Human Factors Methods to Support the Experts’ Review of Automatically Detected Adverse Drug Events

Nicolas LEROY a,1, Brian BJØRN b, Adrian BACEANU c, Marie-Catherine BEUSCART-ZEPHIR a

a Université Lille Nord de France, INSERM CIC-IT-Evalab, Lille; CHU Lille; UDSL EA 2694, Lille, France
b Patient Safety Unit, Region H, Copenhagen, Denmark
c Ideea Advertising®, Bucharest, Romania

The PSIP consortium 2

Abstract. The European project Patient Safety through Intelligent Procedures in Medication (PSIP) aims at semi-automatically identifying and preventing Adverse Drug Events (ADEs). Data mining of the structured hospital data bases provides a list of potential ADEs, along with their frequencies and probabilities. Once a set of ADEs has been detected by data mining techniques, it is necessary to have them validated by human experts. This paper presents the methods used to support the review by clinicians and pharmacologists of these automatically detected ADEs. We use think-aloud methods and portable labs to track and record the experts reasoning and their reviewing cognitive procedures. We present preliminary results obtained with this method, which allows identifying the key data and information used to characterize the ADEs.

Keywords. adverse drug event (ADE), data mining, knowledge elicitation, Think-Aloud protocols, cognitive processes

1. Introduction and Background

In the last ten years, ADEs have become a major public health issue [1]. Healthcare Information and Communication Technology (ICT) applications should help reducing the prevalence of preventable ADEs but their efficiency is impeded by the lack of reliable knowledge about ADEs, and the poor ability of ICT solutions to deliver contextualized knowledge. Given the ever increasing availability of Electronic Health Records (EHR) and Computerized Physician Order Entry (CPOE) systems in the hospital setting it becomes possible to automatically screen or mine these data to identify abnormalities signaling potential ADEs [2, 3]. This is also one of the core objectives of the European project PSIP – Patient Safety through Intelligent Procedures in medication [https://www.psip-project.eu/]. PSIP aims at overcoming the problem of...
ADE detection by searching large repositories of electronic medical records and data in order to detect abnormal cases presenting typical ADE features. In PSIP, a common data model has been created to allow exports of thousands of medical records from different hospitals and different European countries into a common repository. Data mining techniques performed on these medical data allow identifying “abnormal” hospital stays (i.e., suspect of an ADE) along with the association rules statistically linked with these stays. [4]. The main statistical technique applied to the data is decision tree. Decision trees are run on each effect in different contexts (different hospitals / departments). An example of rule issued by a decision tree is “If too high INR (International normalized Ratio) at entry AND hypoalbuminemia AND age>75 → risk of too low INR”. For each context a rule is characterized by (i) its confidence i.e., the number of positive cases for the predictor and (ii) its support, i.e., the number of positive cases vs the number of cases matching the conditions. For the above mentioned rule, in one medical unit of the Danish Region H hospitals 12 stays match the predictor (conditions of the rule) and 7 present the effect (confidence = 7/12 = 58%, support = 7) [5].

When records pointing at a potential ADE are retrospectively automatically identified, it is mandatory that human experts review a sample of normal/abnormal cases to validate their characterization as “normal” vs. “potential ADE” and the clinical relevance of this automatic categorization. In this approach, the experts’ judgment is considered a gold standard. Other authors have described methods to support the experts’ review and validation of electronically screened medical records [6–8]. One of the problems is that having physicians or pharmacologists experts reviewing exhaustive EHR is extremely time consuming and costly. Therefore this review process may involve a pre-reviewing stage by a trained nurse or enrolled physician able to select the most relevant information and collect it into synthesized charts to be reviewed by the experts [6]. Charts are then reviewed independently by two experts, who are asked to characterize each case as (potential) ADE/Non ADE. Judgments concerning ADEs may be supported by structured data forms such as the Adverse Event Analysis form developed by investigators in the Harvard Medical Practice Study (MPS) [9] and used by Murff et al. [6]. This method allows to review a large number of charts and to categorize the corresponding cases as probable ADE or not. It then allows calculating an inter-experts agreement and the accuracy of the automatic detection of ADE (True / False Positives / Negatives). However, the method also presents some limitations. Even when relying on synthesized charts extracted from the complete records, the manual review process remains time consuming. Although the structured data form gives the experts the opportunity to characterize the validated ADEs on a number of variables, it does not provide enough information on the experts reasoning to analyze the reasons why the experts disagreed (or agreed) with each other and/or with the automatic categorization of the cases as ADE / Non ADE.

In the PSIP project, we tried to circumvent these limitations and to take a step further. We took advantage of the existence of the common data model and of the availability of all the clinical data in a common repository to design a query application giving the expert reviewers access to all the data of a given case. We used Human Factors techniques and methods such as Think-Aloud protocols and portable usability labs to monitor and record the experts’ activities (behavioral and cognitive ones) while reviewing the cases and assessing the corresponding rules. The objective is to understand the experts’ reasoning and to identify the parameters or data they rely on to interpret and validate / invalidate the ADE cases.
2. Methods

2.1. Questionnaire

We adapted the Adverse Event Analysis form [5, 8] to accommodate the PSIP context. While the original form covers the entire Adverse Events field, we have restricted it to Adverse Drug Events characterization. The scoring of the outcomes of an ADE or potential ADE has been slightly modified to fit the information available in the PSIP common repository and depending on the PSIP data model.

2.2. Web-Based Query and Reviewing Application

One of the PSIP partners (Ideea-Advertising®) has developed a web-based query application allowing authorized experts to view all the data of any anonymized hospital record exported into the common repository. For each hospital stay, the viewer displays in separate screen pages: (i) the description of the stay (demographics, principal diagnosis, duration of the stay, etc.); (ii) the list of the medical units visited during the stay; (iii) the procedures performed; (iv) the ICD10 (International Classification of Diseases -10) diagnoses; (v) the drugs (and dosage) ordered during the stay (data can be sorted by day of the stay or drug name); (vi) the laboratory results (data can be sorted by day of the stay or biochemical type); (vii) the available documents (ex: discharge letter). Additional interactive functions allow supporting the reviewing process itself, allowing the experts to display the list of cases to be reviewed, to navigate through the data and to document the ADE Analysis form simultaneously.

2.3. Experimental Review Procedure

The first data mining iteration in PSIP has been performed on an export of the data of 2,700 stays from a cardiology unit of one of Region H Hospitals of Copenhagen. To test the feasibility and usability of the procedure and of the reviewing system, we selected at random 10 stays identified as “abnormal” (potential ADE) by the data mining techniques and 10 “normal” stays. Two experts, one physician specialized in patient safety and one clinician pharmacologist independently reviewed this sample of 20 stays. The experts are instructed to “think-aloud” [10] while reviewing the cases and documenting the form. A recording system allows tracking all the expert physicians’ actions with the application along with their comments and verbalized thoughts.

3. Results

3.1. Results of the Reviewing Process

The query application proved easy to learn and easy to use for both experts. Only minor comments were made by the experts (e.g., request for additional sorting possibilities). Table 1 displays the results of the reviewing process.
Table 1. Overview of the results of the reviewing process. The results may be distributed into three categories: (1) both experts (E1 and E2) agree with each other and with the Data Mining (DM) results, which is noted {E1=E2}=DM; (2) both experts agree with each other against the DM results, noted {E1=E2}≠DM; (3) Experts disagree with each other, and therefore one of them agrees with DM results, noted {E1≠E2} ; E1 or E2=DM. For each category, the table displays the number of stays falling in the category and the time spent to review the corresponding cases.

<table>
<thead>
<tr>
<th>Category</th>
<th>DM = Normal</th>
<th>DM = Abnormal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%)</td>
<td>mean time (min – max)</td>
</tr>
<tr>
<td>{E1 = E2} = DM (1)</td>
<td>6 (60%)</td>
<td>3'09&quot; (1'39&quot; – 4'59&quot;)</td>
</tr>
<tr>
<td>{E1 = E2} ≠ DM (2)</td>
<td>3 (30%)</td>
<td>4'36&quot; (2'59 – 8'07&quot;)</td>
</tr>
<tr>
<td>{E1 ≠ E2} ; E1 or E2 = DM (3)</td>
<td>1 (10%)</td>
<td>7'23&quot; (7'09&quot; – 7'37&quot;)</td>
</tr>
<tr>
<td>Total</td>
<td>10 (100%)</td>
<td>4'01&quot; (1'39&quot; – 8'07&quot;)</td>
</tr>
</tbody>
</table>

The reviewers spent about 6 minutes per review, but it took them longer to review the “abnormal” cases (mean time = 8 min 20 sec.) than the “normal” ones (4 min and 1 sec.). “abnormal” stays validated as “ADE” by the experts required the longest reviewing time while “normal” stays categorized as “non ADE” by the experts were the quickest. Inter-experts agreement is higher for DM “normal” stays (9 out of 10) while it is moderate for DM “abnormal” stays (5 out of 10). As for the experts agreement with the DM results at least one expert agreed with DM results for 7/10 “normal” stays and for 6/10 “abnormal” stays.

3.2. Qualitative Analysis of the Experts Reasoning and Reviewing Procedures

The analysis of experts’ verbalizations permitted to clear most of the reasons for inter-experts and/or experts-DM disagreements. Concerning the DM “normal” cases, the main reason for experts-DM disagreement (3 cases) is that the experts identified ADEs that had occurred before the hospitalization and eventually caused that hospitalization. The data mining explicitly exclude these cases.

For the DM “abnormal” cases, the main reasons for experts/DM disagreement are:

- the clinical context (2 cases); ex: the patient is dying, it is impossible to tell an ADE from his critical physiological status
- variability in experts judgments; ex: strict or flexible adherence to biochemical results norms
- error in the data; ex: one lab value is completely out of range
- complexity of DM knowledge; ex: the rule attached to the “abnormal” stay involves a too difficult to understand interaction of drugs

The comments of the experts also permitted to identify useful information currently missing in the data model, such as: Weight, Heart rate, Blood pressure, Blood glucose. The analysis of the protocols shows that both experts adopt the same strategy for reviewing the cases. After a quick glance at the patient’s demographics, they directly look at the discharge letter. Then they review simultaneously (i) the drugs and their dosage, with a special interest for drugs modifications (interruptions, changes of doses, etc.); (ii) the lab results, again with a focus on variations in the lab values and changes in orders (ex: a new biochemical parameter is monitored). The experts would
have appreciated “clinically reliable” diagnoses, but they did not trust ICD10-coded diagnoses that were obviously entered into the hospital information system mostly for billing purposes. Quoting one of the experts: “rule number 1: never trust the diagnosis”.

4. Discussion and Conclusion

Most of the hospitals are now equipped with an electronic patient tracking system (Admission/Transfer/Discharge) along with laboratory and pharmacy (drugs) information coded in electronic format [7], but these systems are rarely linked, therefore impeding the electronic review of the patients’ records. The first result of the PSIP project is the common data model used by French and Danish hospitals. Available upon request for other European hospitals, it allows an easy export in a common database. Combined with the querying application, it provides a quick and easy access to almost all the relevant clinical and paraclinical data of a patient’s stay in a structured way, which eases the process of reviewing of patients’ electronic charts.

As far as ADE reviewing and validation is concerned, the integration of Human Factors methods proves promising in catching the experts’ strategy and reasoning processes. This reviewing procedure is currently applied to a large set of French and Danish stays. It should provide fruitful feedbacks for the improvement of both the data model and the data mining procedures in the PSIP project.

Acknowledgement. The PSIP project has received funding from the European Community Seventh Framework Program (FP7/2007-2013) under grant agreement n°216130.

References