Overview of the CLEF eHealth 2020 Task 2: Consumer Health Search with ad-hoc and Spoken Queries*

Lorraine Goeuriot[1 [0000−0001−7491−1980]],
Hanna Suominen[2,3,4 [0000−0002−4195−1641]], Liadh Kelly[5 [0000−0003−1131−5238]],
Zhengyang Liu[2], Gabriella Pasi[6 [0000−0002−6080−8170]],
Gabriela Gonzalez Saez[1 [0000−0003−0878−5263]],
Marco Viviani[6 [0000−0002−2274−9050]], and Chenchen Xu[2,3]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France,
Lorraine.Goeuriot@imag.fr,
gabriela-nicole.gonzalez-saez@univ-grenoble-alpes.fr
[2] The Australian National University, Canberra, ACT, Australia, {hanna.suominen,
zhengyang.liu, chenchen.xu}@anu.edu.au
[3] Data61/Commonwealth Scientific and Industrial Research Organisation,
Canberra, ACT, Australia
[4] University of Turku, Turku, Finland
[5] Maynooth University, Ireland, liadh.kelly@mu.ie
[6] University of Milano-Bicocca, Dept. of Informatics, Systems, and Communication,
Milan, Italy, {gabriella.pasi, marco.viviani}@unimib.it

Abstract. In this paper, we provide an overview of the CLEF eHealth Task 2 on *Information Retrieval* (IR), organized as part of the eighth annual edition of the CLEF eHealth evaluation lab by the Conference and Labs of the Evaluation Forum. Its aim was to address laypeople's difficulties in retrieving and digesting valid and relevant information, in their preferred language, to make health-centred decisions. The task was a novel extension of the most popular and established task in CLEF eHealth on *Consumer Health Search* (CHS), which makes responses to spoken *ad-hoc* queries. In total, five submissions were made to its two subtasks; three addressed the *ad-hoc* IR task on text data and two considered the spoken queries. Herein, we describe the resources created for the task and evaluation methodology adopted. We also summarize lab submissions and results. As in previous years, organizers have made data, methods, and tools associated with the lab tasks available for future research and development.

* With equal contribution, LG & HS were the co-leaders of the task. LK, ZL, GP, GGS, MV, and CX were the task co-organizers and contributors to the evaluation conceptualization, dataset creation, assessments, and measurements.

Keywords: eHealth · Evaluation · Health Records · Medical Informatics · Information Storage and Retrieval · Speech Recognition · Test-set Generation · Self-Diagnosis · Dimensions of Relevance

1 Introduction

In recent years, *electronic health* (eHealth) content has become available in a variety of forms ranging from patient records and medical dossiers, scientific publications, and health-related websites to medical-related topics shared across social networks. Laypeople, clinicians, and policy-makers need to easily retrieve, and make sense of such medical content to support their decision making. The increasing difficulties experienced by these stakeholders in retrieving and digesting valid and relevant information in their preferred language to make health-centred decisions has motivated CLEF eHealth to organise yearly shared challenges since 2013.

More specifically, CLEF eHealth[7] was established as a lab workshop in 2012 as part of the *Conference and Labs of the Evaluation Forum* (CLEF). Since 2013, it has offered evaluation labs in the fields of layperson and professional health information extraction, management, and retrieval with the aims of bringing together researchers working on related information access topics and providing them with datasets to work with and validate the outcomes. These labs and their subsequent workshops target:

1. developing processing methods and resources (e.g., dictionaries, abbreviation mappings, and data with model solutions for method development and evaluation) in a multilingual setting to enrich difficult-to-understand eHealth texts, support personalized reliable access to medical information, and provide valuable documentation;
2. developing an evaluation setting and releasing evaluation results for these methods and resources;
3. contributing to participants and organizers' professional networks and interaction with all interdisciplinary actors of the ecosystem for producing, processing, and consuming eHealth information.

The vision for the Lab is two-fold:

1. to develop tasks that potentially impact laypeople's understanding of medical information, and
2. to provide the community with an increasingly sophisticated dataset of clinical narrative, enriched with links to standard knowledge bases, evidence-based care guidelines, systematic reviews, and other further information, to advance the state-of-the-art in multilingual *information extraction* (IE) and *information retrieval* (IR) in healthcare.

[7] https://clefehealth.imag.fr/

The eighth annual CLEF eHealth lab, CLEF eHealth 2020 [13], aiming to build upon the resource development and evaluation approaches offered in the previous years of the lab [53, 20, 11, 19, 12, 51, 21], offered the following two tasks:

- *Task 1.* Multilingual IE [28] and
- *Task 2. Consumer Health Search* (CHS).

The CHS task was a continuation of the previous CLEF eHealth IR tasks that ran in 2013, 2014, 2015, 2016, 2017 and 2018 [8, 10, 33, 62, 34, 17, 9], and embraced the *Text REtrieval Conference* (TREC) – style evaluation process, with a shared collection of documents and queries, the contribution of runs from participants and the subsequent formation of relevance assessments and evaluation of participants submissions. The 2020 task used the representative web corpus developed in the 2018 challenge. This year we offered spoken queries, as well as textual transcripts of these queries. The task was structured into a two optional subtasks, covering (1) ad-hoc searchand (2) query variation using the spoken queries, textual transcripts of the spoken queries, or provided automatic speech-to-text conversions of the spoken queries.

The multilingual IE task focused on Spanish. Further details on this challenge are available in [13] and [28].

The remainder of this paper is structured as follows: First, in Section 2, we detail the task, evaluation, and datasets created. Second, in Section 3, we describe the task submissions and results. Finally, in Section 4, we discuss the study and provide conclusions.

2 Materials and Methods

In this section, we describe the materials and methods used in the two subtasks. After specifying our document collection, we address the spoken, transcribed, and speech recognized queries. Then, we describe our evaluation methods. Finally, we introduce our human relevance assessments for information topicality, understandability, and credibility.

2.1 Documents

The 2018 CLEF eHealth Consumer Health Search document collection was used in this year's IR challenge. As detailed in [17], this collection consists of web pages acquired from the CommonCrawl.

An initial list of websites was identified for acquisition. The list was built by submitting queries on the 2018/2020 topics to the Microsoft Bing *Application Programming Interfaces* (APIs), through the Azure Cognitive Services, repeatedly over a period of a few weeks, and acquiring the *uniform resource locators* (URLs) of the retrieved results. The domains of the URLs were then included in the list, except some domains that were excluded for decency reasons.

The list was further augmented by including a number of known reliable health websites and other known unreliable health websites. This augmentation was based on lists previously compiled by health institutions and agencies.

2.2 Queries

Historically, the CLEF eHealth IR task has released text queries representative of layperson medical information needs in various scenarios. In recent years, query variations issued by multiple laypeople for the same information need have been offered. In this year's task, we extended this to also include spoken queries.

These spoken queries were generated by six laypeople in English. All native English speakers. Efforts were made to include a diverse set of accents. Narratives for query generation were those used in the 2018 challenge. These narratives relate to real medical queries compiled by the Khresmoi project [15] which were issued to a search engine by laypeople; full details are available in the CLEF eHealth 2018 IR task overview paper [17]. Spoken transcripts of these narratives were generated for use in query generation in this year's challenge.

To create a spoken query the layperson listened to the narrative; and generated their spoken query associated with the narrative. The layperson then listened to their generated spoken query and created a textual transcript of the query. To ensure accuracy in transcription, they were required to repeat this process of listening to their narrative and textually transcribing it. This allowed us to generate accurate transcripts of the spoken queries. We did not preprocess the textual transcripts of queries; for example, any spelling mistakes that may be present were not removed.The final generated query set consisted of 50 topics, with 6 query variations for each topic.

Ethical approval was obtained to generate the spoken queries, and informed consent obtained from study participants. Spoken queries were downloadable from a secured server for the purpose of participating in this year's CLEF eHealth IR challenge, on completion of a signed use agreement by the participating team.

We also provided participants with the textual transcripts of these spoken queries and automatic speech-to-text translations. This transcription of the audio files was generated using the End-to-End Speech Processing Toolkit (ESPNET), Librispeech, CommonVoice, and Google API (with three models). Speech recognition is assessed using Kaldi [40], an open-source speech recognition toolkit distributed under a free license. We use mel-frequency cepstral coefficient (MFCC) acoustic features (13 coefficients expanded with delta and double delta features and energy : 40 features) with various feature transformations including linear discriminant analysis (LDA), maximum likelihood linear transformation (MLLT), and feature space maximum likelihood linear regression (fMLLR) with speaker adaptive training (SAT).

The speech transcription process is carried out in two passes: an automatic transcript is generated with a GMM-HMM model of 12000 states and 200000 Gaussians. The second pass is performed using DNN (nnet3 recipe in kaldi toolkit) acoustic model trained on acoustic features normalized with the fMLLR matrix.

TEDLIUM dataset [42] was used for training acoustic models. It was developed for large vocabulary continuous speech recognition (LVCSR). The train part of the dataset is composed 118 hours of speech.

The English language model is trained with MIT language model toolkit using following corpora : News commentary 2007-2012 [55], Gigaword version 5 [14], TDT 2-4 [57]. The vocabulary size is 150K based on most frequent words.

2.3 Evaluation Methods

Similar to the 2016, 2017, and 2018 pools, we created the pool using the RBP-based Method A (Summing contributions) by Moffat et al. [29], in which documents are weighted according to their overall contribution to the effectiveness evaluation as provided by the RBP formula (with $p = 0.8$, following Park and Zhang [35]). This strategy, named RBPA, was chosen because it was shown that it should be preferred over traditional fixed-depth or stratified pooling when deciding upon the pooling strategy to be used to evaluate systems under fixed assessment budget constraints [24], as it is the case for this task.

Since the 2018 topics were used in 2020,the pool used in the 2020 CHS task was an extension of 2018's pool. In other words, the merged 2018&2020 pool was used in 2020. For Subtasks 1 and 2, participants could submit up to 4 runs in the TREC format. Evaluation measures were NDCG@10, BPref, and RBP. Metrics such as uRBP were used to capture various relevance dimensions, as elaborated below.

2.4 Human Assessments for Topicality, Understandability, and Credibility

Relevance assessments were conducted on three relevance dimensions: topicality, understandability, and credibility. Topicality referred to a classical relevance dimension ensuring that the document and the query are on the same topic and the document answers the query. Understandability was an estimation of whether the document is understandable by a patient. In assessment guidelines, assessors were required to estimate how readable they think the documents were to a layperson, that is, a person without any medical background. Topicality and understandability have been used as relevance dimensions in the CHS task of CLEF eHealth for several years.

This year, to take into consideration the phenomenon of the spread of disinformation online (especially on health-related topics), we introduced a novel dimension, that is, *credibility*. Over the years, the interest in studying the concept of credibility has gradually moved from *traditional communication environments*, characterized by interpersonal and persuasive communication, to *mass communication* and *interactive-mediated communication*, with particular reference to online communication [26]. In this scenario, retrieving credible information is becoming a fundamental issue [18, 23, 38, 39, 59], also in the health-related context [44].

In general, credibility is described in the literature as a perceived quality of the information receiver [6], and it is composed of *multiple dimensions* that have to be considered and evaluated together in the process of information credibility

assessment [6, 27, 45]. These dimensions usually include the *source* that disseminates content, characteristics related to the *message* diffused, and some *social aspects* if the information is disseminated through a virtual community [56].

For this reason, when evaluating information credibility in the health-related context, assessors were asked in the CLEF eHealth 2020 Task 2 to evaluate the aforementioned multiple aspects by considering, at the same time:

1. any information available about the *trustworthiness* of the source [2, 3, 25] of the health-related information (the fact that information comes from a Web site with a good or bad *reputation*, or the level of *expertise* of an individual answering on a blog or a question-answering system, etc.);
2. *syntactic/semantic characteristics* of the content [7] (in terms of, e.g., completeness, language register, or style); and
3. any information emerging from *social interactions* when available [37] (the fact that the circle of social relationships of the author of a content is reliable or not, the fact that the author is involved in many discussions, etc.).

Obviously, it must be taken into account that the ability to judge the credibility of information related to health depends very much on the sociocultural background of the assessor, on the availability of information about the social network of the source of information, and on the ease versus complexity of identifying in or inferring from the document the different dimensions of credibility.

Assessors considered the three dimensions in assessments (i.e., topicality, understandability, and credibility) on a 3-levels scale:

- *not relevant/understandable/credible*,
- *somewhat relevant/understandable/credible*, and
- *highly relevant/understandable/credible*.

In particular, we added a 4th option for credibility for assessors uncertainty: *I am not able to judge*. This was motivated by the fact that, as illustrated above, the documents to be assessed may actually lack (or it may not be entirely clear) the minimum information necessary to assess their level of credibility.

Assessments were implemented online by expanding and customising the Relevation! tool for relevance assessments [22] to capture our task dimensions, scales, and other preferences. The number of assessors was 30, of which about 12 women (40%) and 18 men (60%). They were based in European countries and Australia. Their expertise ranged from being a medical doctor (in different disciplines) or a nurse to being a layperson with no or limited background in medicine or healthcare. Each assessor was assigned 1 to 5 queries to be evaluated. Each query (concerning a specific domain linked to health) was associated with 150 documents to be evaluated with respect to the three dimensions of relevance mentioned above.

3 Results

CLEF eHealth IR/CHS tasks offered in 2013–2020 have brought together researchers working on health information access topics. The tasks have provided

them with data and computational resources to work with and validate their outcomes. These contributions have accelerated pathways from scientific ideas through influencing research and development to societal impact. The task niche has been addressing health information needs of laypeople (including, but not limited to, patients, their families, clinical staff, health scientists, and healthcare policy makers) — and not healthcare experts only — in a range of languages and modalities — in retrieving and digesting valid and relevant eHealth information to make health-centered decisions [4, 16, 5, 48, 49].

Next, we report on the 2020 participants, method submissions, and their resulting evaluation outcomes. This expands the brief results section of our lab overview [13].

3.1 Participation

In 2020, 24 teams registered to CLEF eHealth Task 2. Of these teams, 3 took part in the task. Registering for a CLEF task consisted of filling in a form on the CLEF conference website with contact information, and tick boxes corresponding to the labs of interest. This was done several months before run submission, which explains the drop in the numbers.

Also, the task was difficult and demanding which is another explanation for the drop.

Although the interest and participation numbers were considerably smaller than before [48–50], organizers were pleased with this newly introduced task, with its novel spoken queries element attracting interest and submissions (Table 1). In addition, every participating team took the offered more-traditional *ad-hoc* task.

3.2 Participants' Method Submissions

Among five submissions to the 2020 CLEF eHealth Task 2, the *ad-hoc* IR subtask was the most popular with its three submissions; the subtasks that used transcriptions of the spoken queries and the original audio files received one submission each. Specifically, the subtask that used transcriptions of the spoken queries had one submission and the subtask where the original audio files were processed had one submission. The submitting teams were from Australia, France, and Italy and had 4, 5, and 3 team members, respectively.[8] Each team had members from a single university without other partner organizations.

The Italian submission to the *ad-hoc* search and spoken queries using transcription subtasks was from the *Information Management System* (IMS) Group of the University of Padua [32]. Its members were Associate Professor Giorgio Maria Di Nunzio, Stefano Marchesin, and Federica Vezzani. The submission to the former task included BM25 of the original query; Reciprocal Rank fusion with BM25, *Query Language Model* (QLM), and *Divergence from Randomness*

[8] Please note that these numbers are corrected from [13], based on the team's finalised working notes.

Table 1. Descriptive statistics about the submitting teams

Subtasks	No. of coauthors	Authors' affliction	Affiliation country
ad-hoc Search & Spoken Queries Using Transcriptions	3	1 university	Italy
ad-hoc Search & Spoken Queries Using Audio Files	5	1 university	France
ad-hoc Search	4	1 university	Australia

(DFR) approaches. Reciprocal Rank fusion with BM25, QLM, and DFR approaches using pseudo relevance feedback with 10 documents and 10 terms (the query weight of 0.5); and Reciprocal rank fusion with BM25 run on manual variants of the query. The submission to the latter task included the Reciprocal Rank fusion with BM25; Reciprocal Rank fusion with BM25 using pseudo relevance feedback with 10 documents and 10 terms (the query weight of 0.5); Reciprocal Rank fusion of BM25 with all transcriptions; and Reciprocal Rank fusion of BM25 with all transcripts using pseudo relevance feedback with 10 documents and 10 terms (the query weight of 0.5).

The French team, LIG-Health, was formed by Dr Philippe Mulhem, Aidan Mannion, Gabriela Gonzalez Saez, Dr Didier Schwab, and Jibril Frej from the Laboratoire d'Informatique de Grenoble of the Univ. Grenoble Alpes [30]. To the *ad-hoc* search task, they submitted runs using Terrier BM25 as a baseline, and explored various expansion methods using UMLS, using the Consumer Health Vocabulary, expansion using Fast Text, and RF (bose-Einstein) weighted expansion. For the spoken queries subtask they used various transcriptions on the same models, opting for the best performing ones based on 2018 qrels. They submitted merged runs for each query.

The Australian team – called SandiDoc – was from the *Our Health In Our Hands* (OHIOH) Big Data program, Research School of Computer Science, College of Engineering and Computer Science The Australian National University [43]. Its members were Sandaru Seneviratne, Dr Eleni Daskalaki, Dr Artem Lenskiy, and Dr Zakir Hossain. Differently from the other two teams, SandiDoc took part in the *ad-hoc* search task only. Its IR method had three steps and was founded on $TF \times IDF$ scoring: First, both the dataset and the queries were pre-processed. Second, TF×ID scores were computed for the queries and used to retrieve the most similar documents for the queries. Third, the team supplemented this method by working on the clefehealth2018_B dataset using the medical skip-gram word embeddings provided. To represent the documents and queries, the team used the average word vector representations as well as the average of minimum and maximum vector representations of the document or query. In documents, these representations were derived using the 100 most frequent words in a document. For each representation, the team calculated the similarity among documents and queries using the cosine measure to obtain the

final task results. The team's aim was to experiment with different vector representations for text.

3.3 Organizers' Baseline Methods

In addition to these participants' methods, we as organizers developed baseline methods that were based on the renowned OKapi BM25 but now with query expansion optimized in a REINFORCE [58] fashion. In this section, we introduce the two steps of query expansion and document retrieval. First, we pre-trained our query expansion model on the generally available corpora. Similar to the REINFORCE learning protocol as introduced in [58], this pre-training step was done by iterations of exploration trials and optimization of the current model by rewarding the explorations. Each time an input query was enriched by the query expansion model from the last iteration, and from where several candidate new queries were generated. The system was rewarded or penalized by matching the retrieved documents from these candidate queries against the ground truth document ranking. To expand the trial of generating new queries and thus provide more training sources, the system used the context words found in these newly retrieved documents to construct queries for the next iteration. For this baseline model, the same datasets from [31], TREC-CAR, Jeopardy, and MSA were used for pre-training.

One key challenge for information retrieval in the eHealth domain is that layperson may lack the professional knowledge to precisely describe medical topics or symptoms. A layperson's input query into the system can be lengthy and inaccurate, while the documents to be matched for these queries are usually composed by people of medical mind and background and thus rigorous in wording. The query expansion phase was added to increase the chance of matching more candidate documents by enriching the original query. With this intuition in mind, we employed similar candidate query construction method and optimization target as introduced in [31]. Given an original query q_0, the system retrieves several ranked documents set D_0. The system constructed new candidate query q_0' by selecting words from the union of words from the original query q_0 and context words from the retrieved documents D_0. The new query was fed back into the retrieval system to fetch documents D_0' as the learning source for the next iteration. The system iteratively apply the retrieval of documents and reformulation of new candidate queries to create the supervision examples $\{(q_0', D_0'), (q_1', D_1'), \ldots\}$. At each iteration, the system memorized the selection operation of words in constructing the new query as *actions* to be judged. The documents retrieved D_k' along with their ranking were then compared with the ground truth document ranking. Correct ranking of the documents becomes *reward* to the new query and thus also the *actions* that generate it. Particularly, the stochastic objective function for calculating the reward was:

$$C_a = (R - \bar{R}) \sum_{t \in T} - \log P\left(t \mid q_0\right),$$

where R and $\bar{R}$ are the reward from the new query and baseline reward, and $t \in T$ are words from the new query. With the *actions* and *reward* being properly defined, the system can be optimized under the REINFORCE learning framework [58]. At inference stage, the system will greedily peek the optimal selection operations to generate a few candidate queries from the input query.

After enriching the input queries by the pre-trained query expansion model, the second step of this baseline model reused the commonly-used BM25 algorithm [41].

3.4 Evaluating Topicality

Table 2 gives the global results of all the teams for the 2 subtasks. This table shows the task metrics (MAP, BPREF, NDCG, uRBP and cRBP) for all the participants runs, and the organizers baselines.

Table 2. Runs evaluation using the *MAP, BPref, nDCG@10, uRBP, cRBP* in the *ad-hoc* search task. Bold cells are the highest scores. Statistical significance tests were conducted but the highest participants runs were not statistically significantly higher than the best baselines.

ad-hoc Search Task						
Team	**Run**	**MAP**	**BPref**	**NDCG@10**	**uRBP**	**cRBP**
Baseline	Bing_all	0.0137	0.0164	0.4856	0.2433	0.3489
	elastic_BM25_QE_Rein	0.1758	0.3071	0.5782	0.3072	0.4542
	elastic_BM25f_noqe	*0.2707*	*0.4207*	***0.7197***	*0.3494*	***0.5424***
	elastic_BM25f_qe	0.1107	0.2110	0.5625	0.2711	0.4399
	indri_dirichlet_noqe	0.0790	0.1807	0.4104	0.1893	0.3299
	indri_dirichlet_qe	0.0475	0.1234	0.3235	0.1504	0.2649
	indri_okapi_noqe	0.1102	0.2227	0.4708	0.2180	0.3738
	indri_okapi_qe	0.1192	0.2394	0.4732	0.2109	0.3703
	indri_tfidf_noqe	0.1215	0.2396	0.4804	0.2139	0.3740
	indri_tfidf_qe	0.1189	0.2344	0.4824	0.2163	0.3799
	terrier_BM25_cli	0.2643	0.3923	0.4963	0.2298	0.3836
	terrier_BM25_gfi	0.2632	0.3922	0.4923	0.2345	0.3828
	terrier_BM25_noqe	0.2627	0.3964	0.5919	0.2993	0.4615
	terrier_BM25_qe	0.2453	0.3784	0.5698	0.2803	0.4516
	terrier_DirichletLM_noqe	0.2706	0.4160	0.6054	0.2927	0.4789
	terrier_DirichletLM_qe	0.1453	0.2719	0.5521	0.2662	0.4456
	terrier_TF_IDF_noqe	0.2613	0.3958	0.6292	0.3134	0.4845
	terrier_TF_IDF_qe	0.2500	0.3802	0.608	0.3013	0.4831
IMS	bm25_orig	0.2482	0.3909	0.6493	0.3204	0.5125
	original_rm3_rrf	**0.2834**	**0.4320**	0.6491	0.3141	0.5039
	original_rrf	0.2810	0.4232	0.6593	0.3225	0.5215
	variant_rrf	0.2022	0.3712	0.6339	**0.3511**	0.4863
LIG	FT_Straight.res	0.2318	0.3669	0.5617	0.2909	0.4555
	Noexp_Straight.res	0.2627	0.3964	0.5919	0.2993	0.4615
	UMLS_RF.res	0.2258	0.3616	0.5918	0.3123	0.4593
	UMLS_Straight.res	0.2340	0.3645	0.5769	0.3062	0.4614
SandiDoc	tfidf	0.0239	0.0536	0.3235	0.2413	0.2281

Table 4 shows the topical relevance results for all the teams and the organizers baselines. The team achieving the highest results on 2 metrics is IMS, although the scores are not statistically significantly higher than the best baseline. Their top run, $original\_rm3\_rrf$ run uses Reciprocal Rank fusion with BM25, QLM,

Table 3. Runs evaluation using the *MAP, BPref, nDCG@10, uRBP, cRBP* in the Spoken search task

ad-hoc Search Task						
Team	**Run**	**MAP**	**BPref**	**NDCG@10**	**uRBP**	**cRBP**
Best ad-hoc baseline	elastic_BM25f_noqe	0.2707	0.4207	0.7197	0.3494	0.5424
IMS	bm25_all_rrf	0.1952	0.3722	0.4478	**0.2981**	0.4622
	bm25_all_rrf_rm3	0.2144	0.3975	0.4711	0.2854	0.4348
	bm25_rrf	0.1962	0.3745	0.4553	0.2923	**0.4636**
	bm25_rrf_rm3	**0.2188**	**0.4036**	**0.4781**	0.2892	0.4493
LIG_Merged	FT_binary_RF	0.1626	0.2964	0.3627	0.2264	0.3592
	Noexp_RF	0.1810	0.3279	0.3963	0.2601	0.4067
	UMLS_binary_RF	0.1582	0.3054	0.3658	0.2557	0.3864
	UMLS_weight_RF	0.1671	0.3085	0.3721	0.2407	0.3722
LIG_Participant_1	DE_FT_binary_RF	0.1565	0.3096	0.3705	0.2441	0.3906
	DE_Noexp_RF	0.1726	0.3192	0.3853	0.2645	0.4051
	DE_UMLS_binary_RF	0.1271	0.2783	0.3238	0.2327	0.3435
	DE_UMLS_weight_RF	0.1416	0.2928	0.3473	0.2333	0.3582
LIG_Participant_2	DE_FT_binary_RF	0.0995	0.2314	0.2764	0.1970	0.2906
	DE_Noexp_RF	0.1206	0.2634	0.313	0.2066	0.3236
	DE_UMLS_binary_RF	0.1017	0.2384	0.2769	0.1998	0.2960
	DE_UMLS_weight_RF	0.1133	0.2575	0.3039	0.2313	0.3338
LIG_Participant_3	VE_FT_binary_RF	0.1274	0.2664	0.3159	0.2106	0.3176
	VE_Noexp_RF	0.1447	0.3023	0.356	0.2106	0.3389
	VE_UMLS_binary_RF	0.1385	0.2984	0.3451	0.2055	0.3164
	VE_UMLS_weight_RF	0.1485	0.3114	0.3637	0.2194	0.3412
LIG_Participant_4	PE_FT_binary_RF	0.1090	0.2582	0.2985	0.1908	0.2987
	PE_Noexp_RF	0.1301	0.2880	0.3382	0.2164	0.3375
	PE_UMLS_binary_RF	0.1246	0.2852	0.3285	0.2020	0.3061
	PE_UMLS_weight_RF	0.1282	0.2877	0.3348	0.2150	0.3317
LIG_Participant_5	PE_FT_binary_RF	0.0952	0.2287	0.2512	0.1424	0.2460
	PE_Noexp_RF	0.1035	0.2412	0.2679	0.1644	0.2765
	PE_UMLS_binary_RF	0.1036	0.2470	0.2741	0.1514	0.2670
	PE_UMLS_weight_RF	0.0917	0.2227	0.2466	0.1535	0.2615
LIG_Participant_6	DE_FT_binary_RF	0.1509	0.2921	0.3496	0.2016	0.3522
	DE_Noexp_RF	0.1744	0.3238	0.3849	0.2220	0.3724
	DE_UMLS_binary_RF	0.1478	0.3019	0.3499	0.2008	0.3233
	DE_UMLS_weight_RF	0.1594	0.3072	0.3618	0.2098	0.3411

DFR approaches using pseudo relevance feedback with 10 documents and 10 terms (query weight 0.5) and achieves 0.28 MAP and 0.43 BPref. The second best run for MAP and BPref is also from IMS, using the same ranking system without PRF. For NDCG, the organizers baseline using ElasticSearch BM25 without query expansion obtains higher results. Interstingly, we observe in that table that BM25 gives very good performances, but its various implementations can give very different results (MAP ranging from 0.11 to 0.26).

Since the ad-hoc task used the same topics and documents as 2018 but intended to extend 2018's pool, we compared teams results for all the ad-hoc task metrics in Figure 1. The figure shows that the extension of the pool had a relatively limited impact on the performances of each submitted systems, except for Bpref measure that shows a consistent decrease. Bpref, correlated to the average precision, is more robust to reduced pools [1] and penalises systems for ranking non-relevant documents above relevant ones. Therefore, this decrease can be explained by the fact that the extension of the pool contained relevant documents.

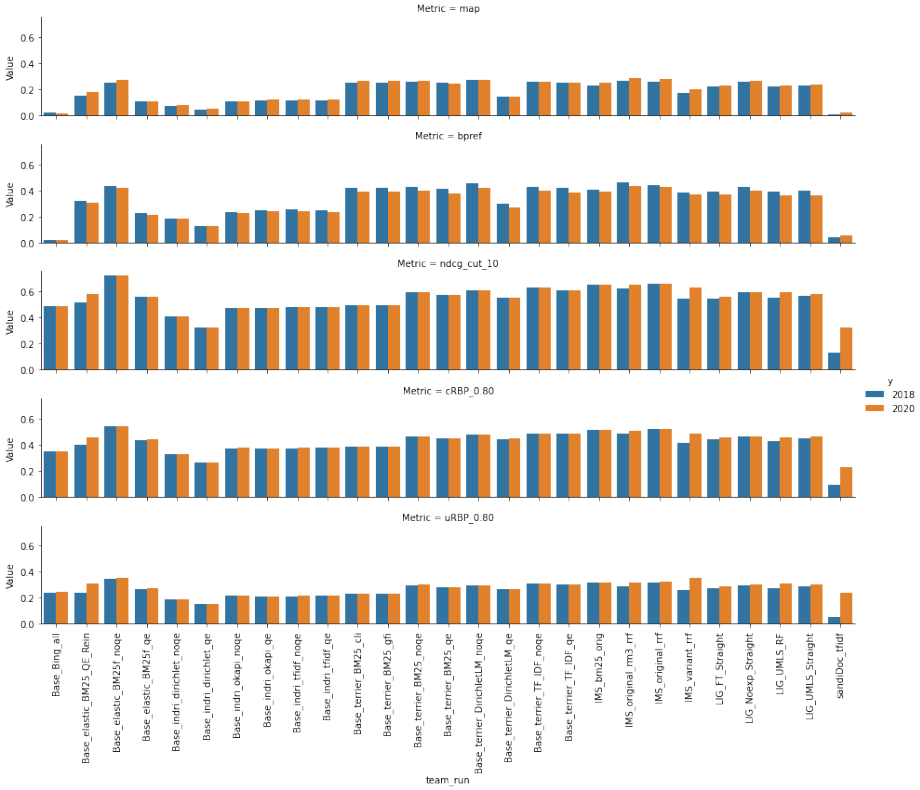

Fig. 1. Comparison of the task metrics with 2018 and 2020 pools. Blue and orange bars represent 2018 and 2018 results respectively

Table 3 shows the results of the two teams taking part in the spoken queries subtask. For comparison purpose, the first line gives the results of the best overall baseline, ElasticSearch BM25. All the participants used the provided transcription. Team IMS, who obtained the highest results, used reciprocal ranked fusion with BM25. They obtained the best results with Reciprocal Rank fusion with BM25 using pseudo relevance feedback with 10 documents and 10 terms (query weight 0.5). Team LIG submitted both merged runs for all participant speaker, using the transcription giving the highest performances. The merged runs gave better results and no expansion method but relevance feedback seemed to help retrieval. This table shows that retrieval from spoken queries is a challenging task and that transcription does not allow to achieve the same results than the original query.

Table 4. Ranking of topicality relevance using *MAP, Bpref and NDCG@10* in the *ad-hoc* search task

	ad-hoc Search Task					
	run	**MAP**	**run**	**BPref**	**run**	**NDCG**
1	IMS original_rm3_rrf	0.2834	IMS original_rm3_rrf	0.4320	Baseline elastic_BM25f_noqe	0.7197
2	IMS original_rrf	0.2810	IMS original_rrf	0.4232	IMS original_rrf	0.6593
3	Baseline elastic_BM25f_noqe	0.2707	Baseline elastic_BM25f_noqe	0.4207	IMS bm25_orig	0.6493
4	Baseline terrier_DirichletLM_noqe	0.2706	Baseline terrier_DirichletLM_noqe	0.4160	IMS original_rm3_rrf	0.6491
5	Baseline terrier_BM25_cli	0.2643	Baseline terrier_BM25_noqe	0.3964	IMS variant_rrf	0.6339
6	Baseline terrier_BM25_gfi	0.2632	LIG Noexp_Straight.res	0.3964	Baseline terrier_TF_IDF_noqe	0.6292
7	Baseline terrier_BM25_noqe	0.2627	Baseline terrier_TF_IDF_noqe	0.3958	Baseline terrier_TF_IDF_qe	0.608
8	LIG Noexp_Straight.res	0.2627	Baseline terrier_BM25_cli	0.3923	Baseline terrier_DirichletLM_noqe	0.6054
9	Baseline terrier_TF_IDF_noqe	0.2613	Baseline terrier_BM25_gfi	0.3922	Baseline terrier_BM25_noqe	0.5919
10	Baseline terrier_TF_IDF_qe	0.2500	IMS bm25_orig	0.3909	LIG Noexp_Straight.res	0.5919

3.5 Evaluating Understandability

The evaluation of understandability have been measure with *understandability-ranked biased precision* (uRBP) [61]. uRBP evaluate IR systems by taking into account both topicality and understandability dimensions of relevance.

Particularly, the function for calculating uRBP was:

$$uRBP = (1-\rho)\sum_{k=1}^{K}\rho^{k-1}r(d@k)\cdot u(d@k), \tag{1}$$

where $r(d@k)$ is the relevance of the document d at position k, $u(d@k)$ is the understandability value of the document d at position k, and the persistent parameter ρ models the user desire to examine every answer, which was set to 0.50, 0.80 and 0.95 to obtain three version of uRBP, according to different user behaviors.

The results for all the participants for understandability evaluation is shown in the second last column of Table 2. Table 5 shows the top ten runs submitted in the ad-hoc Task. The best run was obtained with Reciprocal rank fusion with BM25 run on manual variants of the query by team IMS. The BM25 baseline gives very close performances. The run ranking does not differ from the topicality evaluation. This could be due to the fact than none of the submitted runs included features specifically designed to assess understandability of the results.

These results have been obtained with the binary relevance assessment, and the graded understandability assessment, and `rbp_eval 0.5` as distributed by RBP group. For further details, please refer to `https://github.com/jsc/rbp_eval`.

Table 5. Understandability and Credibility ranking evaluations using the *uRBP and cRBP* in the search task

	ad-hoc Search Task			
	run	**uRBP**	**run**	**cRBP**
1	IMS variant_rrf	0.3511	Baseline elastic_BM25f_noqe	0.5424
2	Baseline elastic_BM25f_noqe	0.3494	IMS original_rrf	0.5215
3	IMS original_rrf	0.3225	IMS bm25_orig	0.5125
4	IMS bm25_orig	0.3204	IMS original_rm3_rrf	0.5039
5	IMS original_rm3_rrf	0.3141	IMS variant_rrf	0.4863
6	Baseline terrier_TF_IDF_noqe	0.3134	Baseline terrier_TF_IDF_noqe	0.4845
7	LIG UMLS_RF.res	0.3123	Baseline terrier_TF_IDF_qe	0.4831
8	Baseline elastic_BM25_QE_Rein	0.3072	Baseline terrier_DirichletLM_noqe	0.4789
9	LIG UMLS_Straight.res	0.3062	Baseline terrier_BM25_noqe	0.4615
10	Baseline terrier_TF_IDF_qe	0.3013	LIG Noexp_Straight.res	0.4615

3.6 Evaluating Credibility

In this section, we report the results produced for the two subtasks: *ad-hoc* search and spoken queries retrieval. In particular, for the *ad-hoc* subtask, only the *ad-hoc* IR subtask is considered (no speech recognition). For each subtask, the results of both baseline methods and team submission methods in the context of credibility assessment related to IR are reported and commented. The measures employed to assess the credibility of the tasks considered are detailed below.

In the literature, *accuracy* is certainly the measure that has been used most frequently to evaluate a classification task, and, as such, it has been usually employed to evaluate the effectiveness of credibility assessment. In fact, to date, the problem of assessing the credibility of information has mostly been approached as a binary classification problem (by identifying credible versus non-credible information) [56]. Some works have also proposed the computation of credibility values for each piece of information considered, by proposing users a credibility-based ranking of the considered information items, or by leaving to users the choice of trusting the information items based on these values [36, 56, 60]. Obviously, once an adequate threshold has been chosen, it would be possible to transform these approaches into approaches that produce a binary classification.

However, our purpose in asking assessors to evaluate the documents from the point of view of their credibility is to be able to generate IR Systems that can retrieve credible information, besides understandable and topically relevant. For this reason, in CLEF eHealth 2020 we have adapted the *understandability-ranked biased precision* (uRBP) illustrated in [61] to *credibility*, by employing the so-called cRBP measure. In this case the function for calculating cRBP is the same used to calculate uRBP (see Equation 1 in Section 3.5 , replacing $u(d@k)$ by the credibility value of the document d at position k, $c(d@k)$:

$$cRBP = (1-\rho)\sum_{k=1}^{K}\rho^{k-1}r(d@k)\cdot c(d@k), \tag{2}$$

As in uRBP the parameter ρ was set to three values, from impatient user (0.50) to more persistent users (0.80 and 0.95).

It is important to underline that, in defining IR approaches implemented by both the organizers and the three groups that submitted runs, no explicit reference was made to solutions for assessing the credibility of documents. Therefore, any potential increase in the evaluation figures must be considered purely coincidental.

Evaluation with Accuracy. The results illustrated in this section were obtained with a binary credibility assessment for 2020 and trustworthiness for 2018 data. A document assessed with a credibility/trustworthiness value $\geq 50\%$ was considered as credible. The accuracy of the credibility assessment was calculated over the top 100 documents retrieved for each query as follows:

$$acc(q) = \frac{\#credible\_retrieved\_docs\_top\_100(q)}{\#retrieved\_docs\_top\_100(q)}.$$

Table 6. Credibility evaluations using the *accuracy* (Acc) in the *ad-hoc* search task

ad-hoc Search Task							
Baseline	**Acc**	**IMS**	**Acc**	**LIG**	**Acc**	**SandiDoc**	**Acc**
Bing_all	0.84	bm25_orig	0.75	FT_Straight.res	0.75	tfidf	0.17
elastic_BM25_QE_Rein	0.63	original_rm3_rrf	0.74	Noexp_Straight.res	0.80		
elastic_BM25f_noqe.out	0.73	original_rrf	0.80	UMLS_RF.res	0.74		
elastic_BM25f_qe.out	0.53	variant_rrf	0.64	UMLS_Straight.res	0.75		
indri_dirichlet_noqe.out	0.67						
indri_dirichlet_qe.out	0.57						
indri_okapi_noqe.out	0.68						
indri_okapi_qe.out	0.63						
indri_tfidf_noqe.out	0.67						
indri_tfidf_qe.out	0.64						
terrier_BM25_cli.out	0.85						
terrier_BM25_gfi.out	0.85						
terrier_BM25_noqe.out	0.80						
terrier_BM25_qe.out	0.79						
terrier_DirichletLM_noqe.out	0.80						
terrier_DirichletLM_qe.out	0.64						
terrier_TF_IDF_noqe.out	0.82						
terrier_TF_IDF_qe.out	0.78						

Table 6, referring to the *ad-hoc* Search subtask, shows that most of the approaches tested presented a good accuracy value when it comes to the credibility of the retrieved documents. However, SandiDoc's submission had, unfortunately, a very low accuracy value. With respect to the spoken queries IR subtask, the results were available only for IMS and LIG, as follows:

With respect to this second subtask, the accuracy values illustrated in Table 7 were lower than those referred to the previous task with respect to all the approaches tested by IMS and LIG. This is most likely due to errors from speech recognition multiplying in IR, similar to what we experienced in the CLEF eHealth 2015 and 2016 tasks on speech recognition and IE to support nursing shift-change handover communication [47, 54].

Table 7. Credibility evaluations using the *accuracy* (Acc) in the spoken queries search task

Spoken Queries Retrieval Task			
IMS	**Acc**	**LIG**	**Acc**
bm25_all_rrf	0.56	FT_binary_RF.res	0.52
bm25_all_rrf_rm3	0.59	Noexp_RF.res	0.52
bm25_rrf	0.56	UMLS_binary_RF.res	0.49
bm25_rrf_rm3	0.59	UMLS_weight_RF.res	0.51

Evaluation with cRBP. These results have been obtained with the binary relevance assessment, and the graded credibility assessment, with the same program referred in section 3.5 (`rbp_eval 0.5`). To obtain cRBP with different persistence values, rbp_eval was ran as follows for credibility:

```
rbp_eval -q -H qrels.credibility.clef201820201.test.binary
runName
```

As for the values obtained by using the cRBP measure, at 0.50, 0.80, and 0.95, it is possible to say that, regardless of the specific approach used, and with respect to the *ad-hoc* search task, they range in the 0.15–0.57 interval for the baseline, with an average of 0.40; in the 0.41–0.57 interval for IMS, with an average of 0.50; in the 0.40 – 0.49 for LIG, with an average of 0.45; in the 0.16–0.23 interval for SandiDoc, with an average of 0.20. These results are pretty coherent with those obtained for accuracy. Considering the spoken queries retrieval subtask, also in this case the results are available only for IMS and LIG. In this case, for IMS they range in the 0.37–0.50 interval, with an average of 0.44, while for LIG they range in the 0.30–0.45 interval, with an average of 0.37.

4 Discussion

This year's challenge offered an ad-hoc search subtask and a spoken query retrieval subtask. In Section 3 we provided an analysis of the results obtained by the 3 teams who took part in these tasks. We showed the results achieved by several baselines provided by the task organizers. We also discussed and compared these results in Section 3. We consider three dimensions of relevance - topicality, understanability, and credibility.

The importance of considering several dimensions of relevance, beyond the traditional topicality measure is highlighted in the results obtained, where we find that different retrieval techniques score higher under each of the relevance dimensions (topicality, understanability, and credibility).

As might be expected, retrieval performance is impacted when the queries are presented in spoken form. Speech-to-text conversion is required before the queries can be used in the developed retrieval approaches. The retrieval performance then is impacted by the quality of the speech-to-text conversion. Future studies will explore this phenomenon in greater detail.

We next look at the limitations of this year's challenge. We then reflect on prior editions of the challenge and the challenges future, before concluding the paper.

4.1 Limitations

As previously illustrated, the concept of credibility has been studied in many different research fields, including both psychology/sociology and computer science. Introducing the concept of credibility in the context of IR is not an easy task. On one hand, it represents a characteristic that can be only perceived by human beings [6]. On the other hand, it can be considered as an objective property of an information item, which an automated system can only estimate, and, as a consequence, the 'certainty' of the correct estimate can be expressed by numerical degrees [56]. For instance, let us consider the extreme example of information seekers who think that Covid-19 does not exist and that, therefore, any measure of social distancing is useless in preventing a non-existent contagion. From the point of topical relevance, documents that affirms that the virus is pure invention should be more relevant to them in a context of personalized search; however, documents claiming that it is useless to take precautions not to get infected should be assessed as non-credible in any case. In this context, the assessment of the global relevance of documents, which should take into account both objective and subjective dimensions of relevance, becomes highly problematic. For the above reasons, it is difficult to measure and evaluate credibility with the traditional measures used in IR.

As illustrated in Section 2.4, measures that have been used so far to evaluate the effectiveness of a systems in identifying credible or not credible information are the traditional ones that are used in machine learning for classification tasks, in particular accuracy. To be able to evaluate IR systems that are capable to retrieve credible information (as well as understandable and topically relevant) one could consider more IR-oriented metrics. In CLEF eHealth 2020, we have adapted uRBP to credibility, by employing the so-called cRBP measure (as illustrated in Section 3.6).

However, this measure considers the credibility related to an information item as subjective, while we believe that we should assess credibility in an objective way. This makes this measure only partially suitable to our purposes (as well as the accuracy was only partially suitable). In this scenario, it becomes essential to develop measures that go beyond taking information credibility into account as a a binary value, as is done in classification systems. This problem calls for advancing IR systems and developing related evaluation measures that factor in the joint goodness of the ranking produced by a search engine with respect to multiple dimensions of relevance. These include, but are not limited to, topicality, understandability, and credibility. Including credibility is critical in order to balance between the subjectivity (of assessors) and the objectivity of a fact reported in an information item. Consequently, its inclusion is certainly one of the most ambitious goals we set for our future work.

4.2 Comparison with Prior Work

The inaugural CLEF eHealth CHS/IR task was organized in 2013 on the foundation set by the 2012 CLEF eHealth workshop. The principal finding of this workshop, set to prepare for future evaluation labs, was identifying laypeople's health information needs and related patient-friendly health information access methods as a theme of the community's research and development interest [46]. The resulting CLEF eHealth tasks on CHS/IR, offered yearly from 2013 to 2020 [53, 20, 11, 19, 12, 52, 21, 13], brought together researchers to work on the theme by providing them with timely task specifications, document collections, processing methods, evaluation settings, relevance assessments, and other tools. Targeted use scenarios for the designed, developed, and evaluated CHS/IR technologies in these CLEF eHealth tasks included easing patients, their families, clinical staff, health scientists, and healthcare policy makers in accessing and understanding health information. As a result, the annual tasks accelerated technology transfers from their conceptualisation in academia to generating societal impact [48, 49].

This achieved impact has led to CLEF eHealth establishing its presence and becoming by 2020 one of the primary evaluation lab and workshop series for all interdisciplinary actors of the ecosystem for producing, processing, and consuming eHealth information [4, 16, 5]. Its niche in CHS/IR tasks is formed by addressing health information needs of laypeople — and not healthcare experts only — in accessing and understanding eHealth information in multilingual, multi-modal settings with simultaneous methodological contributions to dimensions of relevance assessments (e.g., topicality, understandability, and credibility of returned information).

4.3 A Vision for the Task beyond 2020

The general purpose of the CLEF eHealth workshops and its preceding CHS/IR tasks has been throughout the years from 2012 to 2020 to assist laypeople into finding and understanding health information in order to make enlightened decisions concerning their health and/or healthcare [46, 53, 20, 11, 19, 12, 51, 21, 13]. In that sense, the evaluation challenge will focus in the coming years on patient-centered IR in a both multilingual and multi-modal setting.

Improving multilingual and multi-modal methods is crucial to guarantee a better access to information, and to understand it. Breaking language and modality barriers has been a priority in CLEF eHealth over the years, and this will continue. Text has been the major media of interest, but as of 2020, also speech has been included as a major new way of people interacting with the systems.

Patient-centered IR/CHS task has been running since 2013 — yet, every edition has allowed to identify unique difficulties and challenges that have shaped the task evolution [53, 20, 11, 19, 12, 51, 21, 13]. The task has considered in the past, for example, multilingual queries, contextualized queries, spoken queries, and query variants. Resources used to build these queries have also been changed. Further exploration of query construction, aiming at a better understanding of

information seekers' health information needs are needed. The task will also further explore relevance dimensions (e.g., topicality, understandably, and credibility), with a particular emphasis on information credibility and methods to take these dimensions into consideration.

4.4 Conclusions

This paper provided an overview of the CLEF eHealth 2020 Task 2 on IR/CHS. The CLEF eHealth workshop series was established in 2012 as a scientific workshop with an aim of establishing an evaluation lab [46]. Since 2013, this annual workshop has been supplemented with two or more preceding shared tasks each year, in other words, the CLEF eHealth 2013–2020 evaluation labs [53, 20, 11, 19, 12, 51, 21, 13]. These labs have offered a recurring contribution to the creation and dissemination of text analytics resources, methods, test collections, and evaluation benchmarks in order to ease and support patients, their next-of-kins, clinical staff, and health scientists in understanding, accessing, and authoring eHealth information in a multilingual setting.

In 2020 the CLEF eHealth lab offered two shared task. One on multilingual IE and the other on consumer health search. These tasks built on the IE and IR tasks offered by the CLEF eHealth lab series since its inception in 2013. Test collections generated by these shared tasks offered a specific task definition, implemented in a dataset distributed together with an implementation of relevant evaluation metrics to allow for direct comparability of the results reported by systems evaluated on the collections.

These established CLEF IE and IR tasks used a traditional shared task model for evaluation in which a community-wide evaluation is executed in a controlled setting: independent training and test datasets are used and all participants gain access to the test data at the same time, following which no further updates to systems are allowed. Shortly after releasing the test data (without labels or other solutions), the participating teams submit their outputs from the frozen systems to the task organizers, who evaluate these results and report the resulting benchmarks to the community.

The annual CLEF eHealth workhops and evaluation labs have matured and established their presence in 2012–2020 in proposing novel tasks in IR/CHS. Given the significance of the tasks, all problem specifications, test collections, and text analytics resources associated with the lab have been made available to the wider research community through our CLEF eHealth website[9].

Acknowledgements

We gratefully acknowledge the contribution of the people and organizations involved in CLEF eHealth in 2012–2020 as participants or organizers. We thank the CLEF Initiative, Benjamin Lecouteux (Université Grenoble Alpes), João Palotti

[9] https://clefehealth.imag.fr/

(Qatar Computing Research Institute), Harrisen Scells (University of Queensland), and Guido Zuccon (University of Queensland). We thank the individuals who generated spoken queries for the IR challenge. We also thank the individuals at University of Queensland who contributed to the IR query generation tool and process. We are very grateful to our assessors that helped despite the COVID-19 crisis: Paola Alberti, Vincent Arnone, Nathan Baran, Pierre Barbe, Francesco Bartoli, Nicola Brew-Sam, Angela Calabrese, Sabrina Caldwell, Daniele Cavalieri, Madhur Chhabra, Luca Cuffaro, Yerbolat Dalabayev, Emine Darici, Marco Di Sarno, Mauro Guglielmo, Weiwei Hou, Yidong Huang, Zhengyang Liu, Federico Moretti, Marie Revet, Paritosh Sharma, Haozhan Sun, Christophe Zeinaty. The lab has been supported in part by The Australian National University (ANU), College of Engineering and Computer Science, Research School of Computer Science; the Our Health in Our Hands (OHIOH) initiative; and the CLEF Initiative. OHIOH is a strategic initiative of The ANU which aims to transform healthcare by developing new personalised health technologies and solutions in collaboration with patients, clinicians, and healthcare providers. We acknowledge the Encargo of Plan TL (SEAD) to CNIO and BSC for funding, and the scientific committee for their valuable comments and guidance.

References

1. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 25–32 (2004)
2. Carminati, B., Ferrari, E., Viviani, M.: A multi-dimensional and event-based model for trust computation in the social web. In: International Conference on Social Informatics. pp. 323–336. Springer (2012)
3. Damiani, E., Viviani, M.: Trading anonymity for influence in open communities voting schemata. In: 2009 International Workshop on Social Informatics. pp. 63–67. IEEE (2009)
4. Demner-Fushman, D., Elhadad, N.: Aspiring to unintended consequences of natural language processing: A review of recent developments in clinical and consumer-generated text processing. Yearbook of Medical Informatics **1**, 224–233 (2016)
5. Filannino, M., Uzuner, Ö.: Advancing the state of the art in clinical natural language processing through shared tasks. Yearbook of Medical Informatics **27**(01), 184–192 (2018)
6. Fogg, B.J., Tseng, H.: The elements of computer credibility. In: Proc. of SIGCHI (1999)
7. Fontanarava, J., Pasi, G., Viviani, M.: Feature analysis for fake review detection through supervised classification. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 658–666. IEEE (2017)
8. Goeuriot, L., Jones, G.J., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salantera, S., Suominen, H., Zuccon, G.: ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. CLEF 2013 Online Working Notes **8138** (2013)
9. Goeuriot, L., Jones, G.J., Kelly, L., Leveling, J., Lupu, M., Palotti, J., Zuccon, G.: An Analysis of Evaluation Campaigns in ad-hoc Medical Information Retrieval: CLEF eHealth 2013 and 2014. Springer Information Retrieval Journal (2018)

10. Goeuriot, L., Kelly, L., Lee, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Gareth J.F. Jones, H.M.: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In: CLEF 2014 Evaluation Labs and Workshop: Online Working Notes. Sheffield, UK (2014)
11. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2015. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Springer Berlin Heidelberg (2015)
12. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: Clef 2017 ehealth evaluation lab overview. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 291–303. Springer Berlin Heidelberg (2017)
13. Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Gonzales Saez, G., Viviani, M., Xu, C.: Overview of the CLEF eHealth evaluation lab 2020. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020) . Lecture Notes in Computer Science (LNCS) Volume number: 12260, Springer, Heidelberg, Germany (2020)
14. Graff, D., Kong, J., Chen, K., Maeda, K.: English gigaword. Linguistic Data Consortium, Philadelphia **4**(1), 34 (2003)
15. Hanbury, A., Müller, H.: Khresmoi – multimodal multilingual medical information search. In: Medical Informatics Europe 2012 (MIE 2012), Village of the Future (2012)
16. Huang, C.C., Lu, Z.: Community challenges in biomedical text mining over 10 years: Success, failure and the future. Briefings in Bioinformatics **17**(1), 132–144 (2016)
17. Jimmy, ., Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L.: Overview of the clef 2018 consumer health search task. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2018)
18. Kakol, M., Nielek, R., Wierzbicki, A.: Understanding and predicting web content credibility using the content credibility corpus. Information Processing & Management **53**(5), 1043–1061 (2017)
19. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2016. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 255–266. Springer Berlin Heidelberg (2016)
20. Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the ShARe/CLEF eHealth evaluation lab 2014. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 172–191. Springer Berlin Heidelberg (2014)
21. Kelly, L., Suominen, H., Goeuriot, L., Neves, M., Kanoulas, E., Li, D., Azzopardi, L., Spijker, R., Zuccon, G., Scells, H., Palotti, J.: Overview of the clef ehealth evaluation lab 2019. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Heinatz Bürki, G., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 322–339. Springer International Publishing, Cham (2019)

22. Koopman, B., Zuccon, G.: Relevation!: an open source system for information retrieval relevance assessment. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. pp. 1243–1244. ACM (2014)
23. Lewandowski, D.: Credibility in web search engines. In: Online credibility and digital ethos: Evaluating computer-mediated communication, pp. 131–146. IGI Global (2013)
24. Lipani, A., Palotti, J., Lupu, M., Piroi, F., Zuccon, G., Hanbury, A.: Fixed-cost pooling strategies based on ir evaluation measures. In: European Conference on Information Retrieval. pp. 357–368. Springer (2017)
25. Livraga, G., Viviani, M.: Data confidentiality and information credibility in on-line ecosystems. In: Proceedings of the 11th International Conference on Management of Digital EcoSystems. pp. 191–198 (2019)
26. Metzger, M.J.: Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. JASIST **58**(13), 2078–2091 (2007)
27. Metzger, M.J., Flanagin, A.J.: Credibility and trust of information in online environments: The use of cognitive heuristics. Journal of pragmatics **59**, 210–220 (2013)
28. Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., Krallinger, M.: Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020. In: Conference and Labs of the Evaluation (CLEF) Working Notes. CEUR Workshop Proceedings (CEUR-WS.org) (2020)
29. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. **27**(1), 2:1–2:27 (Dec 2008). https://doi.org/10.1145/1416950.1416952, http://doi.acm.org/10.1145/1416950.1416952
30. Mulhem, P., Sáez, G.N.G., Mannion, A., Schwab, D., Frej, J.: LIG-Health at Adhoc and Spoken IR Consumer Health Search: expanding queries using umls and fasttext. In: Conference and Labs of the Evaluation (CLEF) Working Notes. CEUR Workshop Proceedings (CEUR-WS.org) (2020)
31. Nogueira, R., Cho, K.: Task-oriented query reformulation with reinforcement learning. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/d17-1061
32. Nunzio, G.M.D., Marchesin, S., Vezzani, F.: A Study on Reciprocal Ranking Fusion in Consumer Health Search. IMS UniPD ad CLEF eHealth 2020 Task 2. In: Conference and Labs of the Evaluation (CLEF) Working Notes. CEUR Workshop Proceedings (CEUR-WS.org) (2020)
33. Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanburyn, A., Jones, G.J., Lupu, M., Pecina, P.: CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information about Medical Symptoms. In: CLEF 2015 Online Working Notes. CEUR-WS (2015)
34. Palotti, J., Zuccon, G., Jimmy, Pecina, P., Lupu, M., Goeuriot, L., Kelly, L., Hanbury, A.: CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2017)
35. Park, L.A., Zhang, Y.: On the distribution of user persistence for rank-biased precision. In: Proceedings of the 12th Australasian document computing symposium. pp. 17–24 (2007)

36. Pasi, G., De Grandis, M., Viviani, M.: Decision making over multiple criteria to assess news credibility in microblogging sites. In: IEEE World Congress on Computational Intelligence (WCCI) 2020, Proceedings. IEEE (2020)
37. Pasi, G., Viviani, M.: Information credibility in the social web: Contexts, approaches, and open issues. arXiv preprint arXiv:2001.09473 (2020)
38. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 2173–2178 (2016)
39. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 1003–1012 (2017)
40. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF, IEEE Signal Processing Society (2011)
41. Robertson, S.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2010). https://doi.org/10.1561/1500000019
42. Rousseau, A., Deléglise, P., Esteve, Y.: Ted-lium: an automatic speech recognition dedicated corpus. In: LREC. pp. 125–129 (2012)
43. Sandaru Seneviratne, Eleni Daskalaki, A.L., Hossain, M.Z.: SandiDoc at CLEF 2020 - Consumer Health Search : AdHoc IR Task. In: Conference and Labs of the Evaluation (CLEF) Working Notes. CEUR Workshop Proceedings (CEUR-WS.org) (2020)
44. Sbaffi, L., Rowley, J.: Trust and credibility in web-based health information: a review and agenda for future research. Journal of medical Internet research **19**(6), e218 (2017)
45. Self, C.C.: Credibility. In: An integrated approach to communication theory and research, pp. 449–470. Routledge (2014)
46. Suominen, H.: CLEFeHealth2012 — The CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis. In: Forner, P., Karlgren, J., Womser-Hacker, C., Ferro, N. (eds.) CLEF 2012 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1178/ (2012)
47. Suominen, H., Hanlen, L., Goeuriot, L., Kelly, L., Jones, G.: Task 1a of the CLEF eHealth evaluation lab 2015: Clinical speech recognition. In: Online Working Notes of CLEF. CLEF (2015)
48. Suominen, H., Kelly, L., Goeuriot, L.: Scholarly influence of the Conference and Labs of the Evaluation Forum eHealth Initiative: Review and bibliometric study of the 2012 to 2017 outcomes. JMIR Research Protocols **7**(7), e10961 (2018). https://doi.org/10.2196/10961
49. Suominen, H., Kelly, L., Goeuriot, L.: The scholarly impact and strategic intent of CLEF eHealth Labs from 2012 to 2017. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF. pp. 333–363. Springer International Publishing, Cham (2019)
50. Suominen, H., Kelly, L., Goeuriot, L., Krallinger, M.: Clef ehealth evaluation lab 2020. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 587–594. Springer International Publishing, Cham (2020)

51. Suominen, H., Kelly, L., Goeuriot, L., Névéol, A., Ramadier, L., Robert, A., Kanoulas, E., Spijker, R., Azzopardi, L., Li, D., Jimmy, Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2018. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 286–301. Springer Berlin Heidelberg (2018)
52. Suominen, H., Kelly, L., Goeuriot, L., Névéol, A., Ramadier, L., Robert, A., Kanoulas, E., Spijker, R., Azzopardi, L., Li, D., Jimmy, Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2018. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J.Y., Soulier, L., SanJuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 286–301. Springer International Publishing, Cham, Switzerland (2018)
53. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., Jones, G.J., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., Zuccon, G.: Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 212–231. Springer Berlin Heidelberg (2013)
54. Suominen, H., Zhou, L., Goeuriot, L., Kelly, L.: Task 1 of the CLEF eHealth evaluation lab 2016: Handover information extraction. In: CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)
55. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Lrec. vol. 2012, pp. 2214–2218 (2012)
56. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health information—a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **7**(5), e1209 (2017)
57. Wayne, C., Doddington, G., et al.: Tdt2 multilanguage text version 4.0 ldc2001t57. Philadelphia: Linguistic Data Consortium (LDC) (2001)
58. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. In: Reinforcement Learning, pp. 5–32. Springer US (1992)
59. Yamamoto, Y., Tanaka, K.: Enhancing credibility judgment of web search results. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1235–1244 (2011)
60. Yang, K.C., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with artificial intelligence to counter social bots. Human Behavior and Emerging Technologies **1**(1), 48–61 (2019)
61. Zuccon, G.: Understandability biased evaluation for information retrieval. In: Advances in Information Retrieval. pp. 280–292 (2016)
62. Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L., Lupu, M., Pecina, P., Mueller, H., Budaher, J., Deacon, A.: The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred Health Information Retrieval. In: CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (September 2016)