Medical Informatics in a United and Healthy Europe
K.-P. Adlassnig et al. (Eds.)
IOS Press, 2009

doi:10.3233/978-1-60750-044-5-190

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,1], Fleur MOUGIN [a], Michel JOUBERT [b], Frantz THIESSARD [a], Antoine PARIENTE [c], Jean-Charles DUFOUR [b], Gianluca TRIFIRÒ [d], Giovanni POLIMENI [d], Maria Antonietta CATANIA [d], Carlo GIAQUINTO [e], Giampiero MAZZAGLIA [f], Gianluca BAIO [g], Ron HERINGS [h], Rosa GINI [i], Julia HIPPISLEY-COX [j], Mariam MOLOKHIA [k], Lars PEDERSEN [l], Annie FOURRIER-RÉGLAT [c], Miriam STURKENBOOM [m], Marius FIESCHI [b]
[a] *LESIM, ISPED, Université Victor Segalen Bordeaux 2, France (FR)* – [b] *LERTIM, Faculté de Médecine, Université de la Méditerranée, Marseille, FR* – [c] *INSERM U 657, Université Victor Segalen Bordeaux 2, FR* – [d] *IRCCS Centro Neurolesi "Bonino-Pulejo", Messina, Italy (ITA)* – [e] *Pedianet - Societa' Servizi Telematici SRL, ITA* – [f] *Health Search - Italian College of General Practitioners, ITA* – [g] *Department of Statistics – Unit of Biostatistics, Università di Milano-Bicocca, ITA* – [h] *PHARMO Coöperation UA, Netherlands* – [i] *Regional Health Agency of Tuscany, Florence, ITA* – [j] *University of Nottingham, UK* – [k] *London School of Hygiene & Tropical Medicine, UK* – [l] *Aarhus University Hospital, Århus Sygehus, Denmark* – [m] *IPCI – Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands*

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. terminology, UMLS, semantic interoperability, database extraction

1. 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

[1] Corresponding Author: 146 rue Saignat, 33076 Bordeaux, France; E-mail: paul.avillach@isped.u-bordeaux2.fr.

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 permits automated processing of the information and facilitates semantic interoperability between different information systems. Medical information which has been highlighted by appropriate coding can be transmitted, interpreted and processed more easily by different systems and thus enables sharing and reuse of the data between information systems.

In the area of drug safety information sharing could enhance the current spontaneously reported information on adverse drug reactions (ADRs), as reporting is far from optimal. Underreporting is high, and it is estimated that only 4% of ADRs are reported through this channel [3]. Therefore, safety signals may be detected too late, as was recently highly debated after the Vioxx withdrawal. It has been recognized that additional complementary systems are necessary [4, 5], which could profit from the wide availability of health care databases throughout Europe. The use of several medical databases for signal detection will overcome the underreporting problems existing with the current system and may detect signals faster.

From this rationale, the European eu-ADR project[2] 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 initially twenty-three specific events [6]. 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.). 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.

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.

2. Material and Methods

Different terminologies are used to code the clinical events in the eight databases. Thus, a common basis was required in order to harmonize queries. The aim was to provide researchers on the local database with a list of medical concepts and associated terms that they must use to identify the events being investigated. 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. The constituents of this shared foundation are UMLS concepts (grouping together terms from different terminologies with the same medical meaning) and not terms.

[2] http://www.euadr-project.org.

Medical terminologies are structured in the form of lists of concepts[3], generally set out in a hierarchical way. A concept can be defined in many ways since the terms[4] 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) and the READ CODE (RCD) classification. Seven databases use the Anatomical Therapeutic Chemical (ATC) system to code drugs. Only QRESEARCH, codes drugs using the British National Formulary (BNF). Note that the present team has established mapping between the ATC and BNF codes.

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

The Unified Medical Language System® (UMLS®) [7] is a biomedical terminology integration system handling more than 150 terminologies. The four terminologies being 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 notion which can be expressed using different synonyms (terms).

To develop our method, we initially studied « 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 taken for all other twenty-three events that have been identified to be of primary importance [6].

The projection of UMLS concepts in the terminologies method comprises of the following: 1) definition of event, 2) identifying the concepts for the event; 3) discussion about concepts with databases; 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 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

[3] A concept is a unit of thought [ISO 5963].
[4] A term is the designation of a given concept in a language in its linguistic formulation [ISO 1087].

the codes related to the concept in the four terminologies that are used in the project. When this automated search failed to identify a term matching a concept in one of the terminologies studied, we undertook a manual search in the terminology concerned to identify the terms which could be correctly related to the concept.

In step 3, databases were asked to take the 'usual approach' and compare this with the concepts provided. The relevance of each concept was discussed via the eu-ADR consortium Internet forum, conference calls and plenary meetings.

In step 4, terms (codes and free text) in different terminologies and languages were provided to the databases. Every listed concept had necessarily to be present in their query. The list of codes and terms that we provided was non-restrictive. Database administrators were free to add all the terms which they deemed relevant in order to recover the UGIB event from their database providing, however, that these terms 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, which were deemed relevant for retrieval of the information. The codes and strings of the matching terms in the four terminologies of interest were also provided.

3. Results

For the event UGIB, a broad definition was created including the following conditions: Upper gastrointestinal 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 comprised: *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. An example with concept *Haematemesis* is coded « 578.0 » in ICD9-CM, « K92.0 » in ICD10, « D14 » in ICPC and « J680 » in RCD.

4. Discussion

The process we implemented allows the homogeneous identification of events in various European databases. The foundation is the 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 has helped add new concepts to the list making 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 between these different databases.

Several limitations of the process emerged. The first was related to the use of laboratory test results, which are available in some databases, for the identification of certain events (e.g., acute kidney failure). This involves identifying a concept by its biological results and not by its name or its place in a nosological description. Our approach does not provide a solution to this problem. The second entails the differences in coding rules between different classifications. Each database can use codes with the granularity of its terminology. READ 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. These features should be borne in mind when analysing the results extracted from the databases.

5. Conclusion

The projection of UMLS concepts in the terminologies and the manual adjustments were validated 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 databases we exploited. The list of concepts, accompanied by the list of associated codes and their strings in free text, have been used by the database administrators to build queries designed to retrieve information from their database using the appropriate terminology. This method will be used for the remaining events in the project. The extraction of the same medical concepts from the eight databases has enabled biostatisticians working on the project to utilise databases which are comparable with respect to the definition of the events sought, despite the high level of heterogeneity between the databases.

Acknowledgements. 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, J.J. (2007) Collect once, use many. Enabling the reuse of clinical data through controlled terminologies. *Journal of the American Health Information Management Association* 78(2):24–29.

[2] Giannangelo, K. (2006) Making the connection between standard terminologies, use cases and mapping. *The Him Journal* 35(3):8–12.

[3] Begaud, B., Martin, K., Haramburu, F., Moore, N. (2002) Rates of spontaneous reporting of adverse drug reactions in France. *Journal of the American Medical Association* 288(13):1588.

[4] Bates, D.W., Evans, R.S., Murff, H., Stetson, P.D., Pizziferri, L., Hripcsak, G. (2003) Detecting adverse events using information technology. *Journal of the American Medical Informatics Association* 10(2):115–128.

[5] Melton, G.B., Hripcsak, G. (2005) Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association* 12(4):448–457.

[6] Trifirò, G., Pariente, A., Polimeni, G., Miremont-Salame, G., Catania, M.F.S. (2008) Data mining on large health record databases for detecting adverse reactions: Which events to monitor? *Drug Safety* 31(10):885.

[7] Bodenreider, O. (2004) The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research* 32(Database issue):D267–270.