Disponible en ligne sur
ScienceDirect
www.sciencedirect.com
Elsevier Masson France
EM|consulte
www.em-consulte.com
IRBM 37 (2016) 109–115
IRBM

The SYNODOS Project: System for the Normalization and Organization of Textual Medical Data for Observation in Healthcare

C. Bouvry [a,b], N. Tvardik [a,b], I. Kergourlay [c], A. Bittar [d], P. Arnod-Prin [e], F. Segond [e], L. Dini [d], S. Darmoni [c], M.H. Metzger [a,b,f,*]

[a] *Université de Lyon, F-69000 Lyon, France*
[b] *Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France*
[c] *University Hospital of Rouen, CISMeF, Rouen, France*
[d] *Holmes Semantic Solutions, Grenoble, France*
[e] *Viseo Technologies, Grenoble, France*
[f] *Hospices Civils de Lyon, Hôpital de la Croix-Rousse, Unité d'hygiène et d'épidémiologie, F-69317 Lyon, France*

Received 19 January 2016; received in revised form 4 March 2016; accepted 4 March 2016

Available online 29 March 2016

Abstract

Introduction: The electronic health record (EHR) is a very important potential source of data for various areas, such as medical decision support tools, evidence-based medicine or epidemiological surveillance. Much of this data is available in text format. Methods of natural language processing can be used to perform data mining and facilitate interpretation. The purpose of this project was to develop a generic semantic solution for extracting and structuring medical data for epidemiological analyses or for medical decision-support. The solution was developed with the objective of making it as independent as possible from the field of medical application in order to allow any new user to write his or her own expert rules regardless of their area of medical expertise.

Material and methods: SYNODOS offers a modular architecture that makes a clear distinction between the linguistic rules and the medical expert rules. Different modules have been developed or adapted for this purpose: an interface between the multi-terminology server and semantic analyzer during the extraction phase, linguistic rules to extract temporal expressions, expert rules adapted to two areas of application (nosocomial infections, cancer), an interface between the engine and the linguistic knowledge base.

Results: Modular integrations were performed consecutively. The multi-terminology extractor and semantic analyzer were first interfaced during the extraction phase. Output of this data processing was then integrated into a knowledge base. A user interface to access documents and write business rules was developed. Expert rules for the detection of nosocomial infections and for the evaluation of colon cancer management have been developed.

It was necessary to develop an additional module the need for which had not been identified during the drafting of the protocol. This module aims to structure the output of the data processing described above, according to the patient's care pathway. This module is based on the writing of medical expert rules.

Evaluation indicators were obtained at different stages of the process (terminology extraction, semantic relations, data structuring, detection of events of interest).

Discussion: This project helped to highlight the value of combining different technologies (natural language processing, terminology, expert systems integration) to allow for the use of unstructured data in epidemiology. However, the need to develop an additional module of expert rules did not allow a complete and operational solution. Furthermore the multi-terminology extractor (ECMT V2) response time is too long (6 min per report). A change in technology was envisaged at the end of the project to reduce this time.

* Corresponding author at: UMR UCBL – CNRS 5558, HCL, Unité d'hygiène et d'épidémiologie, Hôpital de la Croix-Rousse, 103 Grande-Rue de la Croix-Rousse, F-69317 Lyon cedex 04, France.
E-mail address: marie-helene.metzger@aphp.fr (M.H. Metzger).

http://dx.doi.org/10.1016/j.irbm.2016.03.002

Conclusions: The originality of the SYNODOS project is the development of a single solution that integrates different technologies needed for the production of epidemiological indicators in the context of hospital activity. The project results confirm the interest but certain technological obstacles concerning the processing time need to be resolved in order to render the solution operational in a hospital environment.

Keywords: Epidemiology; Medical records systems, computerized; Natural language processing

1. Introduction

The Electronic Health record (EHR) is a very important potential source of data in fields as varied as assistance in medical decision-making, evidence-based medicine, epidemiological surveillance or data mining. However, very little data is structured and encoded in EHR to allow this type of use. Currently, only the data needed for charging the medical activity are systematically structured and encoded in the hospital medical records. Although significant efforts have been made in recent years to allow the use of the French Diagnosis Related group-based payment (PMSI) for the production of epidemiological data, various studies have shown the difficulties associated with the use of these medico-economic data for epidemiological purposes. More recent work tried to combine this data with other structured data sources such as pathological codes (e.g. ADICAP). However, most of the EHR data is available in unstructured documents (natural language), which requires the development of natural language processing (NLP) tools. The SYNODOS project aims to develop a solution that processes the natural language in medical records for use in epidemiological studies and for the evaluation of the quality of care.

Various works are ongoing to develop phenotyping methods from a secondary use of EHR data. These methods include in particular natural language processing (e.g. EMERGE project [1]). There are currently neither formal language nor standardized approaches to extract these phenotypes [1]. A model was developed in the USA (National Quality Forum's Quality Data Model) which gives a structure describing clinical concepts in a standardized format for the automatic production of quality indicators. This model was evaluated and is used in the eMERGE project [2] or by the SHARPn Consortium which is developing a platform for secondary use of EHR data, based on the use of various resources (terminology, representation models) [3].

Moreover once the phenotype is obtained by these techniques, the second step is to use the data extracted for measuring associations, selecting eligible patients etc. To our knowledge, there is no French language solution integrating all of these steps in a unique solution designed for use in hospitals. The challenge of the SYNODOS project is to successfully assemble all the technologies necessary for the utilization of this data for various purposes in hospitals. So this is a very applied research project.

The objective of the SYNODOS project is to develop a generic solution enabling semantic data mining of electronic health records and organize this medical information so that it can be used for epidemiological studies. The solution will be developed for the French language since the NLP tools and terminology resources are partially dependent on the language.

2. Material and methods

The SYNODOS project (URL: http://www.synodos.fr) brings together two public research laboratories, one expert in the field of research in Medical Informatics (CISMeF) and the other in the field of epidemiology (LBBE), and two industrial partners, one specializing in software development and language resources (Holmes Semantic Solutions) and the other in the integration of business intelligence solutions and web technologies (VISEO). The project was broken down into seven scientific tasks: 1) definition of the general architecture of the solution; 2) development of the semantic processing of medical documents; 3) interface of the multi-terminology server with the semantic analyzer; 4) development of expert rules for structuring the data in the patient care pathway; 5) development of an expert rules generator system; 6) integration of the developed modules of the solution; 7) evaluation of the solution's performances.

2.1. General architecture of the solution

The SYNODOS solution is a web application (SYNODOS-Mediator) which uses the services of two remote servers:
– the existing CISMeF multi-terminology server, which has been enriched with new terminologies related to specific project needs, in particular the SYNODOS terminology, which identified all significant terms extracted from the medical documents and absent of all terminology resources included in the server (containing over a million concepts in French, so, to our knowledge, the richest academic server in terms of health in this language).
– the specific SYNODOS semantic server, based on Holmes Semantic Solutions technologies.

Access to remote servers is done by a "hypertext transfer protocol secure" (HTTPS) so that no apparent data is transmitted. Various software was used for the development of the SYNODOS modules. The project is based on: a programming language (Java), a web framework, an application server, a relational database management system (DBMS) and a Business Rules Management System (BRMS Drools [4]).

2.2. The semantic analyzer

The semantic analyzer is a natural language processing platform developed by Holmes Semantic Solutions that combines

different types of modules implemented incrementally in the process: symbolic modules (rule-based), statistical modules, and modules based on machine learning [5].

2.3. The multi-terminology extractor of concepts

The multi-terminology portal developed by the CISMeF team contains 55 medical terminologies or ontologies, corresponding to one million medical concepts in French and 1,500,000 in English (URL: http://www.hetop.eu) [6]. From this portal, a multi-terminology extractor of concepts (ECMT V2) has been developed to index the medical terms encountered in medical documents. This processing is available in the web service (SOAP or REST). Different medical terminologies were selected to normalize the medical language as part of the project: ICD-10 (International Classification of Diseases, 10th revision), the MESH® thesaurus (Medical Subject Headings) indexing for diagnostic or symptomatological concepts, ATC (Anatomical Therapeutic Chemical Classification) for the coding of drugs, SNOMED v3.5, etc. Different mapping techniques were used in this project: (a) mapping of concepts based on the Concept Unique Identifier based on the UMLS Metathesaurus® (Unified Modelling Language System); (b) mapping based on natural language processing; (c) manual setting or supervised correspondence by CISMeF terminology experts.

2.4. Development of expert rules for structuring the data in the patient care pathway

For the purpose of the SYNODOS solution which is to use data extracted from text documents for epidemiological purposes or decision support, a simple standardization of medical concepts is not enough. Indeed, for appropriate use of extracted concepts, it is necessary to associate a temporal labeling to the concept (is the concept related to the patient's medical history or to the reason for hospital admission or to the evolution during hospitalization?) to reconstitute the patient's care pathway. It is also necessary to create relationships in order to link the concepts together: for example, in the following sentence: "we have seen a drop in gastrinemia levels," it is essential to link the concept of "gastrinemia" to that of "drop" to properly use this information. It is therefore necessary first of all to create a rule that lets you assign a concept like "gastrinemia" to the category "type of biological examination" and the concept "drop" to the category "result of the biological examination" in the fact base. These rules are based on the use of semantic data (e.g. negation like "absence" linked to a medical concept), terminologies (e.g. the UMLS semantic type of the concept) attached to each of these concepts and medical expert rules. This intermediate processing step was not initially identified in the SYNODOS protocol and has been the subject of an additional task. These expert rules were developed and integrated into a web service module (STRAUMED) by the academic laboratory (LBBE). The output format of this processing is a text file adapted to the conceptual model of the fact base [7]. This module has not been integrated into the SYNODOS solution for automatic settlement of the fact base of the SYNODOS solution within the framework of the project.

2.5. Development of a module for writing and running semantic transition rules

In parallel with the development of the STRAUMED solution and in order to compare different approaches, VISEO has developed a transition rules module for structuring medical data in the patient record. Unlike STRAUMED this module does not use expert knowledge and is based solely on metalinguistic rules built from the results produced by the semantic analyzer of HO2S. For example, if in the patient record, it is written that the patient was operated on for appendicitis in 2000 and that we are in 2015, the module uses the time information associated with 2000 to write a rule that will allow it to deduce that appendicitis relates to the "medical history" category. If the word is a subject and the word is "patient" and that the word is in French singular feminine, then the patient is female. Transition rules are used to fill the knowledge base by creating objects (facts) and inserting them into the database. The transition rules are written in pseudocode (in MVEL).

2.6. Development of a reasoning engine above the knowledge base

Once the base is filled with all the information extracted during the different phases (terminological analysis, semantic analysis, transition rules or STRAUMED rules, expert rules) it will be possible to reason about these facts. For this, it is necessary to have a tool capable of reproducing the cognitive mechanisms of an expert in a particular field. This is the role of reasoning engine or inference engine, built by VISEO. This engine performs reasoning from known facts and rules to infer new facts. This reasoning engine is based on Drools Expert tool as an expert system. So we have:

– On one hand, all the facts used to be new facts. This is the fact base. We distinguish the facts known at the beginning, named "known" and deduced facts, named "inferred". The SYNODOS fact base is made up of information known about the patient to be treated (example of "known" fact: the patient has urinary burns).

– On the other hand, all the expert rules called rule base, include expert rules defined by the doctors themselves (example of rule: if the patient has urinary burns then the patient has a urinary tract infection). Reasoning on a patient record requires the resetting of inferred facts, that is to say facts arising from a previous reasoning, if there were any.

2.7. Development of a module for description and interrogation with the web semantic technologies

To evaluate the potential contribution of the Semantic Web technologies in an expert system, VISEO has developed a module comprising the description of the different concepts present in the conceptual model provided by the LBBE and the different

relations linking these concepts, using the OWL format, a set of written rules corresponding to the transition rules in SWRL and a search engine of this knowledge base in SPARQL.

2.8. Development of an expert rules generator system and integration of the different modules

The integration of the different modules (except STRAUMED) was performed by VISEO to obtain a single data access interface. The user interface allows the importing of the data source, terminological semantic processing and accessing of the knowledge base to infer new facts called "inferred facts" by combining "known facts". The solution also handles the access rights of the users and their level of access to the solution.

2.9. Learning and performance evaluation of the SYNODOS solution

Two corpora of textual data are used in the development of the SYNODOS IT solution. The first corpus was formed as part of a previous research project ALADIN-DTH [8] to develop a semantic detection tool of nosocomial infections in medical documents. The method of selection and annotation of these reports has been described in detail in another publication [9]. This corpus of textual data was divided by random selection of files into two sets for the SYNODOS project: a training set ($n = 91$) and a test set ($n = 120$). The annotation aimed to provide a reference for learning and assessment of the settlement of the fact base and to detect the nosocomial infections. A second corpus of textual data was established as part of this new project and concerns a very different medical field: the management of colon cancer diagnosis. The choice of this medical field aimed to assess the generic nature of the solution. The new corpus is made up of 350 medical documents, selected in the regional cancer center of Rhône-Alpes among patients under care for a new colon cancer. The corpus consists of textual documents tracing the management of colon cancer: hospitalization report, the pathology report, imagery report, consultation report, operative reports, multi-disciplinary consultation report). The annotation has provided a reference for learning and assessment of the settlement of the fact base and for the identification of colon cancer (35 cases) and diagnostic elements needed to calculate the management time limits. The annotation is semi-automatic and carried out using an application developed by LBBE on R and MS Access software. This annotation can reconstruct the patient's care pathway. The annotation base of this application is based on the conceptual model of the fact base [7].

Assessment of SYNODOS solution was carried out in various stages:

1) Evaluation of semantic processing of textual documents.
2) Evaluation of terminology processing of textual documents vs. manual annotation.
3) Evaluation of the settlement of the fact base (treatment 1) + 2) + evaluation of the transition rules STRAUMED vs. manual annotation (150 medical files).
4) Evaluation of facts inferred by STRAUMED made in two application areas: healthcare associated infections (100 cases) and time limit of the diagnosis of colon cancer (30 cases).

The results obtained in terms of sensitivity and specificity will be considered the final performance of the tool for the evaluation of semantic relations [10], evaluation of the settlement of the fact base, and evaluation of the detection of nosocomial infections. To estimate the time limit of the management of colon cancer diagnosis, the time estimated by the automatic method will be compared to the time estimated by the manual method. The evaluation indicator used is the average difference in time limit for the period of colon cancer diagnosis reached between SYNODOS solution and manual reference method.

3. Results

3.1. Development of the architecture and modules of the solution

Modular integrations were performed consecutively. The multi-terminology server and semantic analyzer were first interfaced during the extraction phase. Output data of this processing were then integrated into a knowledge base. A user interface for accessing documents and writing expert rules was developed. The SYNODOS solution must be installed in the DMZ of the secure network of the health facility, or on a server of an authorized health data host in order to benefit from any safety-related processing of medical data. The solution is available in a virtual machine, deployable on the institution's hypervisor. A data import module from the hospital information system retrieves the metadata and text data "pushed" by the hospital information system. The module then proceeds to an automatic anonymization of personal data (e.g. personal names, place names, phone numbers, addresses, email address, etc.) to enable exchanges with remote web services and also to allow use of epidemiological data. This module also allows the re-identification of the patient after treatment for certain types of use (e.g. in-hospital nosocomial infection surveillance, where the patient record once spotted by the SYNODOS solution must be validated by the infection control team) (cf. Fig. 1).

3.2. Development of a de-identification tool

To protect the identity of the persons mentioned in medical records, Holmes Semantic Solutions has developed a de-identification system. This system masks sensitive information in the texts that could reveal the identity of those involved, including the full names, dates and addresses. Sensitive information found in the texts are replaced by typed tags (ex. PERSON, PLACE, DATE, etc.). The treatment is based upon a combination of lexical labels and rules in the form of regular expressions applied to sequences of "words".

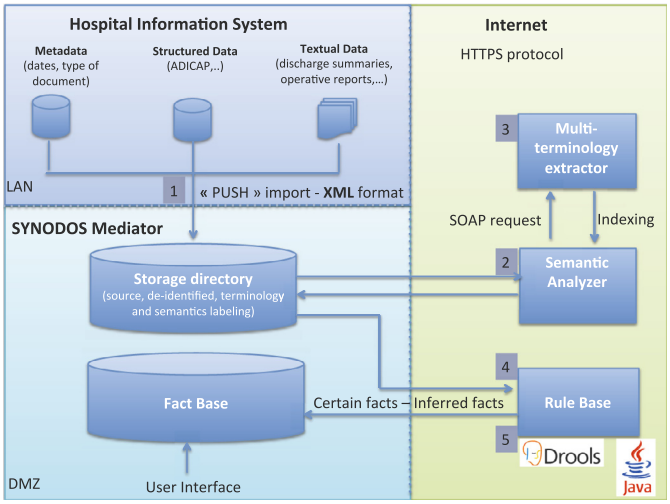

Fig. 1. Architecture of the SYNODOS solution.

3.3. Development of a semantic analysis system

Holmes has also designed and developed a semantic analysis system. This system detects and annotates the semantic relationships between words of the sentence. The output of this system is a graph in which the nodes are the words and the arcs are the semantic relationships between words. Semantic analysis is based on a rich graph transformation grammar, which itself is based on the many linguistic pre-processing performed by the platform Holmes (sentences detection, tokenization, morphosyntactic labeling, stemming, lexical analysis, parsing dependencies) and the labeling provided by MeSH terminology.

3.4. Development of conceptual model of the knowledge base

The LBBE developed the conceptual model of the knowledge base. This conceptual model was developed to meet the generic objective of the solution [7].

3.5. Development of a user interface

Coordination interface of the different modules was produced by VISEO. This interface gives access according to the user's profile, to the patient record (anonymized, analyzed and structured according to the model defined on the interface), visualization and enrichment of patient records, writing and running expert rules to infer conclusions. This interface was developed in Wicket (Java Web Framework), the data is stored in a relational database (PostgreSQL), data access and security is managed via the Spring Framework.

3.6. Development of the STRAUMED web service

This module aims to restructure a posteriori patient care trajectory done for the management of the health event of interest (e.g., surgical management for installation of a hip prosthesis and 12 months postoperative monitoring). The module allows to classify and temporally label the health events in this care pathway (is it a medical history? the reason of hospital admission?...), to distinguish textual repetitions related to the same health event of those corresponding to real repetitions of the same health event in the patient's trajectory.

3.7. Development of the semantic transition rules

This module allows the writing and execution of transition rules. These transition rules link semantic rules produced by HO2S and expert rules provided by medical users. These rules, which are based on non-medical information, are intended to complete the data model provided by the LBBE.

3.8. Development of a reasoning engine

This module allows reasoning from known facts and rules to infer new facts.

3.9. Development of medical expert rules

Expert rules for the detection of nosocomial infections and automatic calculation of time limit for the management of colon cancer diagnosis have been developed by the LBBE. These rules have been incorporated into the STRAUMED solution.

3.10. Evaluation indicators

Evaluation indicators were obtained at different stages of the process (terminology processing [11], semantic relations, structuring, detection of events of interest).

4. Discussion

The progress of computerization of medical records in France is an important element for the development of the NLP tools for epidemiological use. The financial incentive offers to private practitioners by the French Health Insurance (ROSP) allowed an increase of the availability of the "medical summary" in the EHR. This medical summary allows to have a global view of the patient, re-evaluate its management, schedule management in coordination with other health professionals involved in patient monitoring. At the end of 2013, 78% of private practitioners (almost 8 out of 10 practitioners) had this summary in their EHRs [12]. At the hospital level, the High Authority for Health (HAS) measures the quality of the patient record (IPAQSS: Indicators for Improving the Quality and Safety of Care). On a representative sample of patient's records at national level (analysis of 130,934 records of the year 2013), 46% of medical observations and 69% of discharge summaries were computerized [13]. These figures are equivalent to those of the USA where the adoption of "meaningful use" policies has allowed a major increase in the computerization of medical records to 70% [14]. This significant increase of medical records computerization provides a glimpse of the use of NLP technologies for epidemiological use [15]. This type of application is expanding in various domains of epidemiology, such as cancer [16], bodyweight and other vitals

signs data [17], chronic diseases [18], psychiatry [19], nosocomial infection surveillance [8,20], identifying patients for cohort inclusion [21]. The number of publications in the field illustrates the relevance of this new topic. For example requesting on PubMed ("natural language processing" in all fields), the number of publications increased from 50 in the year 2000 to 300 since 2013.

However, this domain of research shows an explosive development mostly in the USA [18], which can be seen as the result of the significant investment in research at national level notably through 4 wide grants called "the Strategic Health IT Advanced Research Projects (SHARP)". These grants are funded by the Office of the National Coordinator for Health Information Technology to facilitate access to data within EHRs [14]. One of these four grants focused on secondary use of EHR Information. The finding was as follows: "Traditionally, a patient's medical information, such as medical history, exam data, hospital visits and physician notes, are inconsistently stored in multiple locations, both electronically and non-electronically. With a vision of solving this issue, the project aims to efficiently leverage EHR data to improve care, generate new knowledge, and address population needs through the secondary use of EHR data" [22]. In addition, the search strategy developed in the USA is based on the open source model, sharable resources and software [23] like the open-source Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) [24]. This strategy promotes the sharing of knowledge between research teams.

In France, the effort sought on secondary use of EHR is still very embryonal and funding sources difficult to obtain. Few research teams are working on these types of use [25,26]. The reasons include: the doubts still important especially in medical environment on the quality of data that could be obtained by such methods and also due to some apprehension for the patient privacy using full clinical text [27]. Removing these concerns depends on the achievement of rigorous evaluations of these experimentations. This assessment is ongoing in the SYNODOS project. The preliminary evaluation results showed that NLP could be a valuable additional source of data for the routine surveillance of nosocomial infections.

The SYNODOS project is to our knowledge the first project financed by a French public research program, that aims to process EHRs for secondary epidemiological use. The strength of the SYNODOS project was to highlight the feasibility of combining different technologies (natural language processing, terminology, expert systems integration) to allow the epidemiological exploitation of unstructured data. However, the project has faced several technical limitations that have failed to consider a routine installation in the pilot healthcare facility. Indeed, as mentioned by Rea et al., the secondary use of EHR data requires robust information models for storing and processing that information [28]. For the development of the SYNODOS information model, it was necessary to build an additional module of expert rules that associates a temporal labeling to the concept. This additional work did not permit to integrate it in the complete data processing during the project duration. Furthermore the multi-terminology server (ECMT V2) response time was too long (6 min per report) for considering the industrialization of the solution. A change in technology was envisaged at the end of the project to reduce this time (NoSQL vs. SQL) and will be part of the future development of the project. Another perspective will be to add in the solution the processing of structured data (DRG codes, microbiology codes, prescriptions) in the solution and enrich the algorithms combining data providing from unstructured and structured sources [29].

5. Conclusion

The originality of the SYNODOS project is the development of a single solution that integrates different technologies needed for the production of epidemiological indicators in the context of hospital activity. The project results confirm the interest but certain technological obstacles need to be resolved to allow the use in a medical environment, in particular the process time limit should be reduced using new technologies for the multi-terminology server.

Acknowledgements

This work was funded by the French National Research Agency, as part of a TECSAN program (SYNODOS Project ANR-12-TECS-0006).

References

[1] Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. Genet Med 2013;15(10):761–71.

[2] Thompson WK, Rasmussen LV, Pacheco JA, Peissig PL, Denny JC, Kho AN, et al. An evaluation of the NQF quality data model for representing electronic health record driven phenotyping algorithms. In: AMIA annu symp proc. 2012. p. 911–20.

[3] Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc 2013;20(e2):e341–8.

[4] Hliaoutakis A, Zervanou K, Petrakis E. The AMTEx approach in the medical document indexing and retrieval application. Data Knowl Eng 2009;68(2):380–92.

[5] Dupuch M, Segond F, Bittar A, Dini L, Soualmia L, Darmoni S, et al. Separate the grain from the chaff: make the best use of language and knowledge technologies to model textual medical data extracted from electronic medical records. In: 6th language & technology conference: human language technologies as a challenge for computer science and linguistics. 2013.

[6] Grosjean J, Soualmia LF, Bouarech K, Jonquet C, Darmoni SJ. An approach to compare bio-ontologies portals. Stud Health Technol Inform 2014;205:1008–12.

[7] Gicquel Q, Tvardik N, Bouvry C, Kergourlay I, Bittar A, Segond F, et al. Annotation methods to develop and evaluate an expert system based on natural language processing in electronic medical records. Stud Health Technol Inform 2015;216:1067.

[8] Proux D, Hagège C, Gicquel Q, Kergourlay I, Pereira S, Rondeau G, et al. ALADIN: Développement d'un outil sémantique d'analyse des documents textuels médicaux pour la détection d'infections associées aux soins IRBM. IRBM 2012;33(2):137–42.

[9] Metzger M, Gicquel Q, Kergourlay I, Cluze C, Grandbastien B, Berrouane Y, et al. Codage standardisé de données médicales textuelles à l'aide d'un serveur multi-terminologique de santé: exemple d'application en épidémiologie hospitalière. In: Degoulet P, Fieschi M, editors. Systèmes d'information pour l'amélioration de la qualité en santé: comptes rendus des quatorzièmes Journées francophones d'informatique médicale. Paris, France: Springer-Verlag; 2011. p. 109–20.

[10] Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 2012;65(3):343.

[11] Tvardik N, Bouvry C, Kergourlay I, Darmoni S, Metzger M. The terminology needs for evaluation of care pathways through electronic medical records. In: WHO-FIC network annual meeting. 2015.

[12] Caisse Nationale de l'Assurance Maladie. La rémunération sur objectifs de santé publique, deux ans après : des progrès significatifs en faveur de la qualité et de la pertinence des soins (The remuneration of public health objectives, two years after: Significant Progress in Favor of the quality and appropriateness of care). Caisse Nationale de l'Assurance Maladie; 2014.

[13] Haute Autorité de Santé. Indicateurs pour l'amélioration de la qualité et la sécurité des soins qualité de la tenue du dossier patient en médecine, chirurgie, obstétrique. Haute Autorité de Santé; 2014.

[14] Chute CG. Invited commentary: observational research in the age of the electronic health record. Am J Epidemiol 2014;179(6):759–61.

[15] Metzger M-H, Durand T, Lallich S, Salamon R, Castets P. The use of regional platforms for managing electronic health records for the production of regional public health indicators in France. BMC Med Inform Decis Mak 2012;12:28.

[16] Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009;42(5):937–49.

[17] Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q. Regular expression-based learning to extract bodyweight values from clinical notes. J Biomed Inform 2015;54:186–90.

[18] Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885.

[19] Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V, et al. Validation of electronic health record phenotyping of bipolar disorder cases and controls. Am J Psychiatr 2015;172(4):363–72.

[20] Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306(8):848–55.

[21] Carrell DS, Halgrim S, Tran DT, Buist DS, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol 2014;179(6):749–58.

[22] Office of the National Coordinator for Health Information Technology – Department of Health. https://www.healthit.gov/policy-researchers-implementers/secondary-use-ehr-data, 27.02.2016.

[23] Wieneke AE, Bowles EJ, Cronkite D, Wernli KJ, Gao H, Carrell D, et al. Validation of natural language processing to extract breast cancer pathology procedures and results. J Pathol Inform 2015;6:38.

[24] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13.

[25] Bouzille G, Sylvestre E, Campillo-Gimenez B, Renault E, Ledieu T, Delamarre D, et al. An integrated workflow for secondary use of patient data for clinical research. Stud Health Technol Inform 2015;216:913.

[26] Pham AD, Neveol A, Lavergne T, Yasunaga D, Clement O, Meyer G, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinform 2014;15:266.

[27] Carrell DS, Halgrim S, Tran DT, Buist DS, Chubak J, Chapman WW, et al. Carrell et al. respond to "Observational research and the EHR". Am J Epidemiol 2014;179(6):762–3.

[28] Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform 2012;45(4):763–71.

[29] Bouzbid S, Gicquel Q, Gerbier S, Chomarat M, Pradat E, Fabry J, et al. Automated detection of nosocomial infections: evaluation of different strategies in an intensive care unit 2000–2006. J Hosp Infect 2011;79(1):38–43.