Evaluation of automatic annotation by a multi-terminological concepts extractor within a corpus of data from family medicine consultations

Charlotte Siefridt, Julien Grosjean, Tatiana Lefebvre, Laëtitia Rollin, Stefan Darmoni, Matthieu Schuers

International Journal of Medical Informatics

Title page

Evaluation of automatic annotation by a multi-terminological concepts extractor within a corpus of data from family medicine consultations

Charlotte Siefridt[a,b,*], Julien Grosjean[b, c], Tatiana Lefebvre[b], Laetitia Rollin[c,d], Stefan Darmoni[b,c], Matthieu Schuers[a,c]

[a] *Department of General Medicine, Rouen University Hospital, Rouen, France*
[b] *Department of Biomedical Informatics, Rouen University Hospital, Rouen, France*
[c] *INSERM, U1142, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, LIMICS, Sorbonne Université, Paris, France*
[d] *Department of Occupational and Environmental Medicine, Rouen University Hospital, Rouen, France*

**Corresponding author (Charlotte Siefridt) at: Department of biomedical informatics, Rouen University Hospital, 37 boulevard Gambetta 76000 Rouen, France*
E-mail address: charlotte.siefridt1@univ-rouen.fr

Abstract

Introduction

Research in family medicine is necessary to improve the quality of care. Current publications remain of heterogeneous quality. Databases from electronic medical records can increase the quality of these works. These data must be coded to be used pertinently. The objective of this study was to assess the quality of semantic annotation by a multi-terminological concept extractor within a corpus of family medicine consultations.

Method

Consultation data in French from 25 general practitioners were automatically annotated using 28 different terminologies. The data extracted were classified into three groups: reasons for consulting, observations and consultation results. The first evaluation led to a correction phase of the tool which led to a second evaluation. For each evaluation, the precision, recall and F-measure were quantified. Then, the inter- and intra-terminological coverage of each terminology was assessed.

Results

Nearly 15,000 automatic annotations were manually evaluated. The mean values for the second evaluation of precision, recall and F-measure were 0.85, 0.83 and 0.84 respectively. The most common terminologies used were SNOMED CT, SNOMED 3.5 and

NCIt. The terminologies with the best intra-terminological coverage were ICPC-2, DRC and CISMeF Meta-Terms.

Conclusion

A multi-terminological concepts extractor can be used for the automatic annotation of consultation data in family medicine. Integrating such a tool into general practitioners' business software would be a solution to the lack of routine coding. Developing the use of a single terminology specific to family medicine could improve coding, facilitate semantic interoperability and the communication of relevant information.

Keywords: Automatic annotation, Databases, Family medicine, Clinical coding, Electronic medical records

Introduction

1.1. Background

The development of research in family medicine is essential to improve the quality of care (1,2). Following an international impetus in the 2000's, the number of publications in the field of family medicine has increased significantly. In France, the creation of an academic course specific to family medicine has enabled research work to flourish. Nevertheless, the number of publications in general medicine remains low on PubMed (3). The use of Electronic Medical Records can increase the number of these publications by increasing the number of data available over a period from birth to death, but also by facilitating access to these data. In addition, these databases have been supplemented (by data matching) with additional data from areas such as laboratory analyses, hospital admissions and mortality statistics (4,5). The Clinical Research Practice Datalink (CPRD) developed in the United Kingdom, has proven its relevance for public health purposes, improving practices and increasing the quality of publications (6,7). In France, such databases are no longer in use, and belong to private companies (CEGEDIM, IMS-Health) or focus on specific diseases (Sentinel network). The French project "Regional Information Platform in General Medicine" is part of the need to build a health data warehouse in family medicine. It has been developed by 11 general practitioners (GPs) from Provence-Alps-French Riviera and 14 GPs from Normandy (8).

To ensure interoperability but also completeness and reliability of data in EMRs, they must be structured and standardized. In France, only 13% of GPs use coding (8,9). This can be explained by the lack of incentives for health professionals and strong academic and governmental support (5,10). Automatic coding with concept extractors could compensate for this low coding rate. In 2006, the CISMeF team created the Multi-Terminological Concepts Extractor (MTCE). This tool extracts concepts from natural language in any biomedical documents in French. It has demonstrated its relevance in the automatic annotation of hospital reports (11–14). It has never been evaluated on data from family medicine consultations before.

1.1. Objectives of the project

The main aim of this work was to evaluate the quality of automatic coding by MTCE within a database from family medicine consultations. The secondary aims were to evaluate the inter- and intra-terminological coverage for this corpus to identify relevant terminologies to use in such contexts.

2. Materials and Methods

2.1. Data description

French EMR data from 25 voluntary GPs were extracted between 2012 and 2015. Among the 25 GPs, 11 worked in the Provence-Alps-French Riviera area and 14 in the multidisciplinary health center in Normandy (France). Twenty-three were internship supervisors associated with a medical school. The corpus consisted of consultations notes divided into three subgroups: 9,182 reasons for consulting, 41,760 observations and 36,508 consultation results. The data was extracted in the formats ".txt" and ".csv" from the GPs' EMRs, using previously installed *ad hoc* extractor software. They were then imported into a MySQL database. Patient data were de-identified. In this study, the data extracted concerned the reasons for consulting, observations and consultation results. The observations included data from the interviews, the clinical examinations and the therapeutic procedures.

2.2. Automatic indexing with MTCE

MTCE is an automatic natural language processing tool. It identifies concepts from biomedical documents using the 75 health terminologies included in the Health Terminology/Ontology Portal (HeTOP) (15).

Among the 75 terminologies present in HeTOP, only 18 are also included in the Unified Medical Language System (UMLS). The other terminologies correspond mainly to French terminologies, such as CCAM for procedures. The number of unique concepts partially translated into French in the UMLS amounts to 158,475 compared to 444,258 in HeTOP.

In this study, we used 28 of the 75 available terminologies. The choice of these terminologies was based on a 2018 study that deemed them to be the most relevant for this work (16). The terminologies natively in French and those partially translated were kept as a priority. The list of terminologies used is available in Table 1. We used all the concepts present in the 28 terminologies. MTCE is based on a "bags-of-words" algorithm coupled with pattern matching (11–13). When a concept was found in several terminologies, only one of them was chosen at random, for the sake of simplification.

Terminologies	Full names
ATC	Anatomical Therapeutic Chemical classification
CCMP	Common Classification of Medical Procedures
CGP	Q codes
CISMeF	Catalogue and Index of French-language Medical Sites
DCR	Dictionary of Consultation Results
FMA	Foundational Model of Anatomy
HPO	Human Phenotype Ontology
HRDO	Human Rare Diseases Ontology
ICD-10	International Classification of Diseases – 10th Revision
ICD-9	International Classification of Diseases – 9th Revision
ICD-O	International Classification of Diseases for Oncology
ICF	International Classification of Functioning, Disability and Health
ICNP	International Classification for Nursing Practice
IUPAC	International Union of Pure and Applied Chemistry
LPS	List of Products and Services
MedDRA	MEDical Dictionary for Regulatory Activities terminologies
MedlinePlus	MedlinePlus
MeSH	Medical Subject Headings
NCIt	National Cancer Institute Thesaurus
PAS	PASCAL
PHA	Medicines
PHT	Public Health Thesaurus
Radlex	Radlex entity Ontology
SNOMED CT	Systematized Nomenclature of MEDicine Clinical Terms
SNOMED int.	Systematized Nomenclature of MEDicine
UTV	Unified Terminology of Vidal

Table 1: List of terminologies used

2.3. Evaluation of the quality of the annotation

The evaluation was carried out by a trained GP (CS), referred to as the MTCE Evaluator (MTCEe), using a dedicated web tool. It displayed on one side the list of identified concepts and on the other side the document concerned. The evaluator assessed the relevance of the annotation according to four options: "valid", "false", "irrelevant" and "to be verified". An

annotation was considered “irrelevant” when the identified concept was correct but did not provide added value when annotated in isolation, such as laterality. An annotation was considered "to be verified" if the evaluator could not evaluate the veracity of the annotation. It was also possible to add an annotation manually, concerning parts of the document already annotated or not. Thanks to his experience as a GP and his knowledge of terminologies, the evaluator identified the data that had not been annotated by the MTCE. There were several ways to add a concept a posteriori: either by overriding errors (spelling, syntax, acronyms, etc.) that prevented the tool from identifying concepts or by manually searching a concept in a classification. To do this, the evaluator selected the word or group of words that was not automatically annotated. Then, the MTCEe proposed a list of corresponding concepts among the 28 terminologies used. When a concept was available, the evaluator assigned it to the unannotated data. An example of adding a concept is explained in Appendix 1.

In case of doubt, the evaluator could refer the matter to two other authors (JG and MS).

From this evaluation, statistics were automatically generated to calculate standard metrics as precision, recall and F-measure.

Precision was defined as the proportion of correct data among all data identified. Precision was calculated as the ratio of the number of annotations validated by the evaluator to the total number of concepts annotated by the MTCE. Recall was defined as the proportion of data identified as correct among all correct data. In this case, recall was calculated as the ratio of the number of concepts validated by the evaluator to the total number of concepts that should have been found. This statistic includes the concepts added manually by the evaluator.

The F-measure is the harmonic mean of precision and recall and assesses the overall performance of the tool. It was calculated according to the formula:

F-measure = 2 * [(precision *recall) / (precision + recall)].

At the same time, the most common errors were identified and classified in a tabulated file. The objective was to improve the performance of the MTCE. The types of errors are listed in table 2. Negations were processed by the MTCE. If the negation was present in the document and not taken into account by the MTCE, the concept was considered false. The error analysis enabled corrections to be made to the tool or to the terminological content. Thus, if the concept was imprecise or irrelevant, like “autre” (other), the recognition of this wording by HeTOP was disabled. If the error was context related, the corresponding terminology was set aside for the annotation of this concept in the second analysis, like for the concept "radio" which was a frequent source of contextual errors. It was most often found as a means of communication rather than as an imaging examination (X-Ray). The expectation of this decision was that another terminology would be used to extract this concept in the second evaluation. When the term corresponded to a false synonym or an acronym, as "symptome" (symptom) for “syndrome" (syndrom), it was then deleted from the HeTOP portal. For stemming errors, the MTCE parameters were modified. Thus, the number of letters constituting a root has been increased. The MTCE first considered compound words as two distinct words. After correction, the MTCE considered them as one word.

Then, a second evaluation was carried out by measuring the same metrics and comparing them to the first evaluation to assess the contribution of the corrections. For this evaluation, the documents present in the first sample were excluded.

To limit the inter-individual variability of the evaluation, a second independent evaluator (TL) conducted an analysis on part of the same sample used for the second evaluation. We then measured the agreement between the two evaluators.

Types of errors	Definition	Action
Bags of words	Association of words that do not go together to extract a single concept	
Stemming	Use of the root of the word	Increase in the number of words composing a root
Adjective/name	Noun used instead of an adjective and vice versa	
Empty word	Neglecting the empty word can change the concept	
Negation	Negation absent or inadequate negation	
Probability	No consideration of uncertainty	
Acronym	Some words are recognized as acronyms and vice versa or some acronyms are false	Removal of the wrong acronym in HeTOP
False synonym	The annotated concept is a false synonym of the initial concept	Removal of the wrong synonym in HeTOP
Context	The word is recognized but is not in the right context	Setting aside the corresponding terminology for the concept
Compound word	Only one part of the word is recognized, or two different concepts are annotated	Recognition of the compound word as a single word
Hypo- or hyperonym	The term identified is a hyponym or hyperonym of the original concept	
Typing error	The initial word is not identified	
Imprecision	The extracted concept is imprecise (no definition, synonym or hierarchy)	Disabling recognition of this label
Abbreviation	The abbreviation is not well recognized	Addition of the abbreviation in the HeTOP portal
Unit Error	Error in the annotation of units	
Proper noun	The proper noun is interpreted as a concept	
Annotated numbers	Error in the annotation of numbers	

Table 2: Types of errors and actions

2.4. Evaluation of terminology coverage

The contribution of each terminology was quantified using two metrics: inter-terminological and intra-terminological coverage. For these measurements, the data was grouped into documents, to facilitate the assessment. A document consisted of a set of ten reasons for consulting or ten consultation results. Faced with a high percentage of false and irrelevant annotations (more than 50% during the first evaluation and 30% during the second), the observations were not retained for the terminology coverage. The inter-terminological coverage was defined as the ratio of the total number of concepts annotated via a terminology to the total number of concepts annotated by all terminologies. This metric

reflects the contribution of each terminology to the annotation process. Intra-terminological coverage is defined as the ratio of the number of unique concepts annotated via a terminology to the number of concepts constituting this terminology. Since the texts were in French, only the number of concepts available in French (translated or native) in the terminology considered were taken into account in the calculation. This metric reflects the level of use of the terminology. It is related to the number of concepts constituting the terminology, as well as its granularity.

3. Results

3.1. Evaluation of the quality of the annotation by MTCE

3.1.1. First evaluation

The analysed sample included 500 reasons for consulting, 500 observations and 500 consultation results.

For the reasons for consulting, 1,033 annotations were generated automatically. Of the annotations, 709 (68.6%) were considered valid, 179 (17.3%) false, 137 (13.3%) irrelevant and 8 (0.8%) to be verified. One hundred and sixty-three annotations were added manually, leading to a total of 1,196 concepts.

Concerning the observations, 5,714 annotations were verified. Of the annotations, 2,432 (42.6%) were considered valid, 1,384 (24.2%) false, 1,726 (30.2%) irrelevant and 172 (3.0%) to be verified. Four hundred and sixty-three were added manually, leading to a total of 6,177 concepts.

Concerning the consultation results, 1,126 annotations were evaluated automatically. Of the annotations, 821 (72.9%) were considered valid, 174 (15.5%) false, 102 (9.1%) irrelevant and 29 (2.6%) to be verified. Eighty-nine annotations were added manually, leading to a total of 1,215 concepts.

The mean values of precision, recall and F-measure were 0.76, 0.85 and 0.80, respectively. These results are summarized in Table 3.

	Precision		**Recall**		**F-measure**	
	1st evaluation	2nd evaluation	1st evaluation	2nd evaluation	1st evaluation	2nd evaluation
Reasons for consulting	0.80	0.82	0.81	0.76	0.81	0.79
Observations	0.64	0.80	0.84	0.84	0.72	0.82
Consultation results	0.83	0.92	0.90	0.89	0.86	0.90
Mean	**0.76**	**0.85**	**0.85**	**0.83**	**0.80**	**0.84**

Table 3: Results of the metrics of the 2 evaluations

The most frequent type of error was the context in the three corpuses with a mean of 30.3%, followed by stemming errors with a mean of 17.1%. One of the most common examples of

context errors was the annotation of the word “radio” as a means of communication and not as an X-ray. For stemming errors, they were related to the use of the root of the word for annotation, as in the case of the annotation of the word “grossesse” (pregnancy) by “grosse” (fat).

The error distribution in each corpus is presented in Table 4.

Type of errors	**Reasons for consulting** n (%)		**Observations** n (%)		**Consultation results** n (%)	
	1st Evaluation	2nd evaluation	1st Evaluation	2nd evaluation	1st Evaluation	2nd evaluation
Context	56 (31.3)	45 (26.9)	334 (24.1)	59 (13.5)	62 (35.6)	23 (27.4)
Stemming	32 (17.9)	23 (13.8)	241 (17.4)	42 (9.6)	28 (16.1)	15 (17.9)
Bags of words	15 (8.4)	24 (14.4)	236 (17.0)	85 (19.4)	9 (5.2)	16 (19.0)
Acronym	21 (11.7)	32 (19.2)	115 (8.3)	95 (21.7)	2 (1.1)	11 (13.1)
Empty words	4 (2.2)	7 (4.2)	196 (14.1)	83 (18.9)	2 (1.1)	3 (3.6)
Compound word	21 (11.7)	3 (1.8)	31 (2.2)	4 (0.9)	27 (15.5)	2 (2.4)
Negation	8 (4.5)	10 (6.0)	95 (6.9)	39 (8.9)	11 (6.3)	9 (10.7)
Wrong synonym	7 (3.9)	7 (4.2)	44 (3.2)	10 (2.3)	11 (6.3)	5 (6.0)
Adjective/ name	1 (0.6)	3 (1.8)	17 (1.2)	1 (0.2)	5 (2.9)	0 (0)
Probability	4 (2.2)	1 (0.6)	15 (1.1)	1 (0.2)	1 (0.6)	0 (0)
Abbreviation	3 (1.7)	0 (0)	21 (1.5)	9 (2.1)	0 (0)	0 (0)
Imprecision	1 (0.6)	1 (0.6)	9 (0.6)	1 (0.2)	11 (6.3)	0 (0)
Hypo or hyperonyme	0 (0)	4 (2.4)	3 (0.2)	0 (0)	4 (2.3)	0 (0)
Typing error	1 (0.6)	0 (0)	7 (0.5)	2 (0.5)	1 (0.6)	0 (0)
Unit error	0 (0)	2 (1.2)	17 (1.2)	0 (0)	0 (0)	0 (0)
Proper noun	3 (1.7)	3 (1.8)	0 (0)	2 (0.5)	0 (0)	0 (0)
Number	1 (0.6)	1 (0.6)	3 (0.2)	0 (0)	0 (0)	0 (0)
Others errors	1 (0.6)	1 (0.6)	2 (0.1)	5 (1.1)	0 (0)	0 (0)
Total	**179 (100)**	**167 (100)**	**1386 (100)**	**438 (100)**	**174 (100)**	**84 (100)**

Table 4: Error distribution in each corpus

3.1.2. Second evaluation

The analysed sample included 500 reasons for consulting, 500 consultation results and 250 observations. The samples were extracted from the same corpus as the first evaluation. Of the reasons for consulting, 1,063 annotations were generated automatically. Of the annotations, 754 (70.9%) were considered valid, 167 (15.7%) false, 133 (12.5%) irrelevant and 9 (0.8%) to be verified. Two hundred and thirty-eight annotations were added manually, leading to a total of 1,301 concepts.

Concerning the consultation results, 1,052 annotations were evaluated. Of the annotations, 901 (85.6%) were considered valid, 84 (8.0%) false, 59 (5.6%) irrelevant and 8 (0.8%) to be verified. One hundred and ten annotations were added manually, leading to a total of 1,162 concepts.

Concerning the observations, 2,753 annotations were verified. Of the annotations, 1,705 (61.9%) were considered valid, 438 (15.9%) false, 575 (20.9%) irrelevant and 35 (1.3%) to be verified. Three hundred and nineteen concepts were added manually, leading to a total of 3,072 concepts.

Finally, the number of manual annotations increased by an average of 3%.

The mean values of precision, recall and F-measure were 0.85, 0.83 and 0.84, respectively. These results are summarized in Table 3.

In the observations, NCIt concepts were found in 20% of the validated annotations and in 74% of the irrelevant or false annotations. For SNOMED CT concepts in the reasons for consultation, they were validated in 42% of the annotations and in 36% of irrelevant or false annotation. For the terminology Anatomical Therapeutic Chemical (ATC) and List of Products and Services (LPP), 100% of the annotated concepts were considered valid.

An increase in “bags-of-words” errors (+7.4%), acronyms (+11%) and empty words (+3.1%) was noted. Thus, the corrections reduced the number of false annotations by an average of 6% and irrelevant annotations by 5%. This was due to an overall decrease in context (-7.7%), stemming (-3.4%) and compound word errors (-8%).

The error distribution of each corpus is presented in Table 4.

CS and TL agreed for 78% of the annotations for the reasons for consulting subgroup, 71% for the observations and 80% for the consultation results subgroups.

3.2. Evaluation of terminology coverage

A total of 636,482 annotations were generated automatically by MTCE from 45,690 reasons for consulting and consultation results. The mean number of annotations per document was 13.9 ± 4.2 with a minimum of 2.4 and a maximum of 32.2. Regarding inter-terminological coverage, the most represented concepts were from SNOMED CT (14.7%), SNOMED Notion (10.4%) and NCIt (10%). The International Classification of Primary Care (ICPC) had the highest intra-terminological coverage (62.8%) then Dictionary of Consultation Results (DCR) (59.4%) and CISMeF Metaterms (57.3%). Conversely, the Foundational Model of Anatomy (FMA) was almost not represented (0.9%). Figure 1 summarizes the results of the inter- and intra-terminological coverage.

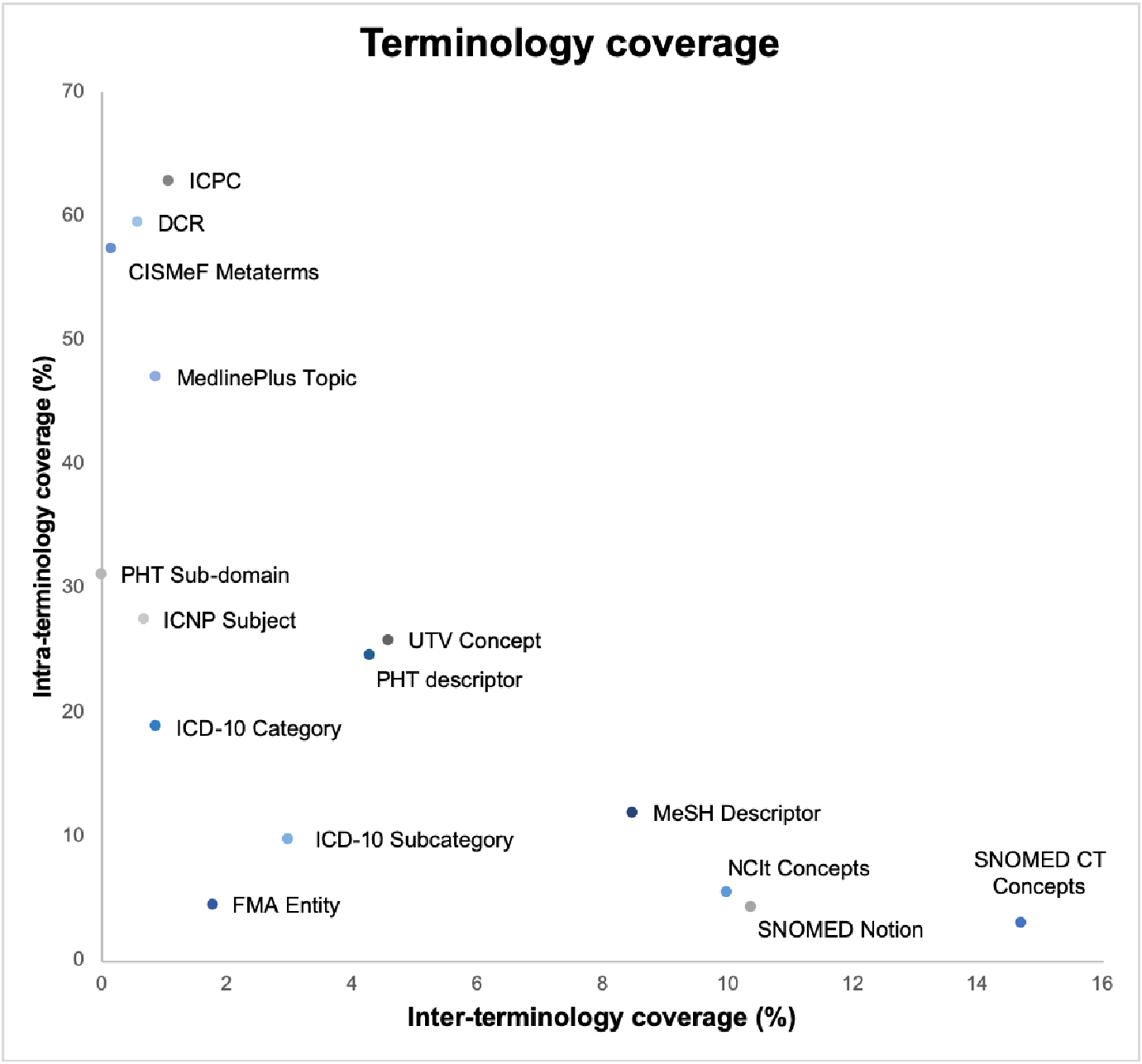

Figure 1: Results of the inter- and intra-terminolgy coverage

4. Discussion

4.1. Main results

With a F-measure of 0.83, 0.84 and 0.85 for the three groups in the second evaluation, MTCE seems to be a powerful tool for the automatic annotation of consultation data in family medicine.

At the time of the first assessment, the precision was essentially similar for the reasons for consulting and consultation results (0.80 and 0.83 respectively). Recall was better for consultation results due to a higher proportion of validated annotations, which is explained by a higher coding of consultation results by GPs in their routine practice (17). Precision and recall were lower for observations (0.72), due to the large amount of free text. Several studies have shown the impact of increasing words in documents which leads to decreasing precision (18,19).

The irrelevant annotations were mostly found in the observations, in particular by the presence of many terms such as "tablet", "during meals", etc.

The frequency of contextual errors in the three corpuses can be explained by the presence of common terms, more prevalent in family medicine consultations. The ambiguity and polysemy of some acronyms explain their frequency, particularly in observations and reasons for consulting.

Inherently, the correction process led to other errors. At the end of this correction phase, the overall performance of the consultation results and observation tool was improved. For the reasons for consulting, the slight increase in precision (0.80 and then 0.82) and the decrease in recall resulted in a moderate decrease in tool performance. This is due to the increase in the number of manual annotations and the increase in acronym errors.

4.2. Comparison with the literature

There have been a number of successes in various applications of biomedical Natural Language Processing (NLP) in English (20).

Several NLP methods have proven their effectiveness in analysing breast radiology reports such as deep-learning based, or Support Vector Machines for drug safety surveillance in English (21,22). These methods can be applied to other uses (retrieval information, diagnosis of speech pathologies and dementias, etc.) (23). The use of the MetaMap web service from the UMLS allows to efficiently extract meta-thesaurus concepts from text (24).

These tools are mainly based on the UMLS. This limits their application to our French-language database, French being poorly represented in the UMLS.

In 2015, MTCE was evaluated on titles from the MEDLINE database and drug description sheets from the European Medicines Agency (11). In 2016, the results obtained were improved thanks to a reduction in the number of terminologies used. The best results obtained were 0.77 for precision and 0.65 for recall (12). Our results were better for all metrics, since drug description sheets and MEDLINE articles are likely more heterogeneous than family medicine notes.

MTCE was also used on the documents included in Rouen University Hospital's data warehouse in 2018. The evaluation found a precision of 0.59, a recall of 0.68 and an F-measure of 0.63 after corrections (16). These results were not as good as these current results. The lack of structure in hospital reports and the large number of words explain these results (19). Contextual errors were more frequent in our corpus (26% versus 17%) due to the more generic terms used in family medicine.

Currently, the F-measure target reference found in the literature is around 0.68 for French, which is lower than the mean F-measure obtained in this study (0.80) (25).

Other tools are used in indexing. Studies have found F-measures at 0.87 and 0.91 for some of these tools (26,27). These tools are only available in English and are not adapted to

French texts due to the low representation of French in the UMLS (28). MTCE is one of the only tools adapted to documents written in French, using the HeTOP terminology server.

4.3. Terminology coverage

The most common terminologies were SNOMED CT, SNOMED int., NCIt and MeSH. They constituted more than half of the annotations. These terminologies were also the most common in the evaluation conducted on the LiSSa corpus and on Rouen University Hospital's database (13,16). The first French terminology was the Public Health Thesaurus (PHT), which represented 4.3% of all annotations. Only the part of the terminologies translated into French was used. These results highlight the importance of continuing efforts to translate terminologies to obtain reliable extraction from biomedical documents.

Intra-terminological coverage of ICPC was high (62.8%) as was that of DCR (59.4%). This result is consistent, as these classifications are specific to family medicine (29). CISMeF metaterms had a high coverage (57.3%). The concepts that make up this terminology are vast, such as the concept of "pain", "prevention" or "gynecology". They are therefore widely used in first-line consultations. In comparison with the study on Rouen University Hospital's data warehouse, the coverage of SNOMED CT was very low (3% versus 30.5%) as for FMA (4.4% versus 48.6%) (16). For FMA, although the coverage was low, the number of annotations was high with more than 740 concepts extracted.

The most common terminologies found were not always the most relevant. In addition, for classifications with the best intra-terminological coverage results, the majority of annotations were considered valid during the evaluations, unlike the results of the inter-terminological coverage.

Finally, given the difficulty of a terminology obtaining relevant results for both inter-terminological and intra-terminological coverage in family medicine data (Figure 1), it is necessary to make a choice when determining the terminology combination according to the expected results.

4.4. Strengths and limitations

The number of annotations evaluated was high, which reinforces the validity of our results. In addition, the data concerned 25 GPs working in different locations with different exercise modalities. The samples were randomly obtained, as was the choice of terminology, which limits selection bias. Given the large number of annotations, the random selection of one terminology did not seem to have any influence on the results of inter- and intra-terminological coverage.

The agreement between the two evaluators was good, which reinforces the validity of our results.

The evaluations were subjective, in particular to assess the relevance of the annotations. It would be interesting to define evaluation rules to limit this subjectivity bias.

The use of several terminologies increases concept recognition and thus increases recall. However, it also provides false and irrelevant concepts that reduce precision. The results obtained would certainly have been different if other terminologies had been used during the annotation process. Thus, it is necessary to ensure the best multi-terminological combination to be used to limit noise (or false positive).

Finally, random selection of a single terminology for each annotated concept can be a source of error. A concept can be evaluated as false for one defined terminology and as true for another and vice versa.

4.5. Perspectives

Based on this study, several solutions to improve automatic annotation by MTCE have been identified. Decreasing the number of words in a “bags-of-words” could limit errors related to it but at the risk of a decrease in recall by a decrease in recognition of long expressions.

It would be interesting to carry out a study to assess the specificity of each terminology and use the best multi-terminological combination for family medicine consultation data. In

addition, another study could evaluate MTCE in family medicine versus a gold standard to limit the subjectivity bias in evaluation.

Based on literature data, improved coding appears to be one of the challenges for improving family medicine research and interoperability (30–33).

In addition, the use of a single classification could also be an area for improvement (4). Since April 2018, the SNOMED CT classification has been used for coding primary care data in the United Kingdom (34). It appears as a pivotal terminology, making it possible to ensure semantic interoperability and thus facilitating a complete and reliable exchange of information (35,36). In France, efforts are necessary to support the development of a single terminology for primary care and for other medical and paramedical specialties.

5. Authors' contributions

All authors have made substantial contributions to the design of the study and to the acquisition of data. CS, MS and JG contributed to the analysis and interpretation of data. CS drafted the first version of the manuscript. All authors have approved the final version to be submitted.

6. Acknowledgements

This work was supported by the ANR ApiApps project, grant ANR-17-CE19-0027 of the French Agence Nationale de la Recherche.

This work was partially funded by the French region Normandy and the European Union (PlaiR2.018 project).

Europe acts in Normandy with the European Regional Development Fund (ERDF).

We are grateful to Nikki Sabourin-Gibbs, from Rouen University Hospital, for her help in editing the manuscript.

7. Statement of conflicts of interest

Conflicts of interest: none.

8. Summary Table

What was already known on the topic	What this study added to our knowledge
• Electronical medical records contain numerous data of great interest that need to be structured and standardized. • Coding remains underused by general practitioners. • Some tools allowing the automatic extraction of concepts from natural language documents have been validated on hospital data corpus.	• A Multi-Terminological Concepts Extractor is a powerful tool for automatic annotation of consultation data in family medicine in French. • Continuing efforts are necessary to translate terminologies to obtain reliable extraction from biomedical documents in any language.

9. References

1. The World Health Report - Primary Health Care Now More Than Ever. World Health Organization; 2008.

2. Weel C van, Rosser WW. Improving Health Care Globally: A Critical Review of the Necessity of Family Medicine Research and Recommendations to Build Research Capacity. Ann Fam Med. mai 2004;2(Suppl 2):s5.

3. Hajjar F, Saint-Lary O, Cadwallader J-S, Chauvin P, Boutet A, Steinecker M, et al. Development of Primary Care Research in North America, Europe, and Australia From 1974 to 2017. Ann Fam Med. janv 2019;17(1):49-51.

4. Chaudhry Z, Mannan F, Gibson-White A, Syed U, Ahmed S, Kousoulis A, et al. Outputs and Growth of Primary Care Databases in the United Kingdom: Bibliometric Analysis. J Innov Health Inform. 17 oct 2017;24(3):284-90.

5. Gentil M-L, Cuggia M, Fiquet L, Hagenbourger C, Le Berre T, Banâtre A, et al. Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature. BMC Med Inform Decis Mak. 25 sept 2017;17(1):139.

6. Kousoulis AA, Rafi I, de Lusignan S. The CPRD and the RCGP: building on research success by enhancing benefits for patients and practices. Br J Gen Pract J R Coll Gen Pract. févr 2015;65(631):54-5.

7. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. janv 2010;69(1):4-14.

8. Lacroix-Hugues V, Darmon D, Pradier C, Staccini P. Creation of the First French Database in Primary Care Using the ICPC2: Feasibility Study. Stud Health Technol Inform. 2017;245:462-6.

9. Dourgnon P, Grandis N, Sourty-Le Guellec M. Apport de l'informatique médicale dans la pratique médicale. Quest Déconomie Santé. 2000;(26):1-6.

10. de Lusignan S. The barriers to clinical coding in general practice: a literature review. Med Inform Internet Med. 2005;30(2):89-97.

11. Soualmia L, Cabot C, Dahamna B, Darmoni. SIBM at CLEF e-Health Evaluation Lab 2015. CEUR-WS Work Notes Conf Labs Eval Forum CLEF 2015. 2015;1391.

12. Cabot C, Soualmia L, Dahamna B, Darmoni S. SIBM at CLEF eHealth Evaluation Lab 2016a: Extracting Concepts in French Medical Texts with ECMT and CIMIND. Orking Notes Conf Labs Eval Forum CLEF 2016. 2016;1609:47-60.

13. Cabot C, Soualmia LF, Grosjean J, Griffon N, Darmoni SJ. Evaluation of the Terminology Coverage in the French Corpus LiSSa. Stud Health Technol Inform. 2017;235:126-30.

14. Sakji S, Gicquel Q, Pereira S, Kergourlay I, Proux D, Darmoni S, et al. Evaluation of a French medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. Stud Health Technol Inform. 2010;160(Pt 1):252-6.

15. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia L, et al. Health Multi-Terminology Portal: a semantics added-value for patient safety. Patient Safety Informatics - Adverse Drug Events , Human Factors and IT Tools for patient Medication Safety. Stud Health Technol Inform. 2011;166:129-38.

16. Ndangang M, Grosjean J, Lelong R, Dahamna B, Kergourlay I, Griffon N, et al. Terminology Coverage from Semantic Annotated Health Documents. Stud Health Technol Inform. 2018;255:20-4.

17. Déborah BUSIDAN, Cittee J, Dumay C. Pratiques du codage des données de consultation par les médecins généralistes des structures pluriprofessionnelles de soins primaires en Ile-de-France : une enquête exploratoire en 2012. Exercer. 2013;(110(suppl3)):92-3.

18. Chebil W, Soualmia LF, Dahamna B, Darmoni SJ. Automatic indexing of health documents in French: Evaluating and analysing errors. Innov Res Biomed Eng. 2012;33:316-29.

19. Hasan KS, Ng V. Automatic Keyphrase Extraction: A Survey of the State of the Art. In Baltimore, Maryland: Association for Computational Linguistics; 2014. p. 1262–73.

20. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semant. 30 mars 2018;9(1):12.

21. Miao S, Xu T, Wu Y, Xie H, Wang J, Jing S, et al. Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches. Int J Med Inf. 2018;119:17-21.

22. Munkhdalai T, Liu F, Yu H. Clinical Relation Extraction Toward Drug Safety Surveillance Using Electronic Health Record Narratives: Classical Learning Versus Deep Learning. JMIR Public Health Surveill. 25 avr 2018;4(2):e29.

23. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. J Biomed Inform. 2018;88:11-9.

24. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 1 janv 2004;32(Database issue):D267-70.

25. Goeuriot L, Kelly L, Suominen H, Névéol A, Robert A, Kanoulas E, et al. CLEF 2017 eHealth Evaluation Lab Overview. In: Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, et al., éditeurs. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer International Publishing; 2017. p. 291-303.

26. Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case reports. J Biomed Semant. 20 déc 2012;3:15.

27. Jonnalagadda S, Cohen T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. J Biomed Inform. févr 2012;45(1):129-40.

28. Névéol A, Grosjean J, Darmoni S, Zweigenbaum P. Language Resources for French in the Biomedical Domain. In 2014.

29. World Organization of National Colleges, Academies, and Academic Associations of General Practitioners/Family Physicians, éditeur. ICPC-2-R: international classification of primary care. Rev. 2nd ed. Oxford ; New York: Oxford University Press; 2005. 193 p. (Oxford medical publications).

30. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Fam Pract. 2006;23(2):253-63.

31. Grobbee DE, Hoes AW, Verheij TJM, Schrijvers AJP, van Ameijden EJC, Numans ME. The Utrecht Health Project: optimization of routine healthcare data for research. Eur J Epidemiol. 2005;20(3):285-7.

32. Van Der Bij S, Khan N, ten Veen P, Bakker D, H D, Verheij RA. Improving the quality of EHR recording in primary care: a data quality feedback tool. J Am Med Inform Assoc. 1 janv 2017;24(1):81-7.

33. Darmon D, Sauvant R, Staccini P, Letrilliart L. Which functionalities are available in the electronic health record systems used by French general practitioners? An assessment study of 15 systems. Int J Med Inf. 1 janv 2014;83(1):37-46.

34. SNOMED CT implementation in primary care [Internet]. NHS Digital. [cité 9 oct 2018]. Disponible sur: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct/snomed-ct-implementation-in-primary-care

35. El-Sappagh S, Franda F, Ali F, Kwak K-S. SNOMED CT standard ontology based on the ontology for general medical science. BMC Med Inform Decis Mak. 31 août 2018;18(1):76.

36. Chniti A, Traore L, Hussain S, Griffon N, Darmoni S. A semantic interoperability framework for Facilitating Cross-hospital exchanges. Stud Health Technol Inform. 2014;205:1255.