Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210836

Are Semantic Annotators Able to Extract Relevant Complexity-Related Concepts from Clinical Notes?

Akram REDJDAL[a,1], Jacques BOUAUD[a,b], Joseph GLIGOROV[c,d] and Brigitte SÉROUSSI[a,d,e]
[a] *Sorbonne Université, Université Sorbonne Paris* Nord*, Inserm, UMRS_1142, LIMICS, Paris, France*
[b] *AP-HP, DRCI, Paris, France*
[c]*Sorbonne Université, Institut Universitaire de Cancérologie, Paris, France*
[d]*AP-HP, Hôpital Tenon, Paris, France*
[e]*APREC, Paris, France*

Abstract. Clinical decision support systems (CDSSs) implementing cancer clinical practice guidelines (CPGs) have the potential to improve the compliance of decisions made by multidisciplinary tumor boards (MTB) with CPGs. However, guideline-based CDSSs do not cover complex cases and need time for discussion. We propose to learn how to predict complex cancer cases prior to MTBs from breast cancer patient summaries (BCPSs) resuming clinical notes. BCPSs being unstructured natural language textual documents, we implemented four semantic annotators (ECMT, SIFR, cTAKES, and MetaMap) to assess whether complexity-related concepts could be extracted from clinical notes. On a sample of 24 BCPSs covering 35 complexity reasons, ECMT and MetaMap were the most efficient systems with a performance rate of 60% (21/35) and 49% (17/35), respectively. When using the four annotators in sequence, 69% of complexity reasons were extracted (24/35 reasons).

Keywords. Information Extraction, Decision Support, Breast Cancer

1. Introduction

In many countries, the treatment of cancer patients must be decided in multidisciplinary tumor boards (MTBs). These meetings have been introduced to provide a collaborative and multidisciplinary approach to cancer care, bringing together surgery, oncology, radiology, and pathology specialists to optimize the decision-making process. Prior to MTBs, physicians in charge of patients whose cases will be discussed prepare a breast cancer patient summary (BCPS) as the basis of the oral presentation of the patient case to all MTB clinicians. However, the benefits of MTBs, which have long been taken for granted, are recently being challenged. Positive outcomes from MTBs depend on the presence of qualified and effective faculty, good preparation of patient cases, efficient leadership, sound discussions, and contributive interactions among MTB clinicians [1].

[1] Corresponding author, Akram REDJDAL, Email: redjdalakram300@gmail.com

Clinical decision support systems (CDSSs) are software components that aim to support clinicians in their decision-making process. CDSSs have proven to increase the compliance of clinician decisions with clinical practice guidelines (CPGs) [2]. DESIREE is a European project which aimed at developing web-based services for the management of primary breast cancer by MTBs. During the evaluation of the guideline-based CDSS of DESIREE, we found that for some patient cases the system did not provide any therapeutic proposals or gave recommendations that were not followed by MTB clinicians [3]. These clinical cases were considered as "complex cases", and we made the assumption that such cases were not correctly handled by guideline-based CDSSs. In the perspective of ultimately building a CDSS able to distinctly support therapeutic decision for complex and non-complex breast cancer cases, the first issue is to identify complex breast cancer cases.

Replicating the mode of operation of MTBs, the aim is to use BCPSs to predict complexity. The first step is to check whether BCPSs do embed complexity-related concepts. As BCPSs are expressed as natural language clinical, non-structured notes, we used different annotators and compared the annotations automatically generated to the reasons of complexity established by a group of clinicians on a sample of BCPSs[2].

2. Material and Methods

2.1. Breast cancer patient summaries

We worked on a sample of 24 BCPSs available as textual unstructured documents. They provide a portrait of patients with all relevant information that MTB clinicians need to know to make the best patient-specific therapeutic decision. BCPSs contain different types of information: reason for presentation, type of tumor, biometric data, personal history, family history, TNM classification, etc. However, unstructured formats make information extraction complicated (e.g., there are many abbreviations, acronyms, and specialized terms.). These 24 BCPSs were manually annotated as "complex" or "non-complex" by a group of seven MTB clinicians of different levels of expertise (from junior to senior) and from different domains (5 oncologists, 2 surgeons). When a clinician considered a clinical case was "complex", (s)he had to explain why and give the reason of the complexity in terms of patient characteristics.

2.2. Annotation tools: ECMT, SIFR, cTAKES, and MetaMap

We implemented four automatic semantic annotators to extract data from BCPSs. Currently, two systems are widely used in the biomedical field for the English language [4], MetaMap and cTAKES. Since we work on a corpus of French BCPSs, we also considered two systems that work for the French language, i.e., ECMT and SIFR [5].

- MetaMap was developed by the National Library of Medicine (NLM) to map biomedical text to concepts in the Unified Medical Language System (UMLS).

[2] This work has been financed by a doctoral grant for AR from the University Institute of Health Engineering (IUIS, Sorbonne University, Paris, France) and received the support of AP-HP health data warehouse.

The tool uses a hybrid approach combining natural language processing (NLP), knowledge-intensive approach, and computational linguistic techniques.

- cTAKES for Clinical Text Analysis and Knowledge Extraction System uses rule-based and machine learning to extract information from clinical text.
- ECMT (*Extracteur de Concepts Multi-Terminologique* http://ecmt.chu-rouen.fr) is a webservice inspired by the CISMef algorithm for information retrieval with Doc'CISMeF. ECMT works for the French language with seven terminologies and supports semantic expansion features.
- SIFR for Semantic Indexing of French Biomedical Data Resources (http://bioportal.lirmm.fr/annotator) annotator is an openly available web service enabling both recognition and contextualization of concepts from 30 medical terminologies and ontologies.

2.3. Pre-treatment of clinical notes

As cTAKES and MetaMap work on English notes, we translated BCPSs from French to English. However, BCPSs contain a lot of acronyms related to the oncological field (e.g., "*HTA*", "*IRM*", "*TEP*"), difficult to translate with a translator. To solve this issue, we created a local dictionary with medical acronyms and their definition based on online available dictionaries. Then, we replaced acronyms in BCPSs by their definition to get a "translatable" text. We finally used a pre-trained Opus-MT translation model. As a result, all BCPSs were available in French and English in textual format (.txt) used as input by the four annotators. For each system, concepts, CUIs (if available), negation, and certainty were extracted. With ECMT, we used the labels of extracted terms to extract CUIs, but we didn't have information about the context (negativity and certainty) [6].

2.4. Evaluation of annotators

From the corpus of BCPSs, we considered that a BCPS described a complex case if it was considered as complex by *at least* one of the seven MTB clinicians. For each of the complex BCPSs, we collected the list of reasons of complexity as provided by MTB clinicians, and we manually checked whether each element of the list was present in the list of extracted annotations.

3. Results

Among the 24 BCPSs, 14 were considered as complex cases, with seven considered as complex *by all* MTB clinicians. We got 35 reasons of complexity. ECMT and MetaMap were the most efficient systems in terms of complexity parameters extraction, ECMT extracted 60% (21/35) of complexity reasons and MetaMap 49% (17/35). SIFR identified 11 complexity parameters (31%) and cTAKES was the less efficient annotator with only 7 parameters (20%). When using the four annotators in sequence, 24 out of the 35 complexity reasons were extracted (69%). Table 1 shows for each BCPS the reasons of complexity and by which annotator they were retrieved.

Table 1. Evaluation of the four annotators on MTB-clinician-provided complexity-related concepts

BCPS	# MTB clinicians	Reason of complexity	ECMT	SIFR	cTAKES	MetaMap
1	7	Pregnancy	yes	yes	yes	yes
		Patient preference (Refusal of recommended treatment)	no	no	no	yes
		Social situation	yes	no	no	yes
2	7	Radio chemotherapy before surgery	yes	yes	no	yes
		No response to standard treatment	yes	no	no	yes
		Inflammatory syndrome	yes	yes	yes	yes
3	7	Patient preference (Refusal of recommended treatment)	yes	no	no	yes
		Incomplete histology	no	no	no	no
4	7	Comorbidities (age, obesity)	yes	no	no	yes
		Incomplete record	no	no	no	no
		Inadequate margins of excision	yes	yes	no	no
		Use of Oncotype DX	yes	no	no	no
5	7	Comorbidities (age)	yes	yes	yes	yes
		Double cancer	yes	no	no	yes
		Polymedication	yes	yes	yes	yes
6	7	Complex surgical decision	no	no	no	no
7	7	Rare situation	no	no	no	no
		Comorbidities (type 2 diabetes)	yes	yes	yes	yes
		Unclear history of the disease	no	no	no	no
8	6	Prophylactic situation	yes	yes	no	yes
		Family antecedents of breast cancer	yes	no	no	no
		Multifocal cancer	no	no	no	no
9	5	Use of Oncotype DX	yes	no	no	no
		Malignancy	yes	yes	yes	yes
10	5	Incomplete record	no	no	no	no
		Use of Oncotype DX	yes	no	no	no
11	3	Complex surgical decision	no	no	no	no
		Complex surgical decision	yes	no	no	yes
		Multiple imaging procedures needed	no	no	no	no
12	3	Use of Oncotype DX	yes	no	no	no
		Discrepancies between ultrasound and MRI	no	no	no	no
		Multiple metastatic lymph nodes and malignancy	yes	yes	no	yes
13	3	Discrepancies between biopsy and excised tissues	no	no	no	no
		Comorbidities (age)	yes	yes	yes	yes
14	2	Patient preference (Refusal of recommended treatment)	no	no	no	no

4. Discussion and conclusion

We implemented four annotators to assess whether they were able to extract relevant complexity-related concepts from BCPSs. All systems are efficient to extract clear medical concepts (pregnancy, inflammatory syndrome, etc.). ECMT and MetaMap were the most efficient systems as they extracted six parameters that were not extracted by SIFR and cTAKES. ECMT was able to identify two parameters ("Use of Oncotype DX" and "Family antecedents of breast cancer") that were not identified by the other annotators, which can be explained by the fact that ECMT is linked to terminologies that contain these concepts. MetaMap was able to detect one parameter related to patient preference ("Refusal of recommended treatment") that was not extracted by the other annotators. However, this parameter was present in two BCPSs and MetaMap only extracted it once. Three parameters were specifically not extracted by cTAKES, which can be explained by the fact that we used the default clinical pipeline of cTAKES. Indeed, studies reported that other pipelines used for extracting cancer-related information showed good results [7]. It is noticeable that one parameter was only extracted by French annotators ("Inadequate margins of excision"), which may be due to a translation problem. Complexity-related concepts not found by the annotators are context or patient-related parameters, e.g., "Refusal of the recommended treatment", "Complex surgical decision", "Discrepancies between ultrasound and MRI". These parameters are interpreted by clinicians during MTBs but are not explicitly written in BCPSs.

Annotation of BCPSs is time-consuming and labor-intensive for MTB clinicians and automatic semantic annotators when used in sequence may help extracting complexity-related structured concepts from non-structured BCPSs. This would allow us to train machine learning algorithms from automatically generated annotations to categorize complex and non-complex cases ahead of MTBs.

References

[1] El Saghir NS, Keating NL, Carlson RW, Khoury KE, Fallowfield L. Tumor boards: optimizing the structure and improving efficiency of multidisci-plinary management of patients with cancer worldwide. Am Soc Clin Oncol Educ Book. 2014:e461-6. doi: 10.14694/EdBook_AM.2014.34.e461. PMID: 24857140.

[2] Bouaud J, Séroussi B, Antoine EC, et al. A before-after study using OncoDoc, a guideline-based decision support-system on breast cancer management: impact upon physician prescribing behaviour Stud Health Technol Inform. 2001;84:420–424.

[3] Redjdal A, Bouaud J, Guézennec G, Gligorov J, Seroussi B. Reusing Decisions Made with One Decision Support System to Assess a Second Decision Support System: Introducing the Notion of Complex Cases. Stud Health Technol Inform. 2021 May 27;281:649-653. doi: 10.3233/SHTI210251. PMID: 34042656.

[4] Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Med Inform Decis Mak. 2018 Sep 14;18(Suppl 3):74. doi: 10.1186/s12911-018-0654-2. PMID: 30255810; PMCID: PMC6157281.

[5] Sakji S, Gicquel Q, Pereira S, Kergourlay I, Proux D, Darmoni S, Metzger MH. Evaluation of a French medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. InMEDINFO 2010 2010 (pp. 252-256). IOS Press.

[6] Redjdal A, Bouaud J, Guézennec G, Gligorov J, Seroussi B. Comparison of MetaMap, cTAKES, SIFR, and ECMT to Annotate Breast Cancer Patient Summaries. Stud Health Technol Inform. 2021, October 2-4, To appear.

[7] Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, Harris D, Hochheiser H, Lin C, Chavan G, Jacobson RS. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. Cancer Res. 2017 Nov 1;77(21):e115-e118. doi: 10.1158/0008-5472.CAN-17-0615. PMID: 29092954; PMCID: PMC5690492.