) targeted biologists' needs (Roberts et al. 2009). The ImageCLEFmed Track (2003–2013) focused on biomedical image retrieval (Kalpathy-Cramer et al. 2011; Müller et al. 2016). The TREC Medical Records Track (2011–2012) (Voorhees and Tong 2011) focused on patient cohort identification. The TREC clinical decision support[3] (CDS) track (Simpson et al. 2014; Roberts et al. 2015), organized for the first time in 2014, focused on patient care. The TREC clinical decision support track collection has also been recently used to evaluate systems for the selection of cohorts to recruit for clinical trial (Koopman and Zuccon 2016).

Most of these evaluation campaigns focus only on medical experts and information needs. Previous research has shown that exposing people with no or scarce medical knowledge to complex medical language may lead to erroneous self-diagnosis and self-treatment and that access to medical information on the web can lead to the escalation of concerns about common symptoms (e.g., cyberchondria) (White and Horvitz 2008; Benigeri and Pluye 2003). Research has also shown that current commercial search engines are still far from being effective in answering such unclear and underspecified queries (Zuccon et al. 2015b).

The CLEF eHealth IR challenges represent the first, and to-date only, evaluation campaigns focusing on evaluating and advancing search engine technologies aimed to support laypeople searching for health information and advice on the web. In this article we analyse the findings and contributions of the 2013 and 2014 labs. The lab has continued in 2015 and 2016 (Palotti et al. 2015; Zuccon et al. 2016); however these newer evaluation campaigns sensibly differ from those in 2013 and 2014:

- Firstly, the topic creation process changed: instead of building queries from medical reports (see Sect. 3.2.1 for details), they were built from images depicting medical conditions, for example image depicting bloodshot eye.[4] This change resulted in a different format and type of query. It also meant a shift in the use case covered: 2013 and 2014

[3] http://www.trec-cds.org/.

[4] Subjects were asked to describe the picture as if it were their own health issue. See Palotti et al. (2015) for details.

topics considered information needs related to the understanding of diseases, conditions and treatments; while 2015–2016 topics focused on information needs related to self-diagnosis and treatment.

- From 2016 onwards, the document collection changed: instead of using a specific medical document collection, we opted for a larger web crawl, closer to the real document collection users are faced with when querying the web.

In order to conduct analysis on a homogeneous and comparable set of runs, this article focuses only on the IR evaluation task in 2013 and 2014.

2.2 Analysis of evaluation campaigns

Establishing a meaningful benchmark task to explore ad-hoc medical IR for lay users requires careful design of the components for the task and use of appropriate techniques to construct these. Construction of an IR test collection requires data collection design, gathering of user information needs, test query construction based on the information needs, and assessment of the relevance of returned results for each information need. In this section we overview relevant existing initiatives which analyze the results of (non-medical) IR tasks.

Probably the best known and most detailed comparative analysis of an IR evaluation was carried out within the Reliable Information Access (RIA) workshop (Harman and Buckley 2004; Soboroff 2009) which examined methods for relevance feedback and their behaviour. Retrieval results were manually examined for different runs and systems to detect weaknesses and system failures. One of the main findings was that most systems suffer from the same errors. Harman and Buckley (2004) concluded that “it may be more important for research to discover what current techniques should be applied to which topics, rather than to come up with new techniques”. While perhaps not an obviously insightful conclusion, this observation was only made possibly based on extensive analysis of very large numbers of experimental results created using many different systems and algorithmic alternatives.

The Robust Track at TREC[5] (Voorhees 2005) focused on queries that are difficult for typical systems in that it is difficult to design an IR method which is able to retrieve relevant documents for these topics, aiming to improve the consistency of retrieval technology. This involved carrying out a very detailed analysis of the document collection, queries and the relevant documents for each query, with the objective of trying to understand why some apparently reasonable queries are in fact very difficult to answer reliably from a collection containing relevant items. This track resulted in considering evaluation metrics such as the geometric mean average precision for IR when consistent IR effectiveness across all queries is important.

In an analysis conducted by Armstrong et al. (2009), an important finding highlighted was that there has in fact been very little improvement over strong baselines for publications describing experiments on established TREC ad-hoc retrieval over a long period of time and the need to compare to the best currently available results for a task. This emphasizes the importance of establishing strong baselines for a task, while seeking to develop and understand the potential contributions of novel methods which might be developed specifically for a specific task. This issue is illustrated clearly for medical IR by results reported by individual teams for the TRECmed search of medical reports task, where many techniques are able to offer improvements over weak baseline methods, but few are able to offer improvement when compared against a strong baseline using established general IR methods (Leveling et al. 2012; Voorhees and Hersh 2012).

[5] http://trec.nist.gov/data/robust.html.

Table 1 CLEF eHealth tasks in 2013 and 2014

Year	Tasks
2013	Named entity recognition in English clinical reports
	Normalization of acronyms/abbreviations
	Patient-centered information retrieval
2014	Visual-interactive search and exploration of eHealth data
	Information extraction from clinical text
	Patient-centered information retrieval
	Subtask: monolingual information retrieval
	Subtask: multilingual information retrieval

Bold entries are the tasks this article focuses on

Other related research on improving IR evaluation has examined minimizing efforts for relevance assessment by dynamically creating the set of pooled documents (Sakai and Mitamura 2010), determining the quality of test collections (Urbano et al. 2013) (which we will also apply to our collections in Sect. 4.2.1), investigating how to automatically predict query performance (Hauff et al. 2010; He and Ounis 2006), and automatic exploitation of this information.

3 CLEF eHealth: information retrieval task

CLEF eHealth has been running as an activity within the benchmark labs of the CLEF Conference since 2013. Each year CLEF eHealth offers IR, information extraction (IE) and information management tasks to volunteer task participants which aim to evaluate systems that support laypeople in searching for and understanding health information (Goeuriot et al. 2015; Kelly et al. 2016, 2014; Suominen et al. 2013). In 2013 and 2014 the tasks were built around an assumed use case of a patient receiving a discharge summary when they leave hospital, and then wishing to find relevant additional information. The discharge summary describes the diagnosis and the treatment that the patient received in hospital. The use case postulates that, given their discharge summary and the diagnosed disorders, patients often have questions regarding their health condition. Table 1 summarizes the tasks organized in 2013 and 2014.

This article focuses on details of the Information Retrieval task offered in 2013 and 2014 which adopted this use case. In this section we provide an overview of the organization of the tasks and of the submissions of the participating groups. More detailed descriptions are available in the 2013 and 2014 task overview papers in the CLEF proceedings (Goeuriot et al. 2013a, 2014c).

3.1 Task description

CLEF eHealth adopts the standard IR evaluation benchmark practice of providing participants with a collection of documents which must be indexed into their IR evaluation system, and a set of queries representative of the user task to be evaluated. In this case the documents covered various health and biomedical topics.

As shown in the Table 1, the task was monolingual in 2013, and had two subtasks in 2014: monolingual and multilingual IR. Figure 1 presents an overview of the data and its use within the task: the provided document collection, described in Sect. 3.2, is used to create an index. The discharge summaries are used to create the topics, as described in

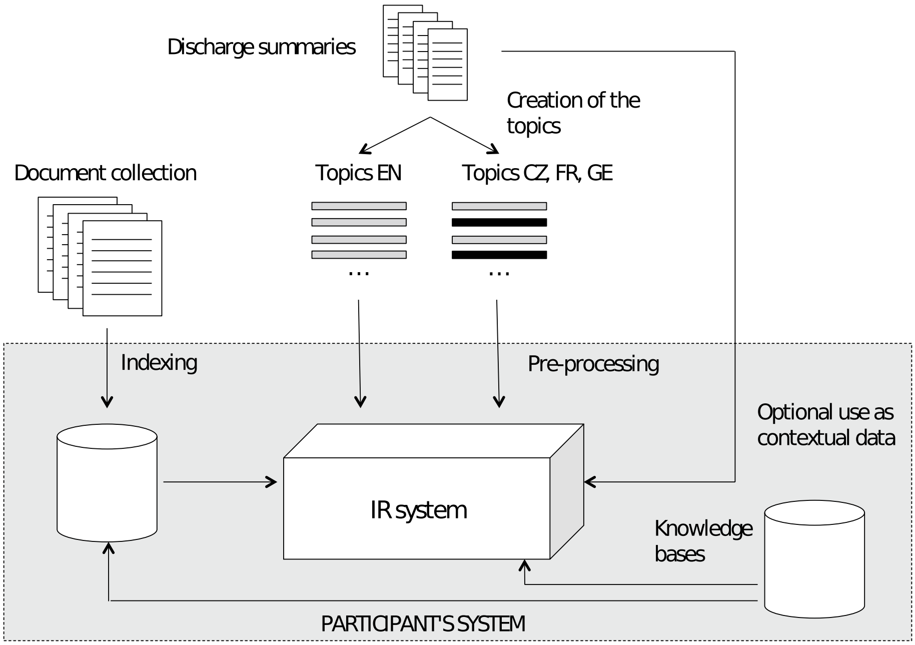


Fig. 1 Summary of the data and its use within CLEF eHealth IR task

Sect. 3.2.1. They can also optionally be used by participants as external data. Similarly, external knowledge bases can be used as an additional source of information (details are provided in Sect. 3.4). The gray box in the figure represents a participant's system: the format and method varies across participating teams and teams' systems.

3.2 Test collection

The task dataset provided to participants in the 2013 and 2014 CLEF eHealth IR challenges comprise a document collection of around one million web pages from medical websites made available through the Khresmoi project (Goeuriot et al. 2013b; Hanbury and Müller 2012). The document collection distributed in 2013 and 2014 are identical, excepting documents excluded because they had incorrectly formatted HTML markup or raised copyright issues identified by the Khresmoi project, see Goeuriot et al. (2013b) for details. The search topics for the test collections were developed by medical experts. Separate sets of 5 training topics and 50 test topics were created in 2013 and in 2014. The topics contain title and description fields as defined by TREC,[6] where: title is a shorter query statement of the user's information need and the description is a larger statement of the same need typically including multiple sentences. The created topic statements also contain additional fields: *discharge-summary*, which contains the discharge summary report which the patient's query stemmed from, and *profile*, containing basic information on the patient. Discharge reports originated from the de-identified MIMIC-II database.[7]

[6] http://trec.nist.gov/.

[7] http://mimic.physionet.org/.

3.2.1 Topic creation

As detailed earlier, the topics[8] used in the task aim to model queries used by laypeople (i.e., patients, their relatives or other non-medical representatives) to find out more about their disorders, once they have examined a discharge summary. Contextual information related to the patient history is contained in the discharge summary which is included in the topic statement. The discharge summaries can automatically be incorporated into the creation of the actual search query used by the IR system. The information contained in the discharge summary can subsequently be used with the query fields of the topic statement to determine the relevance of retrieved information to the specific user. The following example shows an extract from a discharge summary:

```
Admission Date: [**2014-03-28**]
Discharge Date:  [**2014-04-08**]
Date of Birth: [**1930-09-21**]
Sex:  F
Service: CARDIOTHORACIC
Allergies:
Patient recorded as having No Known Allergies to Drugs

Attending:[**Attending Info 565**]
Chief Complaint: Chest pain
Major Surgical or Invasive Procedure:
Coronary artery bypass graft 4.
History of Present Illness:
83 year-old woman, patient of Dr. [**First Name4
(NamePattern1) **] [**Last Name (NamePattern1) 5005**],
Dr. [**First Name (STitle) 5804**] [**Name (STitle)
2275**], with increased SOB with activity, left shoulder
blade/back pain at rest, + MIBI, referred for cardiac
cath. This pleasant 83 year-old patient notes becoming
SOB when walking up hills or inclines about one year
ago. This SOB has progressively worsened and she is now
SOB when walking [**01-19**] city block (flat surface).
[...]

Past Medical History:
arthritis; carpal tunnel; shingles right arm 2000;
needs right knee replacement; left knee replacement
in [**2010**]; thyroidectomy 1978; cholecystectomy
[**1981**]; hysterectomy 2001; h/o LGIB 2000-2001
after taking baby ASA; 81 QOD
[...]
```

[8] A query here is text typed in a search engine. A topic is an enriched query.

Different strategies were used to create topics in 2013 and 2014. In both cases, the topic was manually created by registered nurses, who were also clinical documentation researchers, from a selected disorder in a given discharge summary. This solution has been chosen in place of recruiting patients because of the issues involved with recruitment and privacy. We believe that, being in daily contact with patients receiving treatment and discharge summaries, nurses are familiar with patients' information needs and patient profiles.

- In 2013, a disorder was randomly selected from each discharge summary from among those already annotated. This selected disorder is assumed to be the main aspect of interest to a patient, e.g. a disorder mentioned in the discharge summary that a patient wants to find out more about.
- In 2014, instead of randomly selecting the disorder, we decided to create topics from the main disorder in each discharge summary. This was done using the field "Discharge diagnosis" or "Main diagnosis" in the discharge summary. If several disorders were diagnosed, the medical professionals were free to pick one from the list. When this field did not appear in the report, we asked them to select a disorder that appeared to be the main one in the whole report.

Using selected disorder and the associated discharge summary, the experts developed topics (and the criteria for judging the relevance of documents to the query, for use in the relevance assessment task described in the next section). The following example from 2014 outlines topic structure:

```
<query>
  <title> thrombocytopenia treatment corticosteroids
    length </title>
  <desc> How long should be the corticosteroids treatment
    to cure thrombocytopenia? </desc>
  <narr> Documents should contain information about
    treatments of thrombocytopenia, and especially
    corticosteroids. It should describe the treatment,
    its duration and how the disease is cured using it.
    <scenario> The patient has a short-term disease, or
     has been hospitalised after an accident (little to
     no knowledge of the disorder, short-term treatment)
    </scenario>
    <profile> Professional female </profile>
  </narr>
</query>
```

3.2.2 *Participants run submission*

Participating teams were permitted to submit up to 7 runs:

- Run 1 (mandatory) is a team baseline: only title and description could be used in the query, with no use of external resources such as dictionaries for example.
- Runs 2–4 (optional) any experiment WITH the medical reports.
- Runs 5–7 (optional) any experiment WITHOUT the medical reports.

The runs in each group had to be ranked in order of priority (1, 2 and 5 being the highest priority runs).This ranking allowed us to select the highest priority runs from each team for pool set generation, as detailed in the next section.

3.2.3 *Relevance assessment*

Every query-document pair in the assessment pool was judged by only one assessor. Assessors were domain experts and IR experts in 2013, specifically nursing professionals and researchers at the authors' organizations respectively; and paid professional assessors (but not medical experts) recruited externally in 2014.

Relevance assessment was based on a four point scale, which is mapped to a binary scale:

- {0: non relevant, 1: on topic but unreliable} → non relevant
- {2: somewhat relevant, 3: relevant} → relevant

Relevance assessments for the training queries were formed based on pooled sets created using the Vector Space Model (VSM) (Salton et al. 1975) and Okapi BM25 (Robertson and Jones 1994) for both 2013 and 2014 tasks. Assessments for the training queries were conducted by the domain experts, each document being assessed by one person. In order to investigate the effect of medical expertise on the relevance assessment, in the 2013 task the assessment for the corresponding five training queries was also conducted by an IR expert. A comparison of their assessments and analysis of their agreement is provided later in this article.

Pooling for the 2013 task For the 2013 task, we pooled the top ten documents obtained from the participants' baseline runs (run 1), their top-priority run using discharge summaries (run 2) and their top-priority run not using discharge summaries (run 5).[9] A large number of submissions were received: due to budget constraints the pool depth was limited to the top 10 ranked documents. This resulted in a pool of 6391 documents in total.

Pooling for the 2014 task For the 2014 task, the pool depth was also limited to the top 10 ranked documents. Documents were pooled from the participants' baseline runs (run 1), their top-two priority runs using discharge summaries (runs 2 and 3), and their top-two priority runs not using discharge summaries (runs 5 and 6). Thus, compared to the 2013 assessment pool, the 2014 pool contained two more runs per team. The pool depth was 10: as in 2013, this was mainly dictated by budget constraints. This resulted in a pool of 6,800 documents, in line with the size of the pool for the 2013 task.

[9] Runs are described in the section that analyzes participants' retrieval results.

3.3 Evaluation of the task results

Since the assessment pools were limited to depth ten, we evaluated participants' submissions mainly using metrics at a cut-off of up to 10 documents. This allows us to compare systems using only complete assessments, thus providing a reliable analysis of the difference between systems' performance. In addition, as the task models consumer laypeople using web search engines, it is expected that they rarely go beyond the first page of results (top 10 documents) (Hansen et al. 2003). The evaluation measures that are considered are precision at 5 and 10 document cut off (P@5 and P@10) and normalized discounted cumulative gain (NDCG@5 and NDCG@10). We also considered MAP as an evaluation metric, but we are aware that the MAP values may be unreliable since only the top ten documents have been assessed and submitted runs exhibit little diversity. Nevertheless, we wanted to report a measure covering the full set of up to 1000 retrieved documents. We also report the number of relevant and retrieved documents in the top 1000 results as a more recall-oriented measure.

Nine teams submitted a total of 46 runs in 2013 and 14 teams submitted a total of 62 runs in 2014. Only one team submitted runs for both years.

3.4 Summary of the methods used by task participants

Tables 2 and 3 provide a summary overview of the participating teams' approaches, for each step of the retrieval process: pre-processing of the documents collection, indexing and retrieval. Note that retrieval can involve more than one retrieval pass to enable inclusion of retrieval enhancement techniques such as query expansion via the use of relevance feedback. The tables also show details of any additional external resources used by each team and whether the discharge summaries (DS) were used is also presented. We highlight key features of individual participants' approaches in the next section, and examine the efficacy of these methods for this task. Methods used by each participant are described in full in the Working Notes papers; references provided in Tables 2 and 3.

Most of the external resources are medical thesauri, such as UMLS. The Unified Medical Language System (UMLS) is a metathesaurus gathering various medical knowledge bases and terminologies. It provides for every entry (corresponding to a medical concept) a unique identifier, a definition, semantic types, related concepts, etc. For example, breast carcinoma in UMLS has the identifier C0678222, and as a definition "A malignant neoplasm that develops or arises in breast tissue".[10]

4 Analysis of the results of the evaluation task

The participants' submissions for the CLEF eHealth IR tasks in 2013 and 2014 represent a rich source of information to investigate task design and techniques for Medical IR.

Firstly, in Sect. 4.1, we observe, compare and draw conclusions on participants' runs in the following way: which systems are applied as baselines; how the discharge summaries

[10] A single definition is provided here for example purposes. In reality an exhaustive list of definitions is provided for gathered terminologies.

are integrated in the systems; which external resources are used; if query expansion is integrated and how.

Secondly, in Sect. 4.2, we evaluate the campaigns' datasets in four ways: we first evaluate the reliability of the datasets; secondly, we observe the relevance of documents across queries and across datasets; then we analyze if medical expertise has an impact on the relevance assessments recorded and their quality; finally we investigate the impact of the size of the pool sets by assessing the effect of the relevance of non-assessed documents on the participants' results.

4.1 Analysis of participants results

In this section, we observe and compare the runs and results of the teams participating in the IR task of CLEF eHealth in 2013 and 2014. As the datasets varied from 2013 to 2014, we can only compare results in parallel for each campaign.

4.1.1 Baselines used

Participating teams were required to submit a baseline run (run 1) consisting of a retrieval approach only (e.g. Vector Space Retrieval Model), with no additional information (e.g. discharge summary) or external resources used to boost performance. The organizers also provided baseline runs using BM25 in 2013 and using a variety of retrieval models in 2014 (specifically tf.idf, BM25, language modeling with Jelinek–Mercer smoothing, and language modeling with Dirichlet smoothing).[11] Figure 2 compares the retrieval effectiveness (in terms of P@10) of each team's mandatory baseline system (run 1) with the organizers best performing baseline run and worst performing baseline run. In both years, the effectiveness of the best organizer provided baselines are comparable to those of the participating teams' baseline systems. In particular, in the 2013 task, only two teams achieved higher effectiveness with their baseline than that achieved by BM25, the worst organizers baseline (no statistically significant differences), and no team achieved higher effectiveness than the BM25 with feedback (BM25_FB) baseline provided by the organizers. In 2014 the organizer provided language modeling with Dirichlet smoothing baseline is outperformed by five teams, while the worst organizers baseline (language modeling with Jelinek Mercer smoothing) is outperformed by all participants' baselines excepting team YORKU.

It is interesting to note that the best team baseline effectiveness in 2013 and both the best organizers baseline and team baseline effectiveness in 2014 are obtained using language models with Dirichlet smoothing (but with different pre-processing steps), suggesting that this type of language model forms a consistently strong baseline for system comparison. Furthermore, four of the top five team baselines in the 2013 task and all top five team baselines in 2014 are obtained using language models, while in 2013 team UOG.Tr (4th best team baseline in 2013) used divergence from randomness as implemented in the Terrier Toolkit. Indeed, four main types of baselines can be identified across runs submitted in 2013 and 2014: language models, vector space models, divergence from randomness (only in 2013) and the TOPSIG's document signatures approach (only in 2013). Overall, language models (in particular with Dirichlet smoothing) appear to be obtaining considerably better results than vector space models and its variants, although in 2013 our BM25

[11] Further details on baselines used are provided in the Task overview papers Goeuriot et al. (2013a, 2014c).

Table 2 Overview of the methods used by the participating teams in 2013

Team	Pre-processing	Indexing	Retrieval model	Query expansion	Re-ranking	External resources	DS
AEHRC (Zuccon et al. 2013)	Porter stemmer	Indri	Language modelling	Abbreviations	Readability	Medtex	
	Removed stopwords			Spell correction	Authoritativeness	Google	
						Wordlist from Wikipedia	
Mayo (Zhu et al. 2013)	HTML removal	Indri	Language modelling	Mixture of relevance models	Combined words and concepts	Boilerpipe, MedTagger	✓
	Porter stemmer		Markov random field	PRF with MeSH		TREC2011 MR Track	
	Medical stopwords					TREC2007 Genomics Track	
						Mayo Clinic clinical notes	
						Metamap, UMLS,MeSH	
OHSU (Bedrick and Sheikhshabbafghi 2013)	Removed HTML	Lucene	Language modelling	PRF with MeSH		Metamap	
	Removed non-ASCII		Vector space model			MeSH	
	Removed stopwords						
QUT-TOPSIG (Chappell and Geva 2013)	None	TopSig	TopSig IR model	PRF with DS	TopSig refined mode		✓
SNUMedInfo (Choi and Choi 2013)	Removed stopwords	Indri	Language modelling	Expanding UMLS concepts		Metamap	✓
	Case-fold					UMLS	
	Krovetz stemmer						
THCIB (Zhong et al. 2013)	Removed HTML	Lucene	BM25	Abbreviations	PageRank	HTMLParser	✓
			Vector space model	Expanding UMLS concepts	HITS	Wikipedia	
				Topic each query belongs to	HTML layout	Google	

Table 2 (continued)

Team	Pre-processing	Indexing	Retrieval model	Query expansion	Re-ranking	External resources	DS
KC&RA (Barajas and Akella 2013)	Removed HTML Removed foreign chars Removed numbers Removed small docs Removed stopwords Krovetz stemmer	Indri	Language modelling	Using noun phrases		CTAKES MIMIC II	✓
UOG.Tr (Limsopatham et al. 2013b)	Porter stemmer Removed stopwords	Terrier	Divergence from randomness	PRF using Bo1 Model			
UTHealth (Zhang et al. 2013)	Removed HTML Removed stopwords	Lucene Random index	Vector space model Semantic vector model	Expanding UMLS concepts	UMLS API	Apache Tika UMLS	

Table 3 Overview of the methods used by the participating teams in 2014

Team	Pre-processing	Indexing	Retrieval model	Query expansion	Re-ranking	External resources	DS
CSKU-COMPL (Thesprasith and Jaruskulchai 2014)	Removed HTML Lucene standard analyzer	Lucene	Vector space model	Rocchio-based PRF with genomics data		TREC2004 genomics	
CUNI (Saleh and Pecina 2014)	Different approaches for HTML removal Spell correction with MedlinePlus	Terrier	Hiemstra	PRF using Bo1 model		HTML strip Boilerpipe MedlinePLus JusText	
DEMIR (Ozturkmenoglu et al. 2014)	Removed HTML Porter stemmer Removed stopwords	Terrier	Vector space model	Kullback–Leibler		Weka CLEFeHealth 2013	
ERIAS (Dramé et al. 2014)	Removed stopwords Unigrams and bigrams	Lucene	Vector space model	Synonyms and descendants from MeSH and UMLS		Metamap MeSH thesaurus UMLS	
GRIUM (Shen et al. 2014)	Removed HTML	Indri	Language modelling	Point-wise mutual information		Metamap UMLS	
IRLabDAIICT (Thakkar et al. 2014)	Porter stemmer	Indri	BM25	Linear combination of DS and query terms		Metamap	✓
	Removed stopwords Medical stopwords		Language modelling	MeSH synonyms PRF		MeSH	
KISTI (Oh and Jung 2014)		Lucene	Language Modelling	Expand abbreviations Expand queries based on DS PRF	Clustering-based Centrality-based		✓
MIRACL (Ksentini et al. 2014)	Removed HTML Removed stopwords	Terrier	Vector space model				

Table 3 (continued)

Team	Pre-processing	Indexing	Retrieval model	Query expansion	Re-ranking	External resources	DS
Nijmegen (Verberne 2014)	Regular expressions	Indri	Language modelling	PRF using Ponte Expander		UMLS	✓
	Removed stopwords			Informativeness and phraseness of terms in the DS			
	Case-fold			UMLS synonyms			
RePaLi (Claveau et al. 2014)	Removed HTML	Indri	Language modelling	Expanded synonyms from UMLS		UMLS	
	Krovetz stemmer		Markov random field	Expand abbreviations		Ogmios NLP	
	Removed stopwords			Lexical inclusion based on hierarchical relations		FASTR, YaTeA	
	Replaced broken chars					TreeTagger	
						CLEFeHealth 2013	
SNUMednfo (Choi and Choi 2014)	Removed stopwords	Indri	Language modeling	Intersection of UMLS terms and DS	Learn to rank with random forest	Metamap	✓
	Case-fold					UMLS	
	Porter stemmer					CLEFeHealth 2013	
LABERINTO (Malagon and López 2014)	Removed HTML	Lucene	Vector Space model	Expanded synonyms and concepts from MeSH		Apache Tika	
	Lucene standard analyzer					Metamap	
						MeSH, SKOS	
UIOWA (Yang et al. 2014)	Removed HTML	Indri	Language modelling	Expand abbreviations		Lynx	
	Replaced broken chars		Markov random field	PRF		Genia Sentence Splitter	
	Regular expressions			Medical bigrams for MRF		CLEFeHealth 2013	
	Spell checker					Wikipedia	
YORKU (Wu and Huang 2014)	Not specified	Not specified	Not specified	Not specified	Combination of different learn to rank algorithms	RankLib	

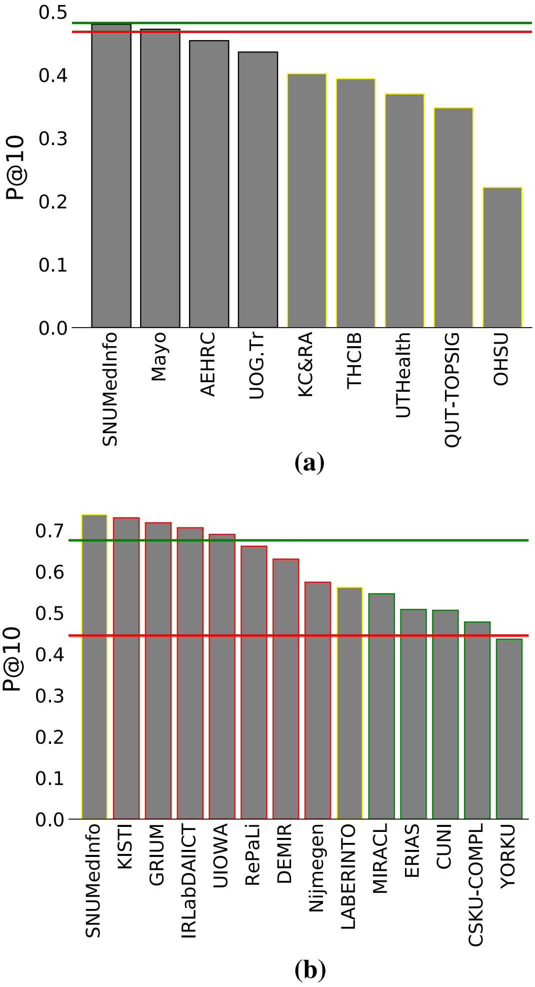


Fig. 2 P@10 values (y-axis) for participants and task organizers (x-axis) provided best and worst performing baselines. **a** 2013 task. Task organizers' baselines: BM25 (red line); BM25 with feedback (green line). Runs that were statistically different from both baselines in a t-test ($p < 0.05$) are marked with a yellow edge color. **b** 2014 task. Task organizers' baselines: language modeling with Jelinek–Mercer smoothing (red line); language modeling with Dirichlet smoothing (green line). Yellow, red and green edge color were used to distinguish run that were statistically different from both, Jelinek–Mercer or Dirichlet baselines, respectively in a t-test ($p < 0.05$) (Color figure online)

task baseline outperformed most participants' baselines. In 2013, divergence from randomness provides effectiveness similar to language models.

It is essential in IR evaluation challenges to provide strong baselines, in order to obtain valuable results and outcomes (Leveling et al. 2012). Even if teams' baseline performance varies, teams' results seem consistent enough, i.e. no team can claim an improvement over a weak baseline.

4.1.2 Use of discharge summaries as contextual information

Figure 3 shows results of teams that used the discharge summaries (DS) in 2013 and 2014. Specifically, the first column is their best run using the DS, the second is their best run without the DS, and the third shows their baseline run. Five teams submitted runs using the DS in 2013, and four teams in 2014. Note that the best runs without discharge summaries

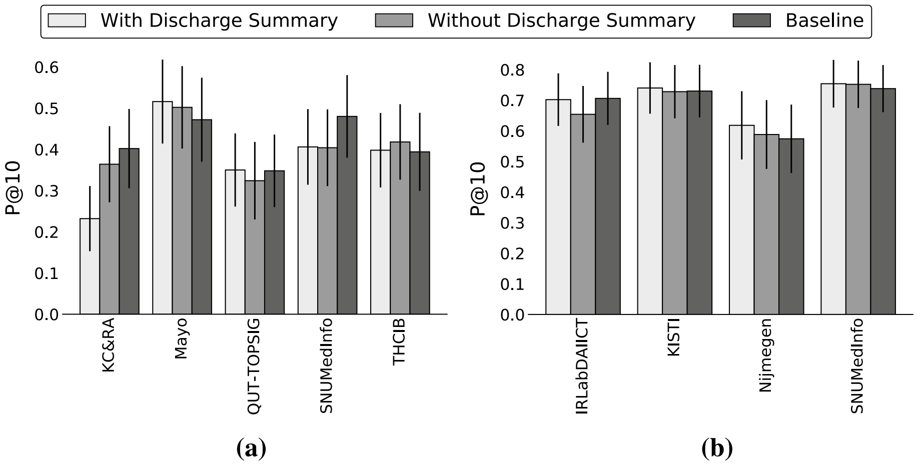


Fig. 3 P@10 for the participants' baselines, best run using discharge summary (DS) and best run without DS in 2013 (left) and 2014 (right). The 95% confidence interval from the mean is represented with error bars. **a** 2013. **b** 2014

are only given for reference, since they do not necessarily have similar experimental settings to the best run with DS.

For 2013, we observe that three teams out of five achieved an improvement over their baseline using the DS. Among these teams, only two obtained better results with the DS than without (QUT-TOPSIG and MAYO). These two teams used the DS as follows: to perform re-ranking based on concepts extracted from documents, queries and DS (MAYO); and to perform query refinement (QUT-TOPSIG). The other teams used DS mainly for query expansion: to filter out expansion terms (Medinfo); for concept-based expansion (KC&RA and THCIB).

As described in Sect. 3.2.1, the disorders used to generate the 2013 topics were selected randomly from within all the disorders identified in each DS. Therefore, the selected disorder was not necessarily the main one mentioned in the DS. This could explain why, for most of the teams, the use of the discharge summaries did not provide useful contextual information to improve retrieval performance. This problem was identified by KC&RA (Barajas and Akella 2013), who tried to identify relevant passages in DS and expand their queries with concepts identified in these passages only.

In 2014, this potential issue was fixed by selecting the main diagnosed disorder mentioned in the DS, hence creating a real link between the DS and the generated topic. Globally we observed much higher performance in 2014 than in 2013, which applies for the runs using DS. Among the four teams, none reported any decrease in their results while including DS information in their systems. However, all obtained results comparable to their baseline or their best runs without DS information, apart from team Nijmegen, who obtained a significant improvement over the baseline.

All teams in 2014 also used the DS to perform query expansion, either to find expansion terms (teams IRLabDAIICT, KISTI and Nijmegen), or to filter expansion terms (team SNUMEDINFO).

Although information in the DS could in theory be ideal for contributing to the selection of relevant and personalized documents, we can see that in an IR environment, refinement needs to be achieved to get more focused and concise contextual information. Further investigation is necessary to fully understand how patient medical information can be used and how DSs can contribute to improving IR (Goeuriot et al. 2014a).

4.1.3 Use of external resources

In this section, we describe the external resources used by the participants, how they were used and the results achieved using them. Table 4 provides an overview of the external resources used in 2013 and 2014. These resources were mainly used in one of three stages of the IR process: indexing, query expansion, or re-ranking. We distinguish three categories of resources:

- Corpora, composed of document collections (generally from the medical domain). The majority of these come from related evaluation campaigns or IR benchmarks;
- Thesauri/lexicons, medically related lexical or semantic resources, very often UMLS or subsets;
- Other types of resources, namely one list of recommended health-consumer websites.

In the medical domain, there are various rich knowledge bases. The Unified Medical Language System (UMLS) is a metathesaurus: it consists of several thesauri and terminologies. MESH and SNOMED belong to UMLS and are often used for text mining applications:

- SNOMED is a clinical health terminology used to process clinical data. It includes terms related to: clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and specimens.
- MESH is a controlled vocabulary thesaurus used for indexing articles on the National Library of Medicine search engine PubMed.

We can see, in Table 4, that most of the resources are used for query expansion. A few teams used them for indexing: teamMayo used UMLS to annotate documents and index the concepts' concept unique identifier (CUI); team KC&RA used SNOMED to identify and index medical noun-phrases; team CUNI used Medline plus for spell-checking during the pre-processing of the documents. Team AEHRC used the list of recommended health-consumer websites, the last resource listed in Table 4, to re-rank retrieval results based on website authoritativeness. We observed very little variation in the resources used in 2013 and 2014. The main one being the use of the 2013 collection and qrels in 2014 to train the systems or predict good expansion terms (teams DEMIR, RePaLi, SNUMEDINFO and UIOWA).

The use of external resources, and in particular thesauri/lexicons, to aid the indexing of noun-phrases or to drive the whole retrieval process (concept retrieval) has shown mixed results. This finding resonate with results from the literature for both the task considered here, and other health related tasks (Koopman et al. 2016; Shen and Nie 2015; Xia et al. 2014). We discuss the use of domain resources for query expansion in greater detail in the next section.

Details of teams' approaches are provided above in the Summary of the Methods Used by Task Participants section, Sect. 3.4.

Table 4 Overview of the external resources and their use

Category	Resource	Indexing	Query expansion	Re-ranking
Corpora	TREC medical records 2011		x	
	TREC genomics 2007		x	
	TREC genomics 2005		x	
	(Wikipedia and Medline)		x	
	TREC genomics 2004		x	
	Mayoclinic clinical notes		x	
	CLEF eHealth 2013		x	
Thesauri,	UMLS	x	x	
lexicon	SNOMED	x		
	MeSH		x	
	Wikipedia list of		x	
	medical abbreviations			
	Medline Plus	x		
Other type of resource	CAPHIS recommended health-consumer sites			x

4.1.4 *Effectiveness of query expansion*

As can be seen from Tables 2 and 3, most teams performed some form of query expansion (QE) in some of their runs.

We distinguish two main approaches: corpus-based QE and concept-based QE. In this section we describe the approaches participants used, and analyze their results from these perspectives.

Corpus-based expansion involves using a document collection to expand the query to add the most salient related terms to the query. The collection used can be the task collection, already indexed for retrieval, or an external collection, often on the same domain or topic.

- In 2013, three teams used corpus-based QE: teams UOG.Tr, QUT-TOPSIG, and MAYO. While QUT-TOPSIG only used the discharge summaries as a resource, the other two teams used the task collection, as well as other related collections such as the TREC Medical Records, TREC genomics, etc. collections (details can be found in Table 2). Teams UOG.Tr and QUT-TOPSIG performed Pseudo Relevance Feedback (PRF), and team MAYO a Mixture of Relevance Model.
- In 2014, five teams performed a PRF QE approach: teams CKSU, CUNI, DEMIR, KISTI and UIOWA. They used the discharge summaries, the task collection and various external collections. Team DEMIR used the Kullback–Leibler divergence approach.

Concept-based expansion involves finding relevant related terms in knowledge bases.

- In 2013, four teams experimented with concept-based QE: teams SNUMEDINFO, THCIB, UTHealth and OHSU. They all use UMLS to select expansion terms, with different selection strategies: preferred terms for identified concepts, top-ranked terms given by the UMLS API concept identification tool, or the sibling entry terms for identified concepts. Team SNUMEDINFO also used the discharge summaries to filter out expansion terms. Team THCIB added to the UMLS preferred terms keywords automatically added after human annotation of the queries.

- In 2014, seven teams performed concept-based QE. They all used UMLS or MeSH to expand queries, with synonyms, preferred terms, descendants, or similarity-based related terms (teams ERIAS, RePaLi, SNUMEDINFO, and GRIUM). Some teams also used various weighting schemes or filtering approaches to rank expansion terms.

In 2013, team AEHRC investigated a QE approach slightly different from the two categories above, with spelling correction and acronym expansion.

Combined approaches involve expanding queries with both corpus- and concept-based approaches. Three teams combined approaches: team MAYO in 2013, and teams NIJMEGEN and ITLabDAIICT in 2014. In combination with the mixture of relevance model, team Mayo also performed some concept-based expansion, adding for each concept identified in the queries its MeSH entry terms and its descendant nodes (they used the discharge summaries to filter out non-relevant expansion terms). Teams NIJM and IRLabDAIICT, combined corpus-based methods (PRF on the DS and a linear combination of DS with query terms) with concept-based methods.

Figure 4a, b present a comparison of P@10 values for the baseline and a run with QE. We chose to compare against the baseline a run with only QE added if available, or with QE among other additions otherwise. When several runs were available for selection, we chose the best performing one. Figure 4a shows that in 2013 three teams improved their baseline using QE, three obtained lower results, and two teams obtained similar results. For the teams who achieved an improvement, team MAYO achieved this by using concept-based QE, and a mixture of relevance models combined, with the CUI indexed as well. As all their runs except the baseline use QE, the improvement cannot be assigned to any individual part of the process. Team AEHRC obtained an improvement over their baseline by expanding the queries with acronyms and spelling errors. Team THCIB improved over their baseline by expanding the queries with the UMLS synonyms and acronyms expanded. Team MEDINFO obtained lower results by adding to the baseline QE with UMLS preferred terms (filtering out the terms not relevant using the discharge summaries). They only used the title of the topics in this experiment, but obtained very similar results when using the description and narrative fields as well. Although there is not much detail in their Working Notes paper, it appears that this is the only addition to their baseline. Team UTHealth added to the VSM retrieval model query expansion using the top-ranked concepts identified by the UMLS API. They used the title, description and narrative fields as a query for the baseline run as well as for the QE run. Team OHSU also used concept-based expansion, adding to the queries MeSH sibling entry terms. Team UOG.Tr and team QUT-TOPSIG, who obtained similar results, used PRF based approaches.

Figure 4b shows that in 2014 almost all teams performance improved with the addition of QE. While this cannot be systematically attributed to QE (as many changes can be made from one run to another), such an attribution appears obvious for some runs. Ten teams achieved better results using QE, which does not discriminate concept-based from corpus-based expansion. Teams CUNI and DEMIR achieved an improvement with corpus-based methods, and teams ERIAS, RePaLi, SNUMEDINFO and GRIUM improved their baseline by expanding their queries with related concepts. Only teams UIOWA and IRLabDAIICT do not obtain any improvement over their baseline using similar techniques. Team NIJM,

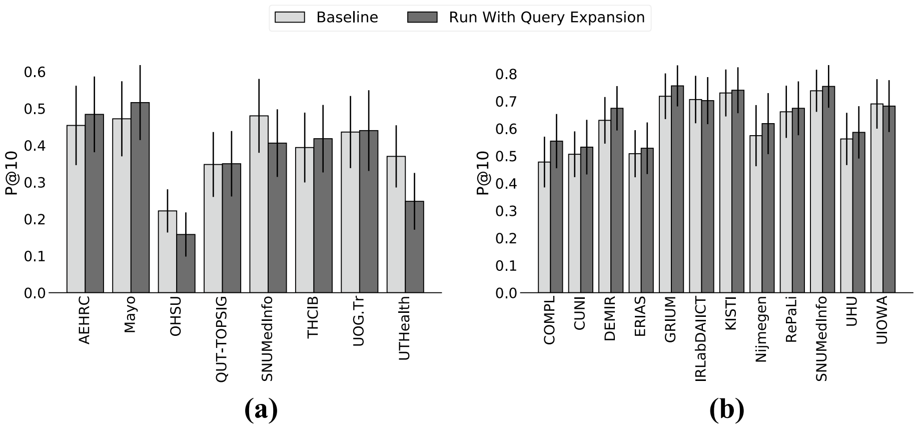


Fig. 4 P@10 for the baseline and the best performing run using query expansion for 2013 (left) and 2014 (right). The 95% confidence interval from the mean is represented with error bars. **a** 2013. **b** 2014

by using both corpus- and concept-based expansion, showed that concept-based expansion on this dataset provided better performance than corpus-based expansion, which was not the case in 2013.

From this set of experiments, concept-expansion appears to introduce noise more so than contributing to retrieval effectiveness on 2013 queries, as most teams observed a decrease in their results using this approach. However, the opposite held true on 2014 queries, with most teams achieving improvement over their baseline with the addition of concept-based expansion. Acronym expansion appears to work well, since two teams obtained an improvement in performance using this technique. While we clearly observed an overall improvement in system performances on the 2014 dataset compared to 2013, it seems difficult to explain why concept-based expansion works better on one set than the other. Possible explanations could be firstly that teams had access to the 2013 dataset in 2014 and therefore could train their systems on much bigger datasets; and secondly that the queries were much simpler, and arguably closer to concepts in UMLS.

Further analysis would be required to generalize this experiment, but it is not possible within the framework of this evaluation task, as we do not have access to the teams' systems, rather only the runs they have submitted.

We note that mixed findings about the effectiveness of query expansion in health information retrieval have been reported in relevant literature. In particular, concept-based query expansion has been shown to be affected by the risk of introducing noise within the reformulated query and that gains are possible if methods are finely tuned; this was found for methods evaluated within the same task considered here (Zuccon and Koopman 2018; Liu et al. 2016; Tibi et al. 2017) and within other health-related tasks, such as cohort selection/health record search (Alsulmi and Carterette 2016; Koopman et al. 2016; Limsopatham et al. 2013a; Zhu and Carterette 2012; Zhu et al. 2014; Zuccon et al. 2012, 2015a) and clinical decision support (Demner-Fushman and Lin 2007; Soldaini et al. 2015).

4.2 Analysis and evaluation of the datasets in the light of the campaign

In this section, we evaluate the task datasets in the light of the campaigns and the participants' results. In particular, we assess the reliability of the collections and the quality of the relevance judgments.

4.2.1 Evaluation of the dataset

Whenever a new test collection is introduced, there is a question about the reliability of the test collection in distinguishing between systems. In general, the more queries one has (i.e. the more test cases) the more confident one is with regard to the reliability of the test collection. In domain specific benchmarking such as CLEF eHealth, the cost of assessment is particularly high, so the number of queries is generally relatively small. In this section therefore, we consider the stability of the 2013 and 2014 CLEF eHealth IR benchmarks. We follow the method recently introduced by Urbano et al. (2013), which is based on Generalizability theory, but also provides information regarding the more common Kendall Tau correlations.

Urbano et al.'s method consists of two steps. First a *G-study* (generalizability study) estimates variance components based on existing data. Second, a *D-study* (decision study) computes reliability indicators. For the 2013 and 2014 collections, Table 5 shows the parameters based on the existing data. The first row shows the sample size, i.e. the number of runs (systems), queries, and run-query pairs (interactions). The following rows show, for each of the above, the variance components.

In calculating these values, for the 2014 collection we only considered the English queries. We also had to eliminate 2 of the 35 runs because they had not provided answers to all the queries.

Figure 5 shows the estimated Kendall Tau correlation and relative stability of the collections, for different sizes of the query set, as well as the 95% confidence intervals (shaded regions). Comparing the two collections we observe that the 2013 collection is more reliable than the 2014 collection. This is explained by the data in Table 5. We know that reliability is related to three components (Lin 2005): query set size, mean effectiveness scores, and variability of scores. We observe that while the query set size is constant for the two years, the mean effectiveness score and the variance are larger in 2013 compared with 2014. That means that in 2014, the collection has a harder time distinguishing between runs, because the runs have smaller overall scores, and are tighter together.

4.2.2 Documents and relevance assessment

Figure 6 shows the number of documents per query that were assessed, along with their graded relevance distribution. For the 2014 task, these documents correspond exactly to those that were pooled from the participants' submissions; for the 2013 task these documents include the pooled ones and those that were identified as duplicates of pooled documents. These figures allow us to analyse the diversity and coverage of the assessed pools of documents. For the 2013 task, this analysis shows that all the queries have a roughly similar amount of documents pooled, except for 2 queries (query 19 and 46), for which the document pool is much larger. A first hypothesis to explain this finding is that for these two queries, participants submitted runs that highly differ in terms of documents that contribute to the pool. A further analysis of the assessments for these queries reveals that this is not

Table 5 G-study for the existing eHealth test collections for NDCG@100

	2013			2014		
	Systems	Queries	Interaction	Systems	Queries	Interaction
Sample size	36	50	1800	33	50	1650
Mean Sq.	0.3214	1.2324	0.0213	0.0834	0.8197	0.0108
Variance	0.0060	0.0336	0.0213	0.0015	0.0245	0.0108
Variance (%)	9.851	55.225	34.924	3.947	66.781	29.372

the case. The large number of documents assessed is explained by the fact that the collection contains duplicate documents: for query 19 and 46, some documents that are largely duplicated in the collection were pooled and assessed, thus producing a large number of documents with relevance labels for these two queries.

For the 2014 task, the analysis highlights that, on average, more documents were assessed than in the 2013 task (on average 124 documents per query in 2013 and 132 documents per query in 2014). The larger pool for the 2014 task suggests that document rankings are more diverse (at least up to the pooled depth) across participants than they were in the 2013 submissions. However, the increase in pool size may also be due to the fact that more runs were submitted (46 in 2013 and 62 in 2014) and that two more runs per participant were pooled in 2014.

The analysis of the 2014 assessments also highlights an increase in number of relevant documents from the 2013 task: in particular it shows an increase in the number of highly relevant documents. This may be due to a number of reasons, for example: (1) the 2014 task considered easier queries (i.e. easier for the retrieval systems to find highly relevant answers); (2) the professional assessors were less stringent than the 2013 assessors in assigning the highly relevant label; or (3) a genuine increase in the effectiveness of the submitted systems. The third hypothesis can be ruled out as (almost all) similar systems deliver different effectiveness in the two years. That is, the higher number of relevant documents in 2014 may not mean that the systems participating in that year were, overall, better than those used in 2013. The remaining two hypothesis may be instead applicable and indeed the findings may be explained by a mix of these two conditions. Changing the procedure used to obtain queries in 2014 may have resulted in queries that refer to more common conditions which may in turn be easier for a system to retrieve highly relevant documents for. Similarly, recent studies have found that medical relevance assessment is hard (Koopman and Zuccon 2014), and thus small changes in assessment conditions, like the use of a different pool of assessors, as in 2014, may affect the results of the assessment exercise.

Figure 7 shows the distribution of binary relevance (relevant/non relevant document) for each query. The majority of queries in the 2013 task had more documents judged as non-relevant than those in 2014. This finding confirms the previous analysis that queries in the 2014 task were less challenging than in the 2013 task, or that the change in type of relevance assessors influenced the distribution of relevant/non relevant documents in the pool, or a mix of these two hypothesis. Figure 8 shows the topics with highest and lowest percentage of relevant documents for both collections.

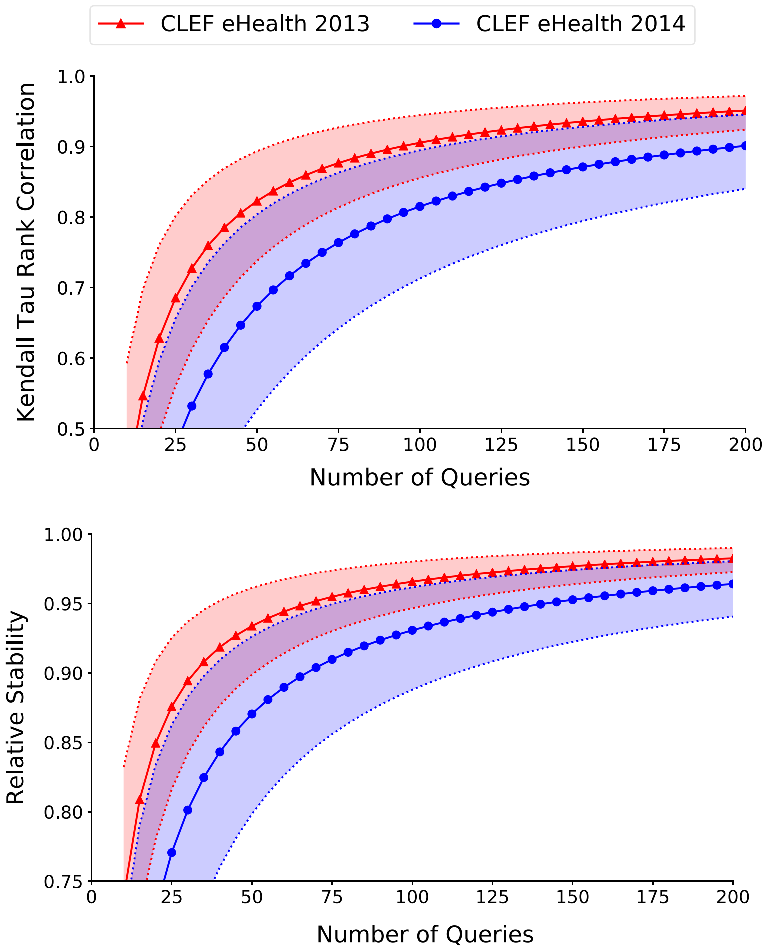
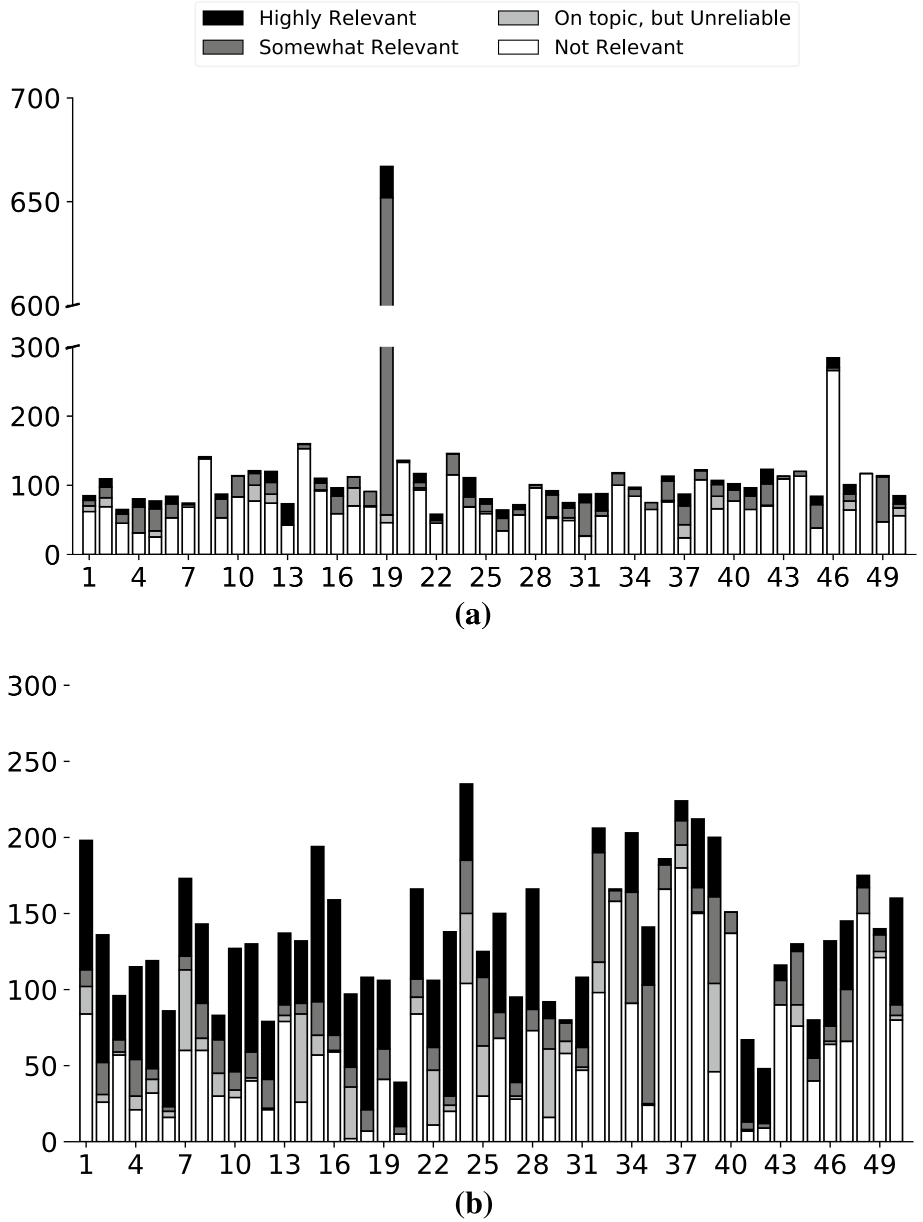


Fig. 5 Kendall Tau (top) and Eρ^2 (bottom) estimates for the 2013 and 2014 eHealth test collection for NDCG@100

4.2.3 Effect of the relevance of non-assessed documents on participant results

The quality of relevance assessments and the depth of the pool of documents assessed can affect how meaningful and reliable system-level IR benchmarking results are. Typical IR evaluation tasks such as the ad-hoc retrieval tasks at TREC and CLEF average between 100–200 assessed documents per topic. For the CLEF eHealth IR task, there are 37.16 assessed documents per topic on average for a pool depth of 10 (1858 relevant documents in total), taken from a subset of the submitted runs.

To investigate if and how the system performance would change with more complete relevance assessments, we conducted experiments which automatically estimate missing relevance assessments. For our experiments, we compare three different strategies of re-assessing documents with missing relevance information for relevance:



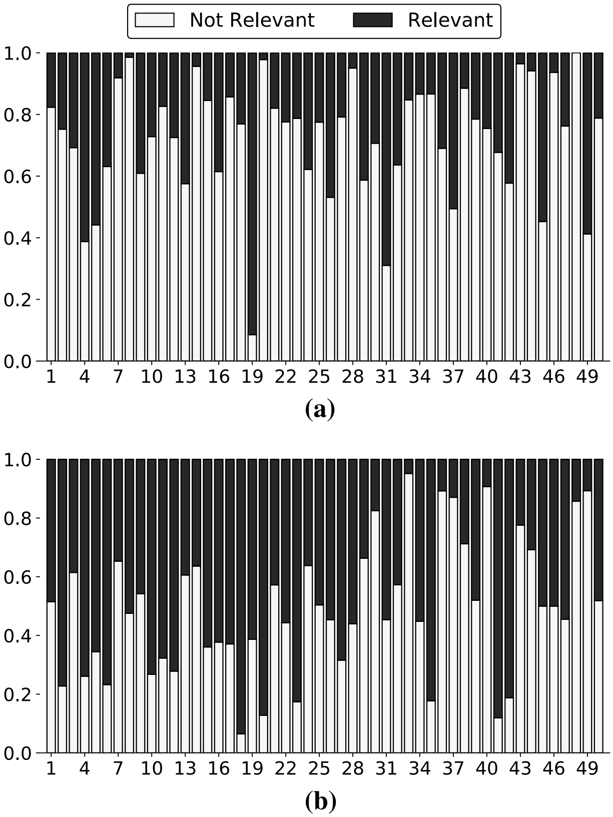
(a)

(b)

Fig. 6 Amount of documents per query in the pool set, and distribution of their graded relevance. **a** 2013 task. **b** 2014 task

- A: All documents with missing relevance information are assumed to be non-relevant. This is the standard approach used in IR evaluation campaigns.
- B: All documents with missing relevance information are assumed to be relevant. This approach corresponds to the worst-case situation arising from incomplete relevance assessment.

Fig. 7 Percentage of relevant and not relevant documents per query. **a** 2013 task. **b** 2014 task



- C: The same number of documents which is known to be relevant for a given topic is assumed to be relevant. This approach corresponds to the observation that for topics with few relevant documents, few additional relevant documents can be found when increasing the pool size, while for topics with many relevant documents, typically many more can be found (Harman and Buckley 2004; Voorhees 2005; Voorhees and Tong 2011).

We compute the extended relevance information (qrels) for a pool depth of 10, 20, 50, 100, 200, and 500 documents and compare the system performance for the three approaches (i.e. corresponding to A10, B10, C10, etc.). Results are presented in Tables 6 and 7. For brevity, only results for the run with the highest MAP for each team are shown. Note that results for method A correspond to the results from the original runs. Also note that for a large pool size, the assumption that all previously unassessed documents are relevant is unrealistic.

It has been noted that the missing relevance information may make the evaluation of sophisticated methods such as automatic query expansion more difficult, as new documents with unknown relevance are found. Even if new proposed methods would in fact significantly improve efficiency over a baseline approach, this increase in effectiveness might not be noted due to missing relevance information in the benchmark data. In particular, it was argued before that experiments with blind relevance feedback or query expansion did not show a significant improvement in performance due to incomplete or missing relevance information (Pecina et al. 2014). Our analysis shows that for all three strategies

Highest Percentage of Relevant Documents for CLEF eHealth 2013 and 2014 Collections

```
<query>
  <id>qtest19</id>
  <discharge_summary>11439-014138-DISCHARGE_SUMMARY.txt</discharge_summary>
  <title>abnominal pain and helicobacter pylori and cancer</title>
  <desc>is abdominal pain due to helicobacter pylori a symptom of cancer</desc>
  <narr>cancer, helicobacter pylori and abdominal pain</narr>
  <profile>A 60-year-old male who knows that helicobacter pylori is causing cancer and
    now wants to know if his current abdominal pain could be a symptom of cancer</profile>
</query>

<topic>
  <id>qtest2014.18</id>
  <discharge_summary>11762-027273-DISCHARGE_SUMMARY.txt</discharge_summary>
  <title>dizziness and hypotension</title>
  <desc>How to prevent dizziness and hypotension?</desc>
  <narr>The document should contain information about hypotension and dizziness.</narr>
  <profile>This 63 year old lady lives in a group home due to her mental illness.
    She fainted and had to visit the hospital. Now that she is back home, her caregivers
    want to know how they could best prevent the fainting occurring again. </profile>
</topic>
```

Lowest Percentage of Relevant Documents for CLEF eHealth 2013 and 2014 Collections

```
<query>
  <id>qtest8</id>
  <discharge_summary>04266-000520-DISCHARGE_SUMMARY.txt</discharge_summary>
  <title>Acidosis and metastasic adeno carcinoma</title>
  <desc>what is the connection between acidosis and metastasic adeno carcinoma</desc>
  <narr>Acidosis and metastasic adeno carcinoma</narr>
  <profile>A 76-year old man who dies from metastatic adeno carcinoma. The family is
  wondering about the asidocis and its connection with carcinoma.</profile>
</query>

<topic>
  <id>qtest2014.33</id>
  <discharge_summary>16994-022078-DISCHARGE_SUMMARY.txt</discharge_summary>
  <title>Repiratory failure and CHF</title>
  <desc>What are the connections between respiratory failure and CHF? </desc>
  <narr>Relevant documents should contain information about respiratory
    failure and CHF.</narr>
  <profile>The patient is a 68 years old woman who has suffered from CHF for
    a long time. Now she was taken to the hospital because of respiratory failure
    and she wants to know about the connection between these two maladies.</profile>
</topic>
```

Fig. 8 Topics *qtest19* and *qtest2014.18* had the highest percentage of relevant documents (91.5 and 93.5%, respectively for 2013 and 2014). Topics *qtest8* and *qtest2014.33* were the ones with lowest percentage of relevant documents (only 1.4 and 4.8%, respectively for 2013 and 2014)

investigated, this is clearly not the case (cf. baseline run with baseline run with blind relevance feedback in the first two lines in Tables 6 and 7).

5 Lessons learned from the evaluation campaign in 2013 and 2014

In this section we gather together the lessons learned from the first two years of the CLEF eHealth IR task. We first detail the lessons learned from the participating teams systems; we then present the lessons learned from the analysis and evaluation of the campaign datasets.

Table 6 Experiments using automatic reassessment of documents with missing relevance information for reassessment strategy A and B (BL = baseline results)

Run ID	BL	A10	A20	A50	A100	A200	A500	B10
eHealth-bl.en_TFB	0.265	0.298	0.216	0.229	0.249	0.255	0.221	0.265
eHealth-bl.en_T	0.304	0.307	0.240	0.256	0.278	0.279	0.234	0.304
MEDINFO.1.3	0.313	0.309	0.244	0.271	0.291	0.294	0.247	0.313
QUT-TOPSIG.1.3	0.201	0.198	0.173	0.187	0.209	0.222	0.202	0.201
THCIB.6.3	0.116	0.108	0.032	0.011	0.006	0.003	0.001	0.116
KC&RA.1.3	0.267	0.265	0.218	0.244	0.266	0.270	0.230	0.267
Mayo.2.3	0.311	0.305	0.248	0.267	0.289	0.296	0.251	0.311
UTHealth.1.3	0.148	0.145	0.097	0.074	0.049	0.021	0.001	0.148
OHSU.5.3	0.100	0.096	0.075	0.069	0.055	0.036	0.017	0.100
teamAEHRC.5.3	0.273	0.267	0.206	0.223	0.239	0.244	0.219	0.273
UOG.Tr.1.3.res	0.244	0.243	0.203	0.229	0.253	0.262	0.227	0.244

Table 7 Experiments using automatic reassessment of documents with missing relevance information for reassessment strategy C (BL = baseline results)

Run ID	BL	C10	C20	C50	C100	C200	C500
eHealth-bl.en_TFB	0.255	0.294	0.207	0.187	0.175	0.163	0.148
eHealth-bl.en_T	0.304	0.307	0.235	0.215	0.196	0.181	0.166
MEDINFO.1.3	0.313	0.309	0.240	0.220	0.206	0.192	0.175
QUT-TOPSIG.1.3	0.201	0.198	0.163	0.145	0.136	0.123	0.114
THCIB.6.3	0.116	0.108	0.058	0.058	0.058	0.058	0.058
KC&RA.1.3	0.267	0.265	0.210	0.191	0.176	0.163	0.146
Mayo.2.3	0.311	0.305	0.245	0.220	0.206	0.191	0.173
UTHealth.1.3	0.148	0.145	0.103	0.093	0.085	0.080	0.075
OHSU.5.3	0.100	0.096	0.075	0.067	0.060	0.056	0.053
teamAEHRC.5.3	0.273	0.267	0.209	0.190	0.176	0.163	0.151
UOG.Tr.1.3.res	0.244	0.243	0.193	0.173	0.163	0.151	0.136

5.1 Lessons learned from the analysis of the runs and teams results

The analysis of the baseline runs submitted by the participants showed that IR using language modelling (LM) with Dirichlet Smoothing gave the best performance overall. This was observed in 2013 and 2014 on both the participating teams baselines and the task baselines. Moreover, observation of the submitted baselines confirms that it is essential for such IR evaluation tasks to provide strong baselines, both for the organizers (to set up a high-level competition) and for the participants.

The discharge summaries (DS) were used by 5 teams out of 9 in 2013 and 4 teams out of 14 in 2014. In 2013, only one team obtained their highest ranked run using the summaries (and even in this case it is not possible to attribute this result to the DS). All other participants obtained similar or lower results when using the DS. In 2014, nearly all teams obtained comparable results with and without the DS. In 2013, the disorders picked to build the queries were selected randomly from within the DS, while in 2014 the main one was selected. The only difference found in the results is that in 2013, the use of DS on average significantly degrades the results while it does not affect them in 2014. It appears that the DS does not bring useful contextual information to the IR task. However, this could be explained by the fact that the dataset itself might not be built properly for personalized/contextual IR: the queries are very short, and the relevance assessment only takes into account minimal elements of the DS.

All teams used external resources in their systems. These resources were used for three purposes: indexing, expansion and re-ranking. The resources are from two main types: corpora (document collection, IR benchmarks) and thesauri/lexicon. Most of the teams used external resources for the query expansion, while a few built concept-based indexes, and filtering or re-ranking systems that proved to give good results.

The participating teams expanded their queries based on two approaches: corpus-based approach (pseudo-relevance feedback) or concept-based approach. The corpus-based approach generally increased the results, especially when used with the same or a very similar document collection. The concept-based approach generally decreased the results in 2013, but increased them on average in 2014 (while the techniques were very close). Methods to filter or weight expansion terms appear to reduce the noise and improve the quality of the expansion. With the vocabulary gaps and the constantly evolving medical terminology, query expansion seems to be essential to medical IR.

Goeuriot et al. (2014b) conducted an analysis on the 2013 dataset and the impact of query complexity on IR performance. The complexity of a query is measured as the amount of specific medical entities it contains. They showed that performance is affected by the query complexity, and that some systems such as language models give better results on complex queries on average.

5.2 Lessons learned from the evaluation of the campaign datasets

We first assessed the dataset reliability, based on the generalizability theory and Kendall Tau correlation. We showed that the 2013 dataset was more reliable than 2014's, the reliability being based on the query set size, the mean effectiveness score and the variability.

The analysis of the relevance judgments showed that 2014 queries had a larger pool on average, which could be explained by the increased number of runs submitted, or the fact that 2 additional runs per team were pooled. We also observed an increase in the proportion of relevant documents in 2014. This could be explained by the fact that the queries are simpler (confirmed by the general higher performance of participant runs in 2014) or that the assessors were less stringent. It is very difficult to achieve the exact same settings from one year to another to build a dataset, which leads to such variations. While they can hardly be avoided, they also make the task more interesting, but ultimately we would hope to see consistent trends in terms of behavior of retrieval algorithms from year to year.

Our 2013 experiment on the effect of medical expertise on relevance judgment showed that the major disagreements were due to the nature and definition of relevance rather than the medical content, and that clearer guidelines could lead to better agreement. The results of this experiment were used to design better guidelines for the relevance judgment within the 2014 evaluation campaign. Moreover, this task targets patients, so it is arguably feasible for the relevance judgments to be made by IR experts, whose medical knowledge is close to patients. This wouldn't be the case with more specialized tasks such as clinical or biomedical tasks.

The analysis of the effect of the relevance of non-assessed documents on 2013 participants' results was conducted by observing the results of the participants when the unassessed documents were considered as non-relevant (standard approach), relevant (worst-case scenario) or partly relevant (a subset). For the three approaches, we observed only minor changes in the results. One could wonder if the amount of unjudged documents could have a negative effect on methods such as query expansion, aimed at finding additional relevant documents. However, this experiment shows that unjudged documents might not be the only reason that these methods degrade performance.

We did not perform an in depth analysis of assessors behavior for the CLEF 2013 and 2014 collections. A follow up study from Palotti et al. (2016) that considered the CLEF 2015 collection (a newer collection stemming from the lessons learned from the CLEF 2013 and 2014 collections investigated here) further analyzed the relationship between assessor type (expert vs. layperson), payment levels (paid vs. unpaid), query variations and relevance dimensions (topicality and understandability). They also indicated the impact of these aspects on system evaluation in the presence of disagreements across assessments obtained in the different settings.

6 Conclusions

This article provides a detailed analysis of the datasets and results of the 2013 and 2014 CLEF eHealth medical IR tasks. These evaluation tasks sought to contribute to the better understanding of the effective evaluation of medical IR for tasks based on realistic user information needs, and also to the identification of the current and potential contributions of established IR methods and medically relevant knowledge resources to effective medical IR techniques. As such, the article provides insights into IR techniques tuned to the support of laypeople in their quest for health advice on the web. In addition, the article documents and crystallizes the methodologies used for the creation of the two CLEF eHealth IR 2013 and 2014 datasets, the main results obtained and the main findings that emerged.

We conducted an analysis of the participating teams results from different angles: which systems are applied as baselines; how discharge summaries are integrated in the systems; which external resources are used; and if query expansion is integrated and how. Then, we analyzed the task dataset in the light of the evaluation campaign: the reliability of the datasets; the relevance of documents across queries and across datasets; the impact of medical expertise on the relevance assessment conducted and its quality; and the effect of the relevance of non-assessed documents on the participants results. This twofold study of the evaluation campaign over two years provides insight on technical aspects of medical IR, such as the effectiveness of query expansion; about the quality and characteristics of the CLEF eHealth IR datasets, such as their reliability; and about other aspects of running an IR evaluation campaign in the medical domain.

References

Alsulmi, M. & Carterette, B. (2016). Improving clinical case search using semantic based query reformulations. In *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 694–698). IEEE.

Armstrong, T. G., Moffat, A., Webber, W., & Zobel, J. (2009). Improvements that don't add up: Ad-hoc retrieval results since 1998. In *CIKM 2009* (pp. 601–610). ACM.

Barajas, K. C., & Akella, R. (2013). Incorporating statistical topic models in the retrieval of health care documents. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Bedrick, S., & Sheikhshabbafghi, G. (2013). Lucene, MetaMap, and language modeling: OHSU at CLEF eHealth 2013. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Benigeri, M., & Pluye, P. (2003). Shortcomings of health information on the internet. *Health Promotion International*, *18*(4), 381386.

Chappell, T., & Geva, S. (2013). Working notes for TopSig at ShARe/CLEF eHealth 2013. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Choi, S., & Choi, J. (2013). SNUMedinfo at CLEFeHealth2013 task 3. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Choi, S., & Choi, J. (2014). Exploring effective information retrieval technique for the medical web documents: SNUMedinfo at CLEFeHealth2014 Task 3. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Claveau, V. (2012). Unsupervised and semi-supervised morphological analysis for information retrieval in the biomedical domain. In *COLING, 2012* (pp. 629–645).

Claveau, V., Hamon, T., Grabar, N., & Le Maguer, S. (2014). RePaLi participation to CLEF eHealth IR challenge 2014: Leveraging term variation. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Demner-Fushman, D., & Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, *33*(1), 63–103.

Dramé, K., Mougin, F., & Diallo, G. (2014). Query expansion using external resources for improving information retrieval in the biomedical domain. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Goeuriot, L., Chapman, W., Jones, G. J. F., Kelly, L., Leveling, J., & Salanterä, S. (2014a). Building realistic potential patients queries for medical information retrieval evaluation. In *Proceedings of the LREC workshop on building and evaluating resources for health and biomedical text processing*.

Goeuriot, L., Jones, G. J. F., Kelly, L., Leveling, J., Hanbury, A., Müller, H., et al. (2013a). ShARe/CLEF eHealth Evaluation Lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *CLEF online working notes*.

Goeuriot, L., Kelly, L., & Leveling, J. (2014b). An analysis of query difficulty for information retrieval in the medical domain. In *Proceedings of the ACM special interest group on information retrieval conference (SIGIR 2014)*.

Goeuriot, L., Kelly, L., Hanlen, L., Suominen, H., Névéol, A., Palotti, J., et al. (2015). Overview of the CLEF eHealth Evaluation Lab 2015. In *Proceedings of CLEF 2015*.

Goeuriot, L., Kelly, L., Jones, G. J. F., Zuccon, G., Suominen, H., Hanbury, A., et al. (2013b). Creation of a new evaluation benchmark for information retrieval targeting patient information needs. In R. Song, W. Webber, N. Kando, & K. Kishida (Eds.), *Proceedings of the 5th international workshop on evaluating information access (EVIA), a satellite workshop of the NTCIR-10 conference*, Tokyo/Fukuoka, Japan, National Institute of Informatics/Kijima Printing.

Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., et al. (2014c). ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In *CLEF online working notes*.

Hanbury, A., & Müller, H. (2012). Khresmoi—Multimodal multilingual medical information search. In *Proceedings of medical informatics Europe 2012 (MIE 2012), Village of the Future*.

Hansen, D. L., Derry, H. A., Resnick, P. J., & Richardson, C. R. (2003). Adolescents searching for health information on the internet: An observational study. *Journal of Medical Internet Research*, *5*(4), e25. https://doi.org/10.2196/jmir.5.4.e25.

Harman, D., & Buckley, C. (2004). The NRRC reliable information access (RIA) workshop. In *SIGIR 2004* (pp. 528–529). ACM.

Hauff, C., de Jong, F., Kelly, D., & Azzopardi, L. (2010). Query quality: User ratings and system predictions. In *SIGIR '10* (pp. 743–744). New York, NY: ACM.

He, B., & Ounis, I. (2006). Query performance prediction. *Information Systems*, *31*(7), 585–594.

Hersh, W. R., Buckley, C., Leone, T.J., & Hickam, D. H. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of SIGIR '94*, (pp. 192–201).

Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A. G. S., & Tsikrika, T. (2011). The CLEF 2011 medical image retrieval and classification tasks. In *Working notes of CLEF 2011 (Cross Language Evaluation Forum)*.

Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., & Zuccon, G. (2016). Overview of the CLEF eHealth Evaluation Lab 2016. In *Proceedings of CLEF 2016*.

Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D. L., et al. (2014). Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In *Proceedings of CLEF 2014*.

Koopman, B., & Zuccon, G. (2014). Why assessing relevance in medical IR is demanding. In *Proceedings of medical information retrieval (MedIR) workshop (SIGIR)*.

Koopman, B., & Zuccon, G. (2016). A test collection for matching patients to clinical trials. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*, SIGIR '16 (pp. 669–672). New York, NY: ACM.

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., & Lawley, M. (2012). An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proceedings of CIKM 2012*.

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., & Lawley, M. (2016). Information retrieval as semantic inference: A graph inference model applied to medical search. *Information Retrieval Journal*, *19*(1–2), 6–37.

Ksentini, N., Tmar, M., & Gargouri, F. (2014). Miracl at CLEF 2014: eHealth information retrieval task. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Leveling, J., Goeuriot, L., Kelly, L., & Jones, G. J. F. (2012). DCU@TRECMed 2012: Using ad-hoc baselines for domain-specific retrieval. In *Proceedings of TREC 2012*. NIST.

Limsopatham, N., Macdonald, C., & Ounis, I. (2013a). Inferring conceptual relationships to improve medical records search. In *Proceedings of the 10th conference on open research areas in information retrieval* (pp. 1–8).

Limsopatham, N., Macdonald, C., & Ounis, I. (2013b). University of Glasgow at CLEF 2013: Experiments in eHealth Task 3 with Terrier. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Lin, J. (2005). Evaluation of resources for question answering evaluation. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2005)* (pp. 392–399).

Liu, X., Nie, J.-Y., & Sordoni, A. (2016). Constraining word embeddings by prior knowledge—Application to medical information retrieval. In *Asia information retrieval symposium* (pp. 155–167). Springer.

Malagon, J. M. C., & López, M. M. (2014). Laberinto at ShARe/CLEF eHealth Evaluation Lab. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Müller, H., Clough, P., Deselaers, T., & Caputo, B. (Eds.). (2016). *ImageCLEF—Experimental evaluation in visual information retrieval* (Vol. 32)., The information retrieval series Berlin: Springer.

Oh, H.-S., & Jung, Y. (2014). A multiple-stage approach to re-ranking clinical documents. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Ozturkmenoglu, O., Alpkocak, A., & Kilinc, D. (2014). Demir at CLEF eHealth: The effects of selective query expansion to information retrieval. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Palotti, J., Zuccon, G., Bernhardt, J., Hanbury, A., & Goeuriot, L. (2016). Assessors agreement: A case study across assessor type, payment levels, query variations and relevance dimensions. In *International conference of the cross-language evaluation forum for European languages* (pp. 40–53). Springer.

Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G. J., et al. (2015). CLEF eHealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *Proceedings of CLEF eHealth Evaluation Lab*.

Pecina, P., Dušek, O., Goeuriot, L., Haji, J., Hlaváčová, J., Jones, G. J. F., et al. (2014). Adaptation of machine translation for multilingual information retrieval in the medical domain. *Journal of Artificial Intelligence in Medicine, Special Issue on Health Document Text Mining and Information*, *61*(3), 165–185.

Roberts, K., Simpson, M. S., Voorhees, E., & Hersh, W. R. (2015). Overview of the TREC 2015 clinical decision support track. In *Proceedings of TREC*.

Roberts, P. M., Cohen, A. M., & Hersh, W. R. (2009). Tasks, topics and relevance judging for the TREC Genomics Track: Five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval*, *12*, 81–97.

Robertson, S. E., & Jones, K. S. (1994). *Simple, proven approaches to text retrieval*. Technical report 356, University of Cambridge.

Sakai, T., & Mitamura, T. (2010). Boiling down information retrieval test collections. In *RIAO 2010* (pp. 49–56). CID.

Saleh, S., & Pecina, P. (2014). Cuni at the ShARe/CLEF eHealth Evaluation Lab 2014. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Shen, W., & Nie, J.-Y. (2015). Is concept mapping useful for biomedical information retrieval? In *International conference of the cross-language evaluation forum for European languages* (pp. 281–286). Springer.

Shen, W., Nie, J.-Y., Liu, X., & Liui, X. (2014). An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM@CLEF2014eHealthTask 3. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Simpson, M. S., Voorhees, E. M., & Hersh, W. (2014). Overview of the TREC 2014 clinical decision support track. In *Proceedings of TREC*.

Soboroff, I. (2009). A guide to the ria workshop data archive. *Information Retrieval*, *12*(6), 642–651.

Soldaini, L., Cohan, A., Yates, A., Goharian, N., & Frieder, O. (2015). Retrieving medical literature for clinical decision support. In *European conference on information retrieval* (pp. 538–549). Springer.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., et al. (2013). ShARe/CLEF eHealth Evaluation Lab 2013: Three shared tasks on natural language processing and machine learning to make clinical reports easier to understand for patients. In *CLEF 2013*. Lecture notes in computer science (LNCS). Springer.

Thakkar, H., Iyer, G., Shah, K., & Majumder, P. (2014). Team IRLabDAIICT at ShARe/CLEF eHealth 2014 Task 3: User-centered information retrieval system for clinical documents. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Thesprasith, O., & Jaruskulchai, C. (2014). CSKU GPRF-QE for medical topic web retrieval. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Tibi, O., Thuma, E., & Mosweunyane, G. (2017). Selective collection enrichment in user-centred health information retrieval. In *2017 1st international conference on next generation computing applications (NextComp)* (pp. 175–181). IEEE.

Urbano, J., Marrero, M., & Martín, D. (2013). On the measurement of test collection reliability. In *SIGIR '13* (pp. 393–402). ACM.

Verberne, S. (2014). A language-modelling approach to user-centred health information retrieval. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Voorhees, E. M., & Hersh, W. (2012). Overview of the TREC 2012 medical records track. In *TREC 2012*. NIST.

Voorhees, E. M., & Tong, R. M. (2011). Overview of the TREC 2011 medical records track. In *Proceedings of TREC*. NIST.

Voorhees, E. M. (2005). The TREC robust retrieval track. *SIGIR Forum*, *39*(1), 11–20.

White, R., & Horvitz, E. (2008). *Cyberchondria: Studies of the escalation of medical concerns in web search*. Technical report, Microsoft Research.

Wu, J., & Huang, J. (2014). York University at CLEF eHealth 2014: A learning-to-rank approach for medical document retrieval. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Xia, Y., Xie, Z., Zhang, Q., Wang, H., & Zhao, H. (2014). *Cannabis*_TREATS_cancer: Incorporating fine-grained ontological relations in medical document ranking. In *Natural language processing and Chinese computing* (pp. 275–285). Springer.

Yang, C., Bhattacharya, S., & Srinivasan, P. (2014). The University of Iowa at CLEF 2014: eHealth Task 3. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Zhang, Y., Cohen, T., Jiang, M., Tang, B., & Xu, H. (2013). Evaluation of vector space models for medical disorders information retrieval. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Zhong, X., Xia, Y., Xie, Z., Na, S., Hu, Q., & Huang, Y. (2013). Concept-based medical document retrieval: THCIB at CLEF eHealth lab 2013 task 3. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Zhu, D., & Carterette, B. (2012). Improving health records search using multiple query expansion collections. In *2012 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 1–7). IEEE.

Zhu, D., Wu, S., James, M., Carterette, B., & Liu, H. (2013). Using discharge summaries to improve information retrieval in clinical domain. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Zhu, D., Stephen, W., Carterette, B., & Liu, H. (2014). Using large clinical corpora for query expansion in text-based cohort identification. *Journal of Biomedical Informatics*, *49*, 275–281.

Zuccon, G., & Koopman, B. (2018). Choices in knowledge-base retrieval for consumer health search. In *ECIR'18*.

Zuccon, G., Koopman, B., & Nguyen, A. (2013). Retrieval of health advice on the web: AEHRC at ShARe/CLEF eHealth evaluation lab task 3. In *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Zuccon, G., Koopman, B., & Palotti, J. (2015). Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In *Advances in information retrieval* (pp. 562–567).

Zuccon, G., Koopman, B., Bruza, P., & Azzopardi, L. (2015). Integrating and evaluating neural word embeddings in information retrieval. In *Proceedings of the 20th Australasian document computing symposium* (p. 12). ACM.

Zuccon, G., Koopman, B., Nguyen, A., Vickers, D., & Butt, L. (2012). Exploiting medical hierarchies for concept-based information retrieval. In *Proceedings of the seventeenth Australasian document computing symposium* (pp. 111–114). ACM.

Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L., Lupu, M., Pecina, P., et al. (2016). The IR task at the CLEF eHealth Evaluation Lab 2016: User-centred health information retrieval. In *Proceedings of CLEF*.