Automatic adverse drug events detection: data agregation and data mining

Emmanuel CHAZARD[a,1], Grégoire FICHEUR[a], Béatrice MERLIN[a], Michael GENIN[a],
Cristian PREDA[b], the PSIP consortium, Régis BEUSCART[a]
[a] *Lille university hospital, EA2694, France*
[b] *Lille 1 sciences and technology university, France*

Abstract. Adverse drug events (ADEs) are a public health issue. The objective of this work is to data-mine electronic health records in order to automatically identify ADEs and generate alert rules to prevent those ADEs. The first step of data-mining is to transform native and complex data into a set of binary variables that can be used as causes and effects. The second step is to identify cause-to-effect relationships using statistical methods. After mining 10,500 hospitalizations from Denmark and France, we automatically obtain 250 rules, 75 have been validated till now. The article details the data aggregation and an example of decision tree that allows finding several rules in the field of vitamin K antagonists.

Keywords. Data mining, adverse drug events, decision trees, association rules, anti-coagulation, vitamin K antagonists, INR.

Introduction

Adverse Drug Events (ADEs) are a public health issue. In some hospitals drugs are prescribed using a computerized provider order entry (CPOE) in the frame of the medication use process. In this case it is possible to couple the CPOE with a clinical decision support system (CDSS). It should be then theoretically possible to prevent ADEs by adding some alert rules, such as “aspirin & vitamin K antagonist => increased risk of bleeding”.

But in usual approaches the alert rules are specified by experts. The knowledge that is used usually relies on 2 main sources:

- Academic knowledge: this knowledge mainly relies on summaries of product characteristics and ADE declarations... but ADE declarations are known to only report a tiny proportion of ADEs [1, 2], mostly rare or grave events where the physician’s responsibility is not involved.
- Staff operated reviews (record reviews, chart reviews): those methods can consider very complex situations mixing diseases, drug characteristics and human factors. But they are time-consuming, mostly because ADEs are rare events so their observation requires the review of many normal cases [2].

Those rules induce another underestimated problem: they are applied in the same way to the medical departments all over the countries although those medical

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

departments vary a lot on several aspects. As a consequence in those classic approaches the alerts are too numerous and of poor accuracy. The physicians often complain of over-alerting and their confidence in the system decreases to such an extent that some of them use to deactivate the CDSS.

The main objective or the PSIP project (Patient Safety through Intelligent Procedures in medication [3]) is to build a CDSS relying on automated rules generation, taking into account the context. The objectives of the present work are:

1. to perform a data aggregation: the aim of this step is to transform complex data into events that could be usable as “causes” and “effects”. This step is very important because of the complexity of the available data (many codes, repeated assessments over time, cardinalities of the data scheme, etc.)
2. to automatically data-mine [4] those data in order:
    - to identify adverse drug events
    - to generate control rules to prevent those ADEs

Two important points must be emphasized:

1. The process has to be able to detect ADEs in routine datasets. In those datasets, the ADEs are not flagged..
2. All the process must be automated so that it could be easily performed on new datasets.

1. Material and methods

1.1. Available data

Electronic Health Records (EHRs) seem to be the best data source in the field of ADEs [5, 6]. For the project we have designed a data model and implemented it in a central repository. This data model contains 8 tables and 92 fields (Figure 1).

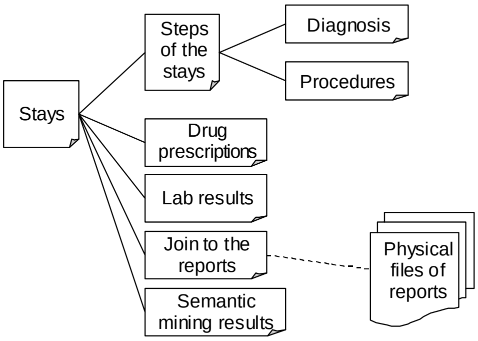


Figure 1. The 8 tables of the data model.

Data extractions are performed to feed the repository. An important point is that no data has to be specifically recorded for the project: we only use routinely collected data from EHRs. Those data include:

- medical and administrative information
- diagnosis encoded using ICD10 [7]
- medical procedures encoded using national classifications
- drug prescriptions encoded using the ATC classification [8]

- laboratory results encoded using the C-NPU classification (IUPAC) [9]

Data are extracted from EHRs provided by the hospitals involved in the project. An iterative quality control of the data is performed in order to get reliable data and to improve the extraction mechanisms. Data extraction is being continued, the present work is performed using 10,500 Danish and French hospital stays of year 2007, mostly from cardiologic or geriatric units:

- Capital Region of Denmark hospitals (Dk): 2,700 hospital stays
- Rouen university hospital (F): 800 hospital stays
- Denain hospital (F): 7,000 hospital stays

1.2. Our hook to fish ADEs

We follow a four-step procedure (Figure 2):

1. Transform the data into events: the native data are complex (thousands of codes, repeated assessments of various lab settings, etc.). They are transformed into binary events. Those events can happen or not. If they happen, they have a start date and a stop date.
2. Qualify the events as “potential cause of ADEs” or “potential events of ADEs”.
3. Automatically find statistical associations between causes and effects. At this step, associations do not necessary mean ADE. For instance we could find “age>90 & renal insufficiency => too long stay”.
4. Filter the associations: associations that contain drugs in their list of causes are kept and are validated against academic knowledge.

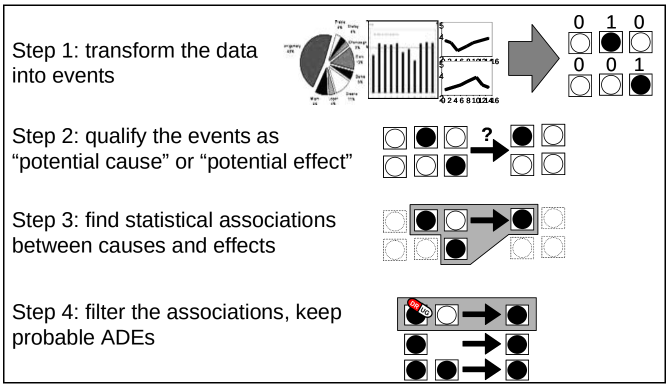

Step 1: transform the data into events
0 1 0
0 0 1
Step 2: qualify the events as “potential cause” or “potential effect”
?
Step 3: find statistical associations between causes and effects
Step 4: filter the associations, keep probable ADEs

Figure 2. Our four-step procedure to fish ADEs.

1.3. Step 1: Transform the data into event (data aggregation)

The extracted datasets fit an 8-tables relational scheme. But such a data repository cannot be used for statistical analysis:

- No statistical method can deal with an 8-tables data scheme
- The encoding systems allow for too numerous classes: about 17,000 codes for ICD10, about 5,400 codes for the ATC, and dozens of different settings for lab results. Many codes have comparable meanings (e.g. the concept of hypoxemia is accessible from the oxygen blood pressure or the oxygen

saturation of hemoglobin) and some concepts result from several settings (e.g. metabolic acidosis).

- Some variables have repeated assessments all the stay along, e.g. lab results (a red cells count can be assessed 20 times during the stay, returning normal, too low or too high values) or drug prescriptions (a specific drug can be prescribed twice per day), etc.

We develop aggregating engines to transform the available data into information described as sets of events (Figure 3). For each kind of data (administrative information, diagnoses, drugs, lab results), a specific aggregating engine is developed and fed with a specific aggregation policy. Each policy is described outside the engine, allowing for several persons to work in the same time to improve the aggregation phase. The aggregating engines allow getting a table that describes the events thanks to many binary fields.

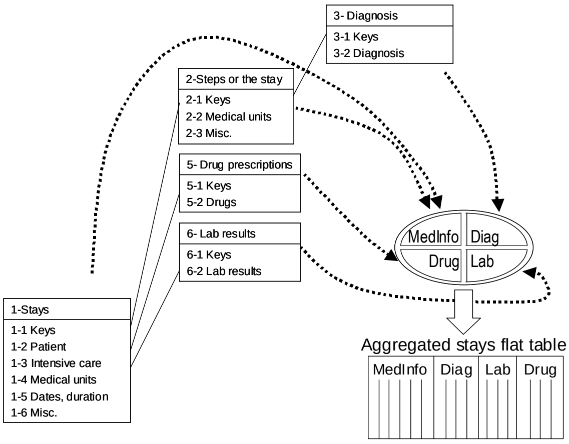


Figure 3. Aggregating engines and mapping policies

Classical statistical analyses rely on associations between different variables that are considered as *stable states*. This is true in some cases (e.g. a patient remains a man or a woman all the stay long) but most often it is false (e.g. a hypoalbuminemia, potential cause of some ADEs, might only exist at days 5, 6 & 7 of a 20-days-long stay). The engines transform data into events. For a given hospital stay, events can have one of the two values (Figure 4):

- 0: the event doesn't occur
- 1: the event occurs at least once. In that case, it is characterized by its start date and end date.

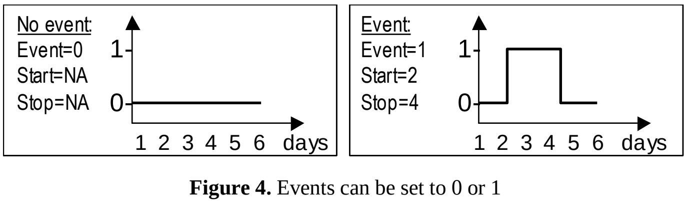


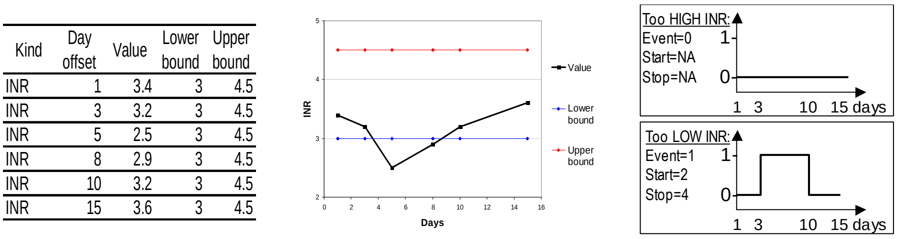
Figure 4. Events can be set to 0 or 1

Example of the Lab Results

As an example, let’s examine the aggregation of the INR (international normalized ratio) values on one example of hospital stay. This setting is interesting for patients under vitamin K antagonists. The expected values are between 3 and 4.5. In this example, INR can assess 2 kinds of risky situations:

- When INR<3, the patient is exposed to a risk of thrombosis.
- When INR>4.5 the patient is exposed to a risk of bleeding.

The lab aggregating engine uses the lab mapping policies and is able to fill two binary variables: too_low_inr and too_high_inr. Those variables are accompanied by start dates and stop dates when their values are set to 1 (Figure 5). LOCF (last observation carried forward) is used to interpolate the available values.



Kind	Day offset	Value	Lower bound	Upper bound
INR	1	3.4	3	4.5
INR	3	3.2	3	4.5
INR	5	2.5	3	4.5
INR	8	2.9	3	4.5
INR	10	3.2	3	4.5
INR	15	3.6	3	4.5

Native data *Chart of the values* *2 binary variables & dates*

Figure 5. Example of transformation: the INR values of a stay

1.4. Step 2: qualify the events as “potential cause” or “potential effect”

We first perform an informal analysis: in the available data, some can be identified as “potential cause of an ADE” and some other as “potential effect of an ADE”. Table 1 shows examples of classifications.

Table 1. Examples of information classification: potential case / potential effect

Kind of information	Ex. of potential ADE cause	Ex. of potential ADE effect
Administrative information	Age, gender	Death, too long stay
Diagnosis	Chronic renal insufficiency	Hemorrhage at the middle of the stay
Lab results	Admission with a too high INR	Hyperkaliemia at the middle of the stay
Drug prescription	Vitamin K antagonist	Specific antidote

Finally thanks to the data-to-events transformation, it is possible to simply consider that all the events that occur after the patient’s admittance are potential effects.

Example of the Lab Results

In this case (Figure 5), a too low INR occurred from the 3rd day (included) to the 10th day (excluded). Those two binary variables can be used as causes and as effects.

- too_low_inr (=1 from day 3 to day 9 in this case)
  - is able to be an effect with value=1. All the other events that occur before day 3 will be candidate to explain that effect
  - is able to be a cause for every effects that occur between day 3 and day 10

- too_high_inr (=0 all along in this case):
  - is able to be an effect with value=0. All the other events will be candidate to explain the absence of effect, whatever their date
  - is able to be a cause with value=0 for every effects, whenever they occur

This approach has two important advantages:

- A statistical association doesn't have any direction. But taking the dates into account prevents from causal relationship inversion. Events that are posterior to the effects cannot be interpreted as causes. Events that are anterior but too far from the effect are not taken into account.
- Effects can become causes in their turn. That approach allows considering an **ADE domino effect**. For instance:
  first drug A & age>70 => acute renal insufficiency
  then acute renal insufficiency & drug B => hemorrhage

1.5. Step 3: automatically find statistical associations between causes and effects

The previous steps allow identifying potential ADE causes and potential ADE effects. The aim of statistical analysis is then to identify some links between (combination of) potential causes and potential effects. Decision trees [10-15] with the CART method were used thanks to the RPART package [16] of R [17]. Decision trees allow identifying several decision rules containing 1 to K conditions such as:

IF(condition_1 & … & condition_K) THEN outcome might occur

Each rule is characterized by its confidence (1: proportion of outcome knowing that the conditions are matched) and its support (2: proportion of records matching both conditions and outcome).

$$Confidence = P(\ outcome \mid condition\_1 \cap \ldots \cap condition\_K) \qquad (1)$$

$$Support = P(\ outcome \cap condition\_1 \cap \ldots \cap condition\_K) \qquad (2)$$

Rules are automatically produced.

1.6. Step 4: filter the associations, keep probable ADEs

The rules are then automatically filtered according to the following criteria:

- The rule must contain at least one of the following events type as a condition:
  - one drug
  - one drug suppression
  - one lab result that is implicitly linked to a drug (e.g. INR for vitamin K antagonist, digoxinemia for digoxin…)
- The rule must increase the prevalence of the effect:
  Confidence > P(outcome)
- The rule must lead to a significant Fisher's exact test for independency between the set of conditions and the outcome.

A theoretical validation of the obtained rules is finally performed by physicians. Only rules that can be explained according to summaries of products characteristics and bibliography are kept. That review uses several drug-related web information portals [18-20], Pubmed [21] referenced papers, and French summaries of products characteristics provided by the Vidal company. In order to be sure the validated rules are reliable, the stays they allow to detect have to be reviewed by experts; this work is currently being processed.

2. Results

2.1. Data aggregation

The figures that are presented here only reflect the current progress status, they are likely to change.

The 18,000 ICD10 codes are aggregated into 48 categories of chronic diseases.

The 5,400 ATC codes are aggregated into 242 drug categories. Those categories are designed to be redundant: they allow for transversal categories such as “hepatic enzyme inhibitors”. The classification has to consider pharmacodynamics and pharmacokinetics although most of the existing classifications are based on therapeutic indications. Drug suppression is also traced as a potential ADE cause.

The laboratory results are aggregated into 35 lab abnormalities.

The various administrative fields are aggregated into 15 different variables.

The data aggregation produces one dataset per medical department. In each dataset up to 538 cause variables can be used to explain or predict 79 effect variables.

2.2. Decision rules

Decision trees are systematically computed in order to explain each effect by all the available potential causes. It allows generating more than 250 decision rules of which 75 have been validated till now.

In the following example we trace the effect “appearance of a too low INR”. When patients are under vitamin K antagonist (VKA) treatment, the international normalized ratio (INR) is traced in order to evaluate the treatment. In case of too high INR, there is a VKA overdose; the patient could present a hemorrhage. In case of too low INR, there is a VKA underdose; the patient is exposed to a risk of thrombosis. A tree is automatically generated.

The first split of the tree shows that the effect is most associated with the admission with a too high INR (risk of bleeding, Figure 6). When a patient enters the department with a too high INR there might be an over-correction of the treatment and a risk of thrombosis in 29% of the cases. Elderly patients admitted with a too high INR and a hypoalbuminemia are over-corrected in 87% cases. Albumin is the blood protein to which VKAs are linked. Only the unlinked fraction of VKAs is biologically active. Hypoalbuminemia was probably the cause of the too high INR but it also increases the effect of VKA correction, which was probably ignored by the physician.

That rule is interesting because it mixes together three kinds of conditions:

- a pharmacokinetics condition: hypoalbuminemia
- an epidemiological condition: the age

- an organizational condition: entry with too high INR

The patients who enter with a normal INR and receive in the same time VKA and a digestive prokinetic drug experience a too low INR in 67% cases (Figure 7a). Digestive prokinetic drugs decrease the bio-availability of VKA.

The patients aged less than 76 that are given VKA and beta lactam antibiotics experience a too low INR in 60% cases (Figure 7b). Several interpretations are possible: the antibiotic indicates an infection; infections may increase hepatic catabolism and decrease VKA bio-availability. Otherwise, antibiotics decrease vitamin K production in the digestive tract, that effect might be known and overbalanced by the physician.

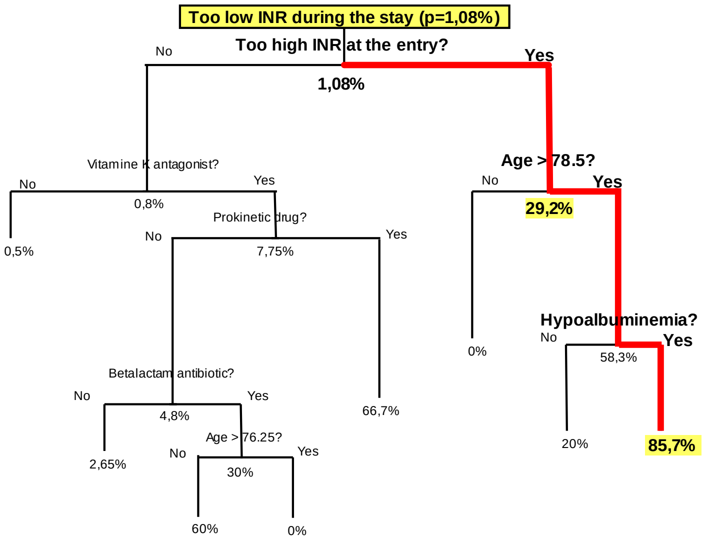


Figure 6. First rule gives p(too low INR during stay)=86% instead of 1%

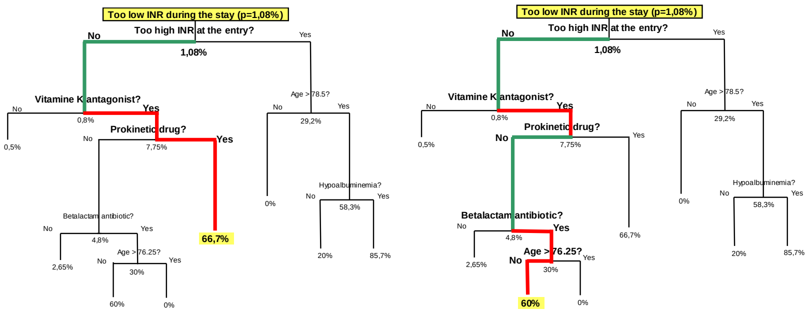


Figure 7. Second and third rules give p(too low INR during stay)= 67% and 80%

3. Discussion and conclusion

Data mining often relies on a simple idea:

1. The observed effect is known: the group with effect (ADE=1) and the group without effect (ADE=0) have already been identified.

2. Several potential causes are available. In epidemiological projects, a restricted list of cause variables is chosen *a priori*. In data mining projects, a large set of potential causes is available; a statistical method is used to find the more significant ones.
3. The appropriate methods are used to explain the effect by the causes

That procedure is not possible in our project:

- mostly because *the effect is not identified*: no one flagged the cases as “normal” (ADE=0) nor “abnormal” (ADE=1), and our objective is to avoid a time-consuming staff operated review
- even most of the causes do not formally exist in the data

For that reason, data aggregation is a very important step. The accuracy of the results essentially relies on that step. However we are aware that despite its advantages, our procedure also suffers from some weaknesses. (1) Only the data that are recorded can be mined. Some clinical events might occur and might not be encoded in the EHR. (2) Diagnosis codes are important to describe acute and chronic diseases. Till now we are only able to take into account chronic diseases and acute diseases that cannot occur during the hospitalization. For instance if an ICD10 code describing a hemorrhage is present in the data, we cannot know if it is the admission ground or an event occurring during the hospitalization.

At the opposite of academic knowledge, the results of the PSIP project allow to sort the knowledge according to the probability of the events. For instance the “contraindication” and “use caution” sections of the French summaries of products characteristics of current VKAs are 3,300 words long. Moreover the knowledge that first appears in the text is already well known by the physicians so that the events that are first described rarely occur. The readers are flooded.

In addition, the rules from the PSIP project are able to take into account “what happened today”. Conditions such as “the patient entered with a too high INR” are typically useful but absent from academic knowledge. Organizational circumstances are probably not enough considered.

First results of the PSIP project are encouraging [22] and announce a new approach in the ADE studies, actual approaches being essentially based on staff operated cases reviews [23] or databases queries [24-26]. The project is still be continued: the mapping policies are improved, the rule discovery is extended to other drugs and diseases, and other data-mining methods are being tried and compared with the decision trees: we are currently working on association rules [27, 28].

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) [29, 30] under Grant Agreement n° 216130 - the PSIP project [3].

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

[4] Adriaans P, Zantinge D, Syllogic (Firm). *Data mining*. Harlow, England ; Reading, Mass.: Addison-Wesley; 1996.

[5] 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.

[6] 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.

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

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

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

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

[11] 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.

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

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

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

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

[16] Therneau TM, Atkinson B, Ripley B. rpart: Recursive Partitioning. 2007.

[17] R_Development_Core_Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.

[18] Pharmacorama. [cited 2009 february 24]; Available from: http://www.pharmacorama.com.

[19] Banque de Données Automatisée sur les Médicaments. [cited 2009 february 24]; Available from: http://www.biam2.org/accueil.html.

[20] Theriaque. [cited 2009 february 24]; Available from: http://www.theriaque.org/InfoMedicaments.

[21] Pubmed. [cited 2009 february 24]; Available from: http://www.ncbi.nlm.nih.gov/pubmed.

[22] Beuscart R, Beuscart-Zéphir MC, the_PSIP_Consortium, editors. Workshop on Patient Safety through Intelligent Procedures in Medication. MIE Conference; 2008 25-28 May; Goteborg, Sweden.

[23] 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.

[24] Honigman B, Lee J, Rothschild J, Light P, Pulling RM, Yu T, et al. Using computerized data to identify adverse drug events in outpatients. *J Am Med Inform Assoc*. 2001 May-Jun;**8**(3):254-66.

[25] Honigman B, Light P, Pulling RM, Bates DW. A computerized method for identifying incidents associated with adverse drug events in outpatients. *Int J Med Inform*. 2001 Apr;**61**(1):21-32.

[26] Seger AC, Jha AK, Bates DW. Adverse drug event detection in a community hospital utilising computerised medication and laboratory data. *Drug Saf*. 2007;**30**(9):817-24.

[27] 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.

[28] Piatetsky-Shapiro G. Discovery, Analysis, and Presentation of Strong Rules. In: Frawley GP-SaWJ, editor. *Knowledge Discovery in Databases*. Cambridge, MA: AAAI/MIT Press; 1991.

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

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