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

doi:10.3233/978-1-60750-588-4-1085

Design and evaluation of a semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EU-ADR project

Paul Avillach[a,b], Michel Joubert[b], Frantz Thiessard[a] , Gianluca Trifirò[c], Jean-Charles Dufour[b], Antoine Pariente[d], Fleur Mougin[a], Giovanni Polimeni[c], Maria Antonietta Catania[c,] Carlo Giaquinto[e], Giampiero Mazzaglia[f], Carla Fornari [g], Ron Herings[h], Rosa Gini[i], Julia Hippisley-Cox[j], Mariam Molokhia[k], Lars Pedersen[l], Annie Fourrier-Réglat[d], Miriam Sturkenboom[m], Marius Fieschi[b]

[a]LESIM, ISPED, Uni. Bordeaux 2, France (FR)
[b]LERTIM, Faculté de Médecine, Uni. de la Méditerranée, Marseille, FR
[c]IRCCS Centro Neurolesi "Bonino-Pulejo", Messina, Italy (ITA)
[d]INSERM U 657, Uni. Victor Segalen Bordeaux 2, FR
[e]Pedianet - Societa' Servizi Telematici SRL, ITA
[f]Health Search - Italian College of General Practitioners, ITA
[g]Centre on Public Health - University of Milano-Bicocca, ITA
[h]PHARMO Coöperation UA, The Netherlands (NL)
[i]Regional Health Agency of Tuscany, Florence, ITA
[j]Uni. of Nottingham, UK
[k]London School of Hygiene & Tropical Medicine, UK
[l]Aarhus Uni. Hospital, Århus Sygehus, Denmark
[m]IPCI - Department of Medical Informatics, Erasmus Uni. Medical Center, Rotterdam, NL

Abstract

The overall objective of the EU-ADR project is the design, development, and validation of a computerised system that exploits data from electronic health records and biomedical databases for the early detection of adverse drug reactions. Eight different databases, containing health records of more than 30 million European citizens, are involved in the project. Unique queries cannot be performed across different databases because of their heterogeneity: Medical record and Claims databases, four different terminologies for coding diagnoses, and two languages for the information described in free text. The aim of our study was to provide database owners with a common basis for the construction of their queries. Using the UMLS, we provided a list of medical concepts, with their corresponding terms and codes in the four terminologies, which should be considered to retrieve the relevant information for the events of interest from the databases.

Keywords:

Drug toxicity, Semantics, Medical records, Indexing, Information storage and retrieval, UMLS.

Introduction

The medical information gathered during clinical follow-up can be reused for a wide variety of related purposes from medico-economic and epidemiological applications to clinical alerts [1, 2]. This information, collected at every stage of the healthcare process, is often registered as free text and is increasingly coded by using one or several specific medical terminologies. Though time-consuming, choosing an appropriate code to describe medical information has the advantage of clarifying unambiguously the significance of the information. Information coding allows automated processing of the information and facilitates semantic interoperability between different information systems. Medical information with appropriate coding can be transmitted, interpreted and processed more easily by different systems and thus enables sharing and reuse of the data among information systems [2, 3].

In the area of drug safety, information sharing could enhance the current spontaneously reported information on adverse drug reactions (ADRs), as reporting rate is far from optimal. Underreporting is high, and it is estimated that only 4% of ADRs are reported through this channel [4]. Therefore, safety signals may be detected too late, as was recently highly debated after the rofecoxib (Vioxx®) withdrawal due to concerns regarding cardiovascular safety. It has been recognized that additional complementary systems are necessary [5, 6], which could profit from the wide availability of health care databases throughout Europe. The use of several medical databases for signal detection could overcome the underreporting problems existing with the current system and may detect signals faster and/or earlier.

From this rationale, the European EU-ADR project has been launched. The aim of this project is to design, develop and validate a computerised system to process data from eight electronic healthcare databases and biomedical knowledge databases for the early detection of safety signals [7]. Each of the eight healthcare databases contains information which is coded according to different terminologies, in different languages, and has its own specific characteristics, depending on its initial objective and local function (administrative, healthcare, medical records, etc.) [2]. Given the structural and semantic heterogeneity of the databases involved in the project, it is impossible to construct a single, completely reusable query system on the different databases, to undertake the same search for each event and drug.

Objective: The aim of this research was to provide a method for extracting relevant information contained in the various databases regarding the event under study and the drugs taken in the population. Our task also entailed a search for greater coherence to enhance our method of extracting information from the different databases. The method described was evaluated through a process using analogy as a logical tool.

Materials and Methods

Concept selection

Different terminologies are used to code the clinical events in the eight databases. Thus, a common basis was required in order to set up queries (adapted to target databases) built-on a shared semantic request. The aim was to provide researchers with a list of medical concepts and associated terms that they must use to identify the events being investigated in their respective databases. A unique query cannot be performed to extract information from the databases used since, intrinsically, different terminologies are used. We built a shared semantic foundation for the eight databases[8]. The constituents of this shared foundation are UMLS[9] concepts (grouping together terms from different terminologies with the same medical meaning) and not terms.

Medical terminologies are structured in the form of lists of concepts[1], generally set out in a hierarchical way. A concept can be defined in many ways since the terms[2] defining it come from different languages and, furthermore, because each language can use distinct synonymous terms to describe the same concept.

The eight databases involved in the EU-ADR project contained information stemming from the medical files of more than 30 million European citizens (Table 1). Four terminologies are used to describe the events: the «international statistical classification of diseases and related health problems» (ICD9-CM and ICD10), the «international classification of primary care» (ICPC) [10] and the READ CODE (RCD) classification[11]. Seven databases use the Anatomical Therapeutic Chemical (ATC) system[12] to code drugs. In the QRESEARCH database, drugs are initially coded using the British National Formulary (BNF) [13], but a mapping between the BNF codes and the ATC classification has been established by the QRESEARCH team. The Unified Medical Language System® (UMLS®) [14] is a biomedical terminology integration system handling more than 150 terminologies. The four terminologies used in the EU-ADR project are integrated in the UMLS. The Metathesaurus® consists of a central vocabulary comprising roughly 1.8 million concepts connected by more than 3.75 million relations. A UMLS concept is identified by a Concept Unique Identifier (CUI) and describes a single medical concept that can be expressed using different synonyms (terms).

To develop our method, we initially studied the event «upper gastrointestinal bleeding» (UGIB) which has a complex medical definition and thus raises difficulties when searching for it in a standardised way in databases. A similar approach is used for the other twenty-three events that have been identified to be of primary importance in the EU-ADR project [7].

Our method is based on the projection of UMLS concepts in the targeted terminologies. The whole method consists of the following: 1) literal definition of event, 2) identifying the UMLS concepts for the event; 3) discussion about concepts with database's administrators; 4) term identification for each concept in each terminology.

Regarding step 1, a «broad» definition approach was initially adopted. The definitions were drawn from clinical reference manuals and were validated by gastroenterology specialists.

Regarding step 2: for each literal expression matching the inclusion criteria listed in the definition of the event, we performed an automated search using Knowledge Source Server (UMLSKS, version 2008AA), in order to identify the UMLS concept and all the terms used to designate the concept in the four terminologies of the project. When this automated search failed to identify terms corresponding to a given concept in one of the terminologies studied, we undertook a manual search in the concerned terminology to identify the potential terms of interest.

In step 3, database's administrators were asked to follow their 'usual approach' to query their databases and compare the criterions they have used with the concepts and terms provided at the step 2 issue. The relevance of each concept, term and corresponding code were discussed via the EU-ADR consortium Internet forum, conference calls and plenary meetings. Thus a consensual list of items (codes, terms and free text expressions) was set up.

[1] A concept is a unit of thought [ISO 5963]

[2] A term is the designation of a given concept in a language in its linguistic formulation [ISO 1087]

Table 1- Description of the eight databases

Database	Terminology		Free text	Type of data*	Patients†
	Event	Drug			
Pedianet – Italia (ITA)	ICD9-CM	ATC	yes (ITA)	EHR	C
Health Search (ITA)	ICD9-CM	ATC	yes (ITA)	EHR	A/C
Lombardy Regional DB (ITA)	ICD9-CM	ATC	no	SDC, D	A/C
Tuscany Regional - ARS (ITA)	ICD9-CM	ATC	no	SDC, D, L, M	A/C
IPCI – Netherlands (NL)	ICPC	ATC	yes (NL)	EHR	A/C
PHARMO (NL)	ICD9-CM	ATC	no	SDC, P,L, M	A/C
QRESEARCH United Kingdom (UK)	RCD	BNF/ATC	no	EHR	A/C
Aarhus University Hospital DB (DK)	ICD10	ATC	no	SDC, D, L, M	A/C

*EHR (Electronic Health Record), SDC (Standardized Discharge Codes), D: Dispensation, L: Laboratory, M: Mortality, P : Prescription. † C: Child, A : Adult

In step 4, the list of items from different terminologies and languages were provided to the databases administrators. Every listed item had necessarily to be present in their query. The list of items that we provided was non-restrictive. Database administrators were free to add all additional criterions/terms that they would consider relevant in order to recover the UGIB event from their database, providing that these criterions offered a new way of describing the selected concepts. Hence, when a given code had "children" (i.e. a more accurate description), the query also had to include the "descendants" of this code that were considered relevant for the retrieval of the information.

Evaluation process

To evaluate this semantic-based method, we needed a knowledge source that could confirm that the events retrieved by the databases administrators correspond really to UGIB. Unfortunately, manual evaluation on a sample of events is a lenghty and expensive process that is not scheduled in the EU-ADR project. On the other hand, we found that sensitivity of the semantic-based method could be estimated in a fast, albeit indirect, way. By exploring the MEDLINE National Library of Medicine's (NLM) database. As in the patient medical records, the full-text versions of the articles indexed in MEDLINE include medical concepts. The MEDLINE notices created by the NLM indexers can thus be considered by analogy as discharge summaries where an effort of selecting an appropriate MeSH code to resume the full content is done. A subset of the MEDLINE database, identified through a classical Pubmed search, can then constitute a validation set. The sensitivity of the search methodology for the UGIB events could be evaluated by performing the search through the MEDLINE notice. The manual examination of the full-text versions of the articles included in the identified subset constitute the "gold standard" for the assessment of the presence of UGIB events in the indexed papers. In order to constitute the validation set using the Pubmed website, we entered the query "upper gastrointestinal bleeding" with the limits: "links to free full text" AND "Humans" AND ("English" or "French"). Using the "Humans" descriptor restricted the selection to notices with MeSH descriptors and then avoids the notices not yet indexed in MEDLINE. Restriction to English and French languages was due to the language expertise of the workgroup performing the manual examination. A random selection of 20% of the notices was done. We then examined the full-text versions of the selected papers to confirm that the notion of UGIB was present.

We compared three methods for retrieving the event UGIB within the notices validation set. The two first ones were initial methods written by the database owners of Lombardy (in ICD9-CM) and Aarhus (in ICD10) and the last one was our proposal.

Because the MEDLINE database is coded in MeSH terms, we first had to translate the ICD9-CM and ICD10 codes used by database owners into MeSH terms. Several steps were conducted: 1) we used the UMLS Metathesaurus to recover the UMLS CUIs associated with the ICD9-CM and the ICD10 codes; 2) we used the tool developed by Bodenreider[3] [15] to obtain only MeSH codes from the resulting CUIs. 3) we extracted the English preferred term for each MeSH code in the Metathesaurus (because there is no MeSH codes in MEDLINE, only the preferred terms); 4) Finally, we checked, for each of the three methods, the ratio of the notices retrieved (within the validation set, our gold standard) when using MeSH terms previously obtained. We then computed the retrieval sensitivity in our subset test of MEDLINE notices.

Results

Concept selection

For the event UGIB, a broad clinical definition was created including the following conditions: Upper gastrointestinal

[3] http://mor.nlm.nih.gov/download/rtm

haemorrage, Oesophagal haemorrage, Gastrointestinal haemorrhage, Bleeding from peptic ulcer, Haematemesis/blood vomiting and Melaena. We then devised a table listing all the UMLS concepts matching the inclusion criteria. Upon evaluation of the usual behaviour of the databases and the provided concepts, the concepts and terms were adapted. These included: *Upper gastrointestinal hemorrhage, Gastrointestinal Hemorrhage, Hematemesis, Melena, Esophageal bleeding, Acute {gastric|duodenal|peptic} ulcer with hemorrhage (and/or) perforation, Acute gastrojejunal ulcer with haemorrhage, without mention of obstruction, Acute gastrojejunal ulcer with hemorrhage and perforation, Acute gastrojejunal ulcer with hemorrhage, Atrophic gastritis, with haemorrhage, Other specified gastritis, with hemorrhage, Unspecified gastritis and gastroduodenitis, with hemorrhage, Acute gastric mucosal erosion.*

Subsequently all codes and terms were provided. As an example, the concept *"Haematemesis"* is coded "578.0" in ICD9-CM, "K92.0" in ICD10, "D14" in ICPC and "J680" in RCD. Some of the corresponding terms (useful for search in the clinical notes that are registered as free text) are as follows: "Ematemesi/vomito sanguinolento" in Italian, "Bloed; braken" in Dutch, "Haematemesis/vomiting blood" in English, etc.

Evaluation process

We performed a broad search on Pubmed, looking for possible citations of a wide set of gastroenterological disorders. From the resulting 1,044 MEDLINE citations, we extracted a random selection of 20% of them (n=208), only 199 of which were working with Pubmed LinkOut (the internet link to retrieve full-text articles). After full-text revision, we classified 151 articles as containing the UGIB notion. So 48 articles did not contain the UGIB notion but other medical notions (lower GIB for example). These 151 notices constitute the test set for our evaluation of the three extraction methods. The number of notices retrieved by each method is described in Table 2. Our proposal of a common semantic base method retrieved 107 notices out of a total of 151 notices with the event UGIB present in the full-text. The sensitivity is the percentage of retrieval in our subset test of MEDLINE (not in all MEDLINE).

Table 2 - Number of notices retrieved for each method

	Gold Standard: presence of UGIB in Full text article	**Sensitivity (%)**
Lombardy's initial method	100	*66.2*
AARHUS's initial method	108	*71.5*
Common Semantic based method	107	*70.9*
Total	**151**	

As a result, we can observe that the common semantic-based method is nearly as sensible as the better between the other two, that is, Aarhus accepting to delete some of its customary concepts on the ground of homogeneity with other databases did not lead to a dramatic fall in sensitivity, whereas Lombardy's sensitivity improved.

Discussion

The process we implemented allowed the homogeneous identification of events in various European databases. It is based on UMLS concepts. This foundation enabled us to propose a list of terms along with their codes and strings in order to standardise queries and, thus, extractions from the eight databases participating in the EU-ADR project. The discussion and harmonisation process led to additional concepts to be included in the list, the definite version included a total of 21 potentially usable concepts for the coding of the UGIB event in the databases. The databases were heterogeneous regarding the terminology used, the presence, or not, of free text data (used in two languages: Italian and Dutch), and the type of data they contain (medical record and claims databases). The UMLS may be helpful to map between these heterogeneous databases and to promote semantic interoperability among these databases. The sensitivity of the retrieval in our validation set is estimated by an analogy method to be around 70%, similar to those of initial queries from the participating databases.

Our process creates an homogeneous set of relevant terms/expressions useful for requesting heterogeneous databases, but does not exhaustively describe the event extraction. First, databases with free-text must perform a local algorithm, based on local information, that avoids ambiguities in the use of free-text. Second, databases with hospital discharge records must agree on whether looking for the UMLS concepts only in primary or also in secondary diagnosis fields. Third, all databases must specify in which health sources they are looking (e.g. only hospitalizations *or* both hospitalizations and deaths). Finally, some health sources contain information that is not corresponding to UMLS concepts: for example, the use of laboratory test results involves identifying a concept by its biological results and not by its name or its place in a nosologic description, and this identification might be crucial for some events (e.g. acute kidney failure). A more detailed terminology mapping instrument must be developed that further describes event extraction.

When common concepts are translated into database-specific codes, it is important to consider when analysing results that each database is confined to the granularity of its terminology. SNOMED CT for instance, can be coded by the user with a high level of granularity whereas ICD is much less granular. Hence, the level of information acquired is not always identical.

Concerning the evaluation process, the analogy between a medical doctor summarizing apathology in a clinical or claim database and a NLM indexer which selects the appropriate code to resume the full content of an article is new and needs confirmation. Secondly, the projection from ICD9-CM or ICD10 codes to MeSH terms could result in some classification bias, according to the numerous steps of the process. In order to compare the three methods, we used the same projec-

tion process. The different biases should then have affected equally the different methods. The EU-ADR workgroup selected 21 concepts for the search of UGIB. To create the selection of citations for the gold standard, we had to work on a small subset of MEDLINE focusing on gastroenterological disorders and to examine the full-text versions of the articles from this subset for the presence of UGIB. We decided to use only the concept "upper gastrointestinal bleeding" and to select only a sample of the relevant papers for the evaluation of the identification of the event UGIB. This does not constitute a major issue as our objective was not to determine the prevalence of the event UGIB in MEDLINE, but to constitute a validation set for our method. .Remark that the same technique cannot be used to estimate the specificity of the common semantic-based method because one can not have the confirmation of absence of the concept UGIB in the full text version of the articles indexed in MEDLINE and not identified by a classical Pubmed search for UGIB.

Conclusion

The projection of UMLS concepts in the terminologies and the additional manual adjustments have been exploited for the four terminologies used in our study. This enabled us to provide a shared semantic basis for the creation of queries adapted to the heterogeneous electronic health record databases we exploited. The list of concepts, accompanied by the list of associated codes, and strings in free text text (where applicable) have been used by the database administrators as a base to build queries designed to retrieve information from their database using the appropriate terminology. We provided evidence that the homogenization of concept selection does not worsen the sensitivity of each database. This method will be used for the other events selected for the EU-ADR project. The extraction of the same medical concepts from the eight databases will enable biostatisticians working on the project to use comparable data from different databases, with respect to the definition of the events sought despite of the high level of heterogeneity between the databases.

Acknowledgments

This research received funding from the European Union Community in the framework of the FP7/2007-2013 convention governing subsidy nº 215847 - the EU-ADR project. The authors also wish to thank the NLM for making UMLS available free of charge and Mr George Morgan for his translation.

References

[1] Cimino JJ. Collect once, use many. Enabling the reuse of clinical data through controlled terminologies. J Ahima. 2007 Feb;78(2):24-9.

[2] Giannangelo K. Making the connection between standard terminologies, use cases and mapping. Him J. 2006;35(3):8-12.

[3] Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform. 2008 May;77(5):291-304.

[4] Begaud B, Martin K, Haramburu F, Moore N. Rates of spontaneous reporting of adverse drug reactions in France. Jama. 2002 Oct 2;288(13):1588.

[5] Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc. 2003 Mar-Apr;10(2):115-28.

[6] Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005 Jul-Aug;12(4):448-57.

[7] Trifiro G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salame G, Catania MA, Salvo F, David A, Moore N, Caputi AP, Sturkenboom M, Molokhia M, Hippisley-Cox J, Acedo CD, van der Lei J, Fourrier-Reglat A. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiology and drug safety. 2009 Sep 15.

[8] Avillach P, Mougin F, Joubert M, Thiessard F, Pariente A, Dufour JC, Trifiro G, Polimeni G, Catania MA, Giaquinto C, Mazzaglia G, Baio G, Herings R, Gini R, Hippisley-Cox J, Molokhia M, Pedersen L, Fourrier-Reglat A, Sturkenboom M, Fieschi M. A Semantic Approach for the Homogeneous Identification of Events in Eight Patient Databases: A Contribution to the European eu-ADR Project. Studies in health technology and informatics. 2009;150:190-4.

[9] Humphreys BL. The 1994 Unified Medical Language System knowledge sources. Health Libr Rev. 1994 Sep;11(3):200-3.

[10] Lamberts H, Wood M, eds. ICPC: International Classification of Primary Care. Oxford: Oxford University Press 1987.

[11] O'Neil M, Payne C, Read J. Read Codes Version 3: a user led terminology. Methods Inf Med. 1995 Mar;34(1-2):187-92.

[12] Miller GC, Britt H. A new drug classification for computer systems: the ATC extension code. Int J Biomed Comput. 1995 Oct;40(2):121-4.

[13] New additions in the BNF 1976-78. Drug Ther Bull. 1976 Nov 19;14(24):93-6.

[14] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.

[15] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.

Address for correspondence

Paul Avillach
paul.avillach@isped.u-bordeaux2.fr
Laboratoire d'Epidémiologie, Statistique et Informatique Médicales (LESIM), ISPED, Université Victor Segalen Bordeaux 2, 146 rue Léo-Saignat, F-33076 Bordeaux cedex, France