Automatic detection of adverse drug events: proposal of a data model

Emmanuel CHAZARD[a,1], Béatrice MERLIN[a], Grégoire FICHEUR[a], Jean-Charles SARFATI[b], the PSIP consortium and Régis BEUSCART[a]
[a] *Lille university hospital, EA2694*
[b] *Oracle France SAS*

Abstract. Our main objective is to detect adverse drug events (ADEs) in former hospital stays. As ADEs are rare, that supposes to screen thousands of electronic health records (EHRs). For that purpose, we need to define a data model that has two main objectives: (1) being able to describe hospital stays from various hospitals (2) being tuned so as to prepare the data mining process: as ADEs are not flagged in the datasets, the data model must be optimized for ADE detection. The article presents the phases of the design and the data model that results from this work. It is compatible with many hospitals. It deals with diagnoses, drug prescriptions, lab results and administrative information. It allows for data mining and ADE detection in EHRs.

Keywords. Adverse drug event detection, data-mining, data model, interoperability.

Introduction

Adverse drug events (ADEs) are a public health issue. Usually, they are detected thanks to non automated methods that encounter several issues:

- Time consuming staff operated reviews: but as ADEs are rare events, the probability of a case to be observed remains low.
- Spontaneous ADE declarations: most often physicians only declare ADEs that are rare or severe and do not result from a reprehensible fault. As a consequence, those declarations only report a low proportion of ADEs [1, 2].

So there is a need for automated ADE detection methods that could allow for large datasets screening. Electronic Health Records (EHRs) are considered to be very useful in the field of ADEs [3, 4] because big amounts of data are routinely collected and allow for a wide retrospective analysis. The objective of our project is to use data mining methods such as decision trees [5-10] and association rules [11, 12] to automatically identify ADEs from EHRs from several French and Danish hospitals. This will help to automatically discover some prevention rules. After a validation step those rules will be implemented into a Clinical decision support system (CDSS).

The first step of that project is to propose a common data model and to feed it with data in respect with the model. In this article we will present the approach we use and

[1] Corresponding Author: Dr Emmanuel Chazard, CHRU de Lille, 2 avenue Oscar Lambret, 59000 Lille, France; E-mail: emmanuel@chazard.org.

the data model we obtain. In our approach, the data model is considered as a strong link in the chain:

- Upstream, the data model has to be compliant with all the available data from French and Danish partners
- Downstream, the data model has to be tuned in order to allow data mining to detect ADEs although they are not explicitly flagged in the data: no field tells “ADE: yes/no”. If such a field was available, it wouldn’t be reliable.

1. Material and methods

1.1. Review of cases

Our aim is to formalize experts’ decision process: assuming a patient encountered an ADE, what part of the available data helped the experts’ advice? A record review is first performed. 90 atypical hospital stays are reviewed by physicians assisted by a computer scientist.

They are asked to answer the following questions:

- *Was there a probable ADE?*
- *If yes, how did you notice there was an ADE?*
- *Is it possible to generalize those criteria as a rule; would new aggregated fields be useful to that?*
- *Does this case inspire you other detectable situations?*

As an example, two clinical cases of ADEs, and the corresponding variables are presented in Table 1 & Table 2. This procedure is followed up and generalized on every available fields (Table 3).

Table 1. First example of ADE case and inferred variable

Clinical case (ADE)	Mr. X had been admitted for phlebitis, treated and discharged. The treatment wasn’t well adapted. The patient bled and had to come back two days later.	
What abnormality is visible in the EHR?	The patient had to come back 2 days later	
What variable(s) could be useful?	*dunh*=delay up to next hospitalization. In present case *dunh*=2.	
Examples of possible uses of the new variable(s)	Binary use:	ifelse(*dunh*< arbitrary_threshold; 1; 0) ifelse(*dunh*< defined_quantile; 1; 0)
	Quantitative use:	1/(*dunh*+1) *equals 0 if the patient never comes back*

Table 2. Second example of ADE case and inferred variable

Clinical case (ADE)		Mr. Y had been admitted in relation with appendicitis and died 4 days later from grand mal status epilepticus.	
Abnormality #1	**What abnormality is visible in the EHR?**	The patient had been admitted for appendicitis but died	
	What variable(s) could be useful?	*Death* {0;1}, in present case *Death*=1. *DRG** {d_1; d_2;…; d_k}, in present case *DRG*=d_i. Expected death knowing the DGR = *exp_death*= P(*Death*| *DRG*=d_i)	
	Examples of possible uses of the new variable(s)	Binary use:	*death* ifelse(*exp_death*<0.05 & *death*=1; 1; 0)
		Quantitative use:	ifelse(*death*=1; log(*exp_death*); 0) ifelse(*death*=1; log(*exp_death*); log(1-*exp_death*))

Abnorm. #2	**What abnormality is visible in the EHR?**	Principal diagnosis of first step of the hospital stay is appendicitis and principal diagnosis of the second step of the stay is epilepsy: those diagnoses concern two different medical specialties.
	What variable(s) could be useful?	Theoretical MDC* of the principal diagnosis of each step *ntmdc*= Number of different theoretical MDCs in the whole stay in present case *ntmdc*=2
	Examples of possible uses of the new variable(s)	Binary use: ifelse(*ntmdc*>1; 1; 0) Quantitative use: *ntmdc* - 1

* DRG: diagnosis related group MDC: major disease category (group of DRGs, medical specialty)

Table 3. Examples of fields' exploitation (ideas)

Native variable	Idea of use	Maybe useful as…
Death	Unlikelihood of the death	ADE detection
ZIP code	Distance from home to the hospital	potential cause
	Does the patient live in the region?	potential cause
	Does the patient live in urban area?	potential cause
Gender	Usable as it	potential cause
External moves	Admittance day of week	potential cause
	Entry by emergency	potential cause
	Transfer to another hospital (acute care only)	ADE detection
Internal moves	Going through ICU*	ADE detection / potential cause
	Back-and-forth patterns	ADE detection
Dates	Age	potential cause
	Unexpected high length of stay	ADE detection
	Short delay up to next hospitalization	ADE detection
Diagnosis	Chronic diseases, admittance grounds	potential cause
	ADE-related diagnosis codes	ADE detection
	Number of different theoretical MDCs*	ADE detection
Lab result	Pre-existing abnormality	potential cause
	Abnormality occurring during the hospital stay	ADE detection
Drug	Drug prescriptions	potential cause
	Specific antidotes, some unexpected prescriptions	ADE detection

* ICU: intensive care unit MDC: major disease category (group of DRGs, medical specialty)

1.2. Review of the available data, cardinalities and encoding systems

A review of the available data is performed to answer the following questions:

- What structured data are available in each partner's EHR?
- What part of those data is mandatory in all the country due to the administrative payment system?
- What part of those data should be available since it is the simplest way to describe information?
- What part of the data could be unreliable or unstable over time?

Then a review of data schemes and cardinality is performed:

- What would be the simplest way to store lab results / drug prescriptions / diagnoses / administrative information?
- Are the available data schemes of the partners able to feed such a relational scheme?

Finally a review of encoding systems allows us to choose common classifications.

1.3. Compromise

We have to reach a compromise from all those considerations in order to define the data scheme. That compromise is reached together by physicians, medical informatics scientists and statisticians over two main axes:

- cardinality of the scheme (number of tables):
  - less relationships: easier data quality control, easier data mining by statisticians
  - more relationships: data closer from the native scheme, easier extraction, less errors, more stability over time
- number of columns (fields):
  - more columns: more data provisions and calculated fields
  - less columns: faster extraction, less errors, compatibility over countries and time

2. Results

2.1. Encoding systems

Diagnoses are encoded using the ICD10 classification [13] (International statistical Classification of Diseases and related health problems, 10th Revision) of the World Health Organization.

DRGs are encoded using the national classifications (French or Danish). The choice of the classification does not have any impact since the groups are only used to compute aggregated statistics that are used in their turn instead of the DRG itself: death frequency, average length of stay, ICU frequency…

Drugs are encoded using the ATC classification [14] (Anatomical and Therapeutic Classification). That classification is not the most precise one but its precision was sufficient for statistical analysis. Moreover it is widely used.

Lab results are encoded using the C-NPU classification (Commission on Nomenclature Properties and Units) of the International Union of Pure and Applied Chemistry [15]. That classification is used by our partners and is chosen because of its ability to take into account units at the opposite of other popular classifications. That point is mandatory to detect abnormal values.

Medical procedures are encoded using local French or Danish classifications. It is not possible to enforce a common system. However the limited uses allows considering partial handmade mappings.

2.2. Data scheme presentation

Figure 1 shows a simplified representation of the data scheme, on which fields are replaced with groups of fields. The data scheme is voluntary not completely normalized as the data are updated from clean transactional databases and are never modified in the repository itself.

Medical and administrative information: the “1- Stays” table contains one row per hospitalization stay. One stay can be made up of one to several steps (emergency, ICU, cardiology…). The “2- Steps of the stay” table contains one row per step. Diagnoses and medical procedures are linked to the steps of the hospital stays.

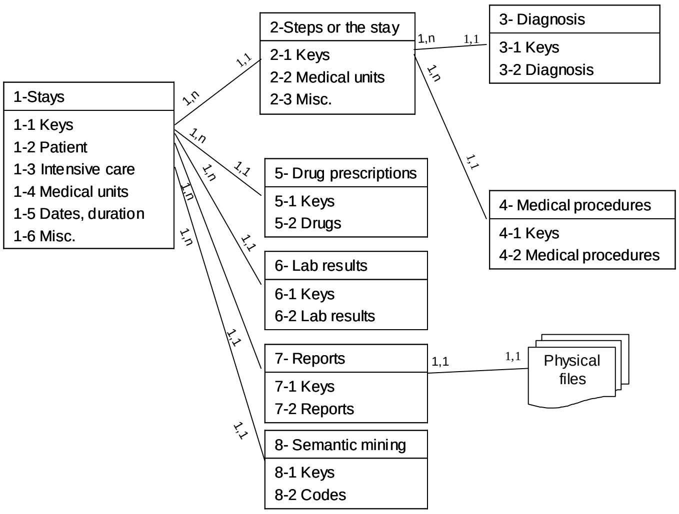

Figure 1. Simplified representation of the data scheme

Drug prescriptions: data-mining is performed stay per stay and details under the day level will be ignored. For a given hospital stay, drug prescriptions must be summed day per day in respect with the administered drug. Drugs corresponding to several ATC codes should be duplicated. Formally speaking, the doses of the drugs are summed and grouped by the {id_hospital, id_stay, date, drug_name, ATC_code} unique quintuplet.

Lab results: one row of the table corresponds to one assessment of one parameter at a given time. If available, each record should contain the normality range (bounds).

Free text data: every report is stored as a physical file linked to its hospital stay thanks to a specific table. Some of our partners use semantic mining to generate ICD10 and ATC codes from reports. A specific table allows registering those codes.

The required data do not contain any nominative nor indirectly nominative data such as birth date, ZIP code or exact dates.

2.3. Fields detailed description

The data scheme contains 8 tables from which 2 are dedicated to reports and 89 fields from which 60 are not identifiers. The field list is shown in Table 4 Table 5 Table 6 Table 7 Table 8 Table 9 Table 10 & Table 11. The original version of the scheme description is completed by detailed description of each field. It is not possible to print it here. That scheme is completed by physical files for the reports.

Table 4. The hospital stay table

Group	Field	Field (long name)	origin	unit
Keys	id_hosp	Hospital ID number	constant	ID number
	id_stay	Stay ID number	database	ID number
	id_patient	Patient ID number	database	ID number
	id_stay_mother	if it was a childbirth, the ID of the mother's stay	database	ID number

	id_stay_newborn	if it was a delivery (childbirth), the ID of the newborn's stay	database	ID number
Patient	age	Age	database	years (float)
	gender	Gender	database	0/1
	drg	Diagnosis Related Group	database	DRG code
	death_01	Death during the stay	database	0/1
	death_exp	Expected frequency of death in this DRG	the proportion in the whole hospital for each DRG	proportion, float between 0 and 1
	geo_state_01	Does the patient come from the hospital's country (state)?	constant	0/1
	geo_region_01	Does the patient come from the hospital's region?	geographic reference	0/1
	geo_dpt_01	Does the patient come from the hospital's department?	geographic reference	0/1
	p_diag	Principal diagnosis	database	ICD10 code
	drg_eff	Number of stays used to compute the various DRG-based statistics (duration_exp, deth_exp, duration_icu_exp, throught_icu_exp)	the number of stays computed in the whole hospital for this DRG	integer
ICU	through_icu_01	Taken care of in intensive care/resuscitation unit?	database	0/1
	through_icu_exp	Expected frequency of stays with intensive care/resuscitation for this DRG	the proportion computed in the whole hospital for each DRG	proportion, float between 0 and 1
	duration_icu	Duration in an intensive care/resuscitation unit	database	days (integer)
	duration_icu_exp	Expected duration in an intensive care/resuscitation unit	the average duration computed in the whole hospital for each DRG	days (float)
	saps	Gravity score	database	integer
	duration_icu_sd	Standard deviation of the duration in an intensive care/resuscitation unit	the std dev of the duration computed in the whole hospital for each DRG	days (float)
	delay_icu	Delay before ICU/resuscitation step	database	integer
Places	nb_mu	Number of medical units visited during the stay	database	integer
	back_forth_01	Back and forth between medical units	database	0/1
	from_emergency_01	Was the patient admitted by an emergency unit?	database	0/1
Dates	duration	Duration of the stay	database	days (integer)
	duration_exp	Expected duration for the stays of this DRG	the average duration computed in the whole hospital for this DRG	days (float)
	delay_next_hosp	Delay up to next hospitalization	database	days (integer)

	duration_sd	Standard deviation of the duration for the stays of this DRG	the std dev of the duration computed in the whole hospital for this DRG	days (float)
Misc	nb_th_mdc	Number of different theoretical MDCs (Major Diagnostic Categories)	table "steps of the stay"	integer
	transfer_entry_01	Transfer from another hospital (whatever the kind)	database	0/1
	transfer_01	Transfer to another acute care hospital	database	0/1
	nb_proc	Number of different medical procedures	database	integer
	nb_diags	Number of different associated diagnosis	database	integer
	weight	weight of the patient	database	float

Table 5. The steps of the hospital stays table

Group	Field	Field (long name)	origin	unit
Keys	id_hosp	Hospital ID number	constant	ID number
	id_stay	Stay ID number	database	ID number
	id_step_stay	StayStep ID number	database	ID number
	id_patient	Patient ID number	database	ID number
Places	mu	Medical unit of the step	database	name
	icu_01	Is it an intensive care unit?	database	0/1
	emergency_01	Is it an emergency room?	database	0/1
Misc	saps	Gravity score	database	integer
	p_diag	Principal diagnosis of step of the stay	database	ICD10 code
	th_mdc	Theoretical MDC of the principal diagnosis	external ICD related table	integer
	weight	weight of the patient during the step	database	float
	step_stay_rank	the rank of that step in the stay (1 for the first step, 2 for the second one, …, k)	database	integer
	duration	Duration of the step of the stay	database	days (integer)

Table 6. The diagnoses table

Group	Field	Field (long name)	origin	unit
Keys	id_hosp	Hospital ID number	constant	name
	id_stay	Stay ID number	database	ID number
	id_step_stay	StayStep ID number	database	ID number
	id_patient	Patient ID number	database	ID number
Diagnosis	diag	Associated diagnosis	database	ICD10 code

Table 7. The medical procedures table

Group	Field	Field (long name)	origin	unit
Keys	id_hosp	Hospital ID number	constant	ID number
	id_stay	Stay ID number	database	ID number
	id_step_stay	StayStep ID number	database	ID number

	id_patient	Patient ID number	database	ID number
Procedures	proc	Medical procedure	database	act code
	delay_proc	delay between the entry and the procedure execution	database	Days (integer)

Table 8. The drug table

Group	Field	Field (long name)	origin	unit
Keys	id_hosp	Hospital ID number	constant	ID number
	id_stay	Stay ID number	database	ID number
	id_patient	Patient ID number	database	ID number
Drugs	name	Commercial name	database	name
	atc	ATC Code	external drugs related table	name
	delay_drug	delay between the entry and the administration	database	Days (integer)
	dose	total drug dose administered during this day	database	number
	unit	Unit used for the total dose	database	name
	route	Route	database	name

Table 9. The lab results table

Group	Field	Field (long name)	origin	unit
Keys	id_hosp	Hospital ID number	constant	ID number
	id_stay	Stay ID number	database	ID number
	id_patient	Patient ID number	database	ID number
Lab	delay_bio	delay between the entry and the sample	database	Days (integer)
	cnpu	C-NPU identifier (IUPAC) of the setting (NPU01685…)	database or external joint	string
	value	value	database	float
	unit	unit used for the value	database	string
	up_bound	upper bound	database	float
	lo_bound	lower bound	database	float

Table 10. The reports table

Group	Field	Field (long name)	origin	unit
Keys	id_hosp	Hospital ID number	constant	ID number
	id_stay	Stay ID number	database	ID number
	id_patient	Patient ID number	database	ID number
Reports	kind	kind of text	database	String
	filename	filename	database	String

Table 11. The semantic mining table

Group	Field	Field (long name)	origin	unit
Keys	id_hosp	Hospital ID number	constant	ID number
	id_stay	Stay ID number	database	ID number
	id_patient	Patient ID number	database	ID number

Codes	terminology	Terminology or nomenclature name	external terminologies	String
	kind	kind of text	database	String
	code	Code of the term	external database	String
	term	Name of the term	external database	String

2.4. Extraction process

Data extraction has been developed in each hospital. A local mechanism is in charge of extracting data directly into tabulated text files. As the same format is used by the statistical software, there is no need for database loading before the data mining process. Nevertheless, a database is used for other needs (Figure 2).

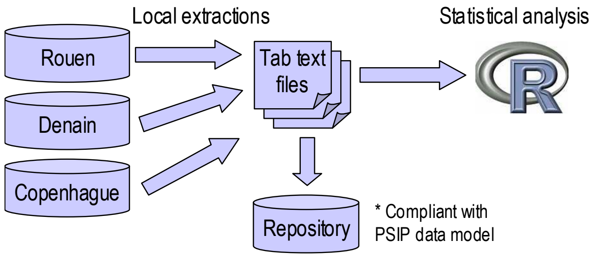

Figure 2. Current data extraction

3. Conclusion

3.1. Perspectives

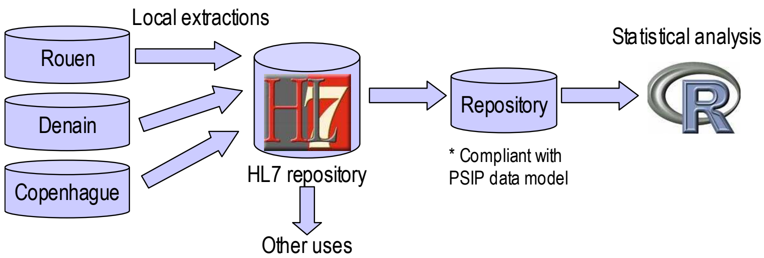

Figure 3. Future data extraction

In the future a central database will be used. That repository will implement an HL7 [16] compatible data scheme. Downstream our data model will still be used to feed the data mining step. The use of HL7 won't improve the process itself but will be useful when new partners join the project. Hospitals that are already able to extract data according to HL7 standards will feed our repository much faster (Figure 3).

3.2. Other uses

The data model has already shown its relevancy since the extraction could be quickly performed and the data mining could be processed and generate interesting results [17].

Moreover, this data model is currently being adapted to another European project. As the study won't concern anymore hospitalizations but long-term follows-up, the data model has to be complemented in order to incorporate more patient-related data and ambulatory care related data.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) [18, 19] under Grant Agreement n° 216130 - the PSIP project [20].

References

[1] Morimoto T, Gandhi TK, Seger AC, Hsieh TC, Bates DW. Adverse drug events and medication errors: detection and classification methods. *Qual Saf Health Care*. 2004 Aug;**13**(4):306-14.

[2] Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. *J Biomed Inform*. 2003 Feb-Apr;**36**(1-2):131-43.

[3] Gurwitz JH, Field TS, Harrold LR, Rothschild J, Debellis K, Seger AC, et al. Incidence and preventability of adverse drug events among older persons in the ambulatory setting. *JAMA*. 2003 Mar 5;**289**(9):1107-16.

[4] Jalloh OB, Waitman LR. Improving Computerized Provider Order Entry (CPOE) usability by data mining users' queries from access logs. *AMIA Annu Symp Proc*. 2006:379-83.

[5] Zhang HP, Crowley J, Sox H, Olshen RA. Tree structural statistical methods. Encyclopedia of Biostatistics. Chichester, England: Wiley; 2001. p. 4561-73.

[6] Breiman L. *Classification and regression trees*. Belmont, Calif.: Wadsworth International Group; 1984.

[7] Fayyad U, Piatetsky-Shapiro G, Smyth P, editors. From data mining to knowledge discovery : an overview. 2nd Int Conf on Knowledge Discovery and Data Mining; 1996.

[8] Lavrac N. Selected techniques for data mining in medicine. *Artif Intell Med*. 1999 May;**16**(1):3-23.

[9] Quinlan JR. Introduction of Decision Trees. *Machine Learning*. 1986;**1**:81-106.

[10] Ripley BD. *Pattern recognition and neural networks*. Cambridge ; New York: Cambridge University Press; 1996.

[11] Agrawal R, Imielinski T, Swami A, editors. Mining Association Rules between Sets of Items in Large Databases. Proceedings of the ACM SIGMOD International Conference on Management of Data; 1993 May. Washington D.C.

[12] Piatetsky-Shapiro G, Frawley W. *Knowledge discovery in databases*. Menlo Park, Calif.: AAAI Press : MIT Press; 1991.

[13] International Classification of Diseases. [cited 2009 february 24]; Available from: http://www.who.int/classifications/icd/en.

[14] Anatomical and Therapeutical Classification. [cited 2009 february 24]; Available from: http://www.whocc.no/atcddd.

[15] International Union of Pure and Applied Chemistry. [cited 2009 february 24]; Available from: http://www.iupac.org.

[16] Health level 7. [cited 2009 April 21]; Available from: http://www.hl7.org/.

[17] Chazard E, Preda C, Merlin B, Ficheur G, Beuscart R. Data-mining-based detection of adverse drug events. *Stud Health Technol Inform*. 2009;**[IN PRESS]**.

[18] European Research Council. [cited 2009 february 24]; Available from: http://erc.europa.eu.

[19] Seventh Framework programme. [cited 2009 february 24]; Available from: http://cordis.europa.eu/fp7/home_en.html.

[20] Patient Safety by Intelligent Procedures in medication. [cited 2009 february 24]; Available from: http://www.psip-project.eu.