Effortless identification of surgical site infections: empowering clinical data warehouses

Francesco Montia, Julien Grosjeana,b, Tristan Petelc,…, Stéfan J. Darmonia,b

**a** Department of Biomedical Informatics, Rouen University Hospital, Rouen, France

**b** Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, U1142, INSERM, Sorbonne Université & Sorbonne Paris Nord, Paris, France

**c** Department of Hygiene, Rouen University Hospital, Rouen, France

Abstract.

*The surveillance of Surgical Site Infections (SSI) contributes to the management of risk in French hospitals. Manual identification of infections is costly, time-consuming and limits the promotion of preventive procedures by the dedicated teams. The introduction of alternative methods using automated detection strategies is promising to improve this surveillance. The present study describes an automated detection strategy for SSI in neurosurgery, based on textual analysis of medical reports stored in a clinical data warehouse. ………*

*Results ………………..*

Keywords…

## 1. Intro

As Electronic Health Record (EHR) systems gain ubiquity, the volume of electronic clinical data continues to burgeon. This proliferation has captivated researchers, healthcare administrators, and clinicians alike, fueling interest in the secondary utilization of EHR data to augment clinical acumen and optimize patient care. Among the myriad applications of EHR data, the targeted detection of specific outcomes and adverse conditions—such as Surgical Site Infections (SSIs)—emerges as a particularly compelling avenue for exploration.

The French Health Authority ardently advocates for the routine surveillance of SSIs, situating it within the broader framework of risk management for Healthcare-Associated Infections (HAIs). While the salutary impact of HAI surveillance on both health outcomes and healthcare efficiency is empirically substantiated, the prevailing identification methodologies remain predominantly manual, executed by local hospital hygienists or direct practitioners. This manual approach is not only resource-intensive but also diverts critical resources away from the conceptualization and monitoring of preventive strategies.

Current SSI detection practices, largely reliant on manual surveillance, are fraught with inconsistencies and limitations—ranging from human error to staff turnover and training gaps. The automation of SSI detection via EHRs offers a more standardized, efficient, and comprehensive modus operandi. It mitigates human error, facilitates real-time monitoring, and seamlessly integrates into extant healthcare IT ecosystems, thereby elevating the caliber of SSI reporting and contributing to superior patient care.

In recent years, Clinical Data Warehouses (CDWs) have emerged as indispensable assets within hospital settings, facilitating the extraction of actionable insights from both structured and unstructured data. These tools predominantly employ Natural Language Processing (NLP) algorithms to sift through clinical narratives for a variety of applications, such as identifying eligible patients for clinical trials and targeted research endeavors. While numerous contemporary approaches pivot on predictive models or machine learning algorithms, our methodology is uniquely anchored in pure information retrieval. This is primarily attributable to the preponderance of relevant information embedded within clinical narratives, which are inherently challenging to manipulate via machine learning techniques due to data constraints.

The principal aim of our study is to evaluate the efficacy of the "Entrepôt des Données de Santé Normand" (EDSaN) in identifying SSIs following spinal surgeries conducted within the neurosurgery or orthopedic wards of Rouen Hospital. The ultimate aspiration and secondary objective of this paper are to expedite and streamline the detection of such events by the Hygiene Department, thereby alleviating the workload engendered by extant surveillance practices.

## 2. Methods

### SSI definition

Surgical Site Infection was defined according to the CDC/NHSN recommendations. SSI secondary to neurosurgery consist of infections that appear to be related to the operative procedure, occurring within 30 days after the operation for non-implant procedures, or within one year if implant is in place. According to this definition, surgery procedure and SSI diagnosis could be reported at different times. Thus, every detected and validated SSI are said to be linked to the first surgical procedure. For the equivalent reason, a surgical event begins with the patient's admission for the originally planned surgery and includes all documents produced by the neurosurgery or orthopaedic surgery department from admission up to one year after the procedure if any kind of material has been implanted.

### Entrepôt des Données de Santé Normand (EDSaN)

EDSaN is an in-house solution to query Rouen University Hospital’s (RUH’s) Clinical Data Warehouse (CDW). EDSaN gathers clinical data since the 1990’s from about 2 million patients. Several data types have been integrated so far from the EHR and various clinical data bases: structured data from biology, virology, diagnoses, procedures and unstructured data such as CN (discharge summaries, letters, procedure results, prescription letters, etc.)[ref].

More precisely, diagnoses and procedures codes are collected from the French Diagnosis Related Groups (DRG), known as PMSI in French language; two main classifications are used to code records: the ICD-10 (International Classification of Diseases- 10th version) for diagnoses and the CCAM (Classification Commune des Actes Médicaux) for procedures.

EDSaN consists in: (a) a query tool that allows to search and/or mine data. For example, it can be used to identify patients from different criteria; (b) a selection tool that allows to filter and explore datasets collected from (a).

The data are pseudonymized or de-identified in EDSaN to preserve patient anonymity. However, for specific purposes, it is possible to re-identify the data. For example, it could be important to contact patients for various reasons; COVID-19 vaccination is one of them[ref].

The EDSaN query software consists in a multilevel search engine that is able to query structured data, unstructured data, and both structured and unstructured data at the same time. This is a very important feature as DRG codes can sometimes lack precision and can sometimes be missing. In these cases, the automatic processing of clinical narrative is used.

It combines several NLP algorithms to ensure that searched keywords are relevant (i.e., not in a negative sentence for example) or present in text specific segments (such as conclusion for example) even if those documents are natively unstructured.

To identify the patients of interest, the following steps were applied:

1. Identification of target data types that should be used for each disease/condition (unstructured or structured data or both);
2. Creation of EDSaN queries;
3. Processing and export of queries;
4. Exclusion of patients not meeting the definition criteria for SSI or that underwent the original spinal surgery somewhere else.

A list of 15 randomly selected and manually validated patients was used as the training set to build the queries.

**BUILDING QUERIES**

Keywords can be entered and enhanced with the following advanced options: wildcards (\*) can be used to deal with variations in spelling (e.g. pso\* for psoriasique, psoriasis…); double quotes (“”) can be used to search exact word sequences (e.g. “allergie au paracetamol”); slope (~) can be used to take in consideration distance between (non-stop) words (e.g. “abcès anal”~2 can have a variation form as abcès de la marge anale); Boolean operators can be used to combine terms (e.g. (paludisme OR “accès palustre” OR palu) AND quinine). A specific module allows to search in one specific segment or to specify if negative or hypothetical clinical concepts are kept in search results or not. Keyword queries can be combined with structured data (age, sex, document type, medical unit, ICD-10 or CCAM codes).

One of the main features of EDSaN CN search is to provide easy and fast access to semantic expansion. The function allows to search synonyms and related terms from the HeTOP server [ref] to leverage the original simple query in order to expand the number of documents (lexical variations, acronyms, etc.).

### Strategies for detection

Four strategies have been put in place : the first one is based upon the exploitation of full-text medical reports while the other three make use of DRG codes for querying the medico-administrative database.

For the first query, based on CN (clinical narratives), an initial list of keywords has been created based on the documents concerning the patients included in the “gold standard list” and later enhanced by common sense, experience and extensive use of the aforementioned semantic expansion utility.

In text-mining analysis three groups of keywords can be identified:

* “rachis institute”: targets relevant activity, mostly conducted between neurosurgery and orthopaedic ward but allows us to not filter by ward directly hence giving us the chance to recover a few more patients hospitalised elsewhere;
* Concepts related to the domain of surgical complications;
* Concepts related to the domain of surgical procedures.

The strategies based on the DRG database use ICD-10 or CCAM codes for querying the medico-administrative database.

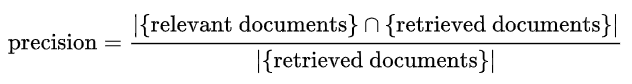
Codes were initially defined based on the “gold standard list” and later further enriched new findings and following a discussion with Medical Information Department physicians.

### Evaluation and performance metrics

The detection strategies were evaluated using recall, precision and F-measure on an independent validation set of patients. A result was said to be a True Positive if at least one of the three detection strategies had identified SSI secondary to spinal surgery for that patient, and if it was confirmed by ………..

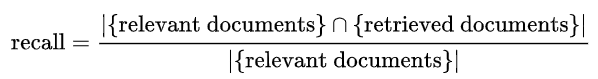
In [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval) contexts, precision and recall are defined in terms of a set of retrieved documents (e.g. the list of documents concerning the patients experiencing a SSI, in other words the true positives) and a set of relevant documents (e.g. the list of all documents retrieved by the queries we built).

Precision expresses the proportion of the data points our model say was relevant actually were relevant:

(temporary)

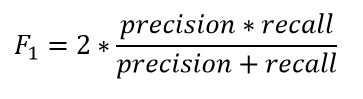
For a text search on a set of documents, precision is the number of correct results divided by the number of all returned results.

While recall is the fraction of the relevant documents that are successfully retrieved:

(temporary)

For a text search on a set of documents, recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is also called sensitivity.

The two measures are sometimes used together in the [F1 Score](https://en.wikipedia.org/wiki/F1_Score) (or f-measure), the harmonic mean of the precision and recall, to provide a single measurement. [F1 Score](https://en.wikipedia.org/wiki/F1_Score) is defined as follows:

 (temporary)

Expliquer l’évaluation sur les 2 « niveaux » document et patient et pourquoi ça varie

NB: Le calcule des courbes ROC et de la Youden’s J statistic n’est pas faisable n’ayant pas le VNs et FNs de notre population entière.

## 3. Results

Overall, this this and that has been found.

Table show complete results of the 4 detections strategies while detailing the queries.

A élaborer une fois comparés les résultats à la liste “gold standard” complete.

*Queries used to identify patients*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data type** | **Ward filter** | **Query** | **Documents** | **Patients nb.** | **T+** | **Precision calculated on patients n°** | **Precision calculated on documents n°** | **Recall patients** | **Recall doc** |
| CN | NO filter | **(**"institut du rachis"~5 OR "chir du rachis" OR "chirurgie du rachis" OR "neurochir rachis" OR "orthopedie rachis" OR sacrectomie OR "Ostéosynthèse interlamaire"~3 OR "Ostéosynthèse transpediculaire"~3 OR "Ostéosynthèse de la colonne"~5 OR "arthrodese lombaire"~5 OR "arthrodese thoraco"~5 OR "arthrodese thoracique"~5 OR "arthrodese thorax"~5 OR "arthrodese dorsal"~5 OR "arthrodese dorsale"~5 OR "arthrodese cervical"~5 OR "arthrodese cervicale"~5 OR "exerese hernie"~5 OR discectomie~ OR Laminarthrectomie~ OR corporectomie OR "decompression medullaire"~3 OR "compression medullaire"~3 OR "correction arthrodese"~10 OR sacrectomie~ OR épiphysiodèse~ OR "reduction cyphose"~5 OR Laminectomie OR Laminoplastie OR Spondyloplastie**)** **AND**  **(**+RECTXT\_TRUE:**(**"infection du site opératoire"~3 OR "infection de plaie opératoire"~3 OR "ISO" OR "infection d'une plaie chirurgicale"~3 OR "infection d'une plaie post-opératoire"~3 OR "infection du site chirurgical"~3 OR "plaie chirurgicale infectée"~3 OR "infection de la cicatrice"~3 OR "sepsis au niveau de la cicatrice"~3 OR "infection au niveau de la cicatrice"~3 OR "infection au site de l'opération"~3**)** **OR** **+RECTXT\_TRUE**:**(**"lavage cicatrice"~5**)** **OR** **+RECTXT\_TRUE**:**(**"lavage cicatricielle"~5**)** **OR** **+RECTXT\_TRUE**:**(**"collection infectieuse"~5**)** **OR** **+RECTXT\_TRUE**:**(**"evacuation peridural"~5**)** **OR** **+RECTXT\_TRUE**:**(**"écoulement purulent"~5 OR pyorrhée~2 OR "écoulement séropurulent"~5**)** **OR** **+RECTXT\_TRUE**:**(**"écoulement séreux"~3**)** **OR** **+RECTXT\_TRUE**:**(**"désunion d'une plaie opératoire"~3 OR "rupture d'une plaie opératoire"~3 OR "cicatrice avec désunion"~3 OR "déhiscence de la cicatrice"~3 OR "disjonction de la cicatrice opératoire"~3 OR "désunion cicatricielle"~3 OR "déhiscence cicatricielle"~3 OR "disjonction de la cicatrice"~3 OR "déhiscence de la cicatrice opératoire"~3 OR "rupture de la plaie"~3 OR "déhiscence d'une plaie opératoire"~3 OR "rupture de la plaie post-opératoire"~3**)** **OR** **+RECTXT\_TRUE**:**(**"drainage chirurgical"~3 OR "incision et drainage"~3 OR "incision et évacuation"~3**))** | 200 | 106 | 52 | 52/106 = **49.05%** | 131/200 = **65.5%** |  |  |
| DRG |  | **AFPA001**\* | 57 | 29 | 29 | 29/29 = **100%** | 57/57 = **100%** |  |  |
| DRG |  | **All CCAM codes concerning spinal acts\*** AND **(**T814 OR T845 OR T846 OR T847**)** | 43 | 29 | 22 | 22/29 = **75.9%** | 32/43 = **74.4%** |  |  |
| DRG | NO filter | **(**G551 **OR** M511 **OR** G952 **OR** M431 **OR** M418 **OR** M480 **OR** M471 **OR** M462 **OR** M463 **OR** M465**)** **AND**  **(**Y831 **OR** T814 **OR** T845 **OR** T846 **OR** T847 **OR** T813**)** | 21 | 10 | 6 | 6/10 = **60%** | 12/21 = **57.14%** |  |  |
|  |  | +RECTXT\_TRUE:**(**"infection du site opératoire"~3 OR "infection de plaie opératoire"~3 OR "ISO" OR "infection d'une plaie chirurgicale"~3 OR "infection d'une plaie post-opératoire"~3 OR "infection du site chirurgical"~3 OR "plaie chirurgicale infectée"~3 OR "infection de la cicatrice"~3 OR "sepsis au niveau de la cicatrice"~3 OR "infection au niveau de la cicatrice"~3 OR "infection au site de l'opération"~3**)** **OR** **+RECTXT\_TRUE**:**(**"lavage cicatrice"~5**)** **OR** **+RECTXT\_TRUE**:**(**"lavage cicatricielle"~5**)** **OR** **+RECTXT\_TRUE**:**(**"collection infectieuse"~5**)** **OR** **+RECTXT\_TRUE**:**(**"evacuation peridural"~5**)** **OR** **+RECTXT\_TRUE**:**(**"écoulement purulent"~5 OR pyorrhée~2 OR "écoulement séropurulent"~5**)** **OR** **+RECTXT\_TRUE**:**(**"écoulement séreux"~3**)** **OR** **+RECTXT\_TRUE**:**(**"désunion d'une plaie opératoire"~3 OR "rupture d'une plaie opératoire"~3 OR "cicatrice avec désunion"~3 OR "déhiscence de la cicatrice"~3 OR "disjonction de la cicatrice opératoire"~3 OR "désunion cicatricielle"~3 OR "déhiscence cicatricielle"~3 OR "disjonction de la cicatrice"~3 OR "déhiscence de la cicatrice opératoire"~3 OR "rupture de la plaie"~3 OR "déhiscence d'une plaie opératoire"~3 OR "rupture de la plaie post-opératoire"~3**)** **OR** **+RECTXT\_TRUE**:**(**"drainage chirurgical"~3 OR "incision et drainage"~3 OR "incision et évacuation"~3**) AND** |  |  |  |  |  |  |  |

*\*See Annexes*

***CN****: « clinical narratives »*

***DRG****: « Diagnosis related groups »****T+****: « True positives »*

## Diagramme de Venn – Patients

## 

## Diagramme de Venn - Documents

## 4. Discussion

Automatic detection of SSI is a very complex task. It relies on multiple data source exploitation and most of the time, even if CN are relevant to search for information, it requires NLP algorithms to maximize the accuracy.

In this study, the automatic detection of SSI reached XXX of f-measure (YYY precision / ZZZ recall) for patients. It means that we have identified NNN new patients compared to the initial list collected manually.

Our approach is based on those NLP features and also uses DRG codes to maximize the recall. This approach does not require prior annotated data such as most of machine learning methods [refs] and can be easily adapted if necessary (addition of DRC codes or CN keywords for example).

Even if building specific queries is a time-consuming task, execution and automatic retrieval is much faster than manual screening. It is a huge time saver. Moreover, once queries are built, they can be reused periodically without any adaptation on a long-term basis.

*Lister les limites :*

*Limites intrinsèques à la qualité du codage PMSI*

*Limites intrinsèques à la qualité des CRs*

*Biais de désirabilité et intérêts perso pourraient induire les chirs à ne pas être exhaustifs*

*Inconsistances dans la façon de rédiger un rapport*

*Impossibilité de définir spécificité et sensibilité n’ayant pas la population globale*

*Algorithme testé que sur le 2020, la même année de la liste gold standard*

*Regarder la prévalence des SSI les années précedentes, si stable ou pas. Ça peut être un point à discuter dans l’évaluation de la performance.*

Expliquer en quoi cette détection va aider le service d’hygiène/les chir/les patients (VM).

Actuellement la surveillance se concentre uniquement sur ortho, les requêtes vont plus loin

## Conclusion

## References

## ANNEXE (work in progress)

### Codes CIM10