Development of an Automated Detection Tool for Healthcare-Associated Infections Based on Screening Natural Language Medical Reports

Marie-Hélène Metzger, MD, PhD[1], Quentin Gicquel, MS[1], Denys Proux, PhD[2], Suzanne Pereira, PhD[3], Ivan Kergourlay, MS[4], Elisabeth Serrot, PhD[3], Frédérique Segond, PhD[2], Stephan Darmoni, MD,PhD[4]

[1]Université Lyon 1 - CNRS UMR 5558, LBBE, Lyon, France [2] XRCE, Meylan, France [3] Vidal, Issy-les-Moulineaux, France
[4] CISMeF, LITIS EA 4108 University of Rouen, France

Introduction

Surveillance of Healthcare-Associated Infections (HAI) is an important activity in the context of control and prevention. However, some issues were addressed concerning the workload and costs generated by this surveillance. Alternative methods based on automation of detection procedures were experimented in different facilities. In this context, the electronic health record (EHR) is a unique opportunity for Infection Control Practitioners (ICP) to automate manual processes. Few experiences of applying text mining techniques for monitoring adverse events are reported in the literature.
The objective of the ALADIN Project is to develop an automated HAI detection tool based on screening French natural language documents of the EHR.

Material and Methods

I First step: development of decision rules for detection of HAI

I.1. Anonymisation of medical reports

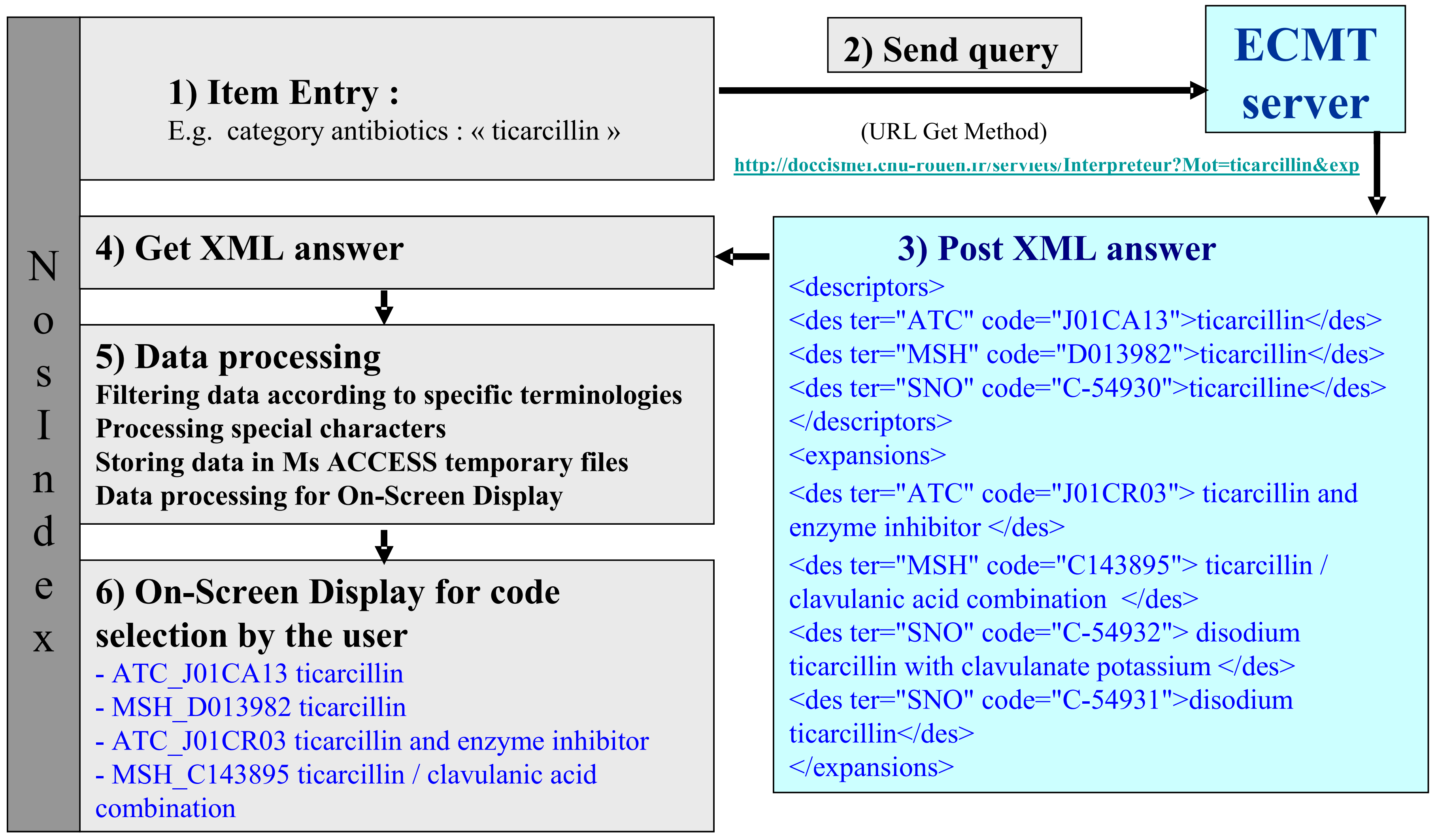
For this project, 2000 medical reports will be extracted from the hospital information system of 4 French University hospitals (Lille, Lyon, Nice, Rouen). Once documents have been retrieved from local databases, they will be anonymized semi-automatically by using the Xerox Incremental Parser (XIP).
XIP has been developed for the last ten years by the Xerox Research Centre Europe. Its combines five linguistic processing layers which are:

- preprocessing (tokenization, morphological analyzer and part of speech tagging),
- named entities,
- chunking,
- dependency extractions between words on the basis of sub-tree patterns over chunk sequences, and
- a combination of those dependencies with boolean operators to generate new dependencies or to modify or delete existing dependencies.

XIP comprises an engine and a metalanguage that allows users to write grammar rules or add words in the lexicon. XIP takes as input any text in plain text format or in XML and gives as output an XML file in which all information that has been found is encoded. This technology has been integrated in a JAVA application which allows to remove all names of persons, localisation, addresses, phone, fax numbers, e-mail addresses and to convert all dates in duration since the date of ward admission. Because a perfect confidentiality is necessary, this process is semi-automatic, the ICP may change terms identifiers that were not detected by the tool**.** The recall and precision of this tool before manual control by ICPs will be soon evaluated.

I.2. Manual annotation of medical reports

The development of decision rules is based on manual annotation of medical reports by ICPs of the four University hospitals. For this step, 600 medical reports reporting HAI and 600 without any report of HAI will be analyzed. These documents will focus on surgical activity (digestive, neurosurgery and orthopaedics) and intensive care units. The ICPs will use for this step a French medical multi-terminology indexing tool (French acronym : ECMT), developed by the CISMeF team. Some of them are integrated in the UMLS metathesaurus; some are not because they are French terminologies (e.g. CCAM, DRC, Orphanet). These main terminologies have their respective objectives : the MeSH is devoted to documentation, SNOMED to describe patient record, ICD10 to epidemiology, CCAM to procedures, ATC to drugs, ICPC2 and DRC to general (or family) medicine, ICF to handicap.
This manual annotation will provide the correspondence between terms used in current medical language and not directly available in standardised terminologies and also it will provide standardised data for building decision rules. For helping ICPs to use these different medical terminologies with which they are not familiar, a MS Access software application (namely, NosIndex) was developed by the CNRS-UMR 5558 team. The figure 1 summarizes the process and exchanges between NosIndex and ECMT. The application exports the request to the ECMT tool and imports the XML file containing all the corresponding codes. The process lasts 2 to 3 seconds before the ICP can visualize all the proposed codes. Then, the ICP can select the most appropriate one. If he does not find any satisfying code, the term will not be coded by the ICP and the term will be in a second step analysed by the semantic experts for decision. A structured questionnaire was developed on the application in order to classify all the medical terms relevant for building detection algorithms of HAI. The ICPs will also enter their conclusion regarding the outcome "suspicion of HAI or not". The manual annotation of each medical report will be conducted independently by two different ICPs. In case of annotation discordance, the two ICPs will meet for consensus procedure. The following categories of terms will be coded: symptoms/diagnosis, bacteriological exams, type of microorganism, biological exams, radiological exams, antibiotics, type of surgical intervention. In order to limit the noise generated by the use of the 9 terminologies, the application filters the results provided by ECMT, depending on the category of medical terms: symptoms/diagnosis: CIM10, SNOMED3.5, MeSH; bacteriological exams : SNOMED3.5, MeSH; type of microorganisms : SNOMED3.5; biological exams : SNOMED3.5, MeSH; radiological exams : SNOMED 3.5, MeSH, CCAM; Antibiotics : ATC, MeSH; Type of surgical intervention: CCAM, MeSH. These filters were chosen manually by the expertise of the ALADIN project members.


N o s I n d e x
1) Item Entry :
E.g. category antibiotics : « ticarcillin »
2) Send query
ECMT server
(URL Get Method)
http://doccismef.chu-rouen.fr/servlets/Interpreteur?Mot=ticarcillin&exp
3) Post XML answer
4) Get XML answer
5) Data processing
Filtering data according to specific terminologies
Processing special characters
Storing data in Ms ACCESS temporary files
Data processing for On-Screen Display
6) On-Screen Display for code selection by the user

```
<descriptors>
<des ter="ATC" code="J01CA13">ticarcillin</des>
<des ter="MSH" code="D013982">ticarcillin</des>
<des ter="SNO" code="C-54930">ticarcilline</des>
</descriptors>
<expansions>
<des ter="ATC" code="J01CR03"> ticarcillin and enzyme inhibitor </des>
<des ter="MSH" code="C143895"> ticarcillin / clavulanic acid combination  </des>
<des ter="SNO" code="C-54932"> disodium ticarcillin with clavulanate potassium </des>
<des ter="SNO" code="C-54931">disodium ticarcillin</des>
</expansions>
```

- ATC_J01CA13 ticarcillin
- MSH_D013982 ticarcillin
- ATC_J01CR03 ticarcillin and enzyme inhibitor
- MSH_C143895 ticarcillin / clavulanic acid combination

Figure 1 : Process and exchanges between NosIndex and ECMT server

This categorization of medical terms was chosen in order to facilitate the expert work between ICPs and linguists for formalizing decision rules regarding HAI suspicion or not when reading the medical report.

II Second step: development of the detection tool

The decision rules defined by common work between ICPs and linguists will then be converted into parsing rules that can be processed by the XIP text parser. The XIP will also use the same multi-terminology approach as this exposed for the manual annotation but by adapting for the project a multi-terminology automatic indexer (F-MTI) developed by CISMeF and Vidal Company. The detection tool should then be able to determine the HAI likelihood according to the elements extracted from text that match with decision rules. The following example illustrates the automated process from extraction to HAI detection which is expected with this detection tool:

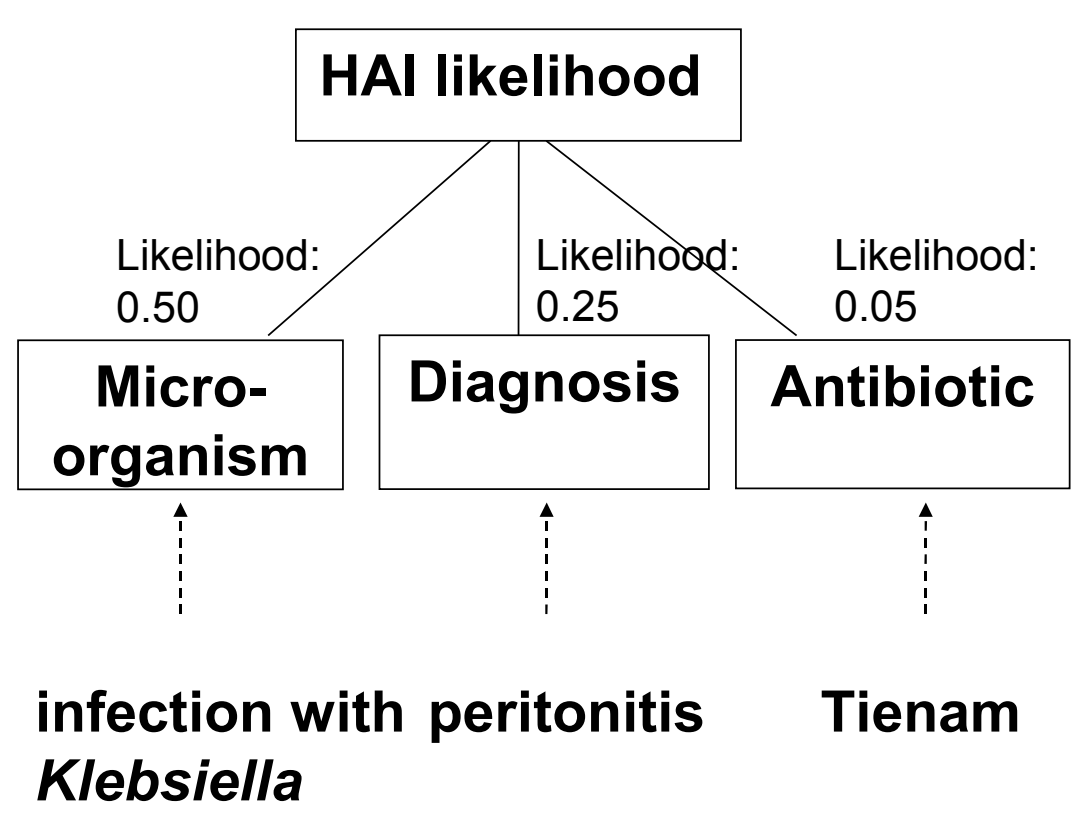
Automated Medical report screening

"The postoperative consequences were marked by abdominal pain and fever, associated with an hyperleucocytosis (53000/mm3) and C-Reactive Protein at 392 mg/l. It was due to multiple intra-peritoneal abscesses and peritonitis without anastomotic dehiscence that required a peritoneal toilet. It was an infection with Klebsiella only sensitive to Tienam…"

Detected facts by XIP/F-MTI

Symptoms/diagnosis
"abdominal pain"; "fever"
"multiple intra-peritoneal abscesses";
"peritonitis";
Biological exams
"hyperleucocytosis (53000 mm3";
"C-Reactive Protein at 392 mg/l)"
Microorganisms
"Infection with Klebsiella"
Antibiotics : '"Tienam"

Detection rule



III Third step: evaluation of the performances of the detection tool

Sensitivity and specificity of the detection tool will be evaluated by using the manual medical analysis as gold standard. For this step, new medical reports will be manually annotated by the ICPs : 400 HAI reports with HAI and 400 HAI reports without HAI.

Conclusion

Healthcare-Associated Infections are an important issue for public health. Epidemiological surveillance and alert system are important methods for the prevention and control of these adverse events. This project aims at developing a detection tool by applying Natural Language Processing Techniques in order to mine medical reports and identify HAI. These type of methods when successful could complete other automated methods of surveillance actually experimented in hospitals and based on the exploitation of bacteriological and antibiotics prescription databases.
Depending on the results, this tool could be used extensively to the detection of other adverse events.

Acknowledgment

This work has been supported by French National Research Agency (ANR) through TecSan program (project ALADIN-DTH n° ANR-08-TECS-001)

Contact : Dr Marie-Hélène Metzger - Hospices Civils de Lyon, SHEP, Hôpital Henry Gabrielle, 20 Route de Vourles BP 57 69 565 Saint-Genis Laval cedex, FRANCE
Tel : 00 33 478 86 49 32 E-mail : marie-helene.metzger@chu-lyon.fr