Overview of the CLEF eHealth Evaluation Lab 2019

Kelly, L.; Suominen, H.; Goeuriot, L.; Neves, M.; Kanoulas, E.; Li, D.; Azzopardi, L.; Spijker, R.; Zuccon, G.; Scells, H.; Palotti, J.

DOI
10.1007/978-3-030-28577-7_26

Publication date
2019

Document Version
Final published version

Published in
Experimental IR Meets Multilinguality, Multimodality, and Interaction

Overview of the CLEF eHealth Evaluation Lab 2019

Liadh Kelly[1(✉)], Hanna Suominen[2,3], Lorraine Goeuriot[4], Mariana Neves[5], Evangelos Kanoulas[6], Dan Li[6], Leif Azzopardi[7], Rene Spijker[8], Guido Zuccon[9], Harrisen Scells[9], and João Palotti[10]

[1] Maynooth University, Kildare, Ireland
liadh.kelly@mu.ie
[2] The Australian National University,
Data61/Commonwealth Scientific and Industrial Research Organisation,
University of Canberra, Canberra, ACT, Australia
hanna.suominen@anu.edu.au
[3] University of Turku, Turku, Finland
[4] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
Lorraine.Goeuriot@imag.fr
[5] German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR), Berlin, Germany
mariana.lara-neves@bfr.bund.de
[6] Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
{E.Kanoulas,D.Li}@uva.nl
[7] Computer and Information Sciences, University of Strathclyde, Glasgow, UK
leif.azzopardi@strath.ac.uk
[8] Cochrane Netherlands and UMC Utrecht,
Julius Center for Health Sciences and Primary Care, Utrecht, Netherlands
R.Spijker-2@umcutrecht.nl
[9] University of Queensland, Brisbane, Australia
{g.zuccon,h.scells}@uq.edu.au
[10] Qatar Computing Research Institute (QCRI), HBKU, Doha, Qatar
jpalotti@hbku.edu.qa

Abstract. In this paper, we provide an overview of the seventh annual edition of the CLEF eHealth evaluation lab. CLEF eHealth 2019 continues our evaluation resource building efforts around the easing and support of patients, their next-of-kins, clinical staff, and health scientists in understanding, accessing, and authoring electronic health information in a multilingual setting. This year's lab advertised three tasks: Task 1 on indexing non-technical summaries of German animal experiments with International Classification of Diseases, Version 10 codes; Task 2 on technology assisted reviews in empirical medicine building on 2017 and 2018 tasks in English; and Task 3 on consumer health search in mono- and multilingual settings that builds on the 2013–18 Information Retrieval tasks. In total nine teams took part in these tasks (six in Task 1 and three

LK, HSu & LG co-chaired the lab. MN; EK, DL, LA & RS; and GZ, HSc & JP led Tasks 1–3, respectively.

F. Crestani et al. (Eds.): CLEF 2019, LNCS 11696, pp. 322–339, 2019.
https://doi.org/10.1007/978-3-030-28577-7_26

in Task 2). Herein, we describe the resources created for these tasks and evaluation methodology adopted. We also provide a brief summary of participants of this year's challenges and results obtained. As in previous years, the organizers have made data and tools associated with the lab tasks available for future research and development.

Keywords: Evaluation · Entity linking · Information retrieval · Health records · High recall · Information extraction · Medical informatics · Self-diagnosis · Systematic reviews · Test-set generation · Text classification · Text segmentation

1 Introduction

Retrieving, digesting, and summarising valid and relevant information to make health-centered decisions has become increasingly difficult in today's information overloaded society. More and more *electronic health* (eHealth) content is becoming available in a variety of forms ranging from scientific papers and health-related websites through patient records and medical dossiers to medical-related topics shared across social networks [27]. Laypeople, clinicians, and policy makers need bespoke systems to retrieve relevant and reliable contents and access them in a clear and concise way to easily judge and make sense of them to support their decision making.

Information retrieval (IR) systems have been commonly used as a means to access health information available online. To illustrate the immense worldwide popularity of going online to consume and produce health information, five years ago, in Australia, 40 per cent of searches were to fulfill health information needs; in Europe, nearly half of the population consider the Internet as a significant source of health information; and in the USA, nearly 70 per cent of people using web search engines want information about diseases, health conditions, or other medical disorders [1]. Based on the "Household Use of Information Technology" survey for 2016–2017 by the *Australian Bureau of Statistics* (ABS)[1], this popularity has grown and stabilised itself to almost 90 per cent of Australian households having access to the Internet (up to 97% for those households that have children aged under 15 years), and approximately 50 per cent of Australians are using it to meet their health or healthcare information needs. However, the information seekers find it difficult to express their health information needs as search queries that find the right information, and also the quality, reliability, and suitability of the information for the target audience varies greatly while high recall or coverage—that is, finding all relevant information about a topic—is often as important as (if not more important than) high precision [24].

CLEF eHealth[2], established as a lab workshop in 2012 as part of the *Conference and Labs of the Evaluation Forum* (CLEF), has offered evaluation labs

[1] Statistics extracted from the ABS pages at https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/8146.0Main+Features12016-17?OpenDocument, titled "8146.0 – Household Use of Information Technology, Australia, 2016–17", on 28 May 2019.

[2] http://clef-ehealth.org/ (last accessed on 28 May 2019).

since 2013 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 data sets to work with and validate the outcomes. More specifically, these labs and their subsequent workshops target (1) developing processing methods and resources in a multilingual setting to enrich difficult-to-understand eHealth texts and provide personalized reliable access to medical information, and provide valuable documentation; (2) developing an evaluation setting and releasing evaluation results for these methods and resources; and (3) contributing to the participants and organizers' professional networks and interaction with all interdisciplinary actors of the ecosystem for producing, processing, and consuming eHealth information.

The CLEF eHealth labs are open for everybody. We particularly welcome academic and industrial researchers, scientists, engineers, and graduate students in natural language processing, machine learning, and biomedical/health informatics to participate. We also encourage participation by multi-disciplinary teams that combine technological skills with biomedical expertise.

This, the seventh year of the evaluation lab (and eight year of the workshop), aiming to build upon the resource development and evaluation approaches by the previous six or seven years of CLEF eHealth [8,9,14,16,26,28,29], offered the following two tasks [15]:

- *Task 1.* Multilingual Information Extraction: *International Classification of Diseases, Version 10* (ICD-10) coding of *non-technical summaries* (NTSs) of animal experiments in German [22] and
- *Task 2. Technology Assisted Reviews* (TAR) in Empirical Medicine in English [13].

In addition, Task 3. Consumer Health Search in Mono- and Multilingual Settings was initially advertised, but unfortunately, due to unforeseen circumstances, it had to be postponed[3].

The *Multilingual Information Extraction* task challenged participants to index German NTSs of animal experiments with the ICD-10 terminology of diseases. A detailed analysis based on the diseases addressed by the NTSs allows more transparency of the animal experiments being carried out by researchers [2]. It could be treated as a text classification or cascaded named entity recognition and normalization task. Even though we only addressed one language (German), we encouraged participants to explore multilingual approaches. The results of high performing systems could be used within the workflow of institutes mandated by the *European Union* (EU) to publish the NTSs approved in their states. The 2019 Task 1 built upon the 2016–2018 information extraction tasks [19–21], which already addressed the ICD-10 terminology to code causes of death from a corpus of death reports in French (2016, 2017, and 2018), English (2017), Hungarian (2018), and Italian (2018). Prior to this, the CLEF eHealth tasks considered *Unified Medical Language System* (UMLS) and *Systematized*

[3] The organizers apologize to the teams that registered their interest in the task for any inconvenience caused by this delay.

Nomenclature of Medicine—Clinical Terms (SNoMed-CT) codification of clinical reports in English in 2013, and UMLS named entity recognition of clinical reports in French in 2015, among others [27].

The *TAR* task was a high-recall IR task in English that aimed at evaluating search algorithms that seek to identify all studies relevant for conducting a systematic review in empirical medicine. The results of the explored approaches in the submitted systems towards generating a clear overview of the current scientific consensus could be informing health care and its policy making in the future. This automated generator might release scientists and policy advisors' time from the currently laborious iterative process of conducting publication searches and revising them in order to retrieve all the documents that are relevant for the purposes of writing reliable systematic reviews; this hard challenge is known in the IR domain as the total recall problem and with the number of published medical papers expanding rapidly, the need for automation in this process becomes of utmost importance.

This year's Task 2, differed from the past two years [11,12] by diversifying the focus across different type of reviews including *Diagnostic Test Accuracy* (DTA), *Intervention*, *Prognosis*, and *Qualitative* reviews. Even though search in the area of DTA reviews is generally considered the hardest [18], this year we wanted to investigate how the technology that has been developed over the past two years would extend to other types of reviews. The typical process of searching for scientific publications to conduct a systematic review consists of three stages: (a) specifying a number of inclusion criteria that characterize the articles relevant to the review and constructing a complex Boolean Query to express them, (b) screening the abstracts and titles that result from the Boolean query, and (c) reading and screening the full documents that passed the Abstract and Title Screening. Building on the 2017 task, which focused on the second stage of the process, that is, Abstract and Title Screening, and same as the 2018 task, the 2019 task focused both on the first stage (*subtask 1*) and second stage (*subtask 2*) of the process, that is, Boolean Search and Abstract and Title Screening.

More precisely, these subtasks of Task 2 were defined as follows:

- *Subtask 1.* Prior to constructing a Boolean Query researchers have to design and write a search protocol that in written and in detail defines what constitutes a relevant study for their review. For the challenge associated with the first stage of the process, participants were provided with the relevant pieces of a protocol, in an attempt to complete search effectively and efficiently bypassing the construction of the Boolean query.
- *Subtask 2.* Given the results of the Boolean Search from stage 1 as the starting point, participants were required to rank the set of *abstracts* (A). The task had the following two goals: (i) to produce an efficient ordering of the documents, such that all of the relevant abstracts are retrieved as early as possible, and (ii) to identify a subset of A which contains all or as many of the relevant abstracts for the least effort (i.e., total number of abstracts to be assessed).

The *Consumer Health Search* task was advertised as a continuation of the previous CLEF eHealth IR tasks that ran every year since the onset of

CLEF eHealth evaluation labs in 2013 [5–7,10,23,25,30], and embraced the *Text REtrieval Conference* (TREC) -style evaluation process, with a shared collection of searchable documents and their search queries, the contribution of runs from participants, and the subsequent formation of relevance assessments and evaluation of these participants' submissions. For the first time, the search queries (and their variants) were intended to not only be in written format but also in spoken format, with automatic speech-to-text transcripts provided. The new document collection introduced in the 2018 Task 3, consisting of over 5 million pages from the *World Wide Web* (WWW) was to be used for this task. This was a compilation of Web pages of selected domains acquired from the CommonCrawl[4]. User stories for search query and query variant generation were those, using the discharge summaries and forum posts, we used in previous years of the task.

The remainder of this overview paper is structured as follows: First, in Sect. 2, we detail for each task its text documents; human annotations, queries, and relevance assessments; and evaluation methods. After this, in Sect. 3, we describe the task submissions and results of the CLEF eHealth 2019 evaluation lab. Finally, in Sect. 4 we conclude the study.

2 Materials and Methods

In this section, we describe the materials and methods used in the two tasks of the CLEF eHealth evaluation lab 2019. After specifying our text documents to process in Sect. 2.1, we address their human annotations, queries, and relevance assessments in Sect. 2.2. Finally, in Sect. 2.3 we introduce our evaluation methods. We also include in Sects. 2.1 and 2.2 a brief description of the document set and its intended query set for Task 3.

2.1 Text Documents

Task 1. The multilingual information extraction task challenged its participants with the fully automated semantic indexing of NTSs of animal experiments using codes from the German version of the ICD-10. The NTPs were short publicly-available summaries[5] written as part of the approval procedure for animal experiments in Germany. The database currently contains more than 10,000 NTPs (as of May/2019).

Task 2. The technologically assisted reviews in empirical medicine task used the PubMed document collection for its Boolean Search challenge and a subset of PubMed documents for its challenge to make Abstract and Title Screening more effective. More specifically, for the Abstract and Title Screening subtask the *PubMed Document Identifiers* (PMIDs) of potentially relevant

[4] http://commoncrawl.org/ (last accessed on 28 May 2019).

[5] The *AnimalTestInfo* database was publicly available at https://www.animaltestinfo.de when the task was launched.

PubMed Document abstracts were provided for each training and test topic. The PMIDs were collected by the task coordinators by re-running the MEDLINE Boolean query used in the original systematic reviews conducted by Cochrane to search PubMed.

Task 3. The document corpus is the same as the corpus used in 2018. It consists of web pages acquired from the CommonCrawl. An initial list of websites was identified for acquisition. The list was built by submitting the CLEF 2018 queries to the Microsoft Bing Apis (through the Azure Cognitive Services) repeatedly over a period of a few weeks, and acquiring the 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, from lists previously compiled by health institutions and agencies.

2.2 Human Annotations, Queries, and Relevance Assessments

Task 1. The task consisted of assigning codes with respect to chapters or groups of the 2016 German Modification of ICD-10[6]. The training and development data set[7] contained a total of 8,386 NTSs of animal experiments recently carried out in Germany (as of September 2018). It was split into training and development sets with 7,544 and 842 NTSs, respectively. For the test set, we released 407 NTSs[8] for which participants should predict the ICD-10 codes. In all data sets, each NTS contained a title, benefits (goals) of the experiments, possible harms caused to the animals, and comments related to the *replacement, reduction and refinement* (3R) principles. All documents were in the German language. The data set included the ICD-10 codes manually assigned by experts. However, some NTSs had no ICD-10 codes assigned to them, since the codes were not applicable to the benefits described in the NTS.

Task 2. In Task 2 Subtask 1, for the No-Boolean-Search challenge as input for each topic participants were provided with

1. a Topic-ID,
2. the title of the review, written by Cochrane experts,

[6] Available at https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kodesuche/htmlgm2016/.

[7] Publicly available on 24 January 2019 at https://www.openagrar.de/receive/openagrar_mods_00046540?lang=en under the *Creative Commons, Attribution-NonCommercial-NoDerivatives 4.0 International* (CC BY-NC-ND 4.0) license as DOI https://doi.org/10.17590/20190118-134645-0.

[8] Publicly available on 6 May 2019 https://www.openagrar.de/receive/openagrar_mods_00049062?lang=en under the *Creative Commons, Attribution-NonCommercial-NoDerivatives 4.0 International* (CC BY-NC-ND 4.0) license.

3. the most important parts of the protocol, written by Cochrane experts, and
4. the entire PubMED database (which was available for downloaded directly from PubMED, through ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline).

In Task 2 Subtask 2, focusing on title and abstract screening, topics consisted of the Boolean Search from the first step of the systematic review process. Specifically, for each topic the following information was provided.

1. a Topic-ID,
2. the title of the review, written by Cochrane experts,
3. the Boolean query, manually constructed by Cochrane experts, and
4. the set of PMIDs returned by running the query in MEDLINE.

Participants were provided with eight topics of DTA reviews, 20 topics of Intervention reviews, one topic of Prognosis, and two of Qualitative reviews, as a test set for both subtasks. The 72 DTA topics (which excludes topics that were reviewed and found unreliable) considered in CLEF 2017 and 2018 TAR tasks were used as training set. Further, we developed 20 Intervention topics that were also provided as training set to participants.

The original systematic reviews written by Cochrane experts included a reference section that listed Included, Excluded, and Additional references to medical studies. The union of Included and Excluded references are the studies that were screened at a Title and Abstract level and were considered for further examination at a full content level. These constituted the relevant documents at the abstract level, while the Included references constituted the relevant documents at the full content level. References in the original systematic reviews were collected from a variety of resources, not only MEDLINE. Therefore, studies that were cited but did not appear in the results of the Boolean query were excluded from the label set for both Subtask 1 and Subtask 2.

Regarding Subtask 2, that is, the Title and Abstract Screening, relevance was assessed at two levels, at abstract level, which expresses the potential of the article to be relevant and included in the review, and hence need to be read in full, and at full content level, after the full article has been read and decided whether to be included or excluded from the study. The following numbers present for each type of study the percentage of relevant document (abstract or content level) in the development set and in the test set, so that the reader can get an idea of the difficulty of the task, the differences across different types of reviews if any, and any changes in the relevance distribution between training and test sets.

Hence, the percentage of relevant document (1) for the DTA studies, (1a) at abstract level, in the training set was 1.7% and in the test set 1.4% of the total number of PMIDs released, while (1b) at content level it was 0.3% in the training set, and 0.8% in the test set. (2) For the Intervention studies, the percentage of relevant documents (2a) at abstract level in the training set was 1.7% and in the test set 0.9%, while at the content level the average percentage was 2.2% in the training set, and 1.2% in the test set. For the Prognosis and Qualitative reviews

no training data was provided. (3) In the test set for the Prognosis, (3a) the percentage of relevant documents is 5.7% at the abstract level and (3b) 2.7% at the content level, while (4) for the Qualitative, (4a) the percentage of relevant documents is 1.7% at the abstract level and(4b) 0.4% at the content level.

All the released data for the 2017 – 2019 CLEF eHealth TAR tasks can be found at https://github.com/CLEF-TAR.

Task 3. With the aim to acquire more relevance assessments and increase the collection reusability, the intent this year was to reuse the same set of 50 query narratives developed in 2018's Task 3 [10]. In 2018, query creators devised 7 query variants from each query narrative. This was accomplished by asking laypeople and medical experts to generate written queries based on the textual narratives. In 2019, in order to increase the variability of generated queries, written narratives were converted into spoken audio. After hearing the narratives, a set of query creators were to generate spoken query variants by speaking their queries aloud. Our intention was to make the generated original spoken queries as well as the output of a speech-recognition software available to the participants.

2.3 Evaluation Methods

Task 1. The training and development sets were released on 24 January 2019, and the test set on 6 May 2019. Teams could submit by 13 May 2019 up to three runs/solutions for the test data set. We evaluated the runs based on the usual metrics of the precision, recall, and F-measure using a publicly-available Python script[9].

Task 2. Teams could submit an unlimited number of runs per task. In addition, participants were also encouraged to submit any number of runs that result from their 2017 and 2018 frozen systems. System performance was assessed using the same evaluation approach as that used for the 2018 TAR challenge [12]. Specifically, (i) similarly to the previous year, runs were evaluated on the basis of identifying the studies to be included (relevant documents), (ii) different from previous years, runs were evaluated on the basis of not only finding the studies to be included, but also finding high quality included studies before low quality included studies.

The assumption behind this evaluation approach (i) was the following: The user of your system is the researcher that performs the abstract and title screening of the retrieved articles. Every time an abstract is returned (i.e., ranked) there is an incurred cost/effort, while the abstract is either irrelevant (in which case no further action will be taken) or relevant (and hence passed to the next stage of document screening) to the topic under review.

[9] https://github.com/mariananeves/clef19ehealth-task1.

Evaluation measures were as follows: Area under the recall-precision curve (i.e., Average Precision); Minimum number of documents returned to retrieve all relevant documents; Work Saved over Sampling at different Recall levels; Area under the cumulative recall curve normalized by the optimal area; Recall @ 0% to 100% of documents shown; a number of newly constructed cost-based measures; and reliability [3].

Evaluation approach (ii) considered not only the relevance but the quality of the articles as well, taking into account indicators such as the risk-of-bias, and the sample size of the trials reported of the studies. This second evaluation approach depended on assessments Cochrane reviewers made manually on aspects of the included studies. Obtaining these assessments turned out to be a difficult task therefore this second evaluation approached was postponed for the future.

The training data set was released at the end of March 2019 and the test data set on 14 May 2019. The relevance labels on the testing data (required by active learning techniques) were provided to participants on 14 May 2019 as well, while the submission deadline was set to 21 May 2019 so that participants could not tune their systems towards the actual labels.

More details on the evaluation are provided in the Task 2 overview paper [13].

3 Results

The number of people who registered their interest in CLEF eHealth tasks was 31 in Task 1 and 36 in Task 2. In total, nine teams submitted to the two shared tasks.

Task 1 received considerable interest with the submission of 14 runs from six teams. We had two teams from Germany (MLT-DFKI and WBI), one from India (SSN_NLP), one from Italy (IMS_UNIPD), one as a collaboration between Spain and UK (TALP_UPC) and one from Turkey (DEMIR). Table 1 summarizes the results obtained by each team.

Participants relied on a diverse range of approaches. WBI utilized the multilingual version of the BERT-Base model [4] and made use of additional resources, such as the *German Clinical Trials Register* (DRKS)[10]. MLT-DFKI utilized Google Translate to convert documents into English and then relied on pretrained BioBERT [17] to perform the prediction of ICD-10 codes. DEMIR utilized ElasticSearch for searching for similar NTSs and selected top documents (NTSs) based on *k-nearest neighbors* (KNN) and on threshold-based methods. SSN_NLP relied on a seq2seq mapping model based on bidirectional *long short-term memory* (LSTM) and experimented with the Normed_Bahdanau and the Scaled Luong attention mechanisms. IMS-UNIPD tried three Naïve Bayes classifiers (Bernoulli, Multinomial and Poisson) based on a 2D representation of the probabilities.

[10] See https://www.drks.de/drks_web/setLocale_EN.do.

Task 2 attracted the interest of 3 teams submitting runs, all from Europe, including one team from The Netherlands (UvA), one team from the UK (Sheffield), and one team from Italy (UNIPD). For Subtask 1, we received no runs. For Subtask 2, we received 36 runs from the three teams. The results on a selected subset of metrics on DTA, Intervention, Prognosis, and Qualitative studies are shown in Tables 2, 3, 4, and 5, respectively. The three teams used a variety of ranking methods including traditional BM25, interactive BM25, continuous active learning, relevance feedback, as well as a variety of stopping criteria to provide a threshold on the ranking.

Table 1. System performance for ICD-10 coding on the test set for German NTSs in terms of Precision (P), recall (R) and F-measure (F). The results are ordered in decreasing order of the scores for F-Measure. We highlight in **bold** the highest scores for P, R, and F.

Team	P	R	FM
WBI-run1	0.83	0.77	**0.80**
WBI-run2	**0.84**	0.74	0.79
WBI-run3	0.80	0.78	0.79
MLT-DFKI	0.64	**0.86**	0.73
DEMIR-run1	0.46	0.50	0.48
DEMIR-run3	0.46	0.49	0.48
DEMIR-run2	0.49	0.44	0.46
TALP_UPC	0.37	0.35	0.36
SSN_NLP-run2	0.19	0.27	0.23
SSN_NLP-run1	0.19	0.27	0.22
SSN_NLP-run3	0.13	0.34	0.19
IMS_UNIPD-run3	0.10	0.05	0.07
IMS_UNIPD-run2	0.009	0.50	0.017
IMS_UNIPD-run1	0	0	0

Table 2. DTA studies with abstract-level QRELs

Run	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Rely	R@k	k
ILPS/DTA/abs-hh-ratio-ilps@uva.out	2420	0.493	0.589	0.682	0.789	0.834	0.406	0.304	0.189	0.815	1132
ILPS/DTA/abs-th-ratio-ilps@uva.out	2676	0.399	0.418	0.536	0.661	0.734	0.312	0.253	0.273	0.744	1558
Padua/DTA/2018_stem_original_p10_t400.out	1190	0.229	0.448	0.634	0.818	0.895	0.662	0.512	0.136	0.963	605
Padua/DTA/distributed_effort_p10_t1500.out	1111	0.229	0.445	0.63	0.814	0.895	0.652	0.513	0.204	0.963	2453
Padua/DTA/2018_stem_original_p10_t1000.out	1141	0.229	0.445	0.63	0.814	0.893	0.658	0.509	0.19	0.986	1195
Padua/DTA/2018_stem_original_p10_t200.out	1282	0.229	0.445	0.634	0.823	0.891	0.66	0.507	0.115	0.877	336
Padua/DTA/2018_stem_original_p10_t500.out	1200	0.229	0.445	0.634	0.818	0.893	0.662	0.509	0.147	0.97	719
Padua/DTA/2018_stem_original_p10_t300.out	1280	0.229	0.452	0.627	0.816	0.893	0.66	0.5	0.113	0.936	477
Padua/DTA/2018_stem_original_p10_t1500.out	1126	0.229	0.445	0.63	0.814	0.895	0.657	0.514	0.228	0.995	1524
Padua/DTA/distributed_effort_p10_t1000.out	1109	0.229	0.445	0.63	0.814	0.895	0.649	0.514	0.129	0.93	1776
Padua/DTA/2018_stem_original_p10_t100.out	2024	0.221	0.418	0.609	0.791	0.868	0.525	0.399	0.291	0.604	180
Padua/DTA/baseline_bm25_t500.out	2470	0.119	0.236	0.402	0.548	0.65	0.39	0.252	0.342	0.638	451
Padua/DTA/distributed_effort_p10_t300.out	1111	0.232	0.445	0.63	0.814	0.886	0.649	0.528	0.117	0.818	802
Padua/DTA/2018_stem_original_p50_t1000.out	1127	0.229	0.445	0.63	0.811	0.893	0.652	0.528	0.235	0.995	1473
Padua/DTA/distributed_effort_p10_t100.out	1271	0.204	0.439	0.614	0.77	0.839	0.61	0.468	0.308	0.572	284
Padua/DTA/2018_stem_original_p50_t200.out	1291	0.229	0.445	0.634	0.82	0.898	0.66	0.499	0.141	0.89	364
Padua/DTA/baseline_bm25_t1000.out	2395	0.119	0.236	0.389	0.543	0.659	0.396	0.26	0.274	0.761	826
Padua/DTA/distributed_effort_p10_t500.out	1116	0.229	0.445	0.63	0.814	0.891	0.634	0.521	0.096	0.874	1083
Padua/DTA/baseline_bm25_t300.out	2493	0.119	0.239	0.405	0.541	0.652	0.391	0.244	0.415	0.499	280
Padua/DTA/baseline_bm25_t100.out	2130	0.12	0.239	0.414	0.564	0.659	0.394	0.295	0.683	0.241	101
Padua/DTA/2018_stem_original_p50_t400.out	1189	0.229	0.448	0.634	0.816	0.891	0.654	0.527	0.154	0.965	672
Padua/DTA/2018_stem_original_p50_t300.out	1272	0.229	0.452	0.627	0.814	0.893	0.656	0.518	0.146	0.945	522
Padua/DTA/2018_stem_original_p50_t100.out	2027	0.222	0.418	0.609	0.786	0.868	0.549	0.394	0.308	0.618	189
Padua/DTA/distributed_effort_p10_t200.out	1194	0.225	0.445	0.632	0.811	0.877	0.663	0.509	0.17	0.735	566
Padua/DTA/baseline_bm25_t400.out	2492	0.119	0.239	0.405	0.539	0.65	0.386	0.246	0.355	0.596	367
Padua/DTA/2018_stem_original_p50_t1500.out	1056	0.229	0.445	0.63	0.814	0.898	0.651	0.537	0.31	1.0	2018
Padua/DTA/2018_stem_original_p50_t500.out	1200	0.229	0.445	0.634	0.809	0.889	0.649	0.524	0.169	0.97	820
Padua/DTA/baseline_bm25_t1500.out	2476	0.119	0.236	0.389	0.541	0.652	0.364	0.254	0.256	0.853	1171
Padua/DTA/baseline_bm25_t200.out	2253	0.12	0.234	0.405	0.55	0.652	0.409	0.278	0.504	0.407	192
Padua/DTA/distributed_effort_p10_t400.out	1116	0.231	0.445	0.63	0.814	0.886	0.634	0.528	0.1	0.856	942
Sheffield/DTA/DTA_sheffield-Chi-Squared.out	1964	0.222	0.305	0.45	0.641	0.73	0.475	0.375	0.479	1.0	3815
Sheffield/DTA/DTA_sheffield-baseline.out	2250	0.175	0.22	0.336	0.525	0.675	0.451	0.338	0.479	1.0	3815
Sheffield/DTA/DTA_sheffield-Odds_Ratio.out	2184	0.248	0.382	0.561	0.707	0.805	0.49	0.347	0.479	1.0	3815
Sheffield/DTA/DTA_sheffield-Log_Likelihood.out	1972	0.234	0.35	0.527	0.668	0.759	0.487	0.381	0.479	1.0	3815

Table 3. Intervention studies with abstract-level QRELs

Run	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Rely	R@k	k
ILPS/Int/abs-hh-ratio-ilps@uva.out	958	0.567	0.518	0.628	0.736	0.813	0.526	0.48	0.213	0.915	773
ILPS/Int/abs-th-ratio-ilps@uva.out	986	0.556	0.478	0.576	0.692	0.774	0.535	0.45	0.197	0.868	555
Padua/Int/2018_stem_original_p10_t400.out	985	0.28	0.307	0.502	0.663	0.744	0.632	0.511	0.334	0.941	487
Padua/Int/distributed_effort_p10_t1500.out	981	0.28	0.306	0.499	0.664	0.745	0.633	0.517	0.247	0.968	1349
Padua/Int/2018_stem_original_p10_t1000.out	977	0.28	0.306	0.499	0.664	0.745	0.63	0.51	0.415	0.973	870
Padua/Int/2018_stem_original_p10_t200.out	1180	0.28	0.312	0.501	0.671	0.775	0.617	0.488	0.267	0.901	301
Padua/Int/2018_stem_original_p10_t500.out	975	0.28	0.306	0.502	0.662	0.742	0.63	0.514	0.353	0.946	560
Padua/Int/2018_stem_original_p10_t300.out	1141	0.28	0.313	0.496	0.665	0.771	0.617	0.494	0.322	0.922	405
Padua/Int/2018_stem_original_p10_t1500.out	952	0.28	0.306	0.499	0.664	0.745	0.63	0.522	0.474	0.984	1117
Padua/Int/distributed_effort_p10_t1000.out	992	0.279	0.306	0.499	0.664	0.745	0.62	0.492	0.157	0.921	975
Padua/Int/2018_stem_original_p10_t100.out	1153	0.274	0.306	0.483	0.639	0.737	0.54	0.474	0.292	0.711	164
Padua/Int/baseline_bm25_t500.out	1233	0.222	0.191	0.282	0.41	0.515	0.435	0.394	0.481	0.741	402
Padua/Int/distributed_effort_p10_t300.out	974	0.276	0.306	0.499	0.664	0.733	0.592	0.481	0.122	0.794	441
Padua/Int/2018_stem_original_p50_t1000.out	836	0.29	0.306	0.498	0.688	0.795	0.643	0.542	0.493	0.988	1139
Padua/Int/distributed_effort_p10_t100.out	1114	0.248	0.315	0.444	0.604	0.704	0.458	0.372	0.402	0.45	156
Padua/Int/2018_stem_original_p50_t200.out	1185	0.29	0.312	0.499	0.693	0.792	0.63	0.481	0.331	0.911	334
Padua/Int/baseline_bm25_t1000.out	1241	0.222	0.191	0.282	0.408	0.524	0.446	0.392	0.471	0.827	682
Padua/Int/distributed_effort_p10_t500.out	991	0.278	0.306	0.499	0.664	0.743	0.606	0.483	0.115	0.842	594
Padua/Int/baseline_bm25_t300.out	1262	0.222	0.187	0.286	0.41	0.523	0.44	0.398	0.506	0.664	270
Padua/Int/baseline_bm25_t100.out	1397	0.223	0.186	0.291	0.429	0.557	0.414	0.368	0.485	0.507	99
Padua/Int/2018_stem_original_p50_t400.out	985	0.29	0.307	0.501	0.685	0.767	0.646	0.514	0.374	0.949	572
Padua/Int/2018_stem_original_p50_t300.out	1144	0.29	0.313	0.495	0.682	0.788	0.639	0.497	0.355	0.933	462
Padua/Int/2018_stem_original_p50_t100.out	1150	0.284	0.306	0.483	0.653	0.752	0.556	0.481	0.362	0.728	188
Padua/Int/distributed_effort_p10_t200.out	965	0.271	0.306	0.482	0.651	0.752	0.56	0.445	0.165	0.714	312
Padua/Int/baseline_bm25_t400.out	1242	0.222	0.191	0.286	0.412	0.523	0.434	0.393	0.485	0.713	337
Padua/Int/2018_stem_original_p50_t1500.out	796	0.29	0.306	0.498	0.688	0.785	0.642	0.553	0.541	0.999	1425
Padua/Int/2018_stem_original_p50_t500.out	1001	0.29	0.306	0.501	0.691	0.779	0.65	0.505	0.395	0.961	677
Padua/Int/baseline_bm25_t1500.out	1203	0.222	0.191	0.282	0.411	0.533	0.453	0.399	0.461	0.933	932
Padua/Int/baseline_bm25_t200.out	1263	0.222	0.189	0.284	0.417	0.535	0.438	0.396	0.466	0.624	191
Padua/Int/distributed_effort_p10_t400.out	981	0.277	0.306	0.499	0.663	0.734	0.595	0.483	0.116	0.822	518
Sheffield/Int/Int_sheffield-Log_likelihood.out	1132	0.293	0.258	0.378	0.583	0.695	0.458	0.381	0.599	1	2100
Sheffield/Int/Int_sheffield-Odds_Ratio.out	1070	0.261	0.267	0.404	0.569	0.7	0.462	0.384	0.599	1	2100
Sheffield/Int/Int_sheffield-baseline.out	1276	0.245	0.22	0.334	0.507	0.653	0.47	0.386	0.599	1	2100
Sheffield/Int/Int_sheffield-Chi_Squared.out	1149	0.262	0.238	0.36	0.537	0.687	0.469	0.415	0.599	1	2100

Table 4. Prognosis studies with abstract-level QRELs

Run	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Rely	R@k	k
ILPS/Pro/abs/abs-hh-ratio-ilps@uva	2885	0.673	0.562	0.714	0.875	0.911	0.591	0.143	0.018	0.948	1221
ILPS/Pro/abs/abs-th-ratio-ilps@uva	2537	0.628	0.521	0.682	0.818	0.927	0.566	0.247	0.014	0.922	867
Padua/Pro/abs/2018_stem_original_p10_t400	2967	0.235	0.214	0.484	0.812	0.901	0.567	0.119	0.035	0.828	735
Padua/Pro/abs/distributed_effort_p10_t1500	2594	0.235	0.214	0.484	0.812	0.896	0.554	0.23	0.049	0.99	2165
Padua/Pro/abs/2018_stem_original_p10_t1000	2644	0.235	0.214	0.484	0.812	0.896	0.554	0.215	0.022	0.943	1332
Padua/Pro/abs/2018_stem_original_p10_t200	2911	0.242	0.214	0.536	0.812	0.901	0.53	0.135	0.162	0.599	398
Padua/Pro/abs/2018_stem_original_p10_t500	2920	0.235	0.214	0.484	0.812	0.891	0.56	0.133	0.027	0.859	832
Padua/Pro/abs/2018_stem_original_p10_t300	2955	0.239	0.214	0.547	0.818	0.891	0.556	0.122	0.054	0.776	597
Padua/Pro/abs/2018_stem_original_p10_t1500	2578	0.235	0.214	0.484	0.812	0.896	0.554	0.234	0.035	0.984	1831
Padua/Pro/abs/distributed_effort_p10_t1000	2563	0.235	0.214	0.484	0.812	0.896	0.554	0.239	0.026	0.974	1566
Padua/Pro/abs/2018_stem_original_p10_t100	2802	0.259	0.286	0.562	0.797	0.891	0.6	0.168	0.411	0.359	198
Padua/Pro/abs/baseline_bm25_t500	3343	0.071	0.057	0.13	0.281	0.422	0.084	0.007	0.621	0.214	501
Padua/Pro/abs/distributed_effort_p10_t300	2964	0.235	0.214	0.484	0.812	0.906	0.567	0.12	0.038	0.818	709
Padua/Pro/abs/2018_stem_original_p50_t1000	2556	0.221	0.214	0.484	0.74	0.87	0.571	0.241	0.041	0.995	1981
Padua/Pro/abs/distributed_effort_p10_t100	2789	0.252	0.25	0.568	0.786	0.875	0.594	0.172	0.288	0.464	248
Padua/Pro/abs/2018_stem_original_p50_t200	2911	0.242	0.214	0.536	0.812	0.901	0.53	0.135	0.162	0.599	398
Padua/Pro/abs/baseline_bm25_t1000	3346	0.07	0.057	0.13	0.276	0.396	0.057	0.006	0.382	0.391	1001
Padua/Pro/abs/distributed_effort_p10_t500	2708	0.235	0.214	0.484	0.812	0.891	0.566	0.196	0.026	0.87	955
Padua/Pro/abs/baseline_bm25_t300	3350	0.071	0.057	0.135	0.276	0.385	0.104	0.005	0.794	0.109	301
Padua/Pro/abs/baseline_bm25_t100	3350	0.066	0.047	0.13	0.255	0.365	0.059	0.005	0.939	0.031	101
Padua/Pro/abs/2018_stem_original_p50_t400	2955	0.231	0.214	0.484	0.807	0.896	0.556	0.122	0.033	0.839	798
Padua/Pro/abs/2018_stem_original_p50_t300	2955	0.239	0.214	0.547	0.818	0.891	0.556	0.122	0.054	0.776	597
Padua/Pro/abs/2018_stem_original_p50_t100	2802	0.259	0.286	0.562	0.797	0.891	0.6	0.168	0.411	0.359	198
Padua/Pro/abs/distributed_effort_p10_t200	2968	0.24	0.214	0.542	0.807	0.906	0.548	0.119	0.079	0.724	501
Padua/Pro/abs/baseline_bm25_t400	3347	0.071	0.057	0.13	0.281	0.417	0.109	0.006	0.696	0.167	401
Padua/Pro/abs/2018_stem_original_p50_t1500	1975	0.219	0.214	0.484	0.74	0.828	0.5	0.413	0.091	1	2966
Padua/Pro/abs/2018_stem_original_p50_t500	2660	0.228	0.214	0.484	0.807	0.891	0.576	0.21	0.022	0.891	993
Padua/Pro/abs/baseline_bm25_t1500	3346	0.07	0.057	0.13	0.276	0.396	0.05	0.006	0.258	0.516	1501
Padua/Pro/abs/baseline_bm25_t200	3350	0.069	0.057	0.125	0.266	0.385	0.111	0.005	0.86	0.073	201
Padua/Pro/abs/distributed_effort_p10_t400	2920	0.235	0.214	0.484	0.812	0.891	0.56	0.133	0.028	0.854	830
Sheffield/Pro/abs/Pro_sheffield-baseline	2990	0.126	0.146	0.255	0.448	0.594	0.247	0.112	0.117	1	3367
Sheffield/Pro/abs/Pro_sheffield-relevence_feedback	2775	0.141	0.151	0.307	0.484	0.646	0.305	0.176	0.117	1	3367

Table 5. Qualitative studies with abstract-level QRELs

Run	L_Rel	MAP	R@5%	R@10%	R@20%	R@30%	WSS95	WSS100	Rely	R@k	k
ILPS/Qual/abs/abs-hh-ratio-ilps@uva.out	1796	0.204	0.478	0.655	0.876	0.929	0.417	0.397	0.326	0.919	1247
ILPS/Qual/abs/abs-th-ratio-ilps@uva.out	2564	0.187	0.487	0.628	0.805	0.92	0.398	0.215	0.341	0.878	1158
Padua/Qual/abs/2018_stem_original_p10_t400.out	2547	0.109	0.496	0.717	0.779	0.894	0.302	0.183	0.568	0.387	704
Padua/Qual/abs/distributed_effort_p10_t1500.out	2544	0.109	0.496	0.743	0.77	0.885	0.268	0.168	0.37	0.745	2098
Padua/Qual/abs/2018_stem_original_p10_t1000.out	2662	0.109	0.496	0.743	0.77	0.885	0.273	0.141	0.29	0.714	1320
Padua/Qual/abs/2018_stem_original_p10_t200.out	2934	0.089	0.478	0.522	0.699	0.805	0.216	0.101	0.627	0.266	397
Padua/Qual/abs/2018_stem_original_p10_t500.out	2535	0.109	0.496	0.743	0.77	0.894	0.301	0.185	0.578	0.396	820
Padua/Qual/abs/2018_stem_original_p10_t300.out	2660	0.103	0.496	0.655	0.752	0.858	0.303	0.159	0.582	0.338	554
Padua/Qual/abs/2018_stem_original_p10_t1500.out	2534	0.109	0.496	0.743	0.77	0.885	0.268	0.17	0.447	0.732	1819
Padua/Qual/abs/distributed_effort_p10_t1000.out	2469	0.109	0.496	0.743	0.77	0.885	0.295	0.199	0.628	0.491	1515
Padua/Qual/abs/2018_stem_original_p10_t100.out	2996	0.071	0.327	0.416	0.637	0.796	0.186	0.09	0.726	0.167	198
Padua/Qual/abs/baseline_bm25_t500.out	2700	0.051	0.274	0.425	0.469	0.611	0.412	0.256	0.683	0.221	501
Padua/Qual/abs/distributed_effort_p10_t300.out	2518	0.109	0.496	0.743	0.77	0.894	0.309	0.193	0.547	0.396	684
Padua/Qual/abs/2018_stem_original_p50_t1000.out	2438	0.116	0.496	0.743	0.92	0.947	0.357	0.194	0.545	0.745	1977
Padua/Qual/abs/distributed_effort_p10_t100.out	2920	0.083	0.416	0.469	0.681	0.814	0.258	0.106	0.659	0.221	244
Padua/Qual/abs/2018_stem_original_p50_t200.out	2934	0.089	0.478	0.522	0.699	0.805	0.216	0.101	0.627	0.266	397
Padua/Qual/abs/baseline_bm25_t1000.out	3040	0.055	0.274	0.425	0.496	0.788	0.239	0.101	0.278	0.601	1001
Padua/Qual/abs/distributed_effort_p10_t500.out	2641	0.109	0.496	0.743	0.77	0.894	0.295	0.162	0.553	0.446	924
Padua/Qual/abs/baseline_bm25_t300.out	2697	0.049	0.274	0.372	0.451	0.628	0.294	0.257	0.726	0.171	301
Padua/Qual/abs/baseline_bm25_t100.out	2700	0.056	0.301	0.389	0.637	0.743	0.399	0.256	0.845	0.086	101
Padua/Qual/abs/2018_stem_original_p50_t400.out	2566	0.109	0.496	0.717	0.779	0.894	0.293	0.174	0.594	0.387	795
Padua/Qual/abs/2018_stem_original_p50_t300.out	2687	0.103	0.496	0.655	0.752	0.858	0.29	0.147	0.591	0.338	595
Padua/Qual/abs/2018_stem_original_p50_t100.out	2996	0.071	0.327	0.416	0.637	0.796	0.186	0.09	0.726	0.167	198
Padua/Qual/abs/distributed_effort_p10_t200.out	2762	0.104	0.496	0.673	0.761	0.867	0.303	0.135	0.56	0.347	486
Padua/Qual/abs/baseline_bm25_t400.out	2700	0.052	0.274	0.434	0.469	0.619	0.417	0.256	0.694	0.203	401
Padua/Qual/abs/2018_stem_original_p50_t1500.out	1970	0.116	0.496	0.743	0.92	0.965	0.356	0.301	0.532	1	2568
Padua/Qual/abs/2018_stem_original_p50_t500.out	2576	0.11	0.496	0.743	0.788	0.894	0.283	0.168	0.624	0.405	991
Padua/Qual/abs/baseline_bm25_t1500.out	3039	0.055	0.274	0.425	0.496	0.779	0.24	0.101	0.382	0.669	1501
Padua/Qual/abs/baseline_bm25_t200.out	2698	0.053	0.274	0.381	0.619	0.726	0.395	0.256	0.764	0.14	201
Padua/Qual/abs/distributed_effort_p10_t400.out	2636	0.109	0.496	0.743	0.77	0.894	0.301	0.165	0.545	0.432	804
Sheffield/Qual/abs/Qual_sheffield-relevance_feedback.out	2940	0.06	0.274	0.549	0.717	0.832	0.185	0.103	0.593	1	3268
Sheffield/Qual/abs/Qual_sheffield-baseline	3031	0.051	0.265	0.451	0.619	0.743	0.135	0.082	0.593	1	3268

4 Conclusions

This paper provided an overview of the CLEF eHealth 2019 evaluation lab. The CLEF eHealth series began its life as a scientific workshop in 2012 with an aim of establishing an evaluation lab [26]. This ambition was realised in 2013, with the running of the first annual CLEF eHealth evaluation lab. Since 2013, this annual lab has run two or more preceding shared tasks each year, in other words, the CLEF eHealth 2013–2019 evaluation labs [8,9,14,16,28,29]. During these past eight years, the CLEF eHealth series has 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.

The CLEF eHealth 2019 lab ran two shared tasks: Task 1 on multilingual information extraction to extend the 2018 task on French, Hungarian, and Italian corpora to German; and Task 2 on technologically assisted reviews in empirical medicine building on the 2018 task in English. In addition, a Task 3 on consumer health search in mono- and multilingual settings was initially advertised, but unfortunately, due to unforeseen circumstances, this task had to be postponed.

Test collections generated by this year's CLEF eHealth 2019 lab offered a specific task definition, implemented in a data set distributed together with an implementation of relevant evaluation metrics to allow for direct comparability of the results reported by systems evaluated on the collections. The CLEF eHealth information extraction task (Task 1) used a traditional shared task model for evaluation in which a community-wide evaluation is executed in a controlled setting: independent training and test data sets 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 are to submit their outputs from the frozen systems to the task organizers, who are to evaluate these results and report the resulting benchmarks to the community. The CLEF technologically assisted reviews task (Task 2) also followed the same setting with independent training and test data sets and all participants gaining access to the test data at the same time; however, labels on the test data were provided to participants to allow for the development of interactive retrieval systems.

Given the significance of the CLEF eHealth tasks over the years, all problem specifications, test collections, and text analytics resources associated with the 2019 and previous years' lab tasks have been made available to the wider research community. They can be found on our CLEF eHealth website[11].

Acknowledgements. We gratefully acknowledge the contribution of the people and organizations involved in the CLEF eHealth 2019 evaluation lab as participants or organizers. The lab has been supported in part by (in alphabetical order) The Australian National University, College of Engineering and Computer Science, Research School

[11] http://clef-ehealth.org (last accessed on 24 May 2019).

of Computer Science; the CLEF Initiative; and Data61/Commonwealth Scientific and Industrial Research Organisation. We thank Dr Benjamin Lecouteux (LIG, Université Grenoble Alpes) for his help in Task 3. We are also thankful to the people involved in the task preparation, data annotation, query creation, and relevance assessment exercise. Last but not least, we gratefully acknowledge the participating teams' hard work. We thank them for their submissions and interest in the lab.

References

1. Adnan, M., Warren, J., Suominen, H.: Patient empowerment via technologies for patient-friendly personalized language. In: Grando, A.M., Rozenblum, R., Bates, D. (eds.) Engaging Patients with Health Information Technology, pp. 147–158. De Gruyter, Berlin (2015)
2. Bert, B., et al.: Rethinking 3R strategies: digging deeper into animaltestinfo promotes transparency in in vivo biomedical research. PLOS Biol. **15**(12), 1–20 (2017). https://doi.org/10.1371/journal.pbio.2003217
3. Cormack, G.V., Grossman, M.R.: Engineering quality and reliability in technology-assisted review. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, pp. 75–84. ACM, New York (2016). https://doi.org/10.1145/2911451.2911510
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
5. Goeuriot, L., et al.: 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)
6. Goeuriot, L., et al.: An analysis of evaluation campaigns in ad-hoc medical information retrieval: CLEF eHealth 2013 and 2014. Inf. Retrieval J. **21**, 507–540 (2018)
7. Goeuriot, L., et al.: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: user-centred health information retrieval. In: CLEF 2014 Evaluation Labs and Workshop, Sheffield, UK, Online Working Notes (2014)
8. Goeuriot, L., et al.: Overview of the CLEF eHealth evaluation lab 2015. In: Mothe, J., et al. (eds.) CLEF 2015. LNCS, vol. 9283, pp. 429–443. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_44
9. Goeuriot, L., et al.: CLEF 2017 eHealth evaluation lab overview. In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 291–303. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_26
10. Jimmy, Z.G., Palotti, J.: 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)
11. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2017 technologically assisted reviews in empirical medicine overview. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings (2017)
12. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2018 technologically assisted reviews in empirical medicine overview. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings (2018)
13. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: CLEF 2019 technology assisted reviews in empirical medicine overview. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings (2019)

14. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2016. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 255–266. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_24
15. Kelly, L., et al.: CLEF ehealth 2019 evaluation lab. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11438, pp. 267–274. Springer, Heidelberg (2019). https://doi.org/10.1007/978-3-030-15719-7_36
16. Kelly, L., et al.: Overview of the ShARe/CLEF eHealth evaluation lab 2014. In: Kanoulas, E., et al. (eds.) CLEF 2014. LNCS, vol. 8685, pp. 172–191. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11382-1_17
17. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. CoRR abs/1901.08746 (2019). http://arxiv.org/abs/1901.08746
18. Leeflang, M.M., Deeks, J.J., Takwoingi, Y., Macaskill, P.: Cochrane diagnostic test accuracy reviews. Syst. Rev. **2**(1), 82 (2013)
19. Névéol, A., et al.: Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Working Notes, CEUR Workshop Proceedings. CEUR-WS.org, ISSN 1613–0073 (2016). http://ceur-ws.org/Vol-1609/
20. Névéol, A., et al.: CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French. In: CLEF 2017 Online Working Notes. CEUR-WS (2017)
21. Névéol, A., et al.: CLEF eHealth 2018 multilingual information extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian. In: CLEF 2018 Online Working Notes. CEUR-WS (2018)
22. Neves, M., et al.: Overview of task 1 in CLEF eHealth 2019: indexing German non-technical summaries of animal experiments. In: CLEF 2019 Online Working Notes. CEUR-WS (2019)
23. Palotti, J., et al.: CLEF eHealth evaluation lab 2015, Task 2: retrieving information about medical symptoms. In: CLEF 2015 Online Working Notes. CEUR-WS (2015)
24. Palotti, J., Zuccon, G., Hanbury, A.: Consumer health search on the web: study of web page understandability and its integration in ranking algorithms. J. Med. Internet Res. **21**(1), e10986 (2019). https://doi.org/10.2196/10986
25. Palotti, J., et al.: 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)
26. 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 (2012). http://ceur-ws.org/Vol-1178/
27. 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 Res. Protoc. **7**(7), e10961 (2018)
28. Suominen, H., et al.: Overview of the CLEF eHealth evaluation lab 2018. In: Bellot, P., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. LNCS, vol. 11018, pp. 286–301. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_26

29. Suominen, H., et al.: Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 212–231. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40802-1_24
30. Zuccon, G., et al.: 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