Original Paper

The Scholarly Influence of the CLEF eHealth Initiative by the Conference and Labs of the Evaluation Forum: Review and Bibliometric Study of the 2012-2017 Outcomes

Abstract

Background: The eHealth initiative of the Conference and Labs of the Evaluation Forum (CLEF) has aimed since 2012 to gather researchers working on health text analytics and to provide them with annual workshop, shared development challenges/tasks, benchmark datasets, and software for processing and evaluation. The overall purpose of this initiative is to ease and support patients, their next-of-kin, clinical staff, health scientists, and healthcare policy makers in accessing, understanding, using, and authoring health information in a multilingual setting.
Objective: This original research paper reports on the outcomes of the first six installations of CLEF eHealth from 2012 to 2017. The focus is on measuring and analysing the scholarly influence by reviewing CLEF eHealth papers, together with relevant citation metrics.
Methods: A review and bibliometric study of the CLEF eHealth proceedings, working notes, and author-declared paper extensions was conducted. Citation data for these publications were collected from Google Scholar. Citation content analysis was used for the publications and their citations.
Results: The large number of registrations, submissions, and citations demonstrate the substantial community interest in the tasks and their resources. In total, 718 teams have registered their interest in the tasks, leading to 130 teams submitting to the 15 tasks. 184 papers using CLEF eHealth data generated 1,299 citations, yielding a total scholarly citation influence of almost 963,000 citations for the 741 co-authors and included authors from 33 countries across the world. The tasks' evaluation outcomes contribute to the knowledge of the difficulty of the research challenges the tasks address and the applicability of particular methods in solving these challenges, with typically statistically significant improvements in processing quality.
Conclusions: These outcomes encourage continuing to develop these technologies to address patient needs. Consequently, data and tools have been opened for future research and development and the CLEF eHealth initiative continues to run new challenges.
Keywords: Evaluation Studies as Topic; Health Records; Information Extraction; Information Storage and Retrieval; Information Visualization; Patient Education as Topic; Speech Recognition; Systematic Reviews; Test-set Generation; Text Classification

Introduction

Health information refers to all health-related content in all data formats, document types, information systems, publication media, and languages from all organisations, states, and countries. The privacy-sensitive, official part of health information consists of data recorded in healthcare services when describing a given patient's health or healthcare (Figure 1). The accessibility of this data is defined as limited (i.e., private or confidential information) and it is recorded either on paper in health records or electronically in *Electronic Health* (eHealth) records. Some common synonyms or related terms include eHealth charts, data, documents, information, letters, notes, reports, and summaries. Typical ways of making the term more specific consist of detailing the record type (e.g., admit/admission document, case sheet, discharge document, or handover form), adding the recording method (e.g., computer based, computerised, digital, electronic, or paper based), and adding the information format (e.g., categorical, free/free-form text, numeric, structured, or textual). The health records term can be generalised over the healthcare professions by using the term clinical records instead of, for example, dental records, medical records, and nursing records for documents authored by dentists, medical doctors, and nurses, respectively.

Figure 1 to be placed approximately here.

Figure 1. Illustration of terminological differences used to refer to health records in the *Medical Subject Headings* (MeSH), created and updated by the US National Library of Medicine from 2006 to 2012 and between MeSH and Australian terminology. The year 2012 has been chosen to reflect the linguistic landscape at the time of introducing CLEF eHealth in 2012.

Figure 2 to be placed approximately here.

Figure 2. Original text, its enrichment, and submission statistics from the CLEF eHealth 2013 evaluation lab. Abbreviations: *Systematized Nomenclature of Medicine — Clinical Terms* (SNOMED CT) and *Unified Medical Language System* (UMLS). The year of 2013 has been chosen as an example here to illustrate the outcomes of the first year of organizing CLEF eHealth evaluation lab.

The requirement to assure that patients can understand their own care epicrises, discharge summaries, and other health records are stipulated by policies and laws [1]. As an international example, *A Declaration on the Promotion of Patients' Rights in Europe* by the *World Health Organization* (WHO) in 1994 states that all patients have the right to be fully informed about their own health status, prognosis, medical conditions, diagnoses, proposed and alternative treatment with potential risks and benefits, effects of non-treatment, treatment progress, and discharge guidelines. It also obligates healthcare workers to give every patient a written summary of this information and communicate in a way appropriate to this patient's capacity for understanding, including minimised use of unfamiliar jargon. Similarly, the *Finnish Act on the Status and Rights of Patients 785/1992*, *Finnish Statute 298/2009 on Patient Documents*, *Swedish Act on the Patient's Right to Information 06:So559/2005*, and *Swedish Patient Data Law 255/2008* obligate the following four constraints on health documentation: First, the health records must adequately detail the patient's conditions, care, and recovery. Second, they have to cover all necessary information. Third, their content must be explicit and comprehensive. Fourth, they can include only generally well-known, accepted terms and shorthand.

However, patients, their next-of-kin, and other laypersons are likely to experience difficulties in understanding the arcane jargon of health records and improving this readability can contribute to *patient empowerment* [2], defined as providing partial control and mastery over health and care which leads to patients having an active role in their healthcare, making better health/care decisions, being more independent from healthcare services, and having decreased costs of care [3]. This could mean replacing jargon words with patient-friendly synonyms; expanding shorthand; and an option to see the original text (Figure 2). The *Medical Subject Headings* (MeSH), *Systematized Nomenclature of Medicine — Clinical Terms* (SNOMED CT), *Unified Medical Language System* (UMLS), and other terminology standards can help in defining synonym replacements, but automated language processing is needed to identify text snippets to be replaced with synonymous snippets.

Patient-friendly language in health records can help patients make informed decisions, but this also depends on their access to consumer leaflets and other *supportive further information* about their health concerns. The Internet is a powerful source of this information; most people will turn to its large range of content that are widely accessible and searchable [4, 5]. However, search for medical information online for layperson can lead to the escalation of these concerns and consequent anxiety [6]. Hence, helping patients retrieve relevant, understandable and reliable information on the Internet is crucial.

Personally Controlled electronic Health Records (PCeHRs) on the Internet can be used to naturally bridge patients' actions of reading their own health records to searching further information. PCeHRs are targeted to both patients and healthcare workers for reading, writing, and sharing information [7]. They are becoming increasingly common and have been open, for example, in Australia since 2012. If combined with the aforementioned record processing, this could mean enriching the health record with hyperlinks to term definitions, care guidelines, and other information on patient-friendly and reliable sites on the Internet (Figure 2) as one way to facilitate patients in understanding their health and healthcare; the current de facto approach is the patients themselves searching the disconnected and sometimes unreliable documents across the Internet [2].

This paper reports on the six installations of CLEF eHealth, organised as part of the *Conference and Labs of the Evaluation Forum* (CLEF) initiative in 2012–2017. In 2012, it ran as a scientific workshop with an aim of establishing an evaluation campaign and in 2013–2017, this annual workshop has been supplemented with three or more preceding evaluation labs each year. An *evaluation lab* (a.k.a. evaluation campaign/initiative, community challenge, computational competition, hackathon, or shared task) is an activity where the participating individuals or teams' goal is to solve the same problem, typically using the same dataset in a given time frame. The 2013 and 2014 labs were organised in collaboration with the *Shared Annotated Resources* (ShARe) and *Knowledge Helper for Medical and Other Information users* (Khresmoi) projects.

The overall purpose of CLEF eHealth is to ease and support patients, their next-of-kin, clinical staff, health scientists, and healthcare policy makers in accessing, understanding, using, and authoring various types of health information in a multilingual setting. In the CLEF eHealth 2013–2017 installations, the aim was to address patient-centric text processing. From 2015, the scope was also extended to ease both patients' understanding and clinicians' authoring of various types of medical content. CLEF eHealth 2017 also introduced a new pilot task on *technology assisted reviews* (TARs) in empirical medicine in order to support health scientists and healthcare policymakers' information access.

Our focus in this article is on measuring and analysing the *scholarly influence* of CLEF eHealth. We review papers on CLEF eHealth problem specifications, evaluation methods, benchmark results, evaluation data releases, software launches, and influence in nurturing real or demonstration systems, together with relevant citation indices and participation counts. It extends a short, invited chapter about the CLEF eHealth problem specifications, data, and citations [8].

Methods

The scholarly influence of the CLEF eHealth installations was measured by conducting a *bibliometric study* — an established method to provide a quantitative and qualitative indication of scientific activities whose use is also emerging in the context of evaluation initiatives [9-11] — of the publications generated as a result of these installations in 2012–2017 and their citations received by 31 October 2017.

This measuring consisted of the following three standard steps: 1. *publication data collection*, 2. *citation data collection*, and 3. *data analysis*. The first two steps were concerned with the collection of materials for the measurement exercise. The last step formed the method of the study.

Publication data was collected from the CLEF proceedings and *Working Notes* (WNs). This publication data collection was limited to the *Conference Papers* (CPs) and WNs relevant to CLEF eHealth in the CLEF Conference Proceedings [12-17] and

CLEF WN Proceedings [18-23]. Its results were supplemented by the known journal extensions of the CLEF eHealth overviews and *previously enlisted papers* that use the CLEF eHealth datasets [8]. WNs were technical reports written by participants describing their participation in the lab.

Citation data for the resulting publication data was collected from *Google Scholar* — one of the most comprehensive citation data sources in general and in particular for computer science, which is the main field of many CLEF eHealth scientists [9-11].

The method used for the data analysis of the third step was the *citation content analysis* [24], founded on the *content analysis* [25] and *grounded theory* [26]. This allowed a systematic, replicable compression of materials from the first and second step as codes and testing of hypotheses about both the quantity and quality of the scholarly influence of CLEF eHealth in 2012–2017. It was chosen over the more established and popular content analysis (Google Scholar had over 15,000 citations of [25]) and grounded theory (Google Scholar had nearly 400 citations of the 2007 revision [26] of the grounded theory – introduced already in 1960s – for health sciences) because it combined these two research techniques for interpreting meaning from the content of text data (incl. the task of reviewing scientific literature, considered by the grounded theory) as one overarching method. All these three methods were based on content coding and analysis of the codified content, and thereby, fundamentally similar.

The method of using Google Scholar for citation data collection in bibliometric studies had at least the following two shortcomings [27-29]: First, paper duplication as a citation entry was frequent in Google Scholar, for example, due to misspellings or incorrectly identified years and would, without manual refinement cause errors in the counts. Another source for counting errors was incorrect automated merging of citation entries by Google Scholar, for example, because of the same or almost the same title of a given conference paper and its journal extension. In order to alleviate counting problems, our citation counts by Google Scholar were reviewed and refined for these two shortcomings by hand. A similar refinement was made in the scholarly influence measurement of CLEF 2000–2009 [10].

As part of the citation content analysis, the included publication data and their respective refined citation data were codified for the following ten content categories: participation (incl. both *expression of interest* (EOI and submission), author, affiliation, problem specification, evaluation method, benchmark result, data release, software launch, demonstration system, and citation. Similarly to [10], attention was paid to not only the number of citations, but also the number of authors, their affiliations, and affiliation countries. In order to illustrate the influence not only to the scholarly community but also to the individual scholars (because most participating teams included graduate students and/or early career academics), the scholarly influence (i.e., 962,559 = 1,299 × 741) was computed by multiplying the number of citations (i.e., 1,299) for the included papers (i.e., 184) by the number of their co-authors (i.e., 741). Hence, the more traditional *scholarly* *impact* [9-11] (i.e., the number of citations that was 1,299 above) could be calculated by dividing the reported scholarly influence (i.e., 962,559) by the number of citations (1,299 above).

Results

Citation Analysis from 2012 to 2017

In 2012, the CLEF initiative introduced eHealth as a workshop that focused on eHealth documents and related analytics with a goal to spin out an evaluation lab [30]. Its program included first three invited talks on collaborative data and software resources; oral talks for eleven papers; and a student mentoring session where PhD students presented their study plans, followed by feedback from their designated mentors. All these talks focused on meeting the needs of healthcare professionals and patients in ease of information recording, access, and understanding via user-centred abbreviation processing, content generation, search engines, and vocabularies, among other tools to support patient–professional interaction across languages and sub-languages. Then, the program continued to verifying this community interest in user-friendly multilingual communication through an expert panel, professional networking session, and a working session for developing a road map for CLEF eHealth 2013.

The topic of patient-friendly multilingual communication formed the focus of the annual CLEF eHealth evaluation labs in 2013–2017 [31-35], generated the total scholarly influence of 962,559 citations for the 184 CLEF eHealth papers, and reached authors from 33 countries across the world (Table 1, Figure 3) [8]. This influence was computed by multiplying the number of co-authors in the 184 papers (i.e., 741) by the number of citations (i.e., 1,299) these papers had received on 26 October 2017. 143 out of the 184 papers (77.7%) had been cited at least once and the maximum, mean, median, and standard deviation of citations per paper were 147, 7, 3, and 15, respectively. The *h-index* (i.e., the number of papers each of which with at least *h* citations) and *i10-index* (i.e., the number of papers with at least 10 citations) were 18 and 35, respectively. The annual number of published papers was 16, 35, 34, 31, 33, and 35, in 2012, 2013, 2014, 2015, 2016, and 2017, respectively. Although a clear 158 majority of the 184 papers were WN publications (85.9%), 22 CP (12.0%) and 4 *Journal Papers* (JPs) (2.0%) were also published.

In accordance with the CLEF eHealth mission to foster teamwork, the number of co-authors per paper was 4 on average, with a maximum, median, minimum, and standard deviation of 15, 3, 1, and 3, respectively. In 47 out of the 184 papers (25.5%), this co-authoring collaboration was international and sometimes even across continents (i.e., Africa — Europe, Asia — Australia, Asia — Europe, Asia — North America, Australia — Europe, Australia — Europe — North America, and Europe — South America).

CLEF eHealth particularly welcomed and attracted multi-disciplinary teams to collaborate and bridge the academy, government, and industrial researchers, scientists, lecturers, and graduate students with engineers, practitioners, and policy makers [31-35]. For example, in the 33 WNs and 1 CP from the CLEF eHealth 2013 evaluation lab, 162 authors from 10 countries highlighted some leading organizations in health information management, extraction, and retrieval, including, The Australian *National Information and Communications Technology Australia* (NICTA), *Commonwealth Scientific and Industrial Research Organisation* (CSIRO), and Health Language Laboratories; Chinese Canon Information Technology (Beijing); French National Center for Scientific Research; Indian RelAgent Private Lt; US National Center for Biotechnology Information, Kaiser Permanente, and Mayo Clinic; and universities from the Australian Capital Territory, New South Wales, Queensland, and Victoria, China, Finland, Ireland, Republic of Korea, Spain, Sweden, UK, and US Alabama, California, Colorado, Maryland, New York, Oregon, Pennsylvania, Texas, Utah, and Virginia [31].

Figure 3 to be placed approximately here.

Figure 3. Map of CLEF eHealth 2012–2017 authors' affiliation countries. Of the total of 33 affiliation 17 were in Europe, 9 in Asia, 2 in Africa, 2 in North America, 1 in Middle East, 1 in Pacific, and 1 in South America.

Table 1. Summary of the bibliometric analysis of CLEF eHealth in 2013–2017. See [8] for extended tabulations with details and citation for each paper. Abbreviations: *conference paper* (CP), *expression of interest* (EOI), *information retrieval* (IR), *journal paper* (JP), and *working note* (WN).

	ask 1No. of EOIs (participating	ask 2No. of EOIs (participating	ask 3No. of EOIs (participating	totalNo. of EOIs (participating	No. of WNs	No. of CPs	No. of JPs	No. of authors	No. of countries	No. of citations
CLEF eHealth 2012										
	-	-	-	-	16	0	0	50	8	20
CLEF eHealth 2013										
	64 (22)	56 (5)	55 (9)	175 (34)	33	1	0	162	10	458
CLEF eHealth 2014										
	50 (1)	79 (10)	91 (14)	220 (24)	28	1	0	107	22	273
CLEF eHealth 2015										
	20 (2)	17 (7)	53 (12)	90 (20)	23	1	0	91	19	138
CLEF eHealth										

2016										
	25 (3)	33 (7)	58 (10)	116 (20)	23	1	0	113	16	110
CLEF eHealth 2017										
	34 (11)	40 (14)	43 (7)	117 (32)	34	1	0	128	22	70
Other papers that use CLEF eHealth data										
	-	-	-	-	0	17	5	90	8	230
In total										
	193 (39)	225 (43)	300 (52)	718 (134)	157	22	5	741	33	1,299

Problem Specifications from 2013 to 2017

The first two installations of the lab, held in 2013 [31] and 2014 [32], focused on text processing, search, and visualisation to ease patients' (or their next-of-kin) ease in understanding their hospital discharge summaries. Each year, three tasks were organised.

The *2013 Tasks 1a and 1b* considered *disorder naming* (e.g., *heartburn* as opposed to *gastroesophageal reflux disease*) by identification of disorder names (e.g., automatically recognising the two text snippets ***bold italics*** in *80 y/o male with 2 yr h/o* ***SOB*** *and* ***GERD***.) and *normalisation of the identified names* by translating them to patient-friendly synonyms (e.g., mapping *GERD* to *heartburn* as opposed to *gastroesophageal reflux disease* in the aforementioned sentence), respectively.

The *2013 Task 2* on *shorthand expansion* aimed at mapping clinical abbreviations and acronyms to patient-friendly synonyms (e.g., automatically expanding and mapping the three text snippets in ***bold italics*** in *80 y/o male with 2 yr* ***h/o SOB*** *and* ***GERD*** to *history of*, *shortness of breath*, and *heartburn*, respectively). Instead of actually writing the disorder names and shorthand expansions in the 2013 Tasks 1b and 2, the respective SNOMED CT and UMLS codes (e.g., *GERD* got the SNOMED CT code *C0017168* in

Task 1b and UMLS code *C0018834* in Task 2) were applied. These coding systems were chosen because they are among the most commonly used in clinical settings.

This challenge continued in the *2014 Task 2* on *template filling*, with the aim of developing attribute classifiers that predict the values of the UMLS *concept unique identifiers* (CUIs) with mention boundaries. The Disease/Disorder Templates consisted of the following ten attributes: Negation Indicator, Subject Class, Uncertainty Indicator, Course Class, Severity Class, Conditional Class, Generic Class, Body Location, DocTime Class, and Temporal Expression.

The *2013 Task 3, 2014 Task 3, and 2014 Task 1* supplemented the processing of health records with information from the Internet, based on the patient's (and next-of-kin's) information needs associated with the records. The 2013 and 2014 Task 3 on *information search* (a.k.a. *information retrieval* (IR)) would, for example, find the *definition of shortness of breath*, *treatment guidelines for heartburn*, and *guidelines on separating the symptoms of heart conditions from heartburn* for the health record with the aforementioned sentence. The challenge also considered in 2014 the problem of an individual expressing their information need in a non-English language, for search on web pages written in *English* (EN). Support of this functionality is important given the large proportion of web medical content written in EN. The 2014 Task 1 on *interactive information visualisation* had the overall goal of designing an effective, usable, and trustworthy environment for navigating, exploring, and interpreting health information as needed to promote understanding and informed decision-making. It was divided into two parts as linkages to the three 2013 tasks with Tasks 1 and 2 on text classification as the first part and Task 3 on IR as the second part. The scenario of the 2014 Task 1 was an EN-speaking, discharged patient (or next-of-kin) in their home in the USA. By reading their discharge document and further information on the Internet on either a networked desktop system or mobile device (e.g., smartphone or tablet), they wanted to learn about their own health and healthcare in general and clinical treatment history, current symptoms/developments, and future implications in particular.

In 2015 [33] and 2016 [34] CLEF eHealth expanded its scope from EN to multilingual text processing, medical web search, and speech-to-text conversion to ease both patients (and their next-of-kin) and clinicians' understanding of various types of medical content. Again, three tasks per year were organised.

The *2015 Task 1* and *2016 Task 1* built on processing tasks, data, and software of [36] by considering its nursing handover report support. In clinical handover between nurses, verbal handover and note taking could lead to loss of information and electronic documentation was seen as laborious, taking time away from patient education. The challenges addressed taking clinical notes automatically by using *Speech Recognition* (SR) to convert spoken nursing handover into digital text and *Information Extraction* (IE) to fill out a handover form, respectively.

The *2015 Task 2* and *2016 Task 2* considered clinical named entity recognition on *French* (FR) texts, which was previously an unexplored language. They aimed to automatically identify clinically relevant entities from FR biomedical articles. In addition, the 2016 task also addressed extracting causes of death from French death reports.

The *2015 Task 3* and *2016 Task 3* considered *cross-lingual medical search* on the web. They focused on trying to retrieve relevant and reliable web pages that meet a given patient's (or their next-of-kin's) general information needs related to their medical complaints (e.g., their need to understand a condition or the cause of a medical symptom). The tasks also considered information needs that were expressed in several non-EN languages.

In 2017, the following three tasks were organised [35] to continue the 2016 Tasks 2 and 3 and introduce a new pilot task: The *2017 Task 1* explored the problem of *multilingual text processing* by considering this year the extraction of causes of death from both FR and EN death reports to ease clinicians understanding of these reports. The *2017 Task 3* developed *medical web search* techniques to address the challenge posed by patients (or their next-of-kin) in locating relevant and reliable medical content on the web. In addition, the *2017 Task 2* considered a new challenge, that of TAR *generation* in empirical medicine to support health care and policy making. Medical researchers and policy-makers while writing systematic review articles (e.g. covering the treatment of a condition) must ensure that they consider all documents relevant to their review. As the size of medical libraries continues to expand, automation in this process is necessary.

Table 2. Timeline of CLEF eHealth Tasks. Abbreviations: *English* (EN), *French* (FR), *information extraction* (IE), *information retrieval* (IR), *technology assisted reviews* (TAR)

<table>
<tr><td></td><td>Task</td><td>2013 data</td><td>2014 data</td><td>2015 data</td><td>2016 data</td><td>2017 data</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td><b>IE</b></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>Named entity recognition and/or normalisation</td><td>EN clinical reports</td><td></td><td colspan="2">FR biomedical articles</td><td></td></tr>
<tr><td></td><td>Extraction</td><td></td><td>EN clinical reports</td><td></td><td></td><td>Multilingual death reports</td></tr>
<tr><td></td><td>Classification</td><td></td><td></td><td></td><td>FR death reports</td><td></td></tr>
<tr><td></td><td>Replication task</td><td></td><td></td><td></td><td colspan="2">Code</td></tr>
<tr><td><b>Information Management</b></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>Visualisation</td><td></td><td>EN eHealth data</td><td></td><td></td><td></td></tr>
<tr><td></td><td>Report generation and management</td><td></td><td></td><td colspan="2">EN nursing handover reports</td><td></td></tr>
<tr><td><b>IR</b></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>Patient-centered IR</td><td colspan="5">Multilingual eHealth data</td></tr>
<tr><td></td><td>Cross-lingual IR</td><td></td><td colspan="4">Multilingual eHealth data</td></tr>
</table>

	TAR in empirical medicine					EN biomedical articles

Evaluation Methods from 2013 to 2017

Evaluation methods for the CLEF eHealth tasks in 2013—2017 (Table 2) were described in detail in [31]–[35], respectively, and are summarised below. The reader is referred to these papers for methodological references.

The evaluation criterion in the 2013 Task 1a on disorder identification was the correctness in identification of disorders text snippets as defined by the *F1 measure* with a non-parametric test called *random shuffling* for the statistical significance assessment on 100 annotated health records for testing. An independent set of 200 annotated health records was provided for training. When computing true positives for the *exact F1*, the snippets by the solution- system and hand-annotation had to be identical whilst an overlap was enough for the *relaxed F1*.

The evaluation criterion in the 2013 Task 1b on disorder normalisation was the correctness in mapping the disorders to SNOMED CT codes as defined by the *accuracy* measure with the aforementioned random shuffling for the statistical significance assessment. The annotated health records and their split between training and testing were the same as in Task 1a. When computing true positives for the *exact accuracy*, the total number of code mappings was computed from the annotated records and the system was penalised for missing codes the same way as for incorrect codes. For the *relaxed accuracy*, the system was only evaluated on annotations that were detected by the system, that is, the total number corresponds to the code mappings with strictly correct text snippet generated by the system.

The evaluation criterion in the 2013 Task 2 on shorthand extension was the correctness in mapping the pre-identified shorthand to UMLS codes. This criterion was formalised using the exact and relaxed accuracy measures with random shuffling for the statistical significance assessment. The annotated health records and their split between training and testing were the same as in Task 1a.

Evaluation of submissions to the 2013 Task 3 on IR was conducted with respect to the relevance of the retrieved documents to the information seeker on 50 test queries and the matching result set. The official primary and secondary measures were the *precision at 10* (P@10) and *normalised discounted cumulative gain at 10* (NDCG@10), respectively. Both measures were calculated over the top ten documents in a run for each query, and then averaged across the whole set of queries. To compute the precision at 10, graded relevance assessments were converted to the two-point scale. The *Wilcoxon test* was used to better compare the measure values for the runs and benchmark.

In the 2014 Task 1 on information visualisation, participants could submit their designs to an optional draft submission to receive comments, followed by the call for final submissions. Final submissions needed to encompass the following three mandatory items: 1) a *design and application report* that highlights the obtained findings, possibly supported by an informal user study or other means of validation, 2) a *demonstration video* illustrating the relevant functionality of the functional design or paper prototype in application to the provided task data when the user knows the functionality, and 3) a *training video* illustrating a novice user being trained to the functionality and using it. Final submissions were judged towards their rationale for the design, including selection of appropriate visual interactive data representations and reference to state-of-the-art techniques by an *expert panel*. For each submission, the panel consisted of four independent experts and one organiser. To be successful, the submission had to demonstrate that the posed problems and information needs are addressed; provide a compelling use-case driven discussion of the work flow supported and exemplary results obtained; and highlight the evaluation approach and obtained findings. Primary judging criteria included the *effectiveness* and *originality* of the proposed design that were further divided to categories for *aesthetics*, *interaction*, *usability*, and *visualisation.*

Evaluations in the 2014 Task 2 on template filling were performed using accuracy and F1. Exact and relaxed versions were defined for both measures. This followed the practices of the 2013 Tasks 1 and 2.

In the 2014 Task 3 on IR, participants were provided with 50 topics, including 5 training topics, with their translation in *Czech* (CS), *German* (DE), and FR. The official primary and secondary measures were similar, and computed in a similar manner to the 2013 Task 2 measures: P@10 and NDCG@10.

The 2015 Task 1 on speech recognition distinguished submissions that developed the SR engine itself from the submissions only considering post-processing methods for the speech-recognised text. Also a separate submission category was assigned to solutions based on both SR and text post-processing. Each participant was allowed to submit up to two systems to the first category and up to two systems to the second category. If addressing both these categories, the participant was asked to submit all possible combinations of these systems as their third category submission. Evaluation was mainly based on the correctness (defined as the number of incorrectly recognised words). Final submission then consisted of the processing outputs for each method on the 100 training and 100 test documents.

The 2015 Task 2 on named entity recognition permitted up to two runs per team for three subtasks that were evaluated separately as follows: 1) For plain entity recognition, raw text was supplied to participants who had to submit entity annotations comprising entity offsets and entity types. 2) For normalised entity recognition, raw text was supplied to participants who had to submit entity annotations comprising entity offsets, entity types, and entity normalisation (UMLS CUIs). 3) For entity normalisation, raw text and plain entity annotations were supplied to participants who had to submit entity normalisation (UMLS CUIs). For each of the subtasks, the system output on the unseen test set was compared to the gold standard annotations and precision, recall, and F1 were computed.

In 2015 Task 3 on IR, teams could submit up to ten runs for the EN queries, and an additional ten runs for each of the multilingual query languages. Teams were required to number runs such as that run 1 was a baseline run for the team; other runs were numbered from 2 to 10, with lower numbers indicating higher priority for selection of documents to contribute to the assessment pool (i.e., run 2 was considered of higher priority than run 3). System evaluation was conducted using similar measures as previous years: P@10 and NDCG@10 were the primary and secondary measures, respectively. A separate evaluation was conducted using both relevance assessments and readability assessments. For all runs, *rank-biased precision* was computed along with its *readability-biased modifications* for the binary readability assessments and the graded readability assessments.

In 2016, the nursing handover support task used precision, recall, and F1 for evaluation. Performance was evaluated first separately in every heading from 1 to 35 and the 36th heading for irrelevant text. Then, we documented the performance in the dominant class of 36 and averaged over the 35 form headings by using *macro-averaging*, because our desire was to perform well in all classes, and not only in the majority classes. This *macro-averaged F1* was used to rank methods with the Wilcoxon test for statistical significance testing.

For the 2016 Task 2 and 2017 Task 1 on IE, teams could submit up to two runs for each subtask. The system output on the unseen test set was compared to the gold standard annotations and precision, recall, and F1 were compared. After submitting their result files, participating teams had one extra week to submit the system used to produce them, or a remote access to the system, along with instructions on how to install and operate the system for the replicability to be tested.

In 2016 and 2017, for the IR task, teams could submit up to three runs for the ad-hoc search on the EN queries, an additional three runs for the query variations challenge on the EN queries, and an additional three runs for each of the multilingual query languages. System evaluation was conducted using P@10 and NDCG@10 as the primary and secondary measures, respectively. Precision was computed using the binary relevance assessments; NDCG was computed using the graded relevance assessments. A separate evaluation was conducted using the multidimensional relevance assessments (topical relevance, readability, and trustworthiness). For all runs, rank biased precision was computed along with its multidimensional modifications for the binary readability assessments, the graded readability assessments, and the binary readability and trustworthiness assessments. In 2017, these measures were parameterised for a given user's expertise.

In the 2017 pilot task on TAR in empirical medicine, teams could submit up to eight official runs. System performance was assessed using a *Simple Evaluation* approach and a *Cost-Effective Evaluation* approach. The assumption behind the former approach is the following: The system user is the researcher that performs the abstract and title screening of the retrieved articles. Every time an abstract is 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. Evaluation measures were *area under the recall-precision curve*, *minimum number of documents returned to retrieve all relevant documents, work saved over sampling at different recall levels, area under the cumulative recall curve normalised by the optimal area*; *recall @ 0% to 100% of documents shown,* a number of newly constructed cost-based measures, and reliability.

Data Releases from 2013 to 2017

Data releases by the CLEF eHealth tasks in 2013—2017 (Table 2) were described in detail in [31]–[35], and references therein, respectively, and are summarised below. The reader is referred to these papers for further information and references.

The CLEF eHealth 2013 tasks used de-identified, annotated health records of the ShARe corpus of the *Multiparameter Intelligent Monitoring in Intensive Care* (MIMIC) II database, Version 2.5.7. These 300 records in total were authored in US intensive care and were in EN. The dataset consisted of discharge summaries and electrocardiogram, echocardiogram, and radiology reports. Their use in the third task was optional.

Although the health records were de-identified, they still needed to be treated with appropriate care and respect. Hence, all participants were required to register to the lab, obtain a US Human Subjects Training Certificate, create an account to a password-protected site, specify the purpose of data usage, accept the data use agreement, and have their account approved. After this, they could access and download the annotated records.

Approximately 181,000 words in these records were expert annotated for the 2013 Task 1 as mentions of disorder names and 7,500 words as mentions of disorder names and the 2013 Task 2 as clinical shorthand. A disorder name was defined as any text snippet, which could be mapped to a SNOMED CT concept, which belongs to the semantic group of Disorder and UMLS semantic type of Acquired Abnormality; Anatomical Abnormality; Cell or Molecular Dysfunction; Congenital Abnormality; Disease or Syndrome; Experimental Model of Disease; Injury or Poisoning; Mental or Behavioral Dysfunction; Neoplastic Process; Pathologic Function; or Signs and

Symptoms. Clinical shorthand was defined as an abbreviation or acronym and was to be codified with one UMLS *Concept Unique Identifier* (CUI) or CUI-less if an appropriate UMLS code was not available.

To enable IR, 55 new search topics were formed specifically for Task 3 and their data was made available to the registered participants for download on the Internet on a secure password-protected server. Each search task was described using a *patient profile* (e.g., A forty year old woman, who seeks information about her condition), *information need* (e.g., description of what type of disease hypothyroidism is), and *query* with separate fields for its *title* (e.g., Hypothyreoidism) and *description* (e.g., What is hypothyreoidism). The profile also allowed the participants to address the task without considering the aforementioned health records. To create result document sets for these search tasks, a large crawl of approximately a million online health resources (saved as *HyperText Markup Language* (HTLM) documents and supplemented with their *Uniform Resource Locator* (URL)) by the Khresmoi project was used. The resources covered a broad range of health topics and targeted both clinicians and laypersons. Over 60 per cent of these documents had been certified by the *Health on the Net Foundation* (HON) as adhering to their code. The returned documents were assessed by experts using the four-point scale of Irrelevant, On topic, but unreliable, Somewhat relevant, and Highly relevant. These relevance assessments were then pooled to the two-point scale of Irrelevant (i.e., the first two points) and Relevant (i.e., the last two points).

The CLEF eHealth 2014 Task 1 built on these 2013 datasets by combining them as a whole in order to address information search and visualisation in a patient-centric way. One mandatory and five optional patient cases were carefully chosen from the 2013 Tasks 1–3 for this task. The mandatory case was as follows: The patient profile was “This 55-year old woman with a chronic pancreatitis is worried that her condition is getting worse. She wants to know more about jaundice and her condition.” The de-identified, 2013 Task 1 and 2 annotated discharge summary served as the respective annotated health record. The information need was “chronic alcoholic induced pancreatitis and jaundice in connection with it.” The query with the title of “chronic alcoholic induced pancreatitis and jaundice” and description of “is jaundice an indication that the pancreatitis has advanced” resulted in for this case 113 returned documents with 26 of these expert-assessed as relevant on the two-point scale. The task data were made available to the registered participants for download on the Internet on a secure password-protected server after the organisers had ensured their de-identification by hand. After the task, the workspace was kept open for registration [37]; by 1 March 2018, access had been granted to 62 people.

Also the 2014 Task 2 on template filling used the 2013 dataset of 300 de-identified health records, supplemented by a test set of 133 unseen discharge documents and new expert-annotations created as part of the ShARe project. The annotations extended the existing disorder annotations from the 2013 Task 1 by focusing on template filling for each disorder mention. As such, each disorder template consisted of ten different attributes, including *Negation Indicator*, *Subject Class*, *Uncertainty Indicator*, *Course Class*, *Severity Class*, *Conditional Class*, *Generic Class*, *Body Location*, *DocTime Class*, and *Temporal Expression*. Each attribute contained two types of annotation values: normalisation and cue detection value with the exception of the *DocTime Class* which did not contain a cue detection value.

To enable IR in the 2014 Task 3, 55 new queries were first formulated by experts from the main disorders diagnosed in discharge summaries provided in the 2014 Task 2 and then associated with result document sets of the aforementioned Khresmoi set. Participants were provided with the mapping between queries and discharge summaries, and were again given an option to use the discharge summaries. Expert-assessments compared the query and its mapping to the content of the retrieved document on the four-point scale of Irrelevant, On topic, but unreliable, Somewhat relevant, and Highly relevant.

The CLEF eHealth 2015 targeted two new tasks as its Tasks 1 and 2, in addition to continuing its established and popular series of IR tasks as its Task 3. This 2015 Task 3, considered the following scenario: A patient or their next-of-kin is first shown images and videos related to medical symptoms and then asked which queries they would issue to a web search engine if they were exhibiting such symptoms and thus wanted to find more information to understand these symptoms or their condition. A total of 266 possible unique queries were collected from volunteers in EN; of these, 67 queries were selected to be used in the task. The queries were also translated by experts to *Arabic* (AR), CS, DE, *Farsi* (FA), FR, *Italian* (IT), and *Portuguese* (PT); these formed the multilingual query sets which were made available to participants for submission of multilingual runs. Along with relevance assessments by expert assessors on the result document sets of the aforementioned Khresmoi dataset, readability judgements were also collected for the assessment pool. Assessments were provided on a four point scale of 0: *It is very technical and difficult to read and understand*; 1: *It is somewhat technical and difficult to read and understand*; 2: *It is somewhat easy to read and understand*; and 3: *It is very easy to read and understand*.

The CLEF eHealth 2015 Task 1 introduced a new problem of supporting handover communication with 300 synthetic patient cases for the SR training, validation, and testing in 2015 and IE training, validation, and testing in the 2016 Task 1. Each case in this *NICTA Synthetic Nursing Handover Data* consisted of a patient profile; a written, free-form text paragraph (i.e., the *written handover document*) to be used as a reference standard in SR; and its spoken (i.e., the *verbal handover document*) and speech-recognized counterparts. The written handover documents were annotated, by a registered nurse using a form with 49 headings (aka classes) to fill out. Irrelevant text was to be classified as 36. *NA* and the annotation task was seen as multi-class classification, that is, each word could belong to precisely one class. In 2015, the first set of 100 cases was used for training and validation and the second, independent set of 100 cases for testing. In 2016, the first and second set were used for IE training and validation, respectively; an independent set of yet another 100 cases were released for testing.

For the 2015 Task 2, two types of biomedical documents were used: a total of 1,668 titles of scientific articles indexed in the MEDLINE database, and six full-text drug monographs published by the *European Medicines Agency* (EMEA). The annotations covered the following ten types of entities of clinical interest, defined by Semantic Groups in the UMLS: *Anatomy*, *Chemicals & Drugs*, *Devices*, *Disorders*, *Geographic Areas*, *Living Beings*, *Objects*, *Phenomena*, *Physiology*, and *Procedures*. The expert annotations marked each relevant entity mention in the documents, and assigned the corresponding semantic type(s) and CUI(s).

In addition to building on the CLEF eHealth 2015 Tasks 1 and 3 on handover communication and IR, respectively, the CLEF eHealth 2016 ran as its Task 2 two separate new challenges, which used two distinct data sets. The first data set was the *QUAERO French Medical Corpus* that was used as a training and validation set in 2016 (two sets of 833 expert-annotated MEDLINE titles and 3 EMEA documents) and a new unseen test set of 833 annotated MEDLINE titles and 4 EMEA documents. The set was annotated for 10 types of clinical entities with normalisation to the UMLS and covered both scientific articles titles and drug inserts. The second data was the *CépiDC Causes of Death Corpus* with free-text descriptions of causes of death as reported by physicians in the standardised causes of death forms. Each document (65,843 death certificates in total) was expert annotated manually coded by experts with the codes from the *International Statistical Classification of Diseases and Related Health Problems*, *Tenth Revision* (ICD-10), as per the international WHO standards. In addition to the training data, the task supplemented its data releases by manually built dictionaries of terms associated with the annotated ICD-10 codes.

In contrast to the previous CLEF eHealth IR tasks, the CLEF eHealth 2016 Task 3 used a new dataset called *ClueWeb12 B13*. This large snapshot of the web (approximately 52.3 million documents), crawled between February and May 2012 had become a prevalent benchmark dataset in IR in 2012–2016. Unlike the dataset used in the previous years, ClueWeb12 was not restricted to health-related pages, making the dataset more in line with the material current web search engines index and retrieve. The task queries extended upon the focus of the 2015 task (self-diagnosis) by considering real health information need expressed by the general public through posts published in a public health web forum called *Reddit*. Forum posts were extracted from its *askDocs* section and presented to query creators, who were asked to formulate queries based on what they read in the initial user post. Six query creators with different medical expertise were used for the task, leading to a set of *query variations* for a fixed number of topics. For the query variations part of the task, participants were told which queries relate to the same information need in order to allow them to produce one set of results to be used as answer for all query variations of a given information need. For the multilingual part of the task, CS, DE, FR, *Hungarian* (HU), *Polish* (PL), and *Swedish* (SW) translations of the queries were provided. Queries were translated by medical experts hired though a professional translation company. Relevance assessments were collected by pooling participants' submitted runs as well as baseline runs. Expert assessment was performed for document relevance, readability/understandability, and reliability. The relevance criteria were drafted considering the entirety of the forum posts used to create the queries; a URL to the forum posts was also provided to the assessors.

Finally, in 2017, the CLEF eHealth 2016 Tasks 1 and 3 were extended and the aforementioned new pilot task with unseen data was introduced as the CLEF eHealth 2017 Tasks 2. The 2017 Task 1 used a corpus of expert-annotated death certificates from France in FR and the USA in EN with respect to the ICD-10 codes. Again, this task supplemented its data releases by manually built dictionaries of terms associated with the annotated ICD-10 codes. The 2017 Task 3 used the same document collection and topics as in 2016, with the aim of acquiring more relevance assessments and improving the collection reusability.

The new TARs in empirical medicine task (i.e., the 2017 Task 2) used a subset of MEDLINE documents for its challenge to make Abstract and Title Screening more effective. More specifically the *PubMed Document Identifier*s (PIDs) of potentially relevant MEDLINE document abstracts indexed by the PubMed search engine were provided for each training and test topic. The PIDs 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. Topics consisted of the Boolean Search from the first step of the systematic review process. Specifically, for each topic the following information was provided: a topic *identifier* (ID); title of the review, written by Cochrane experts; boolean query manually constructed by Cochrane experts; and set of PIDs returned by running the query in MEDLINE. Twenty of these topics were randomly selected to be used as a training set, while the remaining thirty were used as a test set. 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 were 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.

Software Releases and Submission from 2013 to 2017

With an aim to lower the entry barrier and encourage novelty and creativity in problem solutions, CLEF eHealth began providing participants with software and code for method evaluation, record text annotation, and document relevance assessment in 2013 and extended this to also release processing code in 2016 [31]–[35]. The reader is referred to these papers [31], [32], [33], [34], and [35], and references therein, for further information about the software and code releases in 2013, 2014, 2015, 2016, and 2017, respectively.

The software and code releases were motivated by our desire for faster progress, comprehensive benchmarking, and transparency of the CLEF eHealth outcomes. Prior to CLEF eHealth, the progress in eHealth ICT was extremely limited, in comparison to banking, defence, and many other fields that also record big data and benefit from their analytics, because of barriers in limited collaboration in sharing data, processing methods, and evaluation outcomes, together with their common conventions and standards [38].

In the CLEF eHealth 2013 Tasks 1 and 2, we released both a command-line tool and a *Graphical User Interface* (GUI) that the participants could use to compute the values for the official and supplementary evaluation measures and visualise annotations against their method outputs. This *eHOST annotation tool* [39] also supported participants in annotating more data, although methods using teams' own annotations were evaluated separately from those based on the organisers' original annotations alone. In the CLEF eHealth 2013 Task 3, we released the *Relevation! relevance assessment tool* [40] and provided participants with a pointer to an established tool for computing values for the official and supplementary evaluation measures.

The twelve CLEF eHealth 2014–2017 tasks in total continued releasing software and code for computing values for evaluation measures, evaluating statistical significance of their differences between two or more methods, data annotation, and relevance assessment. In addition to releasing purpose-built software and code for the tasks, pointers to such helpful resources by other tasks and groups were also catalogued and provided on the website and overview paper of each task.

As a new initiative of releasing not only evaluation tools, but also processing code, the CLEF eHealth 2016 Task 1 released the organisers' entire software stack as a state-of-the-art solution to the handover IE problem (i.e., both feature generation and IE) [36]. Participants were welcomed, but not mandated, to use the released code and, as intended, the results highlighted all participating teams' methods outperforming this known state-of-the-art baseline.

In parallel to this approach of organisers' releasing software and code for annotation, relevance assessment, evaluation, and processing, CLEF eHealth established its replication track in 2016. The track gave the participants of the CLEF eHealth 2016 Task 2 and CLEF eHealth 2017 Task 1 the opportunity to submit their processing methods to organisers, who then attempted to replicate the runs submitted by participants. In 2016, three participating teams chose this option and submitted a total of seven methods, all of which the organisers were able to replicate perfectly. Two out of the remaining four teams submitted their EOI to submit to this track but ran out of time to finalise their submission; the third team reserved the distribution of their method to commercial use; and the fourth team did not reply to this EOI call. In 2017, five participating teams chose the replication track and submitted a total of 22 methods. The organisers were able to replicate most of them perfectly without contacting the teams. Where team contact was required, replication was achievable after contacting the submitting team for some further technical clarification on system requirements, installation procedure, and practical use. The organisers also reported an overall improvement in method documentation as an outcome of running the track twice.

Participation and Benchmark Results from 2013 to 2017

The CLEF eHealth lab attracted every year in 2013–2017 more than 100 teams to submit their EOI for the task and among them, 20 to 34 teams participated (Table 1). The difference between the number of teams interested and the actual participation was explained by the ease of the registration process versus the substantial amount of work required to actually participate by being able to submit to these difficult tasks. The very high number of EOIs within the first two years was surely related to the novelty of the 2013 and 2014 tasks. The number of participants in 2013–2017 remained stable over the years, despite the regular change and diversity in tasks. The most popular tasks were related to the IR Tasks 3 in 2013–2017. Given that both the number of EOIs and participants were decreasing for the last two years, the task might have to be redefined.

The results of the 15 tasks organised as part of the CLEF eHealth lab in 2013–2017 contributed to the body of knowledge about the difficulty of health information management, extraction, and retrieval (Table 3). In addition, the methodological diversity of the submissions by more than 100 teams all over the world, together with the baselines by the organisers, addressed the applicability of particular methods. Approximately half of the tasks produced statistically significant improvements in processing quality by at least one out of the top-3 methods.

Table 3. Summary of the benchmark results for the CLEF eHealth tasks in 2013-2017. The reader is referred to the CLEF eHealth overviews [31]–[35], and references therein, for further details. Notice that the tasks included some more subtasks – many of which resulted in statistically significant improvements in performance – but for the ease of reading this paper, our tabulation below is limited to the main tasks. Abbreviations: *accuracy* (A), *average precision* (AP), *error* (E = 1 - A), *precision* (P). "*" indicates that the measure value of the method was significantly better than the one for the next best method.

	Task	Performance measure	Measure value range	Best performance	2nd best performance	3rd best performance	cultyWorst performance to	Statistical significance test	n the Significant differences
CLEF eHealth 2013									
	1a	F1	[0%, 100%]	75.0%*	73.7%*	70.7%*	42.8%	Random shuffling with 99% confidence	Yes
	1b	A	[0%, 100%]	58.9%*	58.7%*	54.6%*	0.6%	Random shuffling with 99% confidence	Yes

	2	A	[0%, 100%]	71.9%*	68.3%*	66.4%*	42.6%	Random shuffling with 99% confidence	Yes
	3	P@10	[0%, 100%]	51.8%	50.4%	48.4%	0.6%	Wilcoxon test with 95% confidence	Yes
CLEF eHealth 2014									
	2a	A	[0%, 100%]	86.8%	85.4%	84.3%	76.9%	-	-
	2b	F1	[0%, 100%]	91.3%	67.1%	54.4%	19.0%	-	-
	3	P@10	[0%, 100%]	75.6%	75.5%	75.4%	6.0%	-	-
CLEF eHealth 2015									
	1	E	[0%, 100%]	38.5%*	52.3%	52.8%*	95.4%	Wilcoxon test with 95% confidence	Yes
	2	F1	[0%, 100%]	75.6%	74.1%	70.4%	0.0%	-	-

	3	P@10	[0%, 100%]	53.9%	38.6%	38.0%	25.4%	-	-
CLEF eHealth 2016									
	1	F1	[0%, 100%]	38.2%	37.4%*	34.5%*	0.0%	Wilcoxon test with 95% confidence	Yes
	2 (entity recognition)	F1	[0%, 100%]	74.9%*	70.2%*	69.9%*	12.6%	t-test with 99.9% confidence	Yes
	2 (cause of death)	F1	[0%, 100%]	84.8%*	84.4%*	75.2%*	55.4%	t-test with 99.9% confidence	Yes
	3	P@10	[0%, 100%]	-	-	-	-	95% confidence intervals	-
CLEF eHealth 2017									
	1	F1	[0%, 100%]	85.0%	85.0%	81.9%	0.11%	-	-
	2	AP	[0%, 100%]	31.8%	29.7%	29.3%	4.5%	-	-

	3	P@10	[0%, 100%]	-	-	-	-	95% confidence intervals	-

Discussion

Principal Results

The CLEF eHealth installations have offered shared evaluation challenges in the fields of medical information management, extraction, and retrieval since 2012. Evaluation methods and resources have been developed and shared with the community to support the understanding of and access to medical content by laypeople (or their next-of-kin), clinicians, scientists, and policy-makers. Evaluation results for the methods and resources developed have been released to the community. In so doing the lab has provided an evaluation setting for the progression of research in the multilingual medical ICT. This has facilitated further evaluation into medical system development for information management, extraction, and retrieval, and we postulate that subsequently to aiding the progression of research in these areas. The annual CLEF eHealth lab workshop held at the main CLEF conference provides for the dissemination and discussion of the outcomes of each year’s challenges. This has facilitated discussion among the community, cross-fertilisation of ideas, and further progress in the medical information production, processing and consuming ecosystem. Each year the lab organisers produce lab overview papers describing the challenges offered and participants’ results. These have proven influential, as indicated by their citation indexes.

Comparison with Prior Work

Already twelve years prior to establishing CLEF eHealth in 2012, evaluation labs began addressing limited collaboration as a major barrier that hinders the transfer of ICT for processing free-form text to clinical practice and is evidenced by improvements in developing and sharing data, community conventions, standards, software, and evaluation benchmarks [38]. The other two identified main barriers were absence of user centricity in technology research and development and inabilities to replicable results. By definition as a lab, CLEF eHealth 2012–2017 continued contributing to the barrier of limited collaboration but used the remaining two barriers to distinguish itself from other labs. Namely, it placed layperson patients (as opposed to clinical experts) as targeted technology users to the centre of the shared tasks in 2013 and introduced its replication track in 2016.

The CLEF initiative began in Europe in 2000 and at the same time that the first CLEF eHealth evaluation lab with three shared tasks was launched in 2013, the *CLEF Question Answering for Machine Reading* lab introduced a pilot task on machine reading on biomedical text about Alzheimer's disease [41]. Extending the prior work inclusion criterion from biomedical text to other data modalities, the *ImageCLEF* lab included annual shared tasks on biomedical image processing in 2005–2013 [42-44].

Already before CLEF, The *Text REtrieval Conference* (TREC) was established in the USA in 1992 as an evaluation initiative with evaluation labs of shared tasks leading to annual conferences and workshops. In 2000, the *TREC Filtering* tasks considered user profiling to filter in only the topically relevant biomedical abstracts using the MeSH as topics [45]. In 2003–2007, the *TREC Genomics* tasks ranged from ad-hoc IR to text classification, passage retrieval, and entity-based question answering on data from biomedical papers and eHealth records [46]. In 2011 and 2012, the *TREC Medical Records* tasks targeted building a search engine where the patient cohort's eligibility criteria of a given clinical study can be specified through the search query and then after information search on English eHealth records, the matching population is returned for study recruitment purposes [47].

The *NII Test Collection for Information Retrieval Systems* (NTCIR) was launched in Japan in 1997 as an evaluation initiative and in 2013, its *Medical NLP* lab considered the following three shared tasks on Japanese eHealth records [48]: text de-identification, complaint/diagnosis IE, and an open challenge, where participants were given the freedom to try to solve any other task on the dataset that was used for the first two tasks.

The *Informatics for Integrating Biology and the Bedside* initiative, began in the USA in 2006, addressed clinical text processing through its following shared tasks on English eHealth records in 2006–2012 [49]: text de-identification and identification of smoking status in 2006; recognition of obesity and its co-morbidities in 2008; medication IE in 2009; concept, assertion, and relation recognition in 2010; co-reference analysis in 2011; and temporal-relation analysis in 2012.

The *Medical NLP Challenges*, launched in the USA in 2007, considered automated diagnosis coding of English radiology reports from a US children's radiology department in 2007 and classifying the emotions found in English suicide notes in 2011 [50, 51].

The annual *SemEval/Senseval Workshops*, established in 2004 to address semantic disambiguation, role labelling, IE, IR, frame extraction, temporal annotation, and other multilingual semantic processing tasks, adopted our CLEF eHealth data in 2014 [52]. By supplementing our annotations for the CLEF eHealth 2013 Tasks 1 and 2, it challenged its participants to the same tasks but on a larger test set. A total of 21 participating teams completed this SemEval 2014 Task 1, and 18 of those also participated in the SemEval 2014 Task 2. This resulted in the strict-F1 of 81.3 per cent at its best in the first task, with respective strict-precision and strict-recall of 84.3 per cent and 78.6 per cent. For the second task, the top strict-accuracy of 74.1 per cent was obtained.

Limitations

In this paper, we have presented a bibliometric study of the scholarly influence of CLEF eHealth installations in 2012–2017. The paper and citation data collection has been limited to the CLEF eHealth proceedings and previously catalogued papers. Consequently, other relevant papers and citations are likely to exist, making our citation influence of 1,299 citations in total for the 184 papers by the 741 co-authors from 33 countries rather a modest than exaggerated estimate. This influence of six installations has been computed only two months after the CLEF eHealth 2017 proceedings were published.

In comparison, the scholarly influence of six installations of the TREC Video Retrieval initiative in 2002–2009 has been evaluated retrospectively in 2011, two years after the 2009 installation, as 15,828 citations for the 2,073 papers (of which 319 have been published in the TREC CP or WN Proceedings) [9]. A comparable influence has been achieved within the CLEF initiative by its Image CLEF activity in 2000–2009 [10]: First, seven Image CLEF installations have been evaluated retrospectively in 2013 – four years after the 2009 installation – as having had the influence of 2,018 citations for the 179 papers. Second, the scholarly influence of ten installations of the entire CLEF initiative in 2000–2009 has been evaluated retrospectively in 2013, four years after the 2009 installation, as 9,137 for the 873 papers.

Our average number of citations generated by a paper (i.e., 7) is smaller than this number is for the entire CLEF initiative (i.e., 10) but larger than what many other sub-initiatives achieved (from 0.2 to 35 with 11 for Image CLEF) [10]. CLEF eHealth, established in 2012, is not included in this comparison of sixteen CLEF sub-initiatives with up to ten installations each. Moreover, our numbers for seven installations originate from the year of the last installation as opposed to being collected four years after.

Although the CLEF eHealth installations have attracted substantial community interest, as reflected by the 741 co-authors of the 184 papers from 33 countries, we have not attracted participation from Central America. Also substantially more participation from Africa, South America, and the Middle East should be achievable. However, this problem of insufficient participation has been acknowledged by a recent review of evaluation initiatives in biomedical text mining in 2002–2014 as one of the main conclusions [53]. Fortunately, we have been successful in targeting the coupled problem of insufficient innovation by reaching statistically significant improvements in most CLEF eHealth tasks.

The final limitation of this review [53] remains for both CLEF eHealth and other evaluation initiatives; due to their nature as community efforts with limited time for participation and result announcements with as low entry barrier as possible, the task must simplify or abstract from the real-world problems. Although this creates a gap between the task solutions and their real-world use, the annual instalments provide many significant conclusions to guide future work.

Significance and Future Work

The CLEF eHealth installations with 15 information management, extraction, and retrieval tasks in total uniquely target various layperson (or next- of-kin) information understanding and provision challenges in the medical domain (Table 2). Coupled with this it strives to progress research in the fields of clinician information processing, exchange and understanding support. Finally, for the first time globally it targets challenges towards meeting the needs of policy-makers for TAR generation in empirical medicine. In IE, the lab has targeted named entity recognition and normalisation in clinical reports, and named entity recognition, normalisation, and classification in biomedical articles and in death reports. In information management, the lab has considered medical data visualisation, and nurses' handover report management. Finally, in IR the target has been on patient-centred search, cross-lingual search, and technology assisted reviewing.

The lab has attracted considerable and growing interest from the research community over the years: 34 unique teams participated in the three tasks in 2013, 24 in the three tasks in 2014, 20 in the three tasks in 2015, 20 in the three tasks in 2016, and 32 in the three tasks in 2017. While the lab has yet to become entirely global, it is already far reaching attracting participants from 33 unique countries.

By virtue of the lab series over the first six years of its life, from 2012–2017 inclusive, first, bringing the research community together through the lab series to collaborate and discuss challenges associated with technique development in the biomedical and clinical information management, extraction, and retrieval spaces, second, providing access to shared data, resources, processing methods, and evaluation settings for eHealth system research, development and evaluation, third, offering reproducibility, scalability, and user-centricity, we conjecture that CLEF eHealth has influenced progress in these three spaces. While it is difficult to accurately quantify such influence, the 1,299 citations, with influence of circa 963,000 generated by the lab in its first six years of existence are suggestive. Progress in the areas addressed by the lab has the potential to generate high impact not only on the research field, but more generally influence society, given the importance of health information access to support healthcare as well as to empower people to manage their health.

The CLEF eHealth evaluation lab series runs for the seventh year in 2018. This CLEF eHealth 2018 edition of the lab continues the 2017 IE, TAR, and IR tasks. The 2018

IE task will extend to new European languages with new death reports for developing named entity recognition and normalisation. The 2018 TAR task will build on last year's task to offer a new evaluation framework with new evaluation measures and new data. Finally, the 2018 IR task will offer new queries and new evaluation criteria to support developing new techniques for faceted search and patient-centred IR. Beyond 2018, we envision CLEF eHealth growing further, to extend the scope of IE and IR related challenges offered, to offer new information management challenges, to increase multilingual approaches, and to extend the scope of the lab to consider other challenges relevant to biomedical, clinical, and eHealth content.

Conclusions

Medical content is available electronically 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. In today's information overloaded society it is increasingly difficult to retrieve and digest valid and relevant medical information to make health-centred decisions. The CLEF eHealth lab aims to support the development of techniques to aid laypeople, clinicians and policy-makers in easily retrieving and making sense of medical content to support their decision making.

The successful CLEF eHealth workshop held in 2012, spawned the subsequent CLEF eHealth lab series (2013–2017). Over the year's this lab series has expanded its original goal of supporting patients (or their next-of-kin) in understanding the 'jargon' in their hospital discharge summary, to consider a broader set of medical information needs of both patients (or their next-of-kin), clinicians, scientists, and policy makers. Related to these themes, challenges have been offered in a multilingual setting on the topics of medical information management, extraction, and retrieval. These 15 challenge tasks have obtained much attraction, as evidenced by the annual lab overview papers, participants working notes papers, and external papers using the lab resources, obtaining a combined 184 papers by 741 co-authors from 33 countries across the world with 1,299 citations, totalling a citation influence of circa 963,000. Given the significance of the lab series, all test collections and resources associated with the lab challenges have been made available to the wider research community through the Internet.

The lab has attracted many participants from across the globe since its conception six years ago. In total, 718 teams have registered their interest in the tasks, leading to 130 teams submitting to the tasks. Together we have influenced the progression of health text processing, and in medical IR research. As the lab further progresses we envision its scope and reach extending even further.

Acknowledgements

Adj/Prof Hanna Suominen, Dr Liadh Kelly, and Assoc/Prof Lorraine Goeuriot co-chaired the CLEF eHealth initiative in 2012-2017, led some of its shared tasks, and co-authored the paper. We gratefully acknowledge the contribution of the people involved in the initiative as participants or organisers. The initiative was supported in part (in alphabetical order) by the French National Research Agency (grant No. CABeRneT ANR-13 JS02-0009-01); CLEF Initiative; Data61/CSIRO; European Science Foundation, EvaLuating Information Access Systems (ELIAS) network, Horizon 2020 program (grant No. H2020-ICT-2014-1 644753 KConnect); Khresmoi project, funded by the European Union Seventh Framework Programme/2007–2013 (grant No. 257528); Microsoft Azure for Research Award CRM:0518649; MIMIC II Database; NICTA, funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the Centre of Excellence program; PhysioNetWorks Workspaces; Protégé, supported by the National Institute of General Medical Sciences of the USA National Institutes of Health (grant No. GM10331601); ShARe project, funded by the US National Institutes of Health (grant No. R01GM090187); Swedish Research Council (grant No. 350-2012-6658); Swedish Vardal Foundation; US Department of Veterans Affairs Consortium for Healthcare Informatics Research; and US Office of the National Coordinator of Healthcare Technology, Strategic Health Information Technology Advanced Research Projects (grant No. 90TR0002).

Conflicts of Interest

None declared.

References

1. Suominen H (2012) Towards an international electronic repository and virtual laboratory of open data and open-source software for Telehealth research: comparison of international, Australian, and Finnish privacy policies. In: Smith AC, Armfield NR, Eikelboom RH (eds.) The Second International Conference on Global Telehealth, IOS Press, Amsterdam, The Netherlands, Studies in Health Technology and Informatics, vol 182, pp 153–160. PMID:23138090
2. Adnan M, Warren J, Suominen H (2015) Patient empowerment via technologies for patient- friendly personalized language. In: Grando AM, Rozenblum R, Bates D (eds.) Engaging Patients with Health Information Technology, De Gruyter, Berlin, Germany, vol 10, pp 153–164. NLMID:101646460
3. McAllister M, Dunn G, Payne K, Davies L, Todd C (2012) Patient empowerment: The need to consider it as a measurable patient-reported

outcome for chronic conditions. BMC Health Services Research 12:157. PMID:22694747

4. Lemire M, Sicotte C, Pare G (2008) Internet use and the logics of personal empowerment in health. Health Policy 88(1):130–140. PMID:18436330
5. Ilic D (2010) The role of the internet on patient knowledge management, education, and decision-making. Telemedicine Journal and E-Health 16(6):664–669. PMID:20575610
6. White R, Horvitz E (2009): Cyberchondria: Studies of the escalation of medical concerns in web search. ACM Transactions on Information Systems 27(4). DOI:10.1145/1629096.1629101
7. Andrews L, Gajanayake R, Sahama T (2014) The Australian general public's perceptions of having a personally controlled electronic health record (PCEHR). International Journal of Medical Informatics 83(12):889–900. PMID:25200198
8. Suominen H, Kelly L, and Goeuriot L (2018). The scholarly impact of CLEF eHealth 2012–2017 labs. Submitted to Lecture Notes in Computer Science.
9. Thornley CV, Johnson AC, Smeaton AF, Lee H (2011) The scholarly impact of TRECVid (2003– 2009). Journal of the American Society for Information Science and Technology 62(4):613– 627. DOI:10.1002/asi.21494
10. Tsikrika T, Larsen B, Müller H, Endrullis S, Rahm E (2013) The Scholarly Impact of CLEF (2000– 2009). Lecture Notes in Computer Science 8138:1–12. DOI:10.1007/978-3-642-40802-1_1
11. Angelini M, Ferro N, Larsen B, Müller H, Santucci G, Silvello G, Tsikrika T (2014) Measuring and analyzing the scholarly impact of experimental evaluation initiatives. Procedia Computer Science 38(Supplement C):133–137. DOI:10.1016/j.procs.2014.10.022
12. Catarci T, Forner P, Hiemstra D, Penas A, Santucci G (eds) (2012) Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012), Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany. DOI:10.1007/978-3-642-33247-0
13. Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) (2013a) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the Fourth International Conference of the CLEF Initiative (CLEF 2013), Lecture Notes in Computer Science (LNCS) 8138, Springer, Heidelberg, Germany. DOI:10.1007/978-3-642-40802-1
14. Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds) (2014) Information Access Evaluation – Multilinguality, Multimodality, and Interaction. Proceedings of the Fifth International Conference of the CLEF Initiative (CLEF 2014), Lecture Notes in Computer Science (LNCS) 8685, Springer, Heidelberg, Germany. DOI:10.1007/978-3-319-11382-1
15. Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones GJF, SanJuan E, Cappellato L, Ferro N (eds) (2015) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixth International Conference of the CLEF Association (CLEF 2015), Lecture Notes in Computer Science (LNCS) 9283, Springer, Heidelberg, Germany. DOI:10.1007/978-3-319-24027-5

16. Fuhr N, Quaresma P, Goncalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) (2016) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016), Lecture Notes in Computer Science (LNCS) 9822, Springer, Heidelberg, Germany. DOI:10.1007/978-3-319-44564-9
17. Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, Cappellato L, Ferro N (eds) (2017) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eighth International Conference of the CLEF Association (CLEF 2017), Lecture Notes in Computer Science (LNCS) 10456, Springer, Heidelberg, Germany. DOI:10.1007/978-3-319-65813-1
18. Forner P, Karlgren J, Womser-Hacker C, Ferro N (eds) (2012) CLEF 2012 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org/Vol-1178/. ISSN: 1613-0073.
19. Forner P, Navigli R, Tufis D, Ferro N (eds) (2013b) CLEF 2013 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org/Vol-1179/. ISSN 1613-0073.
20. Cappellato L, Ferro N, Halvey M, Kraaij W (eds) (2014) CLEF 2014 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org/Vol-1180/. ISSN: 1613-0073.
21. Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) (2015) CLEF 2015 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org/Vol-1391/. ISSN: 1613-0073.
22. Balog K, Cappellato L, Ferro N, Macdonald C (eds) (2016) CLEF 2016 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org/Vol-1609/. ISSN: 1613-0073.
23. Cappellato L, Ferro N, Goeuriot L, Mandl T (eds) (2017) CLEF 2017 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org/Vol-1866/. ISSN: 1613-0073.
24. Zhang G, Ding Y, Milojevic S (2013) Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. Journal of the American Society for Information Science and Technology 64(7):1490–1503. DOI:10.1002/asi.22850
25. Hsieh HF, Shannon SE (2005) Three approaches to qualitative content analysis. Qualitative Health Research 15(9):1277–1288. PMID:16204405
26. McGhee G, Marland GR, Atkinson J (2007) Grounded theory research: Literature reviewing and reflexivity. Journal of Advanced Nursing 60(3):334–342. PMID:17908129
27. Rahm E, Thor A (2005) Citation analysis of database publications. ACM Sigmod Record 34(4):48– 53. DOI:10.1145/1107499.1107505
28. Jacso P (2006) Deflated, inflated and phantom citation counts. Online Information Review 30(3):297–309. DOI:10.1108/14684520610675816
29. Bar-Ilan J (2008) Which h-index? — a comparison of WoS, Scopus and Google Scholar. Scientometrics 74(2):257–271. DOI:10.1007/s11192-008-0216-y
30. Suominen H (2014) Text mining and information analysis of health documents. Artificial Intelligence in Medicine 61(3):127–130. PMID:24998391

31. Suominen H, Salanterä S, Velupillai S, Webber Chapman W, Savova GK, Elhadad N, Pradhan S, South BR, Mowery DL, Jones GJF, Leveling J, Kelly L, Goeuriot L, Martínez D, Zuccon G (2013) Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In Forner P, Müller H, Paredes R, Rosso P, Stein B (eds.): Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the Fourth International Conference of the CLEF Initiative (CLEF 2013), Lecture Notes in Computer Science (LNCS) 8138, Springer, Heidelberg, Germany, pp 212–231. DOI:10.1007/978-3-642-40802-1_24
32. Kelly L, Goeuriot L, Suominen H, Schreck T, Leroy G, Mowery DL, Velupillai S, Webber Chapman W, Martinez D, Zuccon G, Palotti JRM (2014) Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds.): Information Access Evaluation – Multilinguality, Multimodality, and Interaction. Proceedings of the Fifth International Conference of the CLEF Initiative (CLEF 2014), Lecture Notes in Computer Science (LNCS) 8685, Springer, Heidelberg, Germany, pp 172–191. DOI:10.1007/978-3-319-11382-1_17
33. Goeuriot L, Kelly L, Suominen H, Hanlen L, Névéol A, Grouin C, Palotti JRM, Zuccon G (2015) Overview of the CLEF eHealth Evaluation Lab 2015. In Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones GJF, SanJuan E, Cappellato L, Ferro N (eds.) (2015): Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixth International Conference of the CLEF Association (CLEF 2015), Lecture Notes in Computer Science (LNCS) 9283, Springer, Heidelberg, Germany, pp 429–443. DOI:10.1007/978-3-319-24027-5_44
34. Kelly L, Goeuriot L, Suominen H, Névéol A, Palotti J, Zuccon G (2016) Overview of the CLEF eHealth Evaluation Lab 2016. In Fuhr N, Quaresma P, Gonc¸alves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds.): Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016), Lecture Notes in Computer Science (LNCS) 9822, Springer, Heidelberg, Germany, pp 255–266. DOI:10.1007/978-3-319-44564-9_24
35. Goeuriot L, Kelly L, Suominen H, Névéol A, Robert A, Kanoulas E, Spijker R, Palotti JRM, Zuccon G (2017) CLEF 2017 eHealth Evaluation Lab Overview. In Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, Cappellato L, Ferro N (eds.): Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eighth International Conference of the CLEF Association (CLEF 2017), Lecture Notes in Computer Science (LNCS) 10456, Springer, Heidelberg, Germany, pp 291–303. DOI:10.1007/978-3-319-65813-1_26
36. Suominen H, Zhou L, Hanlen L, Ferraro G (2015) Benchmarking clinical speech recognition and information extraction: New data, methods, and evaluations. JMIR Medical Informatics 3(2):e19. PMID:25917752
37. CLEF eHealth 2014 Task 1: Visual-Interactive Search and Exploration of eHealth Data. CLEF eHealth 2014. 2018-03-04. URL:https://physionet.org/works/CLEFeHealth2014Task1/. Accessed:

2018-03-04. (Archived by WebCite® at http://www.webcitation.org/6xgJiBBBG)

38. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O (2011) Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association 18(5):540–543. PMID:21846785
39. South BR, Mowery DL, Suo Y, Ferrández O, Meystre SM, Chapman WW (2014) Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. Journal of Biomedical Informatics 50:162–172. PMID:24859155
40. Koopman B and Zuccon G (2014) 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, Association for Computing Machinery (ACM), New York, NY, USA. DOI:10.1145/2600428.2611175
41. Morante R, Krallinger M, Valencia A, Daelemans W (2013) Machine reading of biomedical texts about Alzheimer's disease. In Forner P, Navigli R, Tufin D (eds.): CLEF 2013 Evaluation Labs and Workshop: Online Working Notes, CLEF, Valencia, Spain. DOI:10.1.1.366.3461
42. Müller H, Clough P, Deselaers T, Caputo B (eds.) (2010) ImageCLEF: Experimental Evaluation in Visual Information Retrieval, The Information Retrieval Series, vol 32, Springer, Heidelberg, Germany. DOI:10.1007/978-3-642-15181-1
43. Kalpathy-Cramer J, Müller H, Bedrick S, Eggel I, Garcia Seco de Herrera A, Tsikrika T (2011) Overview of the CLEF 2011 Medical Image Classification and Retrieval Tasks. In Petras V, Forner P, Clough P, Ferro N (eds.): CLEF 2011 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org). DOI:10.1007/978-3-642-15751-6_8
44. de Herrera AGS, Cramer JK, Demner Fushman D, Antani S, Müller H (2013) Overview of the ImageCLEF 2013 medical tasks. In Forner P, Navigli R, Tufin D (eds.): CLEF 2013 Evaluation Labs and Workshop: Online Working Notes, CLEF, Valencia, Spain. URL: http://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-SecoDeHerreraEt2013b.pdf
45. Robertson S, Hull D (2000) The TREC-9 filtering track final report. In: NIST Special Publication 500-249: The 9th Text REtrieval Conference (TREC 9), National Institute of Standards and Technology, Gaithersburg, MD, United States, pp. 25–40. DOI:10.1.1.6.4181
46. Roberts PM, Cohen AM, Hersh WR (2009) Tasks, topics and relevance judging for the TREC genomics track: Five years of experience evaluating biomedical text information retrieval systems. Information Retrieval 12:81–97. DOI:10.1007/s10791-008-9072-x
47. Voorhees E, Hersh W (2012) Overview of the TREC 2012 medical records track. In: NIST Special Publication 500-298: The Text REtrieval Conference 2012, National Institute of Standards and Technology, Gaithersburg, MD, United States. URL: https://www.nist.gov/publications/overview-trec-2012-medical-records-track

48. Morita M, Kano Y, Ohkuma T, Miyabe M, Aramaki E (2013) Overview of the NTCIR-10 MedNLP task. In: Proceedings of the 11th NTCIR Conference, NII Testbeds and Community for Infor- mation access Research (NTCIR), Tokyo, Japan, pp 696–701. DOI:10.1.1.648.5755
49. Uzüner O, South BR, Shen S, DuVall S (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 18(5):552– 556. PMID:21685143
50. Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, Duch W (2007) A shared task involving multi-label classification of clinical free text. In Cohen KB, Demner- Fushman D, Friedman C, Hirschman L, Pestian J (eds.): Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics, Morristown, NJ, USA, pp. 97–104. DOI:10.3115/1572392.1572411
51. Pestian J, Matykiewicz P, Linn-Gust M, South B, Uzuner O, Wiebe J, Cohen K, Hurdle J, Brew C (2011) Sentiment analysis of suicide notes: A shared task. Biomedical Informatics Insights 5(Suppl. 1)):3–16. PMID:22419877
52. Pradhan S, Elhadad N, Chapman WW, Manandhar S, Savova G (2014) SemEval-2014 Task 7: Analysis of clinical text. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, pp. 54–62. DOI:10.1.1.711.8955
53. Huang CC, Lu Z (2016) Community challenges in biomedical text mining over 10 years: Success, failure and the future. Briefings in Bioinformatics 17(1):132–144 . PMID:25935162

Abbreviations

AR: Arabic
CLEF: Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum
CP: conference paper
CS: Czech
CSIRO: Commonwealth Scientific and Industrial Research Organisation
CUI: concept unique identifier
DE: German
eHealth: electronic health
ELIAS: evaluating information access systems
EMEA: European Medicines Agency
EN: English
EOI: expression of interest
FA: Farsi
FR: French
GUI: graphical user interface
HON: Health on the Net
HTML: HyperText Markup Language

HU: Hungarian
ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision
ICT: information and communications technologies
ID: identifier
IE: information extraction
IR: information retrieval
IT: Italian
JP: journal paper
Khresmoi: knowledge helper for medical and other information users
MeSH: medical subject headings
MIMIC: multiparameter intelligent monitoring in intensive care
NDCG@10: normalised discounted cumulative gain at 10
NICTA: National ICT Australia
NLP: natural language processing
NTCIR: NII test collection for IR systems
P@10: precision at 10
PCeHR: personally controlled eHealth record
PID: PubMed document identifier
PL: Polish
PT: Portuguese
ShARe: shared annotated resources
SNOMED CT: systematized nomenclature of medicine — clinical terms
SW: Swedish
TAR: technology assisted reviews
TREC: Text REtrieval Conference
UMLS: unified medical language system
URL: uniform resource locator
WHO: World Health Organisation
WN: working note