Data-mining-based detection of adverse drug events

Emmanuel CHAZARD[a,1], Cristian PREDA[b], Béatrice MERLIN[a], Grégoire FICHEUR[a], the PSIP consortium, Régis BEUSCART[a]

[a] *Medical Information and Records Department EA2694, University Hospital, Lille, France*
[b] *Paul Painlevé Mathematics Laboratory, Sciences & Technology University, Lille, France*

Abstract. Every year adverse drug events (ADE) are known to be responsible for 98,000 deaths in the USA. Classical methods rely on report statements, expert knowledge, and staff operated record review. One of our objectives, in the PSIP project framework, is to use data mining (e.g. decision trees) to electronically identify situations leading to risk of ADE. 10,500 hospitalization records from Denmark and France were used. 500 rules were automatically obtained, which are currently being validated by experts. A decision support system to prevent ADE is then to be developed. The article examines a decision tree and the rules in the field of vitamin K antagonists.

Keywords. data mining, medical informatics, adverse drug events, decision trees, anti-coagulation, vitamin K antagonists.

1. Introduction

Every year, adverse drug events (ADEs) are known to be responsible for 10,000 deaths in France and 98,000 deaths in the USA[1] in both ambulatory care and hospitalization. During hospitalization some ADEs can be prevented when the medication use process is managed by a computerized provider order entry (CPOE) coupled with a clinical decision support system (CDSS). Some alert rules can then be implemented, e.g. when a drug is prescribed despite a contraindication. Those alert rules are usually designed by experts according to academic knowledge taken from summaries of drug characteristics and from ADE report statements. Unfortunately only a tiny proportion of ADEs is known to be reported that way[2,3].

The so produced rules are used in the same way in each medical department although those departments may differ a lot on several aspects: the patient may differ (disease, associated pathologies, age, gender…), the drugs may differ (drug approval, price and availability…), the prescriptions may differ (depending on the physician's specialty and knowledge, risk aversion, scientific beliefs, local procedures…). Therefore the alerts are too numerous and not accurate enough in those typical methods.

Our work aims at using data mining[4]:

1. to electronically identify hospital stays with a suspicion of adverse drug events
2. to generate decision rules that could prevent these ADEs. Those decision rules will be specific to each medical department

[1] Corresponding Author: Emmanuel Chazard, Secteur d'Information et des Archives Médicales, CHRU de Lille, 2 avenue Oscar Lambret, 59000 Lille, France; E-mail: emmanuel@chazard.org

3. to implement the rules in a contextualized CDSS

This work is part of the PSIP project (Patient Safety through Intelligent Procedures in medication)[5], a European project funded by the European Research Council[6,7]. 13 different partners from 6 countries are involved. The project began in January 2008 and is due to last 40 months.

2. Material and methods

2.1. Data model definition, data extraction and control

Data from Electronic Health Records (EHRs) seemed to be the best data source in the field of ADE[8,9]. Available data were extracted from EHR including:

- medical and administrative information
- diagnoses using ICD10 codes[10]
- drug prescriptions using the ATC classification[11]
- laboratory results using the IUPAC classification[12]

A consensual data model containing 8 tables and 92 fields was defined. Its design was based on an informal case review and on an identification of the available data in France and in Denmark. An iterative quality control of the data was performed in order to improve the extraction mechanisms. The extraction process will soon be extended to more records. The present work reviewed 10,500 complete hospital stays over year 2007, mostly from cardiologic and geriatric units:

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

2.2. From data to information: data aggregation

The datasets fit an 8-table relational scheme that cannot be used for statistical analysis:
(1) no statistical method can deal with an 8-table data scheme
(2) classes are too numerous (ICD10: 17,000 codes, ATC: 5,400 codes…)
(3) some variables are collected several times during the same stay, such as lab results (a red cells' count can be assessed 20 times during the stay, with normal, above or below normal results) or drug prescriptions.

Data aggregation processes were defined for each kind of variables. Data aggregation engines were fed with data aggregation policies that had to be defined too. At the moment:

- the 18,000 ICD10 codes are aggregated into 52 categories of chronic diseases.
- the 5,400 ATC codes are aggregated into 244 drug categories.
- lab results are aggregated therefore 31 biological abnormalities can be traced.

The data aggregation produces one dataset per department. In each dataset up to 564 cause variables can be used to explain or predict 48 effect variables.

2.3. From information to statistical association: Data analysis

The aggregation process helps to identify potential ADE causes and potential ADE effects. The aim of statistical analysis is to identify links between (combination of)

potential causes and potential effects. Decision trees[13-18] with the CART method were used thanks to the RPART package[19] of R[20]. Decision trees produce several rules containing 1 to K conditions such as: *"IF(condition_1 & … & condition_K) THEN outcome might occur"*. Each rule is characterized by:

- its confidence: proportion of outcome knowing that conditions are matched
  Confidence = P(outcome | condition_1 ∩…∩ condition_K)
- its support: proportion of records matching both conditions and outcome
  Support = P(outcome ∩ condition_1 ∩…∩ condition_K)

2.4. From association to ADE detection and decision rules: Experts validation

Some physicians performed a theoretical validation of the so obtained associations. They validated only those that looked like ADEs and possible related causes according to various sources. This review used several web information portals[21-23], Pubmed[24] referenced papers, and French summaries of product characteristics.

In order to make sure that the validated rules are reliable, some experts will have to review the hospital stays that the rules singled out. This work is currently being performed.

3. Results

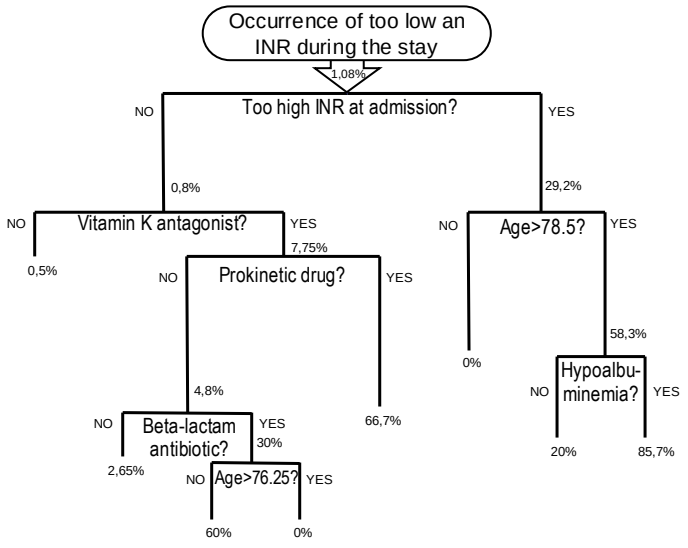

Figure 1. Leaves of the decision tree minimize or maximize P(too low INR during stay)

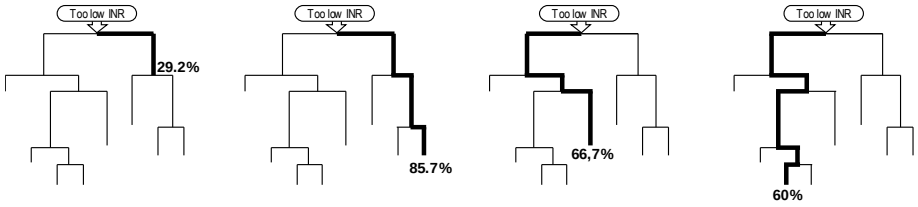

Figure 2. Four examples of decision rules that allow to increase P(too low INR during stay)

Decision trees were automatically generated. The aim was to identify associations between potential ADE effects and potential ADE causes. It was thus possible to get more than 500 association rules. So far 40 out of these 500 association rules have been validated according to academic knowledge and can be used as decision rules.

In the following example the "occurrence of too low an INR" effect is tracked in a medical department (Figure 1). When patients are under vitamin K antagonist (VKA), the international normalized ratio (INR) is monitored in order to evaluate the treatment. In case of too high INR (VKA overdose) the patient is exposed to hemorrhage. In case of too low INR (VKA underdose) the patient is exposed to a risk of thrombosis.

When a patient is admitted in the department with too high an INR (risk of bleeding) there might be an over-correction of the treatment and a risk of thrombosis in 29% of cases (Figure 2a). Elder patients admitted with too high an INR and hypoalbuminemia are over-corrected in 87% cases (Figure 2b). VKA are bound to albumin in the blood. Only the unbound fraction is biologically active. Hypoalbuminemia was probably the cause of the too high INR but it also probably increased the effect of VKA correction.

67% of patients who were admitted with a normal INR and received both VKA and a prokinetic drug, experienced too low an INR (Figure 2c). Prokinetic drugs decrease VKA adsorption.

60% of patients under 76 that are given VKA and beta lactam antibiotics experience too low an INR (Figure 2d). 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: the prescriber may be well aware of this effect and overbalance it by decreasing VKA dose.

4. Discussion and conclusion

This work automatically identifies ADE-prone hospital stays from decision trees using causes-effect statistical associations. Those associations are computed in each department separately. Experts then validate the decision rules that can be used in a CDSS. The rules are specific to each department and refer to situations that have actually occurred. The first results of the PSIP project are encouraging[18] and announce a new method in ADE studies, while current methods essentially rely on time-consuming case reviews[25] or database queries which do not use statistical tools[26-28].

Most of the rules that were validated are already known. But the academic knowledge about drugs presents several problems. First, the rules are too numerous for professionals (around 100 rules in the French drug characteristics summaries for common VKAs). The weighting of the knowledge is based on the severity of the effects. But the most important rules are well known by physicians and the related ADEs seem to occur quite rarely. Moreover the appropriate weights would be different depending on the medical departments. Finally the academic knowledge does not deal with specific circumstances (e.g. "the patient was admitted with too high an INR"). Unfortunately such organizational causes cannot be found in the medical literature yet.

Our rules were first validated using the academic sources. The hospital stays will eventually be reviewed by experts in order to confirm the existence of ADEs. Then human factors will have to be taken into account: when an alert occurs, the system has

to consider what the user is entitled to do, what his profile is and at which step the alert is given.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under Grant Agreement n° 216130 - the PSIP project. Acknowledgments to Mrs Karine Wyndels.

References

[1] Kohn L T, Corrigan J M, Donaldson M S *To err is Human*, Washington DC, National Academy Press, 1999
[2] Morimoto T., Gandhi T.K., Seger A.C., Hsieh T.C., Bates D.W. (2004) Adverse drug events and medication errors: detection and classification methods. *Qual Saf Health Care* **13**: 306-314
[3] Murff H.J., Patel V.L., Hripcsak G., and Bates D.W. (2003) Detecting adverse events for patient safety research: a review of current methodologies. *Journal of Biomedical Informatics* **36**: 131-143
[4] Adriaans P; Zantige *Data mining*. Edingburgh: Addison Wesley, 1996
[5] http://www.psip-project.eu (last visited: January. 16, 2009)
[6] http://erc.europa.eu/ (last visited: January. 16, 2009)
[7] http://cordis.europa.eu/fp7/home_en.html (last visited: January. 16, 2009)
[8] Gurwitz J.H., Field T.S., Harrold L.R. and al. (2003) Incidence and preventability of Adverse Drug Events among older persons in the ambulatory setting. *JAMA* **289**: 1107-116
[9] Jalloh O.B., Waitman L.R. (2006) Improving Computerized Provider Order Entry (CPOE) Usability by Data Mining Users' Queries from Access Logs *AMIA 2006 Symposium Proceedings*: 379-383
[10] http://www.who.int/classifications/icd/en (last visited: January. 16, 2009)
[11] http://www.whocc.no/atcddd (last visited: January. 16, 2009)
[12] http://www.iupac.org (last visited: January. 16, 2009)
[13] Zhang H.P., Crowley J. Sox H. and Olshen R.A. (2001) Tree structural statistical methods. *Encyclopedia of Biostatistics*, 6: 4561-4573, Wiley, Chichester, England.
[14] Breiman L., Friedman J.H., Olshen R. and Stonce C. *Classification and Regression Trees*, Wadsworth, Belmont, California, 1984
[15] Quinlan J R Introduction of Decision Trees, *Machine Learning* **1**: 81-106, 1986
[16] Fayyad U M, Piatetsky-Shapiro, Smyth P, From data mining to knowledge discovery : an overview. *Advances in Knowledge Discovery and Data Mining*, MIT Press : 1-36, 1996
[17] Lavrac N, Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine* **16**: 3-231, 1999
[18] Beuscart R, Beuscart-Zéphir MC, and the PSIP Consortium (2008) Workshop on Patient Safety through Intelligent Procedures in Medication , *MIE Conference*, Goteborg, 25-28 May 2008.
[19] Terry M Therneau and Beth Atkinson. R port by Brian Ripley. (2007). rpart: Recursive Partitioning. R package version 3.1-38.
[20] R Development Core Team. (2006). R: A Language and Environment for Statistical Computing, Vienna, Austria.
[21] http://www.pharmacorama.com (last visited: January. 16, 2009)
[22] http://www.biam2.org/accueil.html (last visited: January. 16, 2009)
[23] http://www.theriaque.org/InfoMedicaments (last visited: January. 16, 2009)
[24] http://www.ncbi.nlm.nih.gov/pubmed (last visited: January. 16, 2009)
[25] Bates D.W., Evans R.S., Murff H., Stetson P.D., Pizziferi L, Hripsack G. (2003) Detecting Adverse Events Using Information Technology *JAMIA* **10**: 115-128
[26] Seger A.C., Jha A.K. and Bates D.W. (2007) Adverse Drug Event Detection in a Community Hospital Utilising Computerised Medication and Laboratory Data. *Drug Safety* **30**: 817-824
[27] Honigman B., Lee J., Rotschild J., Light P., Pulling R.M., Yu T., Bates D.W. (2001) Using computerized Data fo identify Adverse Drug Events in Outpatients. *JAMIA* **8**: 254-266
[28] Honigman B, Light P, Pulling RM, Bates DW. (2001) A computerized method for identifying incidents associated with adverse drug events in outpatients. *International Journal of Medical Informatics* **61**: 21-32