# CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French

Aurélie Névéol[1], Robert N. Anderson[2], K. Bretonnel Cohen[1,3], Cyril Grouin[1], Thomas Lavergne[1,4], Grégoire Rey[5], Aude Robert[5], Claire Rondet[5], and Pierre Zweigenbaum[1]

[1] LIMSI, CNRS, Université Paris-Saclay, Orsay, France
firstname.lastname@limsi.fr
[2] National Center for Health Statistics, USA
RNAnderson@cdc.gov
[3] University of Colorado, USA
[4] Université Paris-Sud, Orsay, France
[5] INSERM-CépiDc, Le Kremlin-Bicêtre, France
firstname.lastname@inserm.fr

**Abstract.** This paper reports on Task 1 of the 2017 CLEF eHealth evaluation lab which extended the previous information extraction tasks of ShARe/CLEF eHealth evaluation labs. The task continued with coding of death certificates, as introduced in CLEF eHealth 2016. This large-scale classification task consisted of extracting causes of death as coded in the International Classification of Diseases, tenth revision (ICD10). The languages offered for the task this year were English and French. Participant systems were evaluated against a blind reference standard of 31,690 death certificates in the French dataset and 6,665 certificates in the English dataset using Precision, Recall and F-measure. In total, eleven teams participated: 10 teams submitted runs for the English dataset and 9 for the French dataset. Five teams submitted their systems to the reproducibility track. For death certificate coding, the highest performance was 0.8674 F-measure for French and 0.8501 for English.

**Keywords:** Natural Language Processing; Entity Linking, Text Classification, French, Biomedical Text

## 1 Introduction

This paper describes an investigation of information extraction and normalization (also called “entity linking”) from French and English-language health documents conducted as part of the CLEF eHealth 2017 lab [1]. The task addressed is the automatic coding of death certificates using the International Classification of Diseases, 10th revision (ICD10) [2]. This is an essential task in epidemiology, as the determination and analysis of causes of death at a global level informs public health policies.

The methodology applied is the shared task model. In shared tasks, multiple groups agree on a "shared" task definition, a shared data set, and a shared evaluation metric. The idea is to allow evaluation of multiple approaches to a problem while minimizing avoidable differences related to the task definition, the data used, and the figure of merit applied [3, 4].

Over the past four years, CLEF eHealth offered challenges addressing several aspects of clinical information extraction (IE) including named entity recognition, normalization [5–7] and attribute extraction [8]. Initially, the focus was on a widely studied type of corpus, namely written English clinical text [5, 8]. Starting in 2015, the lab's IE challenge evolved to address lesser studied corpora, including biomedical texts in a language other than English i.e., French [6]. This year, we continue to offer a shared task based on a large set of gold standard annotated corpora in French with a coding task that required normalized entity extraction at the sentence level. We also provided an equivalent dataset in English.

The significance of this work comes from the observation that challenges and shared tasks have had a significant role in advancing Natural Language Processing (NLP) research in the clinical and biomedical domains [9, 10], especially for the extraction of named entities of clinical interest and entity normalization.

One of the goals for this shared task is to foster research addressing multiple languages for the same task in order to encourage the development of multilingual and language adaption methods.

This year's lab suggests that the task of coding can be addressed reproducibly with comparable performance in French and in English without relying on translation.

## 2 Material and Methods

In the CLEF eHealth 2017 Evaluation Lab Task 1, two datasets were used. The French dataset was supplied by the French CépiDc[1] and the English dataset was supplied by the American CDC[2]. Both datasets refer to the International Classification of Diseases, tenth revision (ICD10),a reference classification of about 14,000 diseases and related concepts managed by the World Health Organization and used worldwide, to register causes of death and reasons for hospital admissions. Further details on the datasets, tasks and evaluation metrics are given below.

### 2.1 Datasets

**The CépiDc corpus** was provided by the French institute for health and medical research (INSERM) for the task of ICD10 coding in CLEF eHealth 2017 (Task 1). It consists of free text death certificates collected from physicians and hospitals in France over the period of 2006–2014 [11].

[1] Centre d'épidémiologie sur les causes médicales de décès, Unité Inserm US10, `http://www.cepidc.inserm.fr/`.

[2] American Center for Disease Control, `https://www.cdc.gov/`

**The CDC corpus** was provided by the American Center for Disease Control (CDC). It consists of free text death certificates collected electronically in the United States during the year 2015. These are all records due to natural causes, i.e., there are no injury-related deaths included.

**Dataset excerpts.** Death certificates are standardized documents filled by physicians to report the death of a patient. The content of the medical information reported in a death certificate and subsequent coding for public health statistics follows complex rules described in a document that was supplied to participants [11]. Tables 1 and 2 present excerpts of the CépiDC and CDC corpora that illustrate the heterogeneity of the data that participants had to deal with. While some of the text lines were short and contained a term that could be directly linked to a single ICD10 code (e.g., "choc septique"), other lines could contain non-diacritized text (e.g., "peritonite..." missing the diacritic on the first "e"), abbreviations (e.g., "DM II" instead of "diabetes mellitus, type 2"). Other challenges included run-on narratives or mixed text alternating between upper case non-diacritized text and lower-case diacritized text.

**Table 1.** A sample document from the CépiDC French Death Certificates Corpus: the raw causes (Raw) and computed causes (Computed) are aligned into line-level mappings to ICD codes (Aligned). English translations for each text line are provided in footnotes

| | line | text | normalized text | ICD codes |
|---|---|---|---|---|
| **Raw** | 1 | choc septique[3] | | - |
| | 2 | peritonite stercorale sur perforation colique[4] | | - |
| | 3 | Syndrome de détresse respiratoire aiguë[5] | | - |
| | 4 | defaillance multivicerale[6] | | - |
| | 5 | HTA[7] | | - |
| **Computed** | 1 | | defaillance multivicerale | R57.9 |
| | 2 | | syndrome détresse respiratoire aiguë | J80.0 |
| | 3 | | choc septique | A41.9 |
| | 4 | | peritonite stercorale | K65.9 |
| | 5 | | perforation colique | K63.1 |
| | 6 | | hta | I10.0 |
| **Aligned** | 1 | choc septique | choc septique | A41.9 |
| | 2 | peritonite stercorale sur perforation colique | peritonite stercorale | K65.9 |
| | 2 | peritonite stercorale sur perforation colique | perforation colique | K63.1 |
| | 3 | Syndrome de détresse respiratoire aiguë | syndrome détresse respiratoire aiguë | J80.0 |
| | 4 | defaillance multivicerale | défaillance multiviscérale | R57.9 |
| | 5 | HTA | hta | I10.0 |

**Table 2.** Two sample documents from the American CDC Death Certificates Corpus

| type | line | text | ICD codes |
|---|---|---|---|
| | | Sample Certificate 1 | |
| Raw causes | 1 | CARDIAC ARREST | - |
| | 2 | ACUTE CORONARY SYNDROME | - |
| | 3 | ACUTE OR CHRONIC KIDNEY DISEASE | - |
| | 4 | DIABETIC NEUROPATHY | - |
| | 6 | PERIPHERAL ARTERIAL DISEASE; DM II | - |
| Computed causes | 1 | | I469 |
| | 2 | | I249 |
| | 3 | | N009 |
| | 3 | | N189 |
| | 4 | | E144 |
| | 6 | | I739 |
| | 6 | | E119 |
| | 6 | | F179 |
| | | Sample Certificate 2 | |
| Raw causes | 1 | STROKE IN SEPTEMBER LEFT HEMIPARESIS | - |
| | 2 | FALL SCALP LACERATION FRACTURE HUMERUS | - |
| | 3 | CORONARY ARTERY DISEASE | - |
| | 4 | ACUTE INTRACRANIAL HEMORRHAGE | - |
| | 6 | DEMENTIA DEPRESSION HYPERTENSION | - |
| Computed causes | 1 | | I64 |
| | 2 | | G819 |
| | 3 | | S010 |
| | 3 | | W19 |
| | 3 | | S423 |
| | 4 | | I251 |
| | 5 | | I629 |
| | 6 | | F03 |
| | 6 | | F329 |
| | 6 | | I10 |

**Descriptive statistics.** Tables 3 and 4 present statistics for the specific sets provided to participants. For both languages, the dataset construction was time-oriented in order to reflect the practical use case of coding death certificates, where historical data is available to train systems that can then be applied to current data to assist with new document curation. For French, the training set

[3] *septic shock*

[4] *colon perforation leading to stercoral peritonitis*

[5] *Acute Respiratory Distress Syndrome*

[6] *multiple organ failure*

[7] *HBP: High Blood Pressure*

covered the 2006–2012 period, and the development set contained death certificates from 2013 and the test set from 2014. For English, data was only available for the year 2015, but the training and test sets were nonetheless divided chronologically during that year. While the French dataset offers more documents spread over an eight year period, it also reflects changes in the coding rules and practices over the period. In contrast, the English dataset is smaller but more homogeneous.

**Table 3.** Descriptive statistics of the CépiDc French Death Certificates Corpus

| | Training (2006–2012) | Development (2013) | Test (2014) |
|---|---:|---:|---:|
| Certificates | 65,844 | 27,850 | 31,690 |
| Aligned lines | 195,204 | 80,899 | 91,962 |
| Tokens[8] | 1,176,994 | 496,649 | 599,127 |
| Total ICD codes | 266,808 | 110,869 | 131,426 |
| Unique ICD codes | 3,233 | 2,363 | 2,527 |
| Unique unseen ICD codes | - | 224 | 266 |

**Table 4.** Descriptive statistics of the CDC American Death Certificates Corpus

| | Training (2015) | Test (2015) |
|---|---:|---:|
| Certificates | 13,330 | 6,665 |
| Non-aligned lines | 32,714 | 14,834 |
| Tokens[9] | 90,442 | 42,819 |
| Total ICD codes | 39,334 | 18,928 |
| Unique ICD codes | 1,256 | 900 |
| Unique unseen ICD codes | - | 157 |

**Dataset format.** In compliance with the World Health Organization (WHO) international standards, death certificates comprise two parts: Part I is dedicated to the reporting of diseases related to the main train of events leading directly to death, and Part II is dedicated to the reporting of contributory conditions not directly involved in the main death process.[10] According to WHO recommenda-

[8] These numbers were obtained using the linux wc -w command

[9] These numbers were obtained using the linux wc -w command applied to the fourth field

[10] As can be seen in the sample documents, the line numbering in the raw causes file may (Table 2) or may not (Table 1) be the same in the computed causes file. In some cases, the ordering in the computed causes file was changed to follow the causal chain of events leading to death.

tions, the completion of both parts is free of any automatic assistance that might influence the certifying physician. The processing of death certificates, including ICD10 coding, is performed independently of physician reporting. In France and in the United States, coding of death certificates is performed within 18 months of reporting using the IRIS system [12]. In the course of coding practice, the data is stored in different files: a file that records the native text entered in the death certificates (referred as 'raw causes' thereafter) and a file containing the result of ICD code assignment (referred as 'computed causes' thereafter). The 'computed causes' file may contain normalized text that supports the coding decision and can be used in the creation of dictionaries for the purpose of coding assistance. We found that the formatting of the data into raw and computed causes made it difficult to directly relate the codes assigned to original death certificate texts. This makes the datasets more suitable for approaching the coding problem as a text classification task at the document level rather than a named entity recognition and normalization task. We have reported separately on the challenges presented by the separation of data into raw and computed causes, and proposed solutions to merge the French data into a single 'aligned' format, relying on the normalized text supplied with the French raw causes [13]. Table 1 presents a sample of French death certificate in 'raw' and 'aligned' format. It illustrates the challenge of alignment with the line 2 in the raw file "péritonite stercorale sur perforation colique" which has to be mapped to line 4 "peritonite stercorale" (code K65.9) and line 5 "perforation colique" (code K63.1) in the computed file.

As can be seen in Table 2 similar alignment challenges can be encountered in the English dataset. In Sample certificate 2, line 1 in the raw file "STROKE IN SEPTEMBER LEFT HEMIPARESIS" has to be mapped to line 1 (code I64, "Stroke, not specified") and line 2 (code G819, "Hemiplegia, unspecified") in the computed file. However, no normalized text was available for English and we were not able to offer an aligned version of the raw and computed files for the American dataset in this edition of the shared task.

**Data files.** Table 5 presents a description of the files that were provided to the participants: training (*train*) and development (*dev*, French only) files were distributed early in the challenge (in January 2017) ; test files (*test*, with no gold standard) were distributed at test time (at the end of April 2017); and the gold standard for test files (*test+g* in aligned format, *test, computed* in raw format) were disclosed to the participants after the text phase (in May 2017) just before the submission of their workshop papers, so that participants could reproduce the performance measures announced by the organizers.

### 2.2 Tasks

**ICD10 coding** The coding task consisted of mapping lines in the death certificates to one or more relevant codes from the International Classification of Diseases, tenth revision (ICD10). For the raw datasets, codes were assessed at the certificate level. For the aligned dataset, codes were assessed at the line level.

**Table 5.** Data files. Files after the dashed lines are test files; files after the dotted lines contain the gold test data. L = language (fr = French, en = English).

| | L. | Split | Type | Year | File name |
|---|---|---|---|---|---|
| **Aligned** | fr | train | aligned | 2006–2012 | corpus/train/AlignedCauses_2006-2012full.csv |
| | fr | dev | aligned | 2013 | corpus/dev/AlignedCauses_2013full.csv |
| | fr | test | aligned | 2014 | aligned/corpus/AlignedCauses_2014test.csv |
| | fr | test+g | aligned | 2014 | aligned/corpus/AlignedCauses_2014_full.csv |
| **Raw** | fr | train | raw | 2006–2012 | corpus/train/CausesBrutes_FR_training.csv |
| | fr | train | ident | 2006–2012 | corpus/train/Ident_FR_training.csv |
| | fr | train | computed | 2006–2012 | corpus/train/CausesCalculees_FR_training.csv |
| | fr | dev | raw | 2013 | corpus/dev/CausesBrutes_FR_dev.csv |
| | fr | dev | ident | 2013 | corpus/dev/Ident_FR_dev_full.csv |
| | fr | dev | computed | 2013 | corpus/dev/CausesCalculees_FR_dev.csv |
| | fr | test | raw | 2014 | raw/corpus/CausesBrutes_FR_test2014.csv |
| | fr | test | ident | 2014 | raw/corpus/Ident_FR_test2014.csv |
| | fr | test | computed | 2014 | raw/corpus/CausesCalculees_FR_test2014_full.csv |
| **Raw** | en | train | raw | 2015 | corpus/CausesBrutes_EN_training.csv |
| | en | train | ident | 2015 | corpus/Ident_EN_training.csv |
| | en | train | computed | 2015 | corpus/CausesCalculees_EN_training.csv |
| | en | test | raw | 2015 | raw/corpus/CausesBrutes_EN_test.csv |
| | en | test | ident | 2015 | raw/corpus/Ident_EN_test.csv |
| | en | test | computed | 2015 | raw/corpus/CausesCalculees_EN_test_full.csv |

**Replication.** The replication task invited lab participants to submit a system used to generate one or more of their submitted runs, along with instructions to install and use the system. Then, two of the organizers independently worked with the submitted material to replicate the results submitted by the teams as their official runs.

### 2.3 Evaluation metrics

System performance was assessed by the usual metrics of information extraction: precision (Formula 1), recall (Formula 2) and F-measure (Formula 3; specifically, we used $\beta$=1.).

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{1}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{2}$$

$$\text{F-measure} = \frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \tag{3}$$

Results were computed using two perl scripts, one for the **raw** datasets (in English and in French) and one for the **aligned** dataset (in French only). The evaluation tools were supplied to task participants along with the training data.

Measures were computed for “ALL” causes in the datasets as our main evaluation reference for the task. In this case the evaluation is performed for all ICD codes. Measures were also computed for “EXTERNAL” causes as our secondary reference for the task. In this case, the evaluation is limited to ICD codes addressing a particular type of deaths, called “external causes” or violent deaths. These causes are of particular interest for two reasons: first, they are considered as “avoidable” and public health policies can target them specifically, e.g., suicide prevention. Second, the context associated with these deaths is often quite different from other deaths in terms of comorbidity, population affected and terminology used to describe the event. In practice, external causes are characterized by codes V01 to Y98.

For the raw datasets, matches (true positives) were counted for each ICD10 full code supplied that matched the reference for the associated document.

For the aligned dataset, matches (true positives) were counted for each ICD10 full code supplied that matched the reference for the associated document line.

The evaluation of the submissions to the **replication** task was essentially qualitative: we used a scoring grid to record the ease of installing and running the systems, the time spent to obtain results with the systems (analysts were committed to spend at most one working day—or 8 hours—to work with each system), and whether we managed to obtain the exact same results submitted as official runs.

## 3 Results and Discussion

Participating teams included between one and twelve team members and resided in Australia (team UNSW), France (teams LIMSI, LIRMM, LITL, Mondeca and SIBM), Germany (teams TUC and WBI), Italy (Team UNIPD) and Russia (team KFU). Teams often comprised members with a variety of backgrounds and drew from computer science, informatics, statistics, information and library science, clinical practice. It can be noted that one team (LITL) participated in the challenge as a master-level class project. One team (LIMSI) was composed of members of the organization team and submitted unofficial runs due to conflict of interest. One team submitted baseline runs.

For the English raw dataset, we received 15 official runs from 9 teams, including one baseline run and one invalid run (due to formatting issues). For the French raw dataset, we received 7 official runs from 4 teams. For the French aligned dataset, we received 9 official runs from 6 teams, including one baseline run.

Five systems were submitted to the replication track, allowing us to attempt replicating a total of 22 system runs.

### 3.1 Methods implemented in the participants’ systems

Participants used a variety of methods, many of which relied on lexical sources including the dictionaries supplied as part of the training data as well as other medical terminologies and ontologies. Some of these knowledge-based methods exploited the gold standard training data as an additional knowledge source.

*IMS-UNIPD.* The UNIPD team submitted official runs for the English dataset and later submitted unofficial runs for the French datasets as well [14]. This team implemented a minimal expert system based on rules to translate acronyms together with a binary weighting approach (run 1) and a tf-idf approach (run 2) to retrieve the items in the dictionary most similar to the portion of the certificate of death. For both configurations, a basic approach was used to select the class with the highest weight.

*KFU.* The KFU team submitted two runs for the English dataset [15]. They used sequence to sequence deep learning models based on recurrent neural networks. As input sequence, the method takes the raw text and outputs sequence of ICD10 codes. Both the supplied corpus and dictionary were used for training, exclusive of any additional data.

*LITL.* The LITL team submitted runs for the French dataset in the raw and aligned formats [16]. The LITL team system was specifically designed by master's students (LITL programme, university of Toulouse) and their teachers for the challenge. The system is based on the search platform SOLR. Training data was indexed using the SolrXML format. The core is organized into ICD codes associated with the corresponding "raw Texts", "diagnostic Texts", ICD headings and SNOMED labels. The raw Texts from the test dataset were automatically transformed into queries and submitted to SOLR. The two runs submitted are based on the same collection and SOLR configuration. For Run 1, raw texts were automatically split into several queries when different causes were detected by using a custom-made rule-based system. For Run 2, each query corresponds to the entire raw text of each CépiDC line.

*LIMSI.* The LIMSI team submitted unofficial runs for all datasets [17]. The starting point for these submissions is their last published system [18], which relied upon dictionary projection and supervised multi-class, mono-label text classification using simple features (bag of normalized tokens, character trigrams, and coding year). They extended this system to multi-label classification and the use of dictionary and token bigram features in the classifier. Character n-grams did not improve the F1-score on the training set and were discarded. Coding year was kept for the French data, but not for the English data, because it only spans year 2015. Because it only relies on the material provided by the task organizers, the same system could be applied to both the French and English datasets. In each case, Run 1 used a supervised machine learning method (multi-label SVM, with unigrams, bigrams and [for French] coding year), and Run 2 used a hybrid method: union of calibrated dictionary and multi-label SVM.

*LIRMM.* The LIRMM team submitted runs for all datasets [19]. They annotated death certificate text through the SIFR Bioportal Annotator (`http:`

`//bioportal.lirmm.fr/annotator`) using different configurations of the web service. For French, Simple Knowledge Organization System (SKOS) was built using ICD10 content from the CISMeF portal, the set of dictionaries provided in the challenge, as well as the training corpus. For the first run, the ontology was generated with a heuristic, where labels that correspond to multiple codes are assigned to the most frequent code only. For the second run, a fall back strategy relaxes the most frequent code heuristic for lines that were not assigned any codes initially. For English, in the first run, the SKOS was built using the American dictionary supplied with training data. In the second run the dictionary was combined with an owl version of ICD10 and ICD10CM (extracted from the Unified Medical Language System).

*Mondeca.* The Mondeca team submitted unofficial runs[11] for all datasets [20]. They approached multilingual extraction of IC10 codes by combining semantic web technology and NLP concepts in four steps: (*i*) transform all the datasets into RDF for a graph-based manipulation; (*ii*) transform the dictionaries for all the years into SKOS for better enrichment across the knowledge-bases; (*iii*) design a GATE workflow to annotate the RDF datasets based on gazetteers extracted from the dictionaries; and (*iv*) work on both French (raw data) and English corpus within a unique workflow, in a multilingual approach thus enabling simultaneous processing of multiple languages.

*SIBM.* The SIBM team submitted runs for all datasets [21]. Their approach of term extraction is performed at the phrase level using natural language processing. The system is built using Python and Python/C extensions and produces the following output for each identified concept: (*i*) the entry text, (*ii*) the offset of the first and the final word contained in the health concept, (*iii*) the ICD10 identifier and (*iv*) the ICD10 term. Three main steps lead to the identification of ICD10 concepts for a given text: During *tokenization*, the input text is sliced into phrases, then words. Stop words are filtered and spell checking is performed using the Enchant library. Next, during *ICD10 candidate selection*, a method based on the phonetic encoding algorithm Double Metaphone (DM) is used for approximate term search. This system relies on a database storing pre-computed DM codes for each word available in the ICD10 dictionaries. Finally, during *candidate ranking*, a combination of the longest common substring and fuzzy match algorithms provides the candidate ranking. The most likely term having the highest score is retained as the matching ICD10 code for the phrase.

*TUC.* The TUC team submitted runs for all datasets [22]. Their approach is focused on the exploration of relevant feature groups for multilingual text classification regarding ICD10 codes. First, a large scale brute-force feature set is constructed using the groups bag of words, bag of bigrams, bag of trigrams, latent Dirichlet allocation, and the ontologies of WordNet and UMLS. In the

[11] One official run was submitted but did not comply with the challenge required format and could not be evaluated.

development phase, three different strategies were evaluated in conjunction with support vector machines for the English and French corpus: each feature group separately, early fusion of all feature groups, and late fusion. For English, early fusion (run 1) and the feature group bag of bigrams (run 2) achieved the best results. For French, average late fusion concerning bag of words and bag of bigrams (run 1), and the feature group bag of bigrams (run 2) performed best.

*UNSW.* The UNSW team submitted runs for the American dataset [23]. They deployed a knowledge-based approach to tackle the task by solely using dictionary lookup. The first step is to index manually coded ICD10 lexicon followed by dictionary matching. Priority rules are applied to retrieve the relevant entity/entities and their corresponding ICD10 code(s) given free text cause of death description. Two priority methods were implemented in the submitted runs: the first one relied on BM25 and the second one on direct term match. The advantages of a knowledge-based method include speed and no need for training data.

*WBI.* The WBI team submitted runs for the English raw dataset and for the French aligned dataset [24]. They combined standard rule-based methods for Named Entity Recognition (NER) with machine-learning approaches for candidate ranking. For NER rule-based dictionary lookup and fuzzy matching using Lucene Sorl was applied. Preference was on generating potential candidates for each match to increase recall. Candidates were then ranked using a machine-learning approach. Based on the hierarchy of the ICD10 terminology (chapters, blocks, sub-chapters) combined with ICD10-Codes and Text available from the provided dictionaries a classifier was developed for ranking candidates.

*Baselines.* To provide a better assessment of the task difficulty and system performance, this year we offer baseline results using two methods: 1/ the *ICD baseline* consisted of exact string matching between the terms in the ICD and the death certificate text. 2/ the *frequency baseline* consisted in assigning to a certificate line from the test set the top 2 most frequently associated ICD10 codes in the training and development sets, using case and diacritic insensitive line matching.

### 3.2 System performance on death certificate coding

Tables 6 to 8 present system performance on the ICD10 coding task for each dataset. Team KFU obtained the best performance in terms of F-measure both overall and for the external causes of death on the English dataset. Team SIBM obtained the best official performance in terms of F-measure both overall and for the external causes of death on the French datasets. It is interesting to note that the participants who obtained the best scores on the French datasets (SIBM and LIMSI) are returning teams who also participated in the coding task in 2016. Team SIBM's performance improved from an F-measure of .680 in 2016 to an F-measure of .804 this year while team LIMSI's performance improved from an

**Table 6.** System performance for ICD10 coding on the **English raw** test corpus in terms of Precision (P), recall (R) and F-measure (F). The top part of the table displays official runs, while the bottom part displays non-official and baseline runs.

| | ALL | | | | EXTERNAL | | | |
|---|---|---|---|---|---|---|---|---|
| | **Team** | P | R | F | **Team** | P | R | F |
| **Official runs submitted** | KFU-run1 | .893 | .811 | **.850** | KFU-run1 | .584 | .357 | **.443** |
| | KFU-run2 | .891 | **.812** | .850 | KFU-run2 | .631 | .325 | .429 |
| | TUC-MI-run1 | **.940** | .725 | .819 | SIBM-run1 | .426 | .389 | .407 |
| | SIBM-run1 | .839 | .783 | .810 | LIRMM-run2 | .233 | **.524** | .323 |
| | TUC-MI-run2 | .929 | .717 | .809 | LIRMM-run1 | .232 | **.524** | .322 |
| | WBI-run1 | .616 | .606 | .611 | TUC-MI-run1 | .880 | .175 | .291 |
| | WBI-run2 | .616 | .606 | .611 | TUC-MI-run2 | **1.00** | .159 | .274 |
| | LIRMM-run1 | .691 | .514 | .589 | UNSW-run1 | .168 | .262 | .205 |
| | LIRMM-run2 | .646 | .527 | .580 | Unipd-run2 | .292 | .111 | .161 |
| | Unipd-run1 | .496 | .442 | .468 | WBI-run1 | .246 | .119 | .160 |
| | UNSW-run1 | .401 | .352 | .375 | WBI-run2 | .246 | .119 | .160 |
| | Unipd-run2 | .382 | .341 | .360 | Unipd-run1 | .279 | .095 | .142 |
| | UNSW-run2 | .371 | .328 | .348 | UNSW-run2 | .043 | .310 | .076 |
| | Mondeca-run1 | *invalid format* | | | Mondeca-run1 | *invalid format* | | |
| | **average** | .670 | .582 | .622 | **average** | .405 | .267 | .261 |
| | **median** | .646 | .606 | .611 | **median** | .279 | .262 | .274 |
| **Non-off** | LIMSI-run2 | .899 | .801 | .847 | LIMSI-run2 | .723 | .373 | .492 |
| | LIMSI-run1 | .909 | .765 | .831 | LIMSI-run1 | .837 | .325 | .469 |
| | Mondeca-run1 | .691 | .309 | .427 | Mondeca-run1 | .042 | .056 | .048 |
| | Frequency baseline | .115 | .085 | .097 | Frequency baseline | 0.00 | 0.00 | 0.00 |
| | ICD baseline | .029 | .007 | .011 | ICD baseline | 0.00 | 0.00 | 0.00 |

F-measure of .652 in 2016 to an F-measure of .867 this year, which also exceeds the best performance of 2016 obtained by team Erasmus with F-measure of .848.[12] This suggests that there is room for improvement on this task, and that iterations of the task are useful to help identify the best ideas and methods to address the task.

To provide a more in-depth analysis of results, this year we also introduced a measure of system performance on the external causes of death, which are of specific interest to public-health specialists, and are also thought to be more difficult to code. This hypothesis was confirmed by the results, as system performance was much lower on the external causes vs. all causes for all systems, both for the English and French datasets. Interestingly, some systems offered very good performance overall, but comparatively quite low performance on external causes, and vice-versa. We also note that the performance of the frequency baseline was much higher on the French aligned dataset, compared to the French raw dataset and English dataset. This suggests that there is value to the alignment

[12] We note that these comparisons are indicative since the data sets used in 2016 and 2017 are not identical; specifically, the 2016 test set was distributed in 2017 as a development set and the 2017 test set consisted of new data (unreleased in 2016).

**Table 7.** System performance for ICD10 coding on the **French raw** test corpus in terms of Precision (P), recall (R) and F-measure (F). A horizontal dash line places the frequency baseline performance. The top part of the table displays official runs, while the bottom part displays non-official and baseline runs.

| | **ALL** | | | | **EXTERNAL** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Team** | P | R | F | **Team** | P | R | F |
| **Official runs** | SIBM-run1 | **.857** | **.689** | **.764** | SIBM-run1 | **.567** | **.431** | **.490** |
| | LITL-run2 | .666 | .414 | .510 | LIRMM-run1 | .443 | .367 | .401 |
| | LIRMM-run1 | .541 | .480 | .509 | LIRMM-run2 | .443 | .367 | .401 |
| | LIRMM-run2 | .540 | .480 | .508 | LITL-run2 | .560 | .283 | .376 |
| | LITL-run1 | .651 | .404 | .499 | LITL-run1 | .538 | .277 | .365 |
| | TUC-MI-run2 | .044 | .026 | .033 | TUC-MI-run2 | .010 | .004 | .005 |
| | TUC-MI-run1 | .025 | .015 | .019 | TUC-MI-run1 | .006 | .005 | .005 |
| | **average** | .475 | .358 | .406 | **average** | .367 | .247 | .292 |
| | **median** | .541 | .414 | .508 | **median** | .443 | .283 | .376 |
| **Non-official** | LIMSI-run2 | .872 | .784 | .825 | LIMSI-run2 | .700 | .594 | .643 |
| | LIMSI-run1 | .883 | .760 | .817 | LIMSI-run1 | .709 | .559 | .625 |
| | TUC-MI-run1-corrected | .883 | .539 | .669 | TUC-MI-run1-corrected | .780 | .290 | .423 |
| | TUC-MI-run2-corrected | .882 | .536 | .667 | TUC-MI-run2-corrected | .767 | .283 | .414 |
| | UNIPD-run1 | .629 | .468 | .537 | UNIPD-run2 | .350 | .381 | .365 |
| | UNIPD-run2 | .518 | .384 | .441 | UNIPD-run1 | .362 | .251 | .296 |
| | Mondeca-run1 | .375 | .131 | .194 | Mondeca-run1 | .335 | .228 | .271 |
| | Frequency baseline | .339 | .237 | .279 | Frequency baseline | .381 | .110 | .170 |

step of data preparation, and to the size of the dataset (the French dataset was significantly larger than the English dataset).

The results show that both knowledge-based and statistical methods can perform well on the task. For English the best performance is obtained from a statistical neural method (team KFU) and the second best is obtained by a machine learning method relying on knowledge based-sources (team LIMSI). For French, the best performance is obtained from a machine learning method relying on knowledge based-sources (team LIMSI), while the second best is obtained with a combination of knowledge based and Natural Language processing methods (Team SIBM). In addition, many teams relied on a system architecture that was the same for both languages and utilized language specific features or knowledge sources, requiring little language adaptation. The results are very encouraging from a practical perspective and indicate that a coding assistance system could prove very useful for the effective processing of death certificates in multiple languages.

### 3.3 Replication track and replicability of the results

Five teams submitted systems to our replication track. Only one of these teams had also participated in the replication track last year. Four systems covered both French and English, and one system only processed English.

**Table 8.** System performance for ICD10 coding on the **French aligned** test corpus in terms of Precision (P), recall (R) and F-measure (F). A horizontal dash line places the frequency baseline performance. The top part of the table displays official runs, while the bottom part displays non-official and baseline runs.

| | **ALL** | | | | **EXTERNAL** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Team** | P | R | F | **Team** | P | R | F |
| **Official runs** | SIBM-run1 | .835 | **.775** | **.804** | SIBM-run1 | .534 | **.472** | **.501** |
| | WBI-run1 | .780 | .751 | .765 | TUC-MI-run2 | **.740** | .318 | .445 |
| | TUC-MI-run2 | **.874** | .611 | .719 | LIRMM-run1 | .412 | .403 | .407 |
| | LITL-run1 | .612 | .550 | .579 | LIRMM-run2 | .412 | .403 | .407 |
| | LIRMM-run1 | .506 | .530 | .518 | LITL-run1 | .482 | .348 | .404 |
| | LIRMM-run2 | .505 | .530 | .517 | LITL-run2 | .534 | .275 | .363 |
| | LITL-run2 | .646 | .402 | .495 | WBI-run1 | .709 | .151 | .249 |
| | TUC-MI-run1 | .426 | .297 | .350 | TUC-MI-run1 | .218 | .119 | .154 |
| | **average** | .648 | .555 | .593 | **average** | .505 | .311 | .366 |
| | **median** | .629 | .540 | .548 | **median** | .508 | .333 | .406 |
| **Non-official** | LIMSI-run2 | .854 | .881 | .867 | LIMSI-run2 | .630 | .674 | .651 |
| | LIMSI-run1 | .865 | .865 | .865 | LIMSI-run1 | .640 | .636 | .638 |
| | TUC-MI-run1-corrected | .875 | .614 | .722 | TUC-MI-run1-corrected | .748 | .323 | .452 |
| | UNIPD-run1 | .604 | .517 | .557 | UNIPD-run2 | .320 | .402 | .356 |
| | UNIPD-run2 | .488 | .418 | .451 | UNIPD-run1 | .376 | .265 | .311 |
| | Frequency baseline | .640 | .470 | .542 | Frequency baseline | .508 | .338 | .406 |
| | ICD baseline | .346 | .041 | .073 | ICD baseline | .000 | .000 | .000 |

In addition, the replication track also used the simple scripts used to produce baseline runs.

Most of the baseline and system runs could be replicated by at least one analyst. However, the analysts still experienced varying degrees of difficulty to install and run the systems. Differences were mainly due to the technical set-up of the computers used to replicate the experiments. Analysts also report that additional information on system requirements, installation procedure and practical use would be useful for all the systems submitted, although documentation was overall more abundant and detailed compared to last year's experiments. In some cases, system authors were contacted for help. They were responsive and contributed to facilitate the use of their system. The results of the experiments suggest that replication is achievable. However, it continues to be more of a challenge than one would hope.

### 3.4 Limitations

**Formatting issues.** In the French dataset, a formatting issue affected the certificates whose narratives contained a semicolon. The data export from IRIS to csv failed to adequately protect the text field with quotes, so that some of the data instances were made difficult to parse. Nonetheless, this problem affected less than 1% of the lines so we believe it had limited impact on the results. The export format will be corrected in future releases of the dataset. However, we would like to note that this type of issue fits within the practical ‘real life’ element of this challenge. While it certainly may have made system development more difficult, it also advocated for systems with strategies for dealing with potentially less-than-perfect data. While unintended, we believe this situation in fact makes for a robust evaluation because this kind of data would also be present in a practical workflow.

**Did smoking contribute to the death?** In the American dataset, the assignment of code F179 “Mental and behavioral disorders due to use of tobacco, unspecified” may be supported by information supplied by the reporting physician either in certificate narrative or in a structured data form. As a result, the gold standard assignment of F179 is sometimes unsupported by text. The prevalence of F179 due to form filling vs. text report is unknown and the two cases are currently indistinguishable in the dataset. The sample document shown in Table 2 illustrates the case of F179 assignment supported by data form and not by text. The prevalence of the code is 4.7% in the training set and 3.9% in the test set, which creates a bias for all evaluated systems. We estimate that the bias could create differences of up to 2% in the overall F-measure. However, we note that the external causes evaluation is not impacted because F179 does not belong to the external cause of death category.

## 4 Conclusion

We released a new set of death certificates to evaluate systems on the task of ICD10 coding in multiple languages. This is the third edition of a biomedical NLP challenge that provides large gold-standard annotated corpora in French. Results show that high performance can be achieved by NLP systems on the task of coding for death certificates in French and in English. The level of performance observed shows that there is potential for integrating automated assistance in the death certificate coding workflow in both languages. We hope that continued efforts towards reproducibility will support the shift from research prototypes to operational production systems. The corpus used and the participating team system results are an important contribution to the research community. In addition, the focus on a language other than English (French) remains a rare initiative in the biomedical NLP community.

## Acknowledgements

We want to thank all participating teams for their effort in addressing new and challenging tasks. The organization work for CLEF eHealth 2017 task 1 was supported by the Agence Nationale pour la Recherche (French National Research Agency) under grant number ANR-13-JCJC-SIMI2-CABeRneT. The CLEF eHealth 2016 evaluation lab has been supported in part by the CLEF Initiative and Data61.

## References

1. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon. CLEF 2017 eHealth Evaluation Lab Overview. CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September, 2017.
2. World Health Organization. ICD-10. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. Volume 2. Instruction manual. 2011.
3. Jones KS, Galliers JR. Evaluating natural language processing systems: An analysis and review. 1995. Springer Science & Business Media:1083
4. Voorhees EM, Harman DK and others. TREC: Experiment and evaluation in information retrieval, vol 1. 2005. MIT press Cambridge.
5. Suominen H, Salantera S, Velupillai S, Chapman WW, Savova G, Elhadad N, Pradhan S, South BR, Mowery DL, Jones GJF, Leveling J, Kelly L, Goeuriot L, Martinez D, Zuccon G. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds), Information Access Evaluation. Multilinguality, Multimodality, and Visualization. LNCS (vol. 8138):212-231. Springer, 2013
6. Goeuriot L, Kelly L, Suominen H, Hanlen L, Névéol A, Grouin C, Palotti J, Zuccon G. Overview of the CLEF eHealth Evaluation Lab 2015. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction. Springer, 2015
7. Kelly L, Goeuriot L, Suominen H, Névéol A, Palotti J, Zuccon G. (2016) Overview of the CLEF eHealth Evaluation Lab 2016. In: Fuhr N. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2016. Lecture Notes in Computer Science, vol 9822. Springer, Cham
8. Kelly L, Goeuriot L, Suominen H, Schreck T, Leroy G, Mowery DL, Velupillai S, Chapman WW, Martinez D, Zuccon G, Palotti J. 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. LNCS (vol. 8685):172-191. Springer, 2014
9. 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. J Am Med Inform Assoc, 18(5):540-3
10. Huang CC, Lu Z (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform, 2015 May 1. pii: bbv024.
11. Pavillon G., Laurent F (2003). Certification et codification des causes médicales de décès. Bulletin Epidémiologique Hebdomadaire - BEH:134-138. http://opac.invs.sante.fr/doc_num.php?explnum_id=2065 (accessed: 2016-06-06)
12. Johansson LA, Pavillon G (2005). IRIS: A language-independent coding system based on the NCHS system MMDS. In WHO-FIC Network Meeting, Tokyo, Japan
13. Lavergne T, Névéol A, Robert A, Grouin C, Rey G, Zweigenbaum P. A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage. Proceedings of the Fifth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing - BioTxtM2016. 2016.
14. Di Nunzio GM, Beghini F, Vezzani F and Henrot G (2017). A Lexicon Based Approach to Classification of ICD10 Codes. IMS Unipd at CLEF eHealth Task 1. CLEF 2017 Online Working Notes. CEUR-WS

15. Miftakhutdinov Z and Tutubalina E (2017). KFU at CLEF eHealth 2017 Task 1: ICD-10 Coding of English Death Certificates with Recurrent Neural Networks. CLEF 2017 Online Working Notes. CEUR-WS
16. Ho-Dac LM, Fabre C, Birski A, Boudraa I, Bourriot A, Cassier M, Delvenne L, Garcia-Gonzalez C, Kang EB, Piccinini E, Rohrbacher C and Séguier A (2017). LITL at CLEF eHealth2017: automatic classification of death reports. CLEF 2017 Online Working Notes. CEUR-WS
17. Zweigenbaum P and Lavergne T (2017). Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. CLEF 2017 Online Working Notes. CEUR-WS
18. Zweigenbaum P and Lavergne T. Hybrid methods for ICD-10 coding of death certificates. In Seventh International Workshop on Health Text Mining and Information Analysis, pages 96-105, Austin, Texas, USA, November 2016. EMNLP 2016.
19. Tchechmedjiev A, Abdaoui A, Emonet V and Jonquet C (2017). ICD10 coding of death certificates with the NCBO and SIFR Annotator(s) at CLEF eHealth 2017 Task 1. CLEF 2017 Online Working Notes. CEUR-WS
20. Atemezing GA (2017). NoNLP: Annotating Medical Domain by using Semantic Techologies. CLEF 2017 Online Working Notes. CEUR-WS
21. Cabot C, Soualmia LF and Darmoni SJ (2017). SIBM at CLEF eHealth Evaluation Lab 2017: Multilingual Information Extraction with CIM-IND. CLEF 2017 Online Working Notes. CEUR-WS
22. Ebersbach M, Herms R and Eibl M (2017). Fusion Methods for ICD10 Code Classification of Death Certificates in Multilingual Corpora. CLEF 2017 Online Working Notes. CEUR-WS
23. Jonnagaddala J and Hu F (2017). Automatic coding of death certificates to ICD-10 terminology. CLEF 2017 Online Working Notes. CEUR-WS
24. Ševa J, Kittner M, Roller R and Leser U (2017). Multi-lingual ICD-10 coding using a hybrid rule-based and supervised classification approach at CLEF eHealth 2017. CLEF 2017 Online Working Notes. CEUR-WS