

Addressing Hospital Acquired Infection Control through Risk Patterns Detection in Medical Reports

Denys Proux, Frédérique Segond, Solweig Gerbier, Marie Hélène Metzger

Abstract—Hospital Acquired Infections (HAI) is a real burden for doctors and risk surveillance experts. The impact on patients' health and related healthcare cost is very significant and a major concern even for rich countries. Furthermore required data to evaluate the threat is generally not available to experts and that prevents from fast reaction. However, recent advances in Computational Intelligence Techniques such as Information Extraction, Risk Patterns Detection in documents and Decision Support Systems allow now to address this problem.

I. INTRODUCTION

PATIENTS' security is a key issue in hospitals. For example, in France, the incidence of adverse events was estimated [1] to 6.6 per 1000 hospital days in 2004, from which 24.1% were Hospital Acquired Infections (HAI). The prevention of HAI is a real challenge and is the responsibility of all health care professionals. Specific prevention programs were developed in most of the European countries, including involvement of Infection Control Teams promoting prevention guidelines, control practices and implementing surveillance systems based on national standards.

Surveillance systems of adverse events are key elements for prevention as it has been demonstrated by various studies ([2], [3], [4] and [5]). An efficient surveillance system should meet several criteria: it should encompass clear definition of targeted infections, be able to detect and react in a very timely effective manner, be sensitive enough to detect small variations in the occurrence rate and should not require too much effort and time investment from the medical staff which is already overworked. Such as system should also be able to take into account a collection of various data dealing with patient's risk factors (morbidity, invasive devices, surgical procedure...). These data have to be gathered from patient's medical reports to be recorded on specific standardized forms for further analysis. However the general workload of the medical staff is so high, with respect

to patient's care (which is the priority), that tracking is often seen as a burden. Furthermore the organization of hospital information systems does not help collecting this information. Medical record system heterogeneity is a nightmare for automatic data extraction. This situation is almost the same in every country and so no matter the state of development of the health care system.

Expertise gained over the last years in Computational Intelligence and more specifically in Risk Patterns detection from the literature allows now to address this problem. The detection of specific combinations of events and underlining relations between symptoms, treatments, drugs, reactions, and biological parameters can allow automatic systems to identify potential adverse events. Alerts could then be sent to risk management teams to help them identifying events that require immediate action and correction measures.

The following paper describes a project aiming to detect HAI by using risk patterns identification methods in textual medical records. The goal is to provide appropriate state of the art technologies included in a global process involving synergies between medical and technical experts to reduce the number of unnoticed cases and shorten time for reaction. The ultimate objective is to reduce the impact of HAI on patients' health as well as to reduce the financial cost for health care systems. To do so Natural Language Processing Techniques, such as a full-fledged robust grammar, will be applied to identify specific terms and sequences of facts in Patient Discharge Summaries. The objective will be to pinpoint drug names, bacteria names, symptoms, prescription and so on, appearing in texts according to a specific configuration that could be related to risk patterns.

II. HOSPITAL ACQUIRED INFECTION

A. Current Status

A Hospital Acquired Infection can be defined as: *An infection occurring in a patient in a hospital or other health care facility in whom the infection was not present or incubating at the time of admission. This includes infections acquired in the hospital but appearing after discharge, and also occupational infections among staff of the facility. If the exact status of the patient is not clearly known when he first came in a medical unit, a period of 48 hours (or superior to the incubation period if it is known) is considered to separate HAI from other kinds of infections coming from outside. As for infections related to surgery a period of 30 days is considered and extended to 12 months in case of*

Denys Proux is with the Xerox Research Centre Europe, France (+33 476 615 107; Denys.Proux@xrce.xerox.com)

Frédérique Segond is with the Xerox Research Centre Europe, France (Frederique.Segond@xrce.xerox.com)

Solweig Gerbier is with the Service d'Hygiène, Epidémiologie et Prévention des Hospices Civils de Lyon, France (solweig.gerbier@chu-lyon.fr)

Marie Hélène Metzger is with the Service d'Hygiène, Epidémiologie et Prévention des Hospices Civils de Lyon, France (marie-helene.metzger@chu-lyon.fr)

implanted device. [6]

Studies [7] show that in Europe, the frequencies of HAI hit 5 to 10% of hospitalized patients. In the extended European Union, there are approximately 3 million identified cases and 50,000 related deaths per year. Mortality related to HAI was estimated in 2005 to 4000 death per year in France and 20 to 30 per cent of these deaths are estimated avoidable by adapted prevention guidelines.

HAI related costs are greatly dependent on the type of infection and on the patient's risk factors. The costs associated to HAI ranged from 500 € for an urinaryinfection up to 40 000 € for a serious bacteraemia in Intensive Care Units (ICU) [8]. In France, estimating that the number of HAI by year is 750 000 and that the preventable part of these cases is 30%, the overcharge for the health system may represent a total cost between 0.11 and 9 billion €. By aiming to reduce the HAI incidence to 10%, the costs savings may be 240 to 600 million € per year, which represent 3 to 6 times the annual budget allocated to HAI prevention in France

B. National and International Organization of HAI surveillance systems

Hospital Hygiene and Infection Control teams encompass the whole health care system and include many players and address many activities in microbiology, quality improvement, risks assessment, clinical audit for nursing and technology assessment. Infection control is based on a multi-disciplinary approach of the problem. Progresses are possible only when all these players collect and share their efforts and experiences [9], [10], [11].

Each European country has its own organization in terms of infection control, antibiotic resistance control, surveillance organization, prevention (advising, auditing, quality improvement) and education. Herewith, organizations involved are mentioned for each EU medical participating country [12], [13], [14], [15].

Surveillance systems are an important part of prevention tools because they allow to quantify the HAI impact, to propose corrective actions and to evaluate the efficiency of these actions.

At the European level, the HELICS group (HELICS: *Hospital In Europe Link for Infection Control through Surveillance*) developed a standardized protocol for HAI monitoring, which allows to collect data from different European countries. Present challenges for the HELICS program are to organize the routine production and dissemination of analysis, to improve tools for collection, quality control, transfer and analysis of data. The review and improvement of data collection methods at a local level should be carried out in the perspective of reducing workloads and improving data quality. Surveillance based on automatic computer systems represents certainly the future for hospitals. Different studies have demonstrated the efficiency of such systems based on the combination of bacteriological data, antibiotic exposure or discharge diagnoses [16] [17]. Several studies have shown a significant

impact on workload reduction when surveillance is based on these automatic computer systems [18], [19].

C. Remaining Problems

However, despite this organization and processes, HAI control is still far from being efficient. At a local level, a large part of the difficulty related to monitoring these events comes from the fact that required information is disseminated through different departments.

Communication between medical units has yet to be improved and technology can help. Data transfer (when a patient moves from one department to another), is still often done using paper based documents. This for example increases the risk of loosing these documents or mixing them with others.

This lack of efficiency and uniformity in documents workflow control is a problem for teams trying to monitor and prevent risks. There is no national guideline or obligation whatsoever to standardize recording systems neither in private clinics nor in public hospitals. However information exists and is organized in a way or another. Main clinical events occurring during a patient's hospitalization are recorded in the patient's record by the medical staff. A summary is written at the hospital discharge which is the main source of communication between the various medical units. But today there is no real standardized summary of this information and medical staff can write this discharge report on the way they want.

In order to overcome this problem a specific event tracking process has been created. When a problem is detected health professionals are now supposed to fill out a reporting form to be sent to the appropriate infection control team. But the efficiency of this process is very dependent on the good will and cooperation of the medical staff. This process has several flaws:

1) First of all it relies on a proactive contribution. But this is a new task added to the staff heavy workload. Therefore it is not systematically done because people do not take the time (or simply do not have the time) to do so (especially if they consider this is a minor event).

2) An HAI can be reported only if it has been detected. This is not always the case because people are not aware of the symptoms, because they are too overworked to take time to notice small signs indicating a possible problem. It can also be because a patient already suffers from several diseases at once which make it difficult to isolate some symptoms that could be related to a specific HAI.

3) Finally it also happens that people may be reluctant to recognize their weaknesses or errors in a report. They prefer to handle the situation by themselves without referring to any other external entity.

This is why this reporting process is not as efficient as it should be. This incomplete information transmission reduces the chances to intervene on time. It stresses therefore the need for an automated tool able to scan daily reports, to mine information written in these texts in order to detect potential risk patterns to send alerts to appropriate people. These

experts should also be able to get access to all necessary data required to conduct their analysis.

III. TEXT MINING AND TERMINOLOGY

A. General Approach

There are several types of systems designed to perform text mining. Some of them are based on keyword search. In that case they are more related to document retrieval than to information extraction (i.e. finding specific facts located in one or several documents possibly related by semantic links). Some other systems use more complex pattern matching mechanisms which allow searching not only for unrelated keywords but also for combinations of keywords. They propose some rules (most of the time defined through regular expressions) to provide flexibility in matching. Finally some other systems target more complex information extraction problems which differ from document retrieval in the fact that not only documents are retrieved but also bricks of information located within the text. These systems require to perform a much deeper analysis of the language than simple pattern matching. In order to be able to identify *who do what to whom how and when*, syntactic dependencies must be identified between these elements. Among these parsers there are still some differences. Those that seems the most promising for processing real texts extracted from scientific or medical papers (which often contains very long and complex sentences) are those so called "Robust Parser" such as the one designed by the Xerox Research Center Europe [20].

Such parsers generally use a layered analysis to perform information extraction. This means that the first step will be to tokenize the text. Tokenization is the process of identifying sentences border in a given bloc of text and then to cut sentences into coherent tokens. This concept of coherence is related to the fact that some words may have more sense as a group rather than taken individually. This is the case for example with "a priori" which should be considered as a single token and not as separated words. Once this tokenization has been performed then comes the time of the Morphological Analysis which purpose is to identify all Part of Speech information related to a given token (gender, plural, etc.). Very efficient and fast POS taggers are generally based on Finite State Transducers such as the one developed by XRCE [21]. This step is generally correlated with the disambiguation process that uses the POS context to disambiguate a word (e.g. is it for example rather a noun or a verb, for instance "can" can be both). One commonly used technology to perform this task applies Hidden Markov Models (HMM). This is based on co-occurrence statistics build for a given language for taking a decision [22]. This disambiguation step is very important as valid POS data are the corner stone to perform syntactic dependencies detection.

At this step detection rules do not apply anymore on words but rather on POS tags. We start to abstract from the lexical level to address the language structure: grammar

rules. Furthermore this is also where a big difference can be made among full parsers and robust parsers. Full parsers adopt a top down strategy and try to build a full tree structure reflecting all possible dependencies between each tokens of a sentence in order to build dependency resolution analysis and decide which links are the correct ones.

On the contrary Robust Parsers adopt a bottom up strategy. They first try to identify chunks [23], which means grouping words related by a same syntactic tag (e.g. Subject, Object, Noun Phrase, Modifier, ...). Then once these chunks have been identified and tagged, new rules apply to detect possible relations between these chunks. At this point some dependencies are easier to identify than other (or at least have little room for ambiguity). This is for example the case for dependencies such as Subject-Verb, Verb-Object, ...

"The escheat law cannot be enforced now because it is almost impossible to locate such property, Daniel declared."

```
DETD(law, The)
MOD_POST_INFINIT(impossible, locate)
MOD_PRE(law, escheat)
MOD_PRE(impossible, almost)
NUCL_VLINK_MODAL(cannot, be)
EMBED_INFINIT(locate, is)
NUCL_VLINK_PASSIVE(be, enforced)
OBJ_N(property, such)
TIME(enforced, now)
SUBJ-N(declared, Daniel)
EMBED(is, enforced)
MAIN(declared)
NUCL_SUBJCOMPL(is, impossible)
SUBJ-N(is, it)
```

Fig 1: Dependency Extraction using a Robust Parser

Once syntactic dependencies have been extracted, the last part of the information extraction process deals with the detection of semantic dependencies between named entities in order to identify Events. These events will be key elements for matching information scenarios that correspond to what is really looked for inside texts.

B. Definition

At his point some formal definitions are necessary to characterize the various elements required by this analysis.

Named Entities: Aside from simply tokenizing a sentence and applying Morphological Analysis, it could be interesting, in order to build a deeper analysis of texts, to characterize some lexical elements through semantic tags. Standard Named Entities (NE) can be people names, company names, places, but also dates, processes, and so on. Terminological resources, and to be even more specific Taxonomies such as SNOMED¹ for instance, are very important at this step. They allow a system to identify these entities with respect to their definition in these dictionaries. These entities can be

¹ <http://www.snomed.org/>

composed of several words (or tokens). In this case specific detection rules apply to regroup all these words under a same semantic tag. In the context of HAI, these entities, taken to the largest extent, can be drug names (e.g. “*Loperamid*”), disease name (e.g. “*Clostridium Difficile Infection*”), exam (e.g. “*abdominal ultrasound*”), symptoms (e.g. “*dehydration*”), etc. At this step disambiguation is also a challenge: “*Apple*” should it be considered as fruit, or as an IT Company? This is where semantic disambiguation can bring some insights.

Named Entity Co-reference: Detection of NE may be much more complex than it seems, because it is very frequent that in a long text, when the same entity occurs several times it is replaced by some naming variations (e.g. *IBM, the computer company, Big Blue, ...*) or just co-reference (e.g. “it is also”). Distance between a co-reference and the nearest NE is often used for disambiguating these cases, but this is not the only method.

Named Entity Metonymy Resolution: Another problem related to NE is the Metonymy resolution [24]. This relates to the use of a common name to refer to a broader domain. For example in the following sentence “*Lebanon still wanted to see the implementation of a UN resolution.*”, this is not the country that wants something but rather the government or inhabitants. NE Metonymy Resolution systems have to identify which is exactly the scope/target related to a given NE. This is necessary for Facts extraction.

Events: These elements refer to the detection of some specific relations between NE generally located in the same sentence or in very close sentences in order to be able to identify clear dependency links. The definition of an Event can change from one domain to another, but one example can be illustrated by the detection of interaction relations between genes and proteins inside scientific papers related to Genomics. Such Event can be characterized by the following example that shows possible variations in the way of describing the same information (Event).

srp acts downstream of hkb.
srp acts as a suppressor of hkb.
srp has an interaction with hkb

Information Scenarios: Once Events have been detected they can be then incorporated inside more complex information scenarios that take into account several events located in one or even several documents that should be related in a certain way. Information scenarios can include probabilities figures that constraint the co-occurrence of several specific events to activate the validity of the scenario. This step is the latest and the more complex in an information extraction system.

C. Previous attempts to use NLP to extract information in Biomedical papers

Even if all this process may appear complex several previous works in this domain have proved to be efficient. This is for example the case of a Ph.D. thesis [25] where the goal was to identify gene interactions in scientific papers about *Drosophila Melanogaster*. The first challenge was to identify gene and protein names [26] as for this organism, at that time dictionaries were not complete, and lot of genes had funny names such as : *zen, act, can, if, gypsy, vamp, zip,...* This was a challenge both for disambiguation and for syntactic dependencies analysis. Nevertheless a shallow parser [27] associated with a Conceptual Graph [28] system proved to provide very interesting results. The use of conceptual graphs seemed to be useful at that time because of the lack of existing hierarchical terminologies. It provided a way to abstract knowledge and to create dependency detection rules not just at a lexical level but rather at a conceptual or semantic level which is much more efficient with respect to the number of rules that have to be written to achieve targeted completeness. However now standardized hierarchical terminologies are available and can be used to feed advanced NLP parsers that also embed the capability to use abstracted level of knowledge or semantic rule descriptions. Relations between key elements detected from text using advanced NLP systems can be an input for standard decision support systems.

Even if this work was not directly related to Risk Assessment for Patient safety it demonstrated the benefit of Natural Language Processing for the detection of key information and complex relations between elements inside real scientific descriptions. In this work, one major problem was directly related to the lack of official nomenclature which created a challenge for customizing the Part of Speech tagger. But now standardization has started to make its way in medical literature. Formal terminologies such as SNOMED are more and more developed and used to allow data sharing between various departments or organisms. This is dictated by the increasing role of computerized information systems deployed in hospitals and above. It is for example mandatory to identify and disambiguate clearly in Patient Discharge Summaries which treatments have been given to a patient in order to establish the bill and send the information to Social Security or medical insurance companies. This standardization will be a key enabler for NLP systems to recognize automatically valuable information from texts.

In that perspective XRCE has been developing for the last ten years a Robust Deep Parser (Xerox Incremental Parser) that can be used in different types of applications. XIP is robust that is to say it has already been used in various projects to process large collections of unrestricted documents (web pages, news, encyclopedias, etc). This engine has been developed by a research team in computational linguistics. It has been designed to follow strict incremental strategies when applying parsing rules. The system never backtracks on rules to avoid falling into

combinational explosion traps which makes it very appropriate to parse real long sentences from scientific texts for example. The analysis is relying on three processing layers which are: Part of Speech Disambiguation, Dependency Extractions between words on the basis of subtree patterns over chunk sequences, and a combination of those dependencies with boolean operators to generate new dependencies or to modify or delete existing dependencies. The following example presents the results of a sentence parsed using XIP. It shows only syntactical dependencies, however it is also possible to apply more semantic rules based on classes of terms and dependency types to identify more complex information such as risk patterns.

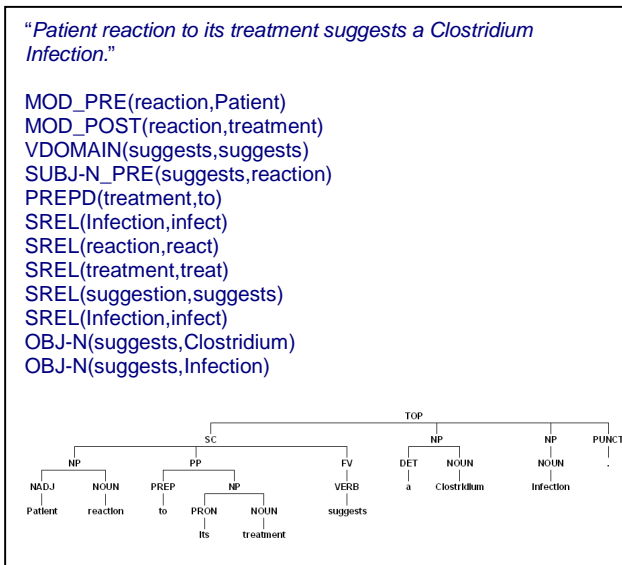


Fig 2: Identifying relations between key factors

Text Mining Technologies seems therefore mature enough to be applied in some real applications and for challenging tasks such as risk patterns detection related to HAI in medical reports.

IV. AUTOMATED SURVEILLANCE SYSTEM

A. Overview

As introduced previously a reporting process has been created in Hospitals to allow a better communication between experts and local departments. These reports are generally based on predefined forms designed to structure information and facilitate storage into databases. Processing predefined forms with automatic systems can be straightforward as information is strongly structured. However it is difficult to cover all possible configurations using predefined templates, this is why they often include a free text area to allow description of complex cases. Moreover, most of hospital reporting is still done using free text documents. The real challenge for automatic systems is to process the content of these free text reports to detect potential problems.

The following sections propose a way to apply a specific combination of text processing components to perform risk assessment through the identification of critical information inside Patient Discharge Summaries. We propose to develop a system that combines a terminology server for Named Entity recognition, then an Incremental Parser embedding Events identification rules, then a decision engine that tries to match these events with predefined threat scenarios designed by experts of HAI surveillance. If risk patterns are detected then alerts can be broadcasted to appropriate people. Furthermore all related data used to conclude to this alert will be recorded into a specific database. This database can then be processed by data mining techniques to perform epidemiological studies in order to discover new risk patterns.

B. Pre-Processing

The HAI surveillance system must not be specific to a given hospital or department but should be operational at a National level. However, as local medical record systems are generally different from one department to another a first step to allow documents access and normalization is necessary. This step implies developing processes and ad-hoc methods to gather required documents from local databases. Interoperability is still an important issue as stressed by EU reports [29]. This is why this part of the system should be customized at local level.

Once documents have been retrieved from local databases the next step is to convert them into a normalized format. This means identifying specific sections inside documents that may indicate timeline (e.g. symptom description, treatments, biological analysis results). Furthermore one important task is also to *anonymize* these data. This means removing all personal information related to people or places (in order to protect patients' privacy). This means that not only people names should be removed but also any information that could lead to an identification such as personal address, phone number, social security number, and so on. This has to be done either by retrieving very specific fields in medical database or by applying natural language processing techniques specifically designed to detect this kind of Named Entities [30].

C. Entity Detection

Once documents have been normalized and *anonymised* they can be processed by a Terminology Server in order to identify and locate all Named Entities that will be useful for the remaining decision process (e.g. drug names, symptoms, processes, dates). This step is done using dictionaries and limited patterns matching methods. List of drug names or Taxonomy such as SNOMED are essential at this step.

However doctors in local hospital generally have their own habits for detailing their work in reports. Therefore the system must be able to cope with these specificities and merge this local terminology into a standardize taxonomy model.

The purpose of this taxonomy is to unify lexical variation of Named Entities into a Unique Concept definition. This allows decision rules to work at the concept level and not at the lexical level. Furthermore, in a foreseeable future of interconnected systems developed for other languages this taxonomy will be the corner stone to allow data exchange.

Once all key concepts have been located inside documents the system can start detecting possible dependencies.

D. Event Extraction

Natural Language Processing techniques are used at this step to identify, disambiguate [31] and index each Events in texts [32]. This is the starting point of a more complex semantic analysis which aims at detecting causality between events.

Incremental Parsers such as XIP can provide robust parsing capabilities to perform this task. This parser will use rules designed for several common languages to identify syntactic and semantic relations between these events (who, when, where, how, ...). But it also must be customized to cope with specificities found in medical reports and related to HAI description.

This customization has to be done in close collaboration with experts from HAI surveillance groups. They are those who can truly define what is important in a report to characterize an HAI, or a potential threat.

The output of this step is the detection of specific semantic dependencies between Named Entities (e.g. symptoms, treatment, conditions) to characterize Events. This information can then be used to perform matching with predefined scenarios describing specific risk patterns.

Report extract:

“...In ICU, beyond dehydration signs, a light left iliac fossa pain was noticed. He presented a hyperleucocytosis (53000/mm3) and inflammation (C-Reactive Protein at 392 mg/L). An abdominal ultrasound showed colitis pictures. Because of a suspicion of Clostridium Difficile Infection (CDI), the patient was treated by parenteral metronidazole...”

Detected Facts:

- Location + Clinical Parameter + Clinical Parameter
- Biochemical Parameter + Biochemical Parameter + Biochemical Parameter + Biochemical Parameter
- Risk pattern
- Risk Pattern + Treatment

Fig 3: Fact Extraction sample

Several works have already shown strong evidence of efficiency in this domain. We can for instance stress the work done by A. Sandor [33] to detect risk patterns in texts related to human food. The process that was used is based on a preliminary detection of named entities and their disambiguation. A semantic tag is then attached to these entities, and then semantic rules are applied to detect specific semantic tag patterns as well as direct syntactical connections between these entities. Once a pattern has been

detected then an alert is triggered. The example in figure 3 illustrates the identification of some key elements. This information will be used at next step to match HAI scenarios.

E. Decision Support System

Once Events have been identified the next step is to search for matches within pre-defined risk scenarios. These patterns are coherent elements that make experts conclude to HAI events. These scenarios involve the occurrence of specific Events in a given timeline. Flexibility is also important at this level because not all information necessary to characterize an HAI scenario may be present in a report and so for various reasons. It can be because the system was not able to detect it, because, it was not noticed by doctors or reported, ... Therefore it is important to add to these scenarios weights to stress the importance of some elements versus others that, if detected, are enough to trigger an alert. Furthermore the weighting system can also allow the system to compute confidence scores for HAI identification.

Other information than just dependencies between Events may also be part of the reasoning process. According to input documents metadata and localization (sections where the information is detected) can also be taken into account to make the decision. Paragraphs ordering as well as dates inside these paragraphs provide useful data to build a timeline that helps classifying information between actions and consequences.

Scenarios used at this step must be designed with experts from surveillance groups. They formalize techniques or reasoning methods they already apply to identify HAI.

However it is difficult to be fully exhaustive while creating these rules for several reasons. First of all there are so many ways to describe and characterize HAI that it is almost impossible to provide completeness at a first attempt. This is an incremental process that should be revised and improved on a regular basis. Furthermore, as medicine evolves, new drugs are introduced on the market, new diseases and infections appear, processes change and so on. Therefore detection rules should be also able to keep close to this evolution. This is why at this point building a database to record all detected events is necessary. It will allow applying data-mining and machine learning techniques in order to discover new relations between facts that may also characterize HAI. Some of these relations may only appear at a macro level through statistical trends. A database enriched at each step of the analysis will be a key element for epidemiological studies.

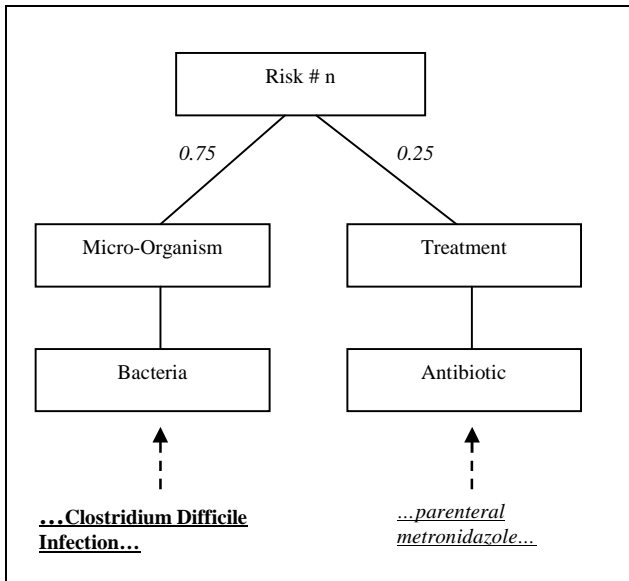


Fig 4: HAI scenario

V. PLAN AND NEXT STEPS

The system detailed in section IV is part of a collaboration that is starting up between the Lyon University Hospital and the Xerox Research Centre Europe to apply state of the art Computational Intelligence Techniques to address a problem jeopardizing public health. This project will be developed in close collaboration between HAI surveillance experts and Linguistic and Knowledge Management experts in order to design the necessary set of rules to identify HAI from medical reports.

On a first hand only some departments and infections will be targeted. These are those related to the highest risks and that have the highest impact on human health. These departments are:

- Intensive Care Unit
- Surgery

A consistent set of reports containing HAI cases will be selected and highlighted by experts for system design and validation.

We are currently building the first version of the system architecture focusing on the detection of the following events:

- Context (e.g. “re-entry”, “Surgery”, “HAI”, ...)
- Clinical Parameters (e.g. “fever”, “inflammatory trace”, “pulmonary secretion”, “cough”, ...)
- Biological Parameters (e.g. “bacteriological exam”, ...)
- Biochemical Parameters (e.g. “white-cell > 10 000/mm³”, ...)
- Treatments (e.g. “specific anti-biotic drugs”, ...)

The real impact of such a project remains to be

determined by experiments. In particular, we need to evaluate how much relevant information is present in the patient record. Also, as mentioned in section II.A., a governmental study indicates that almost 30% of the HAI are estimated avoidable with appropriate processes or surveillance mechanisms. The system described in the previous section might not be able to catch up with all of these cases for various reasons and one of them is related to the constraint that enough information should be present in Patient Discharge Summaries to take efficient decisions. However, on a pure text analysis and information extraction stand point, previous experiments have demonstrated that an overall recall of 50% is achievable for a precision around 90 to 95%. Therefore this could lead to a detection of 15% of these HAI which is in a way a significant benefit for those 15% of people impacted by these infections. Of course the system could also be tuned to increase the recall to the detriment of precision (which means in this case generating more false alerts) according to the objectives.

VI. CONCLUSION

Providing tools to shorten the time and effort necessary to discover and react against HAI is crucial to reduce the impact on public health and therefore the cost for the health care system.

In this paper we have shown how fine grained information extraction relying on robust natural language processing and medical taxonomies could analyze real life data in order to address risk management issues in the case of Hospital Acquired Infections. In that context XRCE is developing a collaborative project between HAI surveillance experts from the Lyon University Hospital and Knowledge management experts to design a control and prevention system able to analyze medical reports for HAI detection.

REFERENCES

- [1] Michel P, Quenon JL, Djihoud A, Tricaud-Vialle S, de Sarasqueta AM: a French national survey of inpatient adverse events prospectively assessed with ward staff : Qual Saf Health Care. 2007 Oct;16(5):369-77
- [2] R.W. Haley, J.W. White et al. "The efficacy of infection surveillance and control programs in preventing nosocomial infections in US hospitals." Am J Epidemiol 121. 1985, pp.182-205
- [3] R. Condon, W. Schulte, et al. (1983). "Effectiveness of a surgical wound surveillance program." Arch Surg 1983; 118: 303-7.
- [4] S. D. Bärwolff, C. Geffers, C. Brandt, R.P. Vonberg, et al. "Reduction of surgical site infections after caesarean delivery using surveillance." J Hosp Infect 2006; 64: 156-161.
- [5] P. Gastmeier, C. Brandt, I. Zuschneid, D. Sohr et al. "Effectiveness of a nationwide nosocomial infection surveillance system for reducing nosocomial infections." J Hosp Infect 2006; 64 : 16-22.
- [6] Garner JS, Jarvis WR, Emori TG et al. CDC definitions for nosocomial infections,1988. Am J infect Control 1988;16 : 128-40.
- [7] H, S. E. Humphreys. "Prevalence surveys of healthcare-associated infections : what do they tell us, if anything?" Clin Microbiol Infect 2006; 12: 2-4.
- [8] Senat. 2005. <http://www.senat.fr/rap/r05-421/r05-4211.pdf>.
- [9] B.S. Cooper, S.P. Stone, C.C. Kibbler, B.D. Cookson BDet al. Systematic review of isolation policies in the hospital management of methicillin-resistant *Staphylococcus aureus* : A review of the

- literature with epidemiological and economic modelling. *Health Tech Assess* 2003; 7: pp. 1-194.
- [10] R. Plowman, N. Graves, M. Griffin, J. Roberts et al. *Socio-Economic Burden of Hospital Infection*, PHLS, 1999, ISBN number 0 901144 487.
- [11] T. Kunori, B. Cookson, C. Kibbler, Stone, S., J. Robert. *The cost effectiveness of different MRSA screening methods*. *J Hosp Infect* 2002; 51; pp. 189-200.
- [12] R. Masterton, B. Cookson, A. Pallett, A. Mifsud et al. *Review of Hospital Isolation and Infection Control Related Precautions: Report of the Joint Working Group*. <http://www.amm.co.uk/documents/HIPFIN.DOC> and *Leader J Hosp Infect* 2003;54; pp. 171-173.
- [13] M.A. Borg, B. Cookson, E. Scicluna and the ARMed Project Steering Group. *Survey of infection control infrastructure in selected southern and eastern Mediterranean countries* *Clin Micro Infect* 2007;13:344-346.
- [14] M. Struelens, D. Wagner, J. Bruce, F.M. MacKenzie, B. Cookson et al. *Status of Infection Control Policies and Organisation in European Hospitals*, 2001: The ARPAC Study. *Clinical Microbiology and Infection* 2006;12: 729-37.
- [15] R.J. Pratt, H.P. Loveday, C.M. Pellowe, P.H. Harper, S. Jones et al. *A comparison of international practices in the management and control of hospital acquired infections*. A component of the VFM Follow-Up Study of the National Audit Office Report: The Management and control of Hospital-Acquired Infections in Acute Hospitals in England HC 876 2003-2004 ISBN: 0102929157.
- [16] L. Pokorny, et al., *Automatic detection of patients with nosocomial infection by a computer-based surveillance system: a validation study in a general hospital*. *Infect Control Hosp Epidemiol*, 2006. **27**(5): p. 500-3
- [17] L.R. Hirschhorn, J.S. Currier, and R. Platt, *Electronic surveillance of antibiotic exposure and coded discharge diagnoses as indicators of postoperative infection and other quality assurance measures*. *Infect Control Hosp Epidemiol*, 1993. **14**(1): p. 21-8
- [18] S. Bouam, E. Girou, et al. "An intranet-based automated system for the surveillance of nosocomial infections: prospective validation compared with physicians' self-reports." *Infection Control and Hospital Epidemiology* (2003) **24**(1): pp. 51-55.
- [19] H.M. Glenister, L.J. Taylor, C.L. Bartlett, E.M. Cooke et al. *An evaluation of surveillance methods for detecting infections in hospital inpatients*. *J Hosp Infect*; (1993) **23**: pp. 229-42
- [20] S. Ait-Mokhtar, J.P. Chanod, C. Roux. Robustness beyond Shallowness: Incremental Dependency parsing. Special Issue of the *NLE Journal* (2002)
- [21] A. Schiller. Multilingual Part-of-Speech Tagging and Noun Phrase Mark-up. Proceedings of the 15th European Conference on Grammar and Lexicon of Romance Languages (ECGLRL96). University of Munich, Germany, September 1996
- [22] J. Kupiec. Robust Part-of-Speech Tagging Using a Hidden Markov Model. *Journal of Computer Speech and Language*. Vol. 6., 1992
- [23] S. Abney. parsing by Chuks. In Robert Berwick, Steven Abney, and Carol Tenny (eds). *principle-Based parsing*. Kluwer Academic Publishers. 1991
- [24] C. Brun, M. Ehrmann, G. Jacquet. *A Hybrid System for Named Entity Metonymy Resolution*. Proceeding of 4th International Workshop on Semantic Evaluations (ACL-SemEval), Prague, June 2007, pp. 23-24.
- [25] D. Proux. *Muninn: une strategie d'extraction d'information dans des corpus specialisés par application de methodes d'analyse linguistique de surface et de representation conceptuelles des structures semantique*. PhD, Thesis. Universite de Bourgogne. 2001
- [26] D. Proux, J Laurent, F. Rechenmann. *Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction*, *Genome Informatics*, 9: pp.72-80, 1998
- [27] S. Ait-Mokhtar, J.P. Chanod, (1997) *Incremental Finite-State Parsing*. In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97), Washington March 31st to April 3rd, 1997, pp.72-79
- [28] J.F. Sowa. (1984). *Conceptual Structures*. Information Processing in Mind and Machine. Reading, Mass, Addison-Wesley Publishing Comp
- [29] eHealth for Safety. Impact of ICT on Patient Safety and Risk Management. European Commission. Information Society and Media. ISBN 13 978 92 79 06841-6
- [30] C. Brun, C. Hagège. Intertwining deep syntactic processing and named entity detection. ESTAL 2004, Alicante, Spain, October 20-22, 2004
- [31] B. Jacquemin, C. Brun, C. Roux. *Enriching a text by semantic disambiguation for information extraction*. Conference Proceedings LREC, Las Palmas, Spain, June 2, 2002.
- [32] D. Proux, L. Julliard, F. Rechenmann. *A Pragmatic Information Extraction Strategy for Gathering Data on Genetic Interactions*. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. 2000. ISBN:1-57735-115-0, pp 279 - 285.
- [33] A Sandor. *Using the author s comments for knowledge discovery*. Semaine de la connaissance, Atelier texte et connaissance, Nantes, June 29, 2006.