MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-491

A full-text information retrieval system for an epidemiological registry

Marc Cuggia[a], Sahar Bayat[a], Nicolas Garcelon[a], Lauren Sanders[a], Florence Rouget[c], Arnaud Coursin[b], Patrick Pladys[c]

[a]*Unité Inserm U936, IFR 140, Faculté de Médecine, University of Rennes 1, France*
[b]*DEFSI – CHU de Rennes*
[c]*Réseau périnatal Bien Naitre en Ille et Vilaine, Rennes, France*

Abstract

Case finding for epidemiologic registries still relies mainly on a manual process. In this paper, we show that retrieval information tools could be a complementary way to identify cases for a pediatric malformation registry. We developed a full-text and metadata search engine plugged to a clinical documents repository and used it to identify Epi/Hypospadia and Spina bifida cases. The queries were enriched with Snomed terminologies. We compared the performances of this prototype versus the hospital DRG database (classical method). The best precisions of prototype for identification of Spina bifida and Epi/Hypospadia were respectively 73% and 87%. The prototype overlap with the DRG system was 83% and 97%. Compared to DRG, 13 new not referenced and 2 miscoded cases were detected. This free full-text retrieval system prototype allows efficiently reusing clinical documents for case finding for an epidemiologic purpose.

Keywords:
Registries, Epidemiology, Information Storage and Retrieval

Introduction

The conventional method of case finding used by many institutions to feed an epidemiologic registry is still largely to manually scan and read printed reports or manuscript notes generated during the health care process. These documents are most of the time archived in the paper patient record. This task aims to identify cases in a comprehensive way. It is a painstaking, time consuming and costly work [1]. With the development of Electronic Medical Records (EMR), patient data are more accessible as ever, however these data remain widely coded as free text. Clinicians either dictate or type relevant patient data into the EMR without formal structure or a controlled medical vocabulary.

Coding information is difficult and the reasons why clinicians are so reluctant to do it are well identified [2]. So, clinicians still code little information, either because it's mandatory (e.g. for the billing process), or because they derive a direct benefit (e.g. in term of scientific research or patient safety). Hence, electronic free text medical reports are a mine of information for the registry feeding or cohort enabling, but they are also difficult to exploit. Numerous works have been carried out to develop NLP methods for automatically extracting information from patient data but they are still rarely applied outside of the laboratory [3]. An alternative way to achieve this goal is to provide to clinicians, information retrieval tools (that rely on methods of indexing, searching, and recalling data, particularly text or other unstructured forms). These tools, derived from the web technologies are today broadly used for searching accurate information amongst terabytes of data for personal information management as well as on the Internet (e.g. Google Desktop or Apache Lucene). Hence, "Finding a needle in a haystack" becomes a reality.

For example Schulz & al. [4] have developed a retrieval system to support search across patient discharge letters. They used a linguistic transformation method to deal with the morpho-semantic wealth of German language. They have shown the very good acceptance among the physicians of a WWW-like querying. Few works using these technologies have been carried out in the epidemiologic field. For example, Hanauer & al [5] have defined a custom-made list of terms, phrases and Snomed codes intended to build free text query. Their system was designed to populate a cancer registry database.

Rosier & al [6] applied regular expressions to extract relevant information from surgical reports in order to populate a cardiologic registry.

In this paper, we describe an information retrieval platform combining a searching engine "plugged" to a clinical documents repository automatically fed by our Hospital Information System (HIS). We study the performance of this retrieval system in order to identify cases for an epidemiological pediatric registry. We compare our searching method using free-text clinical documents and their related metadata versus a classical query method, based on the DRG (Diagnosis related group) coded database. We study also the interest in our system of using SNOMED (versions 3.5 and CT) terminologies for a semantic enrichment of the full text queries. As the coverage of the DRG coded database is structurally lower than our EMR repository (as it concerns only inpatients) we measure the added value of our system to identify cases that are not reachable through the DRG coded system.

Materials and Methods

The pathologies studied:

This work fits within a broader feasibility study for the future neonatal malformations registry of Brittany (France). This study focuses on Epi/Hypospadia, Spina bifida malformations.

According to the MeSH definitions, Epi and hypospadia *are* birth defects due to malformation of the urethra in which the urethral opening is below its normal location. Spina bifida is a congenital defect of closure of one or more vertebral arches, which may be associated with other malformations.

Population of interest:

We searched cases amongst children aged 1 year old or lower, born between 11/1/2006 and 11/1/2008, and who had a contact with the University Hospital of Rennes until March 1rd 2009.

System design:

We developed a repository containing the medical reports corpus extracted from our HIS (see Figure 1). The repository is automatically fed by an Extract Transform and Load (ETL) process. ETL is a process in database usage and especially in data warehousing that involves: Extracting data from outside sources; transforming it to fit operational needs (which can include quality levels); loading it into the end target. Each document extracted from the EMR is associated with a XML notice containing all the relevant metadata available from the EMR such as patient information (birth date, sex, weight) or administrative information (such as ward id, dates of admission and discharge). Then the documents and the XML notices have been indexed. We used Apache Lucene [7] as the indexing and search engine tool.

Lucene is an open-source, high-performance, full-featured text search engine library written entirely in Java. It relies on the tri-grams indexing method, allowing phrases, wildcard and proximity queries and returns ranked search results. All the developments were based on web 2.0 technologies (PHP, Ajax). The system provides statistical analysis: e.g. occurrence of expression present in each document, and in the whole corpus. Accessing the document is immediate, and the searched expressions are highlighted in the text to make the reading easier (see Figure 2).

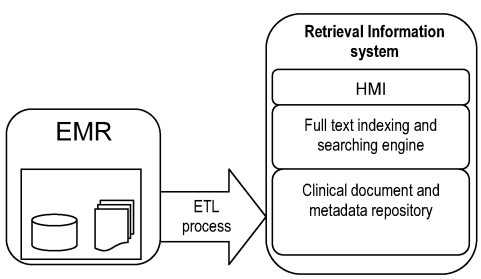

Figure 1-System design

Free Text queries building:

For each pathology, we asked 2 experts (a pediatric surgeon, and a pediatric physician) to provide ten of the most meaningful literal expressions usually employed in clinical notes and that characterize these pathologies. In order to avoid a too high level of noise, a third expert split these expressions in two sets:

The "Major set" consists in expressions relative to the diagnosis and a procedure used in these pathologies (e.g. "neural tube defect" or "myelomeningocel" for *spina bifida* malformation).

The "Context set" corresponds to expressions not directly explicit, but used often with one or more major expressions: for example the anatomical term "penis" or "prepucial" for *Hypospadia* malformation.

As a registry is supposed to be exhaustive, we tried to semantically enrich the expressions given by the experts, by building the "extended major set" containing expressions with synonym and subsumed terms found both in Snomed (version 3.5 and CT). Amongst these candidate terms, we selected only those containing any redundant terms (regarding to the other expressions) e.g: for *spina bifida,* [76916001:Spina bifida occulta] was ruled out, [61819007:Rachischisis] was ruled in.

Then, for the two pathologies, we built 6 queries, combining with Boolean expressions the expressions from respectively the Major set, Context set, and Extended Major Set.

In order to avoid searching in the narrative text different morpho-syntactical forms of a same expression, we transformed them in morphemes. Below is an example of full text query for *spina bifida*. The contextual expressions are underlined.

Expert expressions : *Spina bifida, myéloméningocèle, méningocèle, Fermeture du tube neural, syndrome d'Arnorld Chiari, paraplégie, agrandissement de vessie, cystosomie continente, intervention de Mitranoff, intervention de Malone, sphincter artificial.*

"Major set" query :*(spina) (bifida) (myelomeningocele*) (+fermeture +tube +neural)(meningocele*)*

"Context_set" query:*(+Arnold +chiari) (paraplegie)(+agrandissement +vessie)(cystostomie continente) (mittranoff)(malone)(+sphincter +artificiel)*

"Extended set" query: *(spina) (bifida) (myelomeningocele*) (+fermeture +tube +neural)) (meningocele*) (+rachischisis +aperta) (syringomyelocele*) (meningocele*) (meningomyelocele*) (myelocystocele*) (hydromeningomyelocele*) (+hydromeningocele* +spinal) (+spina +fissure*) (hemimyelocele*) (lipomeningocele*) (rachischisis) (holorachischisis*) (myelocele*) (hydromyelocele*).*

Assessment:

We compared the outcomes of the full text method with the information available in the DRG system (called the PMSI in France). The PMSI aims to provide medico-economic statistics for the hospital budget.

As the PMSI information is supposed to be exhaustive, we used it as a gold standard. It contains the entire coded discharge summaries for all inpatients. Each summary is coded

with ICD10 diagnosis and CCAM procedure classifications. In French hospitals, this coding concerns only inpatients. In our hospital, this coding is centralized and manual. We searched all the patients having a coded discharge report in the DRG System, with ICD10 codes relative to one or more of both pathologies studied.

DRG query building:

For each pathology, a DRG expert selected the codes from ICD10. Two medical experts reviewed these codes. Then we queried the DRG database, using SQL language. For example, the query for spina bifida concerned all the codes starting by *Q05*, which includes of course *spina bifida*, but also codes detailing specific anatomical types e.g : *Q05.1:Thoracic spina bifida with hydrocephalus, etc.*

Comparison method

Two authors reviewed manually each case found by the two searching methods. We searched for the wrongly coded cases and looked at the pertinence of the expressions found with the search engine. For example, for the free text, if the expressions were used with negation, or written in the list of familial histories, the case was considered as False Positive (FP).

For the DRG coded query method, we checked if the DRG were in adequation wih the content of the medical record.

We tried to figure out the reason of non-agreement between the both methods: if for example a case was found by the ICD10 method and there was no track of information of this diagnosis in the medical record, then we considered the case as being "wrongly coded" and false positive.

We also characterized the contributions of the extended set. For each query, we calculated the percentage of added cases owed to the Snomed enrichment.

In France, the coverage of the DRG system only concerns the inpatients. As our clinical documents repository gathers inpatient and out patient's medical reports, it represents a source of information broader than the DRG, and contains potentially precious information about newborns who have been seen only in consultation. Therefore, we measures the number of cases we spotted with our system and amongst them, the number of cases that are true and false positive.

Results

Information sources features:

The clinical reports repository contained 787.000 documents such as surgery reports, clinical notes, or discharge summaries. A subset of these documents, corresponding to the period of study was selected thanks to the "birth" metadata extracted from the HIS. This subset contained 16.512 documents . The full text queries have been applied on this corpus. The run time to queries never exceed 6 seconds.

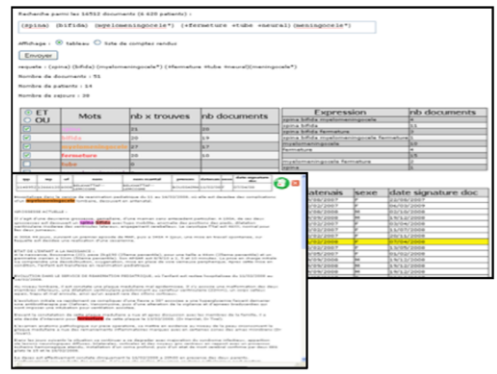

Figure 2-Screenshot of the Retrieval Information System

Cases found with the full-text queries

For the *spina bifida*,
With the "Major set" query, using only the expression given by the two experts, we found 51 documents related to 14 different children and corresponding to 38 contacts in the hospital (in and outpatients). Among them, after checking each document, it turned out that 11 cases were true positives (precision = 73%) i.e. patients for whom there was one or several reports containing an explicit reference to spina bifida. 3 cases have been considered as false positives (FP) and corresponded to the use of related terms in the familial history section of the documents.

The "extended set" query contained 14 supplementary expressions derived from the Snomed terminology but none of these expressions has been found in the reports. So the cases found with the "extended set" query were strictly the same as for the "Major query".

The "context set" query which contained "major" and "contextual" expressions, returned 65 documents related to 20 patients for 47 hospital contacts . Obviously, these results included the cases of the "major set" query. The six new cases were actually five false positive, so the precision was 55%: for example , documents for three cases containing the expression "Malone" which is both the name of a surgery procedure, and the first name of the three patients; one case related to a paraplegia thought it was not caused by a spina bifida. One case corresponded to the maternal familial history, another case was considered by the physicians as a possible spina (a MRI was ordered to confirm the diagnosis), but we didn't consider this case as true positive.

For the epi/hypospadia, with the "Major query", we found 318 documents related to 116 different children and corresponding to 274 contacts in the hospital (in and outpatients). Among them, after checking each document, it turned out that 102 cases were true positives (precision= 87%). 14 cases have been considered as false positives (FP). In 4 cases, we found a negation associated to the expression (e.g : "*no argument for an hypospadia*"), 6 cases were related to an abnormality of the penis with close similarity with an hypospadia (e.g : chordee is a condition strongly associated with hypospadia) but, eventually, the surgeon or the pediatrist did not consider that this

was an actual case of hypospadia. Two cases were related to a penis surgery, but for another problem. Two case contained "hypospadia" in the familial history section. Let's note that we only found one case of epispadia (this pathology is very rare).

The "Context set" query has brought out 45 FP cases, 95 % concerned another diagnosis, 5% a normal clinical exam.

The "Extended set "query (with Snomed) for this pathology was exactly the same as the expert's one. Indeed, after having applied the non-redundancy rule (defined in the method), no new expression was found.

So, for both pathologies, the method using the Major set query (i.e. using only the expressions given by the experts) turned out to be the best one.

The "Extended set" queries gave the same results either because the added expressions were not present in the clinical reports, or because, according to our method of semantic extension, no new expressions were added in the query.

Concerning the "context set" query, for both pathologies, the new generated cases were only FP that is noise.

Comparison with the DRG system

We compare here the best method (i.e. "the major set" using only the expression provided by the experts) versus the DRG query method.

In order to analyze comparable populations, we only consider the hospitalized patients (inpatients).
We calculated the number of patients corresponding to the following sets (Figure 3 illustrates the relationships between these subsets):
Set A =C+D: cases retrieved by the full-text method amongst the inpatients
Set B =C+E: cases retrieved by the DRG methods
Set C: cases retrieved by both DRG and full-text methods
Set D: cases retrieved amongst the inpatients and only by full-text but not by DRG method
Set E: cases retrieved only by the DRG method
Ratio C/B: % of coverage of full-text / DRG method

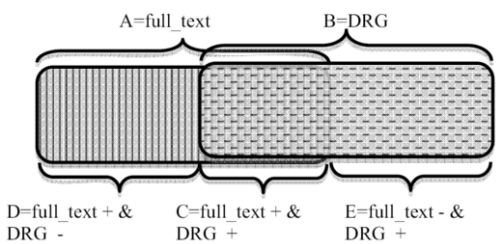

Figure 3-Set definition

The results (in number of patient terms) are presented in the following table for each pathology. The number of true positives and false positives are given for the discordant sets (sets D and E)

Table 1 -Comparison of full text vs DRG searching methods for spina bifida and hypospadia.

Sets	Spina bifida	Hypospadia
A	10	98
B	6	88
C	5	85
D	5 (3TP and 2 FP)	13 (9 TP and 4 FP)
E	1 (1 FP)	3 (2TP and 1FP)
C/B	83%	96,5 %

Concerning the set D *(full text + and DRG -)* and for both pathologies, our system found 12 new true positive cases. These cases weren't retrieved in the DRG database, because they were either not coded, or the DRG summaries didn't contain any code relative to the malformations.
Inversely, our system found 6 false positive cases. The reason was the same as the one quoted above (e.g.: expressions associated with a negation, or found in the familial history, etc).

Concerning the set E *(DRG + and full text -)* and for both pathologies, the DRG method found 2 True Positive cases. It turned out that these cases were coded from manuscript notes, or from the paper-based medical record, so no electronic clinical document about these children was present in the hospital information system.

Inversely, 2 cases were false positives and where clearly due to miscoding.

Added value of the full text method:

Compared to the DRG system, which is focused only on inpatients population, our system, gives in addition access to the outpatient population. The new cases found from this population are summarized in the following Table 2.

Table 2-New cases found with the Full text method from the outpatient population.

	Cases found from outpatients population	TP	FP
Spina bifida	2	2	0
Hypospadia	19	13	6
Total	21	15 (68%)	6 (27%)

Discussion

Our system aims to help researchers retrieve relevant information amongst a huge number of documents. The results are quite good according to the very good overlap between the full-text and the DRG coded sources (83% to 96,5%).

As we dealt with very rare pathologies, we were not able to calculate classical measures of the search engine efficiency (especially recall, Fall-out or F-measure). Indeed, this would have supposed to check a documents sample large enough for containing a significant number of cases, i.e. probably an enormous sample. For this study, we used the same methodology as Li and al. [8] who had carried out a similar study for searching eligible patients for clinical trials.

As the related previous study carried out by Moskovitch and al [9], our outcome tends to show clearly that a combined strategy, using both full text and structured methods, can improve the search of clinical cases for epidemiologic registries. Indeed, the use of the "birth date" metadata for selecting a subset of the clinical reports has clearly helped us to define a context (newborns aged =< 1 year old). This first selection allowed avoiding the adult cases. We plan to include in our system more metadata or structured information in order to reinforce this mechanism.

The noise is the main problem of the full-text searching method. We got a low noise level, probably because the expressions used in the queries were very specific and because we had started by selecting a subset of documents, using the metadata extracted from the HIS.

We assumed that the semantic enrichment of the queries could improve the search of news cases. It seems here that is not the case. The expressions given by the experts were sufficiently comprehensive and specific. Despite this, a terminological system could be useful to help users for designing queries, especially if the user is not a domain expert.

Unlike Moskovitch [9] we didn't find better performance by adding contextual expressions to the queries. On the contrary, these expressions have caused noise.

We were rarely confronted with lexical variations, even if the step of morphemes transforming was manual. Our system could certainly be improved by reusing methods [4] of automatically transforming natural expressions into "sub words" (or lexical units) straightly useable by the search engine.

Finally, from a technical point of view, we were careful to take into account the usability of the system. The system was developed with the Web Ajax technology and so allowed to improve the system accessibility while keeping good ergonomics, as it has been already shown [10].

For example, users could instantly open the reports and check their pertinence at a glance (the keywords were highlighted like Google cached pages [11]). Statistics were also very useful to assess if the query or a document was relevant.

We designed the system in order to be integrated to the HIS but we deliberately chose to build an independent data repository for several reasons.

First, querying such a huge set of documents would be overloading the HIS if the search engine was directly connected to it. This would compromise the health care process in an inacceptable way.

Second, we wanted to control the ETL process, which is a critical step in data warehouse building. Indeed, some information has to be cleaned or corrected before loading (e.g.: outliers, birth date, missing data, incoherent identities).

Third, we can consider extending the repository to other data sources (such as biologic tests or genomic data) and for other purposes such as searching patients eligible for a study.

Conclusion

This prototype allows to efficiently reuse clinical documents for case finding for epidemiologic purposes. As a perspective, we plan to create a regional repository that gathers information from different HIS.

Acknowledgments

We thank Delphine Rossille, Denis Delamarre, Laure-Anne Haumont, and the "Réseau Périnat 35" members for their inestimable help.

References

[1] Hutchinson CL and Menck H, Cancer Registry Management: Principles & Practice, Kendall Hunt, 2004.

[2] Rector AL, "Clinical terminology: why is it so hard?," Methods of Inf in Med, 38, Déc. 1999, pp. 239-252.

[3] Meystre SM, Savova GK, Kipper-Schuler KC, and Hurdle JF, "Extracting information from textual documents in the electronic health record: a review of recent research," Yearbook of Med Informatics, 2008, pp. 128-144.

[4] Schulz S, Daumke P, Fischer P, and Müller ML, "Evaluation of a document search engine in a clinical department system," AMIA Annual Symp Proc 2008, pp. 647-651.

[5] Hanauer DA, Miela G, Chinnaiyan AM, Chang AE, and Blayney DW. "The registry case finding engine: an automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes," J American College of Surgeons, 205, Nov. 2007, pp. 690-697.

[6] Rosier A., Burgun A., and Mabo P., "Using regular expressions to extract information on pacemaker implantation procedures from clinical reports," AMIA Annual Symposium Proceedings 2008, pp. 81-85.

[7] Cutting D., "Apache Lucene, a High-Performance, Full-featured Text Search Engine Library Written Entirely in Java," Apache Software Foundation, 2006.

[8] Li L, Chase HS, Patel CO, Friedman C, and Weng C, "Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study," AMIA Annual Symp Proc 2008, pp. 404-408.

[9] Moskovitch R, Martins SB, Behiri E, Weiss A, and Shahar Y, "A comparative evaluation of full-text, concept-based, and context-sensitive search," JAMIA, 14, Avr. 2007, pp. 164-174.

[10] Kluge J, Kargl F, and Weber M, "The Effects of the AJAX Technology on Web Application Usability," WEBIST 2007 Int Conf on Web Information Systems and Technologies, 2007, pp. 289–294.

[11]"Google Cached Pages: What Are Cached Pages? - Google Guide." Accessed Sept 2009 (url www.googleguide.com/cached_pages.html)

Address for correspondence

Dr Marc CUGGIA
UMR 936 – Faculté de Médecine de Rennes
Rue du Pr Léon Bernard 35000 Rennes France
email : marc.cuggia@univ-rennes1.fr